Excess Risk Transfer and Learning Problem Reduction towards Reliable Machine Learning
（信頼性の高い機械学習を目指した 剰余リスク転移と学習問題の帰着）

by

Han Bao
包 含

A Doctor Thesis
博士論文

Submitted to
the Graduate School of the University of Tokyo
on December 3, 2021
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Information Science and Technology
in Computer Science

Thesis Supervisor: Masashi Sugiyama　杉山 将
Professor of Computer Science

ABSTRACT

Owing to the massive numbers of data collected through the emergence of the Internet and the rapid developments in computing hardware, machine learning has become one of the indispensable components for data-driven knowledge discovery in modern society. Based on statistical machine learning, inductive inference, in particular, has been a spectacular success, replacing classical artificial intelligence based on deduction. This approach has significantly improved the applicability and plasticity of intelligent systems in domains where abundant data can be automatically accumulated, and will eventually push science and engineering into new frontiers at a pace never seen before.

Despite its success in practical applications, in contrast to deduction, induction can only provide probable consequences based upon enumerated evidence. When the results of statistical inference are applied in risk-sensitive domains, it becomes crucial to estimate how likely it will be for the predictions on unseen patterns to be correct given the limited number of observations. This brings up the concept of statistical learning theory of *generalization*. The central aim of the generalization is to bridge the gap between the *empirical risk* and *population risk*—the former being a discrepancy between the predictions and the observations computed using collected data, and the latter being computed using unseen patterns—to guarantee that the learner is sufficiently confident even with predictions on unseen patterns. Over the past few decades, this field has been extensively developed with the aid of statistics, computer science, and information theory. Roughly speaking, a learner is capable of predicting expected outcomes correctly in common cases given a sufficient number of observations. That being said, is it sufficient to funnel as many data as possible into learning algorithms to achieve reliable learning systems?

This dissertation endeavors to establish a conceptually orthogonal axis to generalization in learning theory, i.e., a theory of *excess risk transfer*, to convince readers that the design of risk functionals plays an important role in seeking reliable learning algorithms and a better understanding of learning mechanisms. At the beginning of the 21st century, the importance of excess risk transfer was first described in studies on consistent surrogate losses, which scrutinizes the gap between learning criteria optimized by popular machine learning methods and the ultimate evaluation metrics, referred to as the *surrogate risk* and *target risk*, respectively. Previous studies on consistent surrogate losses found a monotonic relationship between the surrogate and target risks given well-manufactured surrogate risks. This monotonicity implies excess risk transfer—minimizing the excess of a target risk shall be transferred to minimize the excess of a surrogate risk. As a result, by optimizing a well-crafted surrogate risk with its optimization-friendly characteristics, learners can escape from the intractability of a target risk often arising from its discrete nature.

Although existing research has thus far targeted the analysis of classical learning problems such as binary and multi-class classification, modern learning scenarios involve more intricate structures. In this dissertation, the theory of excess risk transfer is extended beyond the classical setup from two perspectives. First, excess risk transfer is incorporated with a variety of user-specified constraints and properties on predictions, which is in sharp contrast to the previous theory focusing purely on the prediction accuracy. Second, excess risk transfer is further utilized to draw a link between two apparently irrelevant learning problems, different from the previous perspective in which a target risk is regarded as an ultimate and a surrogate risk is merely an auxiliary tool. This viewpoint provides us a tool for comparing the difficulty of two learning problems. To summarize, an excess risk transfer enables us to encompass more learning problems with diverse constraints. Consequently, the theory of excess risk transfer tells us about structures of a broad class of learning problems and what learners can elicit from them, which in turn helps us to design a minimally sufficient learning problem that aligns with our desideratum. Recall that the current direction investigating excess risk transfer is conceptually independent of generalization theory; hence, both can be integrated. The remainder of this dissertation is organized as follows.

In Chapter 1, the history of machine learning and statistical learning theory is introduced, followed by the research questions and the contributions of this dissertation.

In Chapter 2, background knowledge of supervised learning and its theory is provided. After formulating supervised classification, the generalization theory and a classical anal-

ysis of the surrogate excess risk transfer are reviewed. In addition, several relevant notions including classification-calibrated losses and proper losses are introduced along with recent related studies.

In Chapter 3, we describe the design of a consistent surrogate objective for the training of a classifier evaluated based on complex metrics such as the F-measure and Jaccard index—metrics that have been frequently applied in information retrieval and semantic segmentation to deal with class imbalance. These performance metrics belong to a family called linear-fractional metrics, which has a non-decomposable nature, and thereby hampers the application of the classical design of surrogate losses. By carefully designing a tractable surrogate bound of the target evaluation metric, we derive sufficient conditions ensuring an excess risk transfer relationship, i.e., optimization of the surrogate bound implies optimizing the target evaluation metric. A simulation study on benchmark datasets suggests that a classifier optimizing the derived surrogate bound outperforms the plug-in classifier, particularly when the sample size is small. This result not only demonstrates the efficacy of the proposed surrogate objective but also the importance of surrogate optimization over the plug-in classifier.

In Chapter 4, we focus on adversarially robust classification, in which we look for a classifier that is insensitive to adversarial perturbations of the test patterns. This learning problem is often formulated through minimax optimization, where the target risk is the worst-case value of the classification risk subject to a bound on the size of the perturbations. While significant effort has been devoted to making this optimization tractable, it remains unclear whether the relaxations can be justified in light of the target risk of robust classification. For this reason, an excess risk transfer analysis is applied to inspect which surrogate risk is aligned with the robust classification risk, and it was eventually determined that no convex surrogate losses can lead to the optimally robust solutions under the assumption of linear models. In addition, useful insight into the design of nonconvex surrogate losses is provided herein. These results are interesting from two perspectives. First, introducing an adversary makes the consistency of convex surrogate losses impossible to achieve. Second, the theory of consistent surrogate losses is shown to be applicable not only to the prediction accuracy but also to other desirable properties such as robustness.

In Chapter 5, we investigate the underlying connection between similarity learning and classification. Similarity learning is a general framework applied to elicit useful representations of data by predicting the relationship between a pair of patterns, which includes a number of preprocessing tasks such as metric learning and contrastive learning. Although a classifier built upon the learned representations is expected to perform well on downstream classification, little theoretical insight in this area is known thus far. To describe how similarity learning supports downstream classification, we reveal that a specific formulation of similarity learning is tightly related to the classification risk. This link indicates that minimizing the excess risk of similarity learning leads to minimizing the excess classification risk from an excess risk transfer perspective. Consequently, we discover that similarity learning is essentially capable of eliciting the underlying binary decision boundary. From the viewpoint of excess risk transfer, we take one step further by regarding two distinct learning problems, i.e., binary classification and similarity learning, as target and surrogate problems, respectively, which opens a new possibility of examining interconnections among learning problems and characterizing whether one problem can be reduced to solving the other problem.

In Chapter 6, we conclude this dissertation with future prospects. A perspective of learning theory is investigated in light of excess risk transfer, which enables us to analyze whether a learned classifier satisfies the desired properties, and to reduce one learning problem to the other problem.

論文要旨

現代社会ではウェブの出現や計算資源の急速な発達によって莫大なデータの収集が可能になり，データ駆動型の知識発見において機械学習が重要な要素技術の一つとなりつつある．特に統計的機械学習に基づく帰納推論は目覚ましい成功を収めており，演繹推論に基づく旧来の人工知能に置き換わっている．データが大量にかつ自動的に蓄積されるような応用領域では，帰納推論に基づく知識システムが従来に比べてより柔軟で応用しやすくなっており，科学や工学などの水準が急速に向上している．

実用的には成功しているものの，演繹推論とは異なり帰納推論は実例を集めて結果の正しさを蓋然的に保証することしかできない．仮に統計的推論がリスクの大きい応用領域にて用いられる場合，有限の観測が与えられた下で未知の入力に対して予測がどの程度正しいかを知ることは肝要である．汎化の統計的学習理論はこの課題意識に根ざしている．汎化理論の中心的な目的は，経験リスク（有限のデータから計算された予測と観測の差）と期待リスク（未知のデータから計算された予測と観測の差）の違いを調べることである．過去数十年の間，当該分野は統計学，計算機科学，情報理論の各分野に支えられて大きな発展を遂げてきた．大まかに言えば，汎化理論によって，一般的には十分量の観測さえ与えられれば学習器は期待される出力を正しく予測することが可能であることがわかっている．それでは，信頼できる学習システムを構築するためには，できるだけ多くのデータを学習アルゴリズムに対して流し込むだけで果たして十分なのであろうか．

本博士論文では，汎化理論とは趣を異にする剰余リスク転移の学習理論を構築し，信頼性の高い学習アルゴリズムの設計や学習のメカニズムの解明においてリスク関数の設計が重要であるという見方を示す．剰余リスク転移の重要性は，二十一世紀初頭に行われてきた代理リスクに関する一連の研究を通して認知されはじめた．**代理リスク**は一般的な機械学習手法が最適化する学習基準であり，最終的な評価指標である**目的リスク**と異なっていたとしても，損失が適切に設計されていれば両者には単調な関係が存在することが明らかにされた．この単調性によって剰余リスク転移，すなわち**代理リスクの剰余量の最小化によって目的リスクの剰余量の最小化が実現されるという関係性**を得ることができる．結果的に，一般的に離散構造を持つが故に最適化が容易ではない目的リスクを用いる代わりに，最適化しやすい代理リスクを用いることが正当化される．

これまでの研究が主に二値分類や多値分類といった古典的な学習問題の解析を対象としてきた一方で，近年の学習問題はより複雑な構造を持つことが多い．本博士論文では，次の二つの観点から剰余リスク転移の理論を古典的な設定から拡張する．第一に，剰余リスク転移を予測に対するより多様な制約や性質と関連付ける．この関連付けは，従来理論が予測の正しさにのみ着目してきた点と大きく異なる．第二に，一見異なる二つの学習問題の間に関連性を見出すために剰余リスク転移を利用する．従来は目的リスクが最終的な対象であり代理リスクはあくまで補助的な量とみなされてきたが，ここでは二つの学習問題の困難性を比較するのに剰余リスク転移が利用できることを確認する．まとめると，剰余リスク転移によってより多種多様な制約を含む多くの学習問題が扱えるようになり，さらに学習問題の困難性に基づいて問題同士の関連性を明らかにすることができるようになる．したがって，より広いクラスの学習問題の構造や学習問題からどのような知識を獲得できるかが明らかになり，我々の最終目的に対して必要最小限な学習問題を設計する指針を得ることができる．補足すると，ここで考察している剰余リスク転移の理論は汎化理論とは

概念的に独立しており，両者を統合することが可能である．以下，本博士論文の構成の詳細を示す．

第一章では，機械学習と統計的学習理論の歴史を振り返り，本学位論文の貢献をまとめる．

第二章では，教師付き学習の背景知識とその理論をまとめる．教師付き分類を定式化した後，汎化理論と代理誤差の古典的な解析方法に触れる．また，分類適合的損失や proper loss といった関連する概念を最近の関連研究とともに紹介する．

第三章では，F 値や Jaccard 指標といった複雑な評価指標で分類器を評価する際に適切な代理目的関数の設計を行う．これらの複雑な評価指標は，近年情報検索やセマンティック・セグメンテーションにおいてクラス不均衡に対処するためによく用いられているが，線形分数型指標と呼ばれるクラスに属しており，古典的な代理損失を適用しにくい分解不可能性という性質を持つ．ここでは目的となる評価指標の代理となる目的関数を適切に設計することで，代理目的関数を最適化した際に評価指標が最適化されるための十分条件を導出する．ベンチマークデータを用いた数値実験では，特にデータ数が小さい状況下において，クラス事後確率に基づくプラグイン分類器と比較して設計した代理目的関数が良い性能を発揮することが確認できた．この結果は提案手法の有用性を検証するだけでなく，プラグイン分類器に対する優位性を知る上でも重要な手がかりとなる．

第四章では，敵対的攻撃に対して頑健な分類，すなわち敵対者がテスト入力に対して加える摂動に影響されにくい分類器の学習を考える．この学習問題は，摂動の大きさに対する制約下での最悪の分類誤差値を目的誤差とする，ミニマックス最適化として最適化されることが多い．従来の研究ではこの最適化を緩和することに主眼が置かれてきたが，目的リスクの観点から緩和問題が正当化できるかどうかは知られていない．そのため，剰余リスク転移解析を用いて頑健な分類誤差に対して適合している代理損失を調べ，結果として線形モデルの仮定の下では真に頑健な解が得られる凸な代理損失が存在しないことを明らかにしたと同時に，非凸損失を設計するために有用な指針が得られた．この結果は，敵対者の導入によって凸代理損失が適合的にならなくなるという事実に意外性があり，また適合的代理損失の理論が予測の正しさ以外に予測が備えるべき性質に対しても適用可能であるということが明らかになった点が重要である．

第五章では，類似度学習と分類の間の関係性を議論する．類似度学習は二つのデータ間の関係性を予測することでデータの有用な表現を得るための枠組みであり，距離学習や対比学習といった前処理問題を包摂する．学習した表現を用いて構成した分類器によって分類性能が向上されることが期待されているが，これまで理論的な背景はほとんど知られてこなかった．類似度学習がどのように分類性能を向上させるかを解明するため，我々は類似度学習の特定の定式化が分類誤差と密接に関係していることを示した．剰余リスク転移の観点から，この関係性によって類似度予測リスクの剰余量を最小化することで分類剰余リスクが最小化できることが説明できる．その結果，類似度学習は内部的には二値分類境界を学習しているということがわかった．剰余リスク転移の考え方を用いると，例えば二値分類と類似度学習のように，二つの異なる学習問題をそれぞれ目的問題と代理問題とみなすことができ，学習問題間の関係性を調べたり，ある問題が別の問題に帰着することができるかどうかを特徴づけることができるようになる可能性が秘められている．

第六章では，本学位論文の結論とこれからの課題を述べる．本論文では剰余リスク転移の見方から学習理論の新しい展望を提示し，学習器が期待される性質を満たしているかどうかを解析したり，学習問題を相互に帰着できるようになった．

In memory of my grandmother

Acknowledgements

First of all, I am deeply grateful to Professor Masashi Sugiyama for advising me over the past five years. In 2015, when I was a bachelor student, he kindly invited me into the fantastic research community of machine learning and provided me numerous opportunities, including Machine Learning Summer School 2015 Kyoto, a short-term visit at CMU in 2016, France-Japan Machine Learning Workshop 2017, and Japan-Israel Machine Learning Meeting (JIML) 2018. These experiences shocked me a lot—what many interesting topics and studies there are outside the world around me! In addition, I could intensely focus on my research and travel to many places to have discussions with researchers freely under the current environment. All of these experiences and environments owe to him.

I would like to express my gratitude to the dissertation committee members, Professor Seiya Imoto, Professor Yusuke Miyao, Professor Shinya Takamaeda, Professor Shinpei Kato, and Professor Hidetoshi Shimodaira, for providing valuable feedback on my dissertation.

My fruitful graduate school life must be significantly owing to Nontawat, who had a seat next to me for more than five years. We had a lot of common research interests in loss functions, which could be partly owing to our reading group held in the summer of 2018. I don't know how many hours we had discussions since then. Not only did we learn a lot of concepts in the field of loss functions together but also I was inspired by Nontawat's fresh research ideas. Because we had an unspoken consensus that we don't mind having research discussions anytime, I could polish my immature ideas and listen to his unreserved opinions. I hope I have a chance to collaborate with Nontawat in near future!

During my three-year Ph.D. course, I spent four months at University of Michigan, Ann Arbor (A2), in late 2019. This stay was highly impressive for me because I have never spent so long time in schools outside Japan until then. I could have this opportunity because of Clay's passionate support. He spent a large amount of time on research discussion with me, and we eventually arrived at a very interesting insight into adversarial machine learning. The ideas obtained through discussions with Clay have been the basis of my research philosophy since then. I would like to thank many other people who supported my life in A2, including Yutong for taking me around A2 and discussing many ideas with me; Alex for having a life together in small attic rooms of a shared house; Saki for kindly bringing me into a Japanese community in A2; Takuma, Keiichi, Hideaki, and Kaito for enjoying many beers and whiskeys! A2 was a perfect place for concentrating on research and drinking beers. I should acknowledge Ashley's and HopCat for letting me enjoy many beers.

My research activity in the Ph.D. course be fulfilled without brilliant collaborators. Apart from Clay and my supervisor Professor Masashi Sugiyama, I had many inspiring discussions with Takuya, Liyuan, and Professor Issei Sato in the similarity learning project. In particular, Takuya is so dedicated a person that his analysis and experiments were always thorough. I learned many things from his frank opinions. I also had a chance to work on contrastive learning with Yoshihiro and Kento in the last year of my life in the graduate school. This project was pretty amazing. At first, Yoshihiro and I were asking Kento about a recent theoretical study on contrastive learning and its validity, and we three had a strong belief that there should be a better theory, which consequently led to a new research project! In this project, every member contributed significantly from the theory and experiments, and hence this research could not be accomplished without any of us. I believe that this project was one of the ideal forms of collaborative research. Although I did not collaborate in my Ph.D. course, I had many professional comments from Professor Junya Honda. I would like to appreciate him for much advice based on his expertise.

Tokyo is an exciting city from the perspectives of both culture and study. Many friends come and go from all over the world. Hence, I could also have many friends at the University of Tokyo, which was precious memory for me. Although I would like to appreciate any individuals I met in our lab, here I particularly express my appreciation to Takeshi for having inspired me a lot by fantastic research ideas and thoughts; to Ikko for supporting our lab a lot, sharing a hotel room in NeurIPS2017, and enjoying beers together; to Henry for being a nice colleague with me for many years; to Jongyeong for playing boardgames and mahjong; to Masahiro Fujisawa for giving me new insights into many probabilistic divergences; to Taira for studying together and having dinner several times in Kyoto; to Yivan for being a good lab mate to talk research and daily stuff; to Shintaro for working with me on optimal transport, which was a new field to me; to Yuko for taking me out to dinner many times! I also would like to mention En Cafe at Kuramae (I could not finish my AISTATS2021 paper without this comfortable cafe) and Hitch x Kakeru (a beer bar in Ueno and make me feel relaxed from work stress).

One of the great advantages of academic life is that we have many chances to travel abroad to have discussions with wonderful researchers across the world, but this is impossible without the help of others. In my case, I could luckily have many people's support; Mathurin took me around, let me stay in his flat, and organized my talk at Inria Saclay when I visited Paris in the summer of 2019; Pascal and Benjamin organized my stay in Inria Lille and had dinner together in the summer of 2019; Sanmi kindly discussed with me and introduced me to students in UIUC when I dropped by at Urbana-Champaign in early 2020. I would like to appreciate all of these friends. I will definitively host researchers from overseas sincerely if I have a chance.

Before finishing, I need to give a special thanks to Motoya and Ryuichiro, who I met and became friends in Tel Aviv in late 2018. For some reason, we get along pretty well and discuss a lot of topics including not only research but also politics, society, culture, and philosophy. I was significantly influenced by them in my view of philosophy. Surprisingly, we still actively communicate with each other, although Motoya and Ryuichiro are in Seattle and Genova, respectively, as of December 2021.

My graduate research was financially supported by the Japan Society for the Promotion of Science (JSPS) Fellowship Program (the grant number 19J21094) and the Japan Science and Technology Agency (JST) ACT-I Program (the grant number JPMJPR18UI). Without these financial supports, I could not have the above chances to have research collaboration, overseas visits, and preparing research environments.

Last but not least, I would like to express my biggest gratitude to my parents for continuously supporting me. Without their supports, I could not even have an opportunity to study in Tokyo.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> *For a large class of cases—though not for all—in which we employ the word meaning it can be explained thus: the meaning of a word is its use in the language.*
>
> — Ludwig Wittgenstein, *Philosophical Investigations*

From antiquity, we human beings have been continuously seeking mechanisms of perception, reasoning, learning, comprehension, and communication, which seem to constitute the basis of our intelligence. Modern philosophers and scientists have used a powerful tool, reductionism, in an attempt to understand intelligent systems as automata [Descartes, 1985]. Correspondingly, a part of human intelligence has been implemented on machines to imitate human behaviors, and modern artificial intelligence has achieved astonishing successes and even a performance superior to that of human intelligence in certain areas [Krizhevsky et al., 2012, Silver et al., 2016] owing to the rapid development of electronic computer systems and architectures. At the same time, computer scientists face numerous obstacles that current artificial intelligence has yet to resolve [Turing, 1950]: How intelligent can computer systems become reliable? What is the fundamental limit on the amount of knowledge that artificial intelligence can acquire? In this dissertation, we aim to provide a clue to these questions from the perspective of machine learning by focusing on the gap between learning and evaluation criteria.

1.1 Machine Learning and its History

Learning is one of the foundational elements that seem to support human intelligence. We repeatedly encounter many new concepts, and to make the world perceivable by reducing an infinite number of concepts to reasonable finite number of categories, we must try to determine what is common and what is different within such concepts. This is the origin of classification in the field of philosophy, and this idea can even be found in Aristotle's metaphysics [Cohen and Reeve, 2020]. Human beings are constantly exposed to a large number of observations and attempt to classify them into a few classes by grasping their underlying law. After the emergence of computers, researchers believed that this procedure of classification could be implemented on computers and automated. Thus, research on artificial intelligence and pattern recognition has begun [McCarthy et al., 1955].

Research into artificial intelligence was partly initiated by psychologists, who aimed to develop plausible models of biological neurons, including the McCulloch-Pitts model [McCulloch and Pitts, 1943] (see Figure 1.1). These neuro-inspired models have been conventionally called neural networks. Stimulated by neural network research, Frank Rosenblatt introduced one of the oldest formulations of

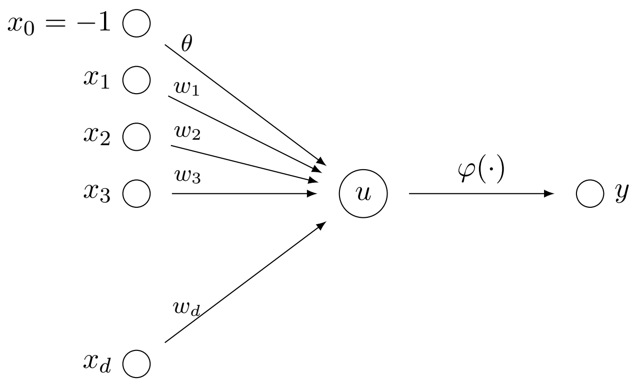


Figure 1.1: Illustration of the McCulloch-Pitts model. For d inputs $\{x_1, \ldots, x_d\}$, the hidden unit returns $u = \sum_{i=1}^{d} w_i x_i - \theta$, which is activated by the Heaviside step function $\varphi(u) = \mathbb{1}_{\{u>0\}}$. This model simplifies the activation mechanism of biological neurons. The hidden unit u is also called the membrane potential.

pattern recognition, called the perceptron [Rosenblatt, 1957]. The perceptron algorithm models a classifier by a linear function and exposes the model to binary labeled examples sequentially, trying to predict labels of future patterns. The model is updated by a local correction of the coefficients when the classifier misclassifies an input. If a given sequence of data is linearly separable, the perceptron is guaranteed to converge with a finite number of updates [Novikoff, 1963]. The perceptron was successful in providing a novel formulation to learn from examples, and the framework has had a significant influence in the subsequent research on pattern recognition [Valiant, 1984, Kearns et al., 1994, Blum et al., 1998, Kalai and Sastry, 2009]. One of the drawbacks of the perceptron algorithm is that it cannot handle linearly inseparable data points. Later, the perceptron was extended to the multi-layer perceptron [Rosenblatt, 1961], which consists of a hidden layer and a nonlinear activation function. These newly introduced components have made artificial neural networks nonlinear and capable of learning more complex target functions, and it has even been proven that a Turing machine can be simulated on a multi-layer perceptron [Minsky and Papert, 1972]. However, Minsky and Papert [1972] showed that some functions including the XOR logical function cannot be represented by a single-layer perceptron (see Figure 1.2), which was misunderstood to be a limitation of the perceptron algorithm, despite a nonlinear perceptron being able to learn such functions. This eventually led to a hiatus in neural network research until the 1980s.[1] Since the 1980s, however, continuous efforts have been made by researchers in this area. Cybenko [1989], Funahashi [1989], and Hornik et al. [1989] independently showed that a multi-layer perceptron is a universal function approximator of every continuous function. In addition, LeCun et al. [1989] used backpropagation [Rumelhart et al., 1986] to successfully train convolutional neural networks (as illustrated in Figure 1.4) applied to digit recognition. A sophisticated activation function called a rectified

[1] Whereas Minsky and Papert's book "Perceptrons" has a large impact on suspending neural network research, there has been a long dispute between connectionism and classicism in psychology and cognitive science: The former models semantics by units and connections among them, leading to neural network research, and the latter does by atomic logical and syntactic expressions. Fodor and Pylyshyn [1988] criticized connectionism for it presupposes that all semantics are represented by units nonhierarchically, which disagrees to that logical reasoning entails combinatorial semantics. Fodor and Pylyshyn supposed that connectionism is useful as an implementation principle but not as a cognitive model. Though neural networks have been improved dramatically, it still remains an open question whether connectionism aligns with our cognition or not [Lake and Baroni, 2018, Geiger et al., 2019, Yanaka et al., 2020].

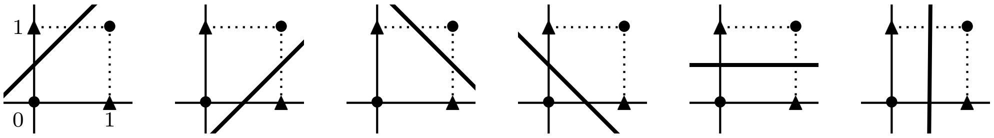


Figure 1.2: XOR data points and inability of linear separation. ● indicates an output value of 0 and ▲ indicates an output value of 1. As can be seen, no linear functions can separate all data points.

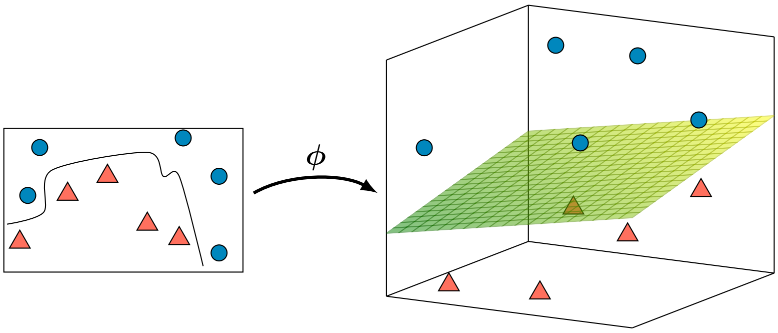
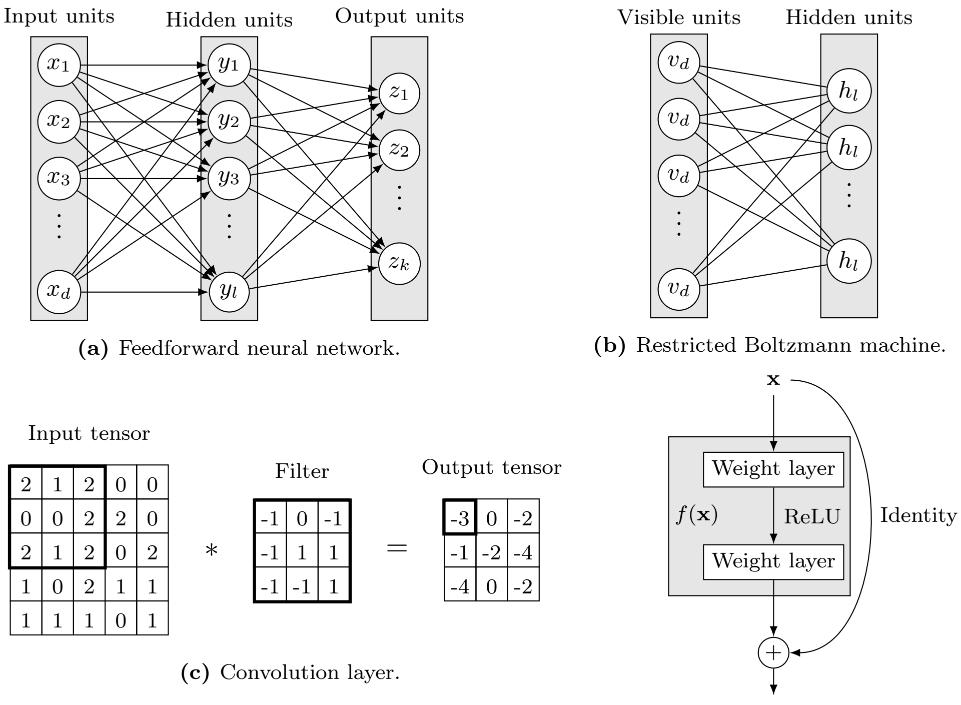


Figure 1.3: Illustration of support vector machines. Given a set of data points that is not linearly separable in the original feature space (left), feature maps ϕ based on a kernel function lift the features into a high-dimensional space (right), where the data points are linearly separable. Importantly, the kernel trick enables us to cheaply compute the inner product in a lifted space.

linear unit (ReLU) was used to overcome the difficulty of neural network training [Fukushima and Miyake, 1982]. More variations of neural networks have been proposed, including Hopfield networks [Hopfield, 1982], restricted Boltzmann machines [Smolensky, 1986] (illustrated in Figure 1.4), autoencoders [Hinton and Zemel, 1994], and long short-term memory [Hochreiter and Schmidhuber, 1997].

During the 1990s, support vector machines (SVM) emerged and rapidly surpassed neural networks of those days in many application domains [Cortes and Vapnik, 1995]. SVMs introduce nonlinearity to classifiers by using nonlinear kernel functions as feature maps, such as the polynomial, Gaussian, and Laplacian kernels [Shawe-Taylor et al., 2004]. Their successes can mainly be attributed to the following two factors. The first is a kernel trick, which enables us to conduct arithmetic operations on high-dimensional vectors implicitly without manipulating the lifted feature vectors by encapsulating all data manipulations into low-cost inner product operations. See Figure 1.3 for an illustration. This trick largely increases the training speed without sacrificing the representation performance. The second is various types of kernel functions that can be applied to complex data structures. For example, graph kernels [Vishwanathan et al., 2010] operates on graphs; string kernels [Lodhi et al., 2002], on sequences and sentences; multi-instance kernels [Gärtner et al., 2002], on sets; tree kernels [Collins and Duffy, 2001], on tree structures; and Fisher kernels [Jaakkola et al., 1999], on probability distributions. For these reasons, SVMs can be applied to various data structures in a scalable manner, and their application led to machine learning successes during the 2000s.

During the late 2000s, connectionists, or neural network researchers, continuously developed neural network architectures and their training methods. Although the idea of deep learning, i.e., the stacking of multiple hidden layers in a feedforward neural network, appeared during the 1980s [Dechter, 1986], the training of deep neural networks has yet to be successful until the late 2000s. Hinton et al. [2006] proposed the use of deep belief networks and showed that pre-training based on autoencoders and supervised fine-tuning through backprop-



(a) Feedforward neural network.

(b) Restricted Boltzmann machine.

(c) Convolution layer.

(d) Residual block.

Figure 1.4: Various neural networks and their components. **(a)** All information moves in only one direction from the input layer to the output layer. **(b)** In contrast to feedforward networks, the units form a bidirectional bipartite graph. The network is trained by minimizing an energy function defined using the Ising model [Smolensky, 1986]. **(c)** A convolution layer consists of filter parameters, which are applied to the input tensor. We slide the filter across the width and hight of the input tensor. **(d)** For input $\mathbf{x}$, the block returns $f(\mathbf{x}) + \mathbf{x}$, where f consists of two weight layers with the ReLU activation in the middle. The weight layer consists of convolution and batch normalization [Ioffe and Szegedy, 2015] layers. The operation $+\mathbf{x}$ is called a skip connection, which prevents the vanishing gradient problem [He et al., 2016].

agation [Rumelhart et al., 1986] can result in a successful deep learning operation. Further, Krizhevsky et al. [2012] proposed a deep convolutional network called AlexNet and won the first place at the computer vision competition, ImageNet Large Scale Vision Recognition Challenge (ILSVRC) 2012, by a large margin [Russakovsky et al., 2013], stimulating further research into neural networks. Many novel architectures have been proposed including generative adversarial networks [Goodfellow et al., 2014], variational autoencoders [Kingma and Welling, 2014], normalizing flows [Rezende and Mohamed, 2015], ResNet [He et al., 2016] (illustrated in Figure 1.4), and Transformer [Vaswani et al., 2017], to mention a few. This deep learning revolution has achieved significant successes in the field, such as game AI [Silver et al., 2016], predictions of protein structures [Jumper et al., 2021], and autonomous driving [Grigorescu et al., 2020].

As can be seen, machine learning has been developed and supported through many aspects, including novel model architectures, optimization, theoretical understanding, training techniques, and large-scaled datasets, which when combined constitute the current success of machine learning and deep learning.

1.2 Paradigms of Machine Learning

The machine learning framework can be largely classified into three categories: supervised learning, unsupervised learning, and reinforcement learning. We briefly look at an overview of the three frameworks and then dive into the details of supervised learning, which is a central framework that we focus on in this dissertation.

1.2.1 Supervised, Unsupervised, and Reinforcement Learning

Supervised learning [Vapnik, 1998] considers a setup where a learner is exposed to a set of an input $\mathbf{x}$ associated with an output y and asked to elicit the underlying relationship between $\mathbf{x}$ and y.[2] This is a type of inductive reasoning. An input $\mathbf{x}$ is typically represented by a real-valued vector representation, which is called a feature. This learning framework contains many problems including regression (y is real-valued), classification (y is categorical), and structured prediction (y has certain structures such as graphs). Some studies have been motivated by the fact that supervised learning requires a high cost in terms of label acquisition, leading to research on semi-supervised learning [Chapelle et al., 2006], weakly-supervised learning [Zhou, 2018], and self-supervised learning [Jaiswal et al., 2021].[3] In these newly proposed setups, a learner cannot necessarily have direct access to the complete form of label y. We provide more details on supervised learning in the next subsection.

Unsupervised learning attempts to find useful structures in data from inputs $\mathbf{x}$ only. Although there are several other problems, the most important problem in unsupervised learning is density estimation, which aims at estimating the underlying probability density function of the data. The following problems are a few example problems.

- Dimensionality reduction: The transformation of data into a low-dimensional space while retaining important characteristics of the original data [van der Maaten et al., 2009].
- Clustering: The grouping of a set of data into several clusters according to their similarity [Xu and Wunsch, 2005].
- Independent component analysis: Recovery of the original signals given multiple mixed signals [Hyvärinen, 2013].
- Anomaly detection: Identification and segregation of some rare events from the majority of data [Chandola et al., 2009].

Reinforcement learning [Sutton and Barto, 2018] is a different framework, in which an agent can interact with the environment by taking actions and receiving rewards accordingly. The environment is typically modeled by a Markov decision process, and reinforcement learning aims to train an agent so that it can receive as high a reward as possible. Whereas reinforcement learning can be solved through dynamic programming if the dynamics of the environment are known, sophisticated learning methods such as Monte Carlo simulations and temporal difference learning are needed under typical situations in which the dynamics are unknown. To deal with a case in which even the reward function is not known, inverse reinforcement learning has been proposed to estimate the reward function from

[2] In this dissertation, a *learner* indicates a learning machine that elicits useful knowledge and the underlying law from a finite number of observations. Our algorithmic procedure for exposing a learner to observations is called *training*. A *model* is occasionally used in place of a learner in formal contexts. In reinforcement learning, a learner is often called an agent.

[3] It is difficult to categorize self-supervised learning into either supervised or unsupervised learning. Some studies have attempted to solve supervised learning problems by automatically generating supervision without label information, whereas others have attempted to acquire generic data representations from unsupervised data [Jaiswal et al., 2021].

expert behaviors [Pomerleau, 1991, Ng et al., 2000]. Occasionally, bandit problems [Cesa-Bianchi and Lugosi, 2006], i.e., problems of finding an optimal strategy for an agent with no state transition to sequentially take an action that maximizes the expected reward, are considered to be a part of reinforcement learning.

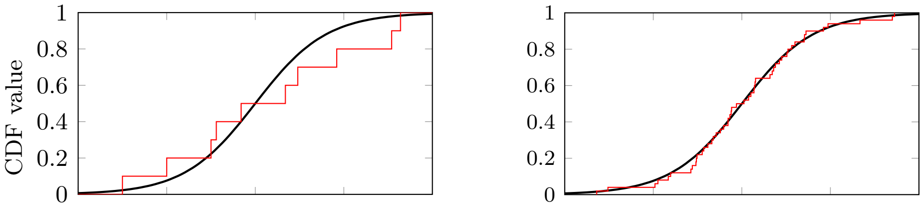
1.2.2 Supervised Learning and Learning Models

As we reviewed, supervised learning asks a learner to elicit an underlying law between an input $\mathbf{x}$ and an output y from a large number of labeled training data. The trained learner will predict the output for the future test data. Many machine learning algorithms such as a perceptron, SVMs, and deep learning are categorized as supervised learning. Because a learner only has access to a finite number of observations, supervised learning is essentially a type of inductive inference. In contrast to deductive inference, inductive inference cannot be justified without any assumptions.[4] Hence, we commonly adopt an assumption that both training and test data are distributed from the identical probability distribution [Vapnik, 1998]. The ability of the learner to predict outcomes of future data well is called generalization, which is a significantly important concept in the machine learning community.

Given a formal concept of the underlying probability distribution, we are interested in *learnability*, namely, whether a good learner can be trained under a specific learning setup, and what algorithm can successfully let a learner acquire the expected predictive performance. There are mainly two streams of research tackling this question: *computational learning theory* and *statistical learning theory*.

The community of computational learning theory has actively discussed the definition of learnability, and several definitions of learnability have been proposed. One of the most foundational formulations is *probably approximately correct* (PAC) learning proposed in Valiant [1984]. Within the framework of PAC learning, we are given an error tolerance ε and probability parameter δ. We then draw n data points, where n is a polynomial in ε and δ, and ask whether we can find a hypothesis in polynomial time that agrees to an underlying law with an error rate of less than ε (approximately correct), with a probability of at least $1 - \delta$ (probably). This formulation was innovative in that machine learning and computational complexity theory are connected, and it thereby became possible to discuss the learnability of certain target functions from the perspective of computational complexity. Although PAC learning has a limitation that labels are generated through a deterministic target function, this has been relaxed by several studies: Kearns et al. [1994] proposed agnostic learning, where a target function does not necessarily belong to the space in which we seek a hypothesis. Eventually, it was discovered that even simple target functions such as halfspaces are not agnostically learnable within a polynomial time. Blum et al. [1998] introduced the random classification noise model, where labels are allowed to be flipped with a fixed probability after being generated by a deterministic target function. Kearns [1988] discussed weak learnability, asking whether we can obtain a good classifier from a weak learner that only slightly correlates to the underlying law. This research question led to boosting [Schapire, 1990], an algorithm used to aggregate multiple weak learners and obtain a good classifier. More noise models have also been proposed, including the Massart noise model [Massart and

[4] Historically, David Hume discussed the validity of reasoning in the modern era and formulated the uniformity principle as a minimally required assumption to justify induction [Henderson, 2020].



(a) Sample size $n = 10$.

(b) Sample size $n = 50$.

Figure 1.5: Illustration of the Glivenko-Cantelli theorem. The cumulative distribution function (CDF) is a logistic distribution in both cases, in which the numbers of observations differ. As the number of the observations increases, the empirical distribution function (red lines) becomes closer to the cumulative distribution function (black lines).

Nédélec, 2006]. Overall, computational learning theory has attempted to answer learnability, i.e., what kind of target functions can be efficiently learned from the perspective of both computational and sample complexities.

Statistical learning theory [Vapnik, 1998] is devoted more to sharply analyzing how many samples are needed for generalization, whereas computational learning theory is mostly interested in learnability. In statistical learning theory, machine learning problems are formulated by a cost function defined over a hypothesis space, so-called the *risk*, which measures the discrepancy between the predictions of a learner and actual observations. By minimizing the risk, we expect to obtain a good hypothesis. Although we are interested in the underlying law, the risk function is approximated with finite observations that we have access to; hence, there is a gap between the expected value of the risk and the risk over the observations, which is called the empirical risk. The central topic of statistical learning theory is to fill in this gap and characterize how fast the empirical risk converges to the expected risk. The learning approach used to minimize the empirical risk is called *empirical risk minimization* (ERM). When we view the difference between the expected and empirical risks as a stochastic process, this is often referred to as an empirical process [Billingsley, 2008]. The history of empirical processes dates back to Glivenko [1933] and Cantelli [1933], who independently proved the Glivenko-Cantelli theorem, or the so-called the Fundamental Theorem of Statistics, stating that an empirical distribution function uniformly converges to a cumulative distribution function (as illustrated in Figure 1.5). Its convergence rate was later quantified by Dvoretzky et al. [1956]. Uniform convergence plays a key role in learning theory to handle the dependency of an empirical minimizer on the training data. Dudley [1967] introduced the metric entropy to characterize the complexity of the index set of an empirical process and showed that the empirical process can be bounded by the metric entropy (see Figure 1.6 for an illustration). The theory of empirical processes was specialized for supervised learning in Vapnik and Chervonenkis [1971], which introduced the VC-dimension (Vapnik-Chervonenkis-dimension) to characterize the complexity of the hypothesis space and showed that the VC-dimension can be used to bound the generalization error, i.e., the gap between empirical and expected risks. A distribution-dependent complexity measure called the Rademacher complexity was recently introduced to better characterize the convergence rate [Bartlett and Mendelson, 2002]. Thus, statistical learning theory seeks better characterizations of the generalization error from the viewpoint of a uniform convergence and its rate. Note that several alternative approaches to the ERM have been proposed, including an adversarial prediction [Asif et al., 2015], invariant risk minimization [Arjovsky et al., 2019], and learning using statistical invariants [Vapnik and Izmailov, 2019]; yet, the

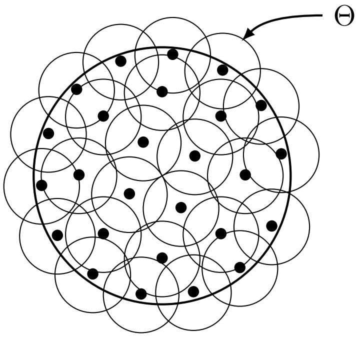


Figure 1.6: Illustration of the covering number of a set Θ, which is conceptually the smallest number of balls needed to fully cover Θ. The metric entropy is the logarithm of the covering number [van de Geer, 2000]. These concepts characterize the "volume" of a set with the (possibly) infinite cardinality by a finite number of representative points.

ERM has remained the most common approach in statistical machine learning.

Although both computational and statistical learning theories have thus far been successful at formulating learning problems and understanding algorithms, researchers have raised several issues and attempted to overcome them. Herein, we briefly summarize two issues. The first issue is a problem of a distributional assumption. The traditional learning theory assumes that training and test data follow the same underlying law, as we mentioned. Although this is a key assumption in justifying an inductive inference, we often encounter data that are collected in different environments. Hence, mitigating an environmental shift is an important problem in machine learning. This problem is called transfer learning [Pan and Yang, 2009], which subsumes many subproblems and algorithms such as a covariate shift adaptation [Shimodaira, 2000], domain adaptation [Ben-David et al., 2007, Redko et al., 2020], continual learning [Delange et al., 2021], and multi-task learning [Caruana, 1997], to mention a few. The second issue is the sharpness of learning theory analyses. Although the Rademacher complexity is believed to provides a distribution-dependent generalization bound that is sharper than the traditional VC generalization bound, complex function classes such as deep neural networks have been observed to have prohibitively large Rademacher complexity values, and some researchers are therefore skeptical of learning theory based on the uniform convergence [Nagarajan and Kolter, 2019, Zhang et al., 2021]. Recently, PAC-Bayes analysis [McAllester, 1999, Langford and Shawe-Taylor, 2003] has received attention and has been applied to the generalization analysis of deep learning [Neyshabur et al., 2017] owing to its empirical sharpness.

Throughout this dissertation, we primarily focus on the framework and formulation of machine learning from the viewpoint of statistical learning theory.

1.3 Limitation of Current Learning Theory

Despite the success of learning theory in formulating supervised learning and providing a unified view of learning algorithms, the current learning theory is not sufficiently capable of capturing the behavior of machine learning algorithms. As mentioned in the last section, this is partly because the current learning theory does not fully address the problem of distribution shifts or the sharpness issue. However, there is another factor creating a gap between the perspective of learning theory and practice: the gap between learning criteria and the evaluation metrics.

1.3.1 Gap between Excess Risks

Let us consider supervised binary classification as an initial step. From the perspective of statistical learning theory, our goal is to train a binary classifier achieving as low a misclassification error as possible for a fixed underlying probability distribution. Mathematically, this problem is formulated as a minimization problem of the classification risk over a hypothesis space [Vapnik, 1998]. However, it is known that the direct minimization of the classification risk is infeasible even if a simple hypothesis space such as a linear model is adopted.[5] By contrast, the current practical machine learning algorithms bypass this computational hurdle by introducing some approximations. Through the lens of statistical learning theory, SVMs can be regarded as minimizing the risk defined by the hinge loss [Cortes and Vapnik, 1995], whereas logistic regression and its boosting extension, LogitBoost [Friedman et al., 2000], can be regarded as minimizing the risk defined by the logistic loss. These classification methods do not explicitly minimize the classification risk in their formulations but achieve a good performance in practice. During the early 2000s, Lin [2004], Zhang [2004a], and Bartlett et al. [2006] have attempted to fill in this gap between the disagreement of the objective functions and the practical performance from a theoretical perspective. Noting that the objective functions optimized by the popular machine learning algorithms differ from the classification risk, let us label them surrogate risk.[6] The authors then ask the following important research question.

Research question (classification-calibration).

When we minimize the surrogate risk (learning criterion), can we obtain a good solution in terms of the classification risk (evaluation criterion)?

Bartlett et al. [2006] answered this question systematically by providing sufficient conditions on a surrogate loss, which justifies the use of common loss functions such as the hinge loss and logistic loss in terms of minimization of the classification risk. The loss functions that satisfy the sufficient conditions are called *classification-calibrated*.[7] Technically, an *excess* of the classification risk, namely, the difference between the classification risk of a learned hypothesis and the optimal risk, is shown to have a monotonic relationship with an excess of the surrogate risk if a surrogate loss is classification-calibrated. We refer to this risk bound as an *excess risk transfer bound*.

As we can see in the case of classification-calibrated losses, there often exists a gap between the learning criterion optimized through learning algorithms and

[5] Such results were obtained mainly in the field of computational learning theory. For example, Kearns et al. [1994] showed that the agnostic learning of halfspaces is NP-hard, whereas Feldman et al. [2012] showed that the agnostic learning of monomials is NP-hard (under the unique games conjecture). Because the common setup of binary classification in statistical learning theory is equivalent to agnostic learning [Shalev-Shwartz and Ben-David, 2014], it is intractable to minimize the classification risk with a common hypothesis space.

[6] In this dissertation, we consistently use the term a *loss (function)* to indicate a cost function defined over a single input-output pair, whereas a *risk (functional)* is used as the expected loss function over the underlying probability distribution. A risk functional is defined over a hypothesis space.

[7] In addition to classification-calibrated losses, the following terminologies indicate the same concept: infinite-sample consistent losses [Zhang, 2004a], admissible losses [Steinwart, 2005], and Fisher consistent losses [Lin, 2004]. Although the word "consistency" is often used for the same purpose, it must be clearly distinguished from the concept of consistent estimators in classical statistics.

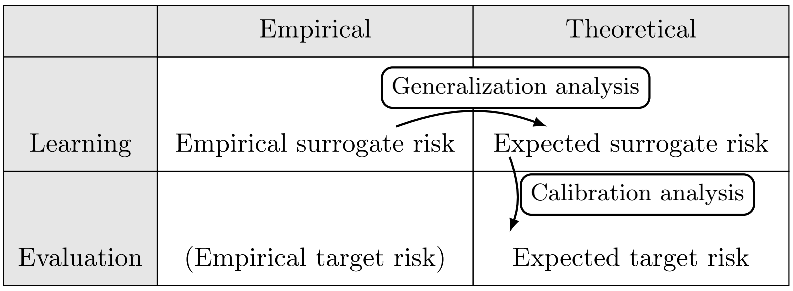


Figure 1.7: Relationship between risk functionals. Generalization analysis concerns the gap between empirical and expected risks, whereas calibration analysis concerns the gap between surrogate and target risks. Overall, we are interested in whether minimizing an empirical surrogate risk may lead to minimizing an expected target risk. We use the terminology surrogate/target risks and learning/evaluation criteria interchangeably.

the evaluation criterion that we are ultimately interested in. Importantly, this gap cannot be filled by the aforementioned learning theory regarding the generalization error because a generalization analysis only concerns the gap between empirical and expected risks. Unless we are aware of the gap between learning and evaluation criteria, the trained learners will not behave as we expect. In contrast to generalization analysis, learning theory aiming to investigate the gap between learning and evaluation criteria is called calibration analysis in this dissertation. The conceptually orthogonal relationship between generalization analysis and calibration analysis is illustrated in Figure 1.7.

Since the seminal studies by Lin [2004], Zhang [2004a], and Bartlett et al. [2006], calibration analysis has gradually received attention from the community of learning theory. Follow-up studies have provided an analysis on a variety of machine learning problems, including multi-class classification [Zhang, 2004b, Tewari and Bartlett, 2007, Long and Servedio, 2013, Ávila Pires and Szepesvári, 2016], multi-label classification [Gao and Zhou, 2011, Zhang et al., 2020], partial label classification [Cid-Sueiro et al., 2014, Cabannnes et al., 2020], cost-sensitive classification [Scott, 2011, 2012], bipartite ranking [Dembczynski et al., 2012, Gao and Zhou, 2015], and listwise ranking [Ravikumar et al., 2011], to mention a few. This line of work has posed the following question.

Research question (surrogate consistency).

Given a learning problem and its evaluation criterion, what is an appropriate surrogate loss (learning criterion) leading to a good prediction performance in terms of the evaluation criterion?

In contrast to binary classification, these learning problems are often much more complicated,[8] and the design of appropriate surrogate measures is not trivial. Thus, calibration analysis has been an important tool for understanding the learning mechanism and designing good learning algorithms.

Although calibration analysis has been successful at drawing the connection between surrogate and target risks, the existing research has to date mainly fo-

[8] For example, it is known that smoothness plays a key role in multi-class classification [Tewari and Bartlett, 2007]. That is to say, smooth loss functions such as the (variants of binary) logistic loss are classification-calibrated, whereas non-smooth loss functions such as the (variants of binary) hinge loss are not. This result differs from binary classification, where both the logistic and hinge losses are classification-calibrated [Bartlett et al., 2006].

cused on analyzing the prediction performance, such as the accuracy of classification. As machine learning has been applied more in real applications, it becomes insufficient to only guarantee the prediction performance. Rather, other perspectives such as security [Kurakin et al., 2016], fairness [Mehrabi et al., 2021], and sensible confidence [Guo et al., 2017] have become increasingly crucial. These new perspectives are indispensable for reliable machine learning and real-world deployment. In addition, we still do not know much about what knowledge a learner acquires through training in traditional machine learning, which merely seeks a better prediction performance. Specifically, we pose the following questions to the community of learning theory.

- Reliability: Can we verify whether a learning algorithm can output a reliable classifier?[9]

- Transferability: Can we know what a learner has essentially learned after solving a machine learning problem?

This dissertation provides answers to these questions based on the idea of excess risk transfer. Subsequently, we will see detailed discussions on these two questions.

1.3.2 Reliable Machine Learning Predictions

It was observed that even a predictor with a high prediction performance can often have reliability issues. Some examples are as follows.

- Vulnerability to adversarial attacks [Kurakin et al., 2016]: Although the recent deep neural networks have achieved an incredible prediction performance in many learning tasks, it has been reported that we can handcraft an imperceptible perturbation to a test input to manipulate the prediction result arbitrarily. Such a perturbation is called an adversarial attack, which has a large impact on the security of machine learning deployment.

- Fairness of predictions [Mehrabi et al., 2021]: In 2016, computer software used in the US courts to assess potential recidivism risk aroused criticism because it was claimed that the algorithm estimates the risk based on human race.[10] After this incident, the fairness of decisions made by computer software and artificial intelligence received significant attention. In modern society, assessing and verifying whether a predictor is intrinsically biased based on sensitive attributes such as gender and race is important.

- Out-of-distribution (OOD) generalization [Rahimian and Mehrotra, 2019, Shen et al., 2021]: Although classical machine learning methods assume that training and test data follow the same distribution, this assumption

[9] At a first glance, it may remind one of *formal methods* [Baier and Katoen, 2008], which verifies behaviors of computer systems based on mathematical specifications. Although formal methods have been actively applied to analyze machine learning models recently, the main focus is to verify if a *learned* model behaves as we expected [Seshia et al., 2016]. Seshia et al. [2016] discussed the necessity to develop a new design process called "correct-by-construction", which means that a machine learning model is refined interactively by repeated verification steps, yet this remains open. Our perspective is slightly different from the above ones: We are interested in whether a *learning algorithm* behaves as we expected. We believe that this meta approach may lead to a correct-by-construction design principle.

[10] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

rarely holds in real-world scenarios. Toward reliable machine learning deployment, the robustness to out-of-distribution data is necessary, and several approaches such as distributionally robust optimization and invariant learning were considered.

- Model calibration [Guo et al., 2017]: Although modern deep neural networks are often extremely confident in their predictions owing to their flexible architectures, overconfidence can often be an issue because humans cannot assess how much we can rely on the predictions. Model confidence is particularly important in making critical decisions such as medical diagnosis and political decisions. For this reason, model calibration, a procedure used to achieve a reasonable level of confidence of a predictor, has been actively studied.[11]

Because traditional learning theory was mainly designed to deal with the generalization error in terms of the classification risk, these reliability issues have not been sufficiently considered. Hence, we ask the following question.

Research question (reliability).

When we minimize a surrogate risk (learning criterion), can the learner successfully achieve a specific reliability property? Can we design an appropriate surrogate risk leading a learner to reliable predictions?

Although some existing research was devoted to designing algorithmic tricks to achieve a reliable prediction, the effectiveness was often observed only empirically, or the theoretical guarantee was given only on some relaxed measurements. For instance, Goodfellow et al. [2015] proposed a popular method for adversarial training of robust models; however, this method essentially minimizes a local approximation of the classification risk when considering an adversarial attacks [Shaham et al., 2018]. Within the context of a model calibration, a convenient trick called temperature scaling has been commonly used [Guo et al., 2017], although it is still unclear whether temperature scaling provably leads to good solutions in terms of the evaluation criteria of the model calibration, such as the expected calibration error (ECE). We will seek a theoretically guaranteed learning procedure based on calibration analysis and excess risk transfer. Our approach is similar to that of Narasimhan [2018], who designed a learning algorithm that can provably satisfy complex constraints such as fairness.[12]

1.3.3 Knowledge Transfer between Learning Problems

Once a learner is trained, it is used for predicting future outcomes. However, because machine learning problems are formulated as minimization problems of some cost functions, it is not straightforward to understand what knowledge and

[11] Be aware that model calibration is a completely different notion from calibrated loss functions, although they share the same terminology.

[12] Despite the fact that our idea is similar to Narasimhan [2018], a crucial difference is that Narasimhan [2018] focused on the plug-in classifier, i.e., a classifier based on the class-posterior probability estimation. As we will discuss later, the plug-in classifier is often inferior to the direct ERM approach because the class-posterior probability estimation imposes relatively strong assumptions on the underlying distribution.

information the learner is acquiring during the training. Interpretability and explainability have recently become hot topics within the machine learning community [Carvalho et al., 2019], for which a black-box model is investigated through local linear approximations, as an example. Although this line of research has contributed to a certain extent to understanding local behaviors of a black-box model, what a learner learns from the data remains unclear.[13] This perspective will have a significant impact when we want to extract knowledge from a trained learner and apply or transfer it to other situations.

Thus, what type of formulations can provide meaningful insights into learned knowledge? Although the extraction of knowledge from pre-trained language models has achieved success to a certain extent and recently received attention in the community of natural language processing [Petroni et al., 2019], learned knowledge should not be limited to knowledge explainable by natural language.[14] From a machine learning perspective, we believe that learned knowledge is useful when it can help solve other problems. Technically, this idea can be formulated through the idea of excess risk transfer, i.e., by noting that a learning problem is defined based on an evaluation metric, *we can treat an evaluation metric of one problem as the learning criterion and another evaluation metric as the evaluation criterion.* We therefore pose a second research question.

Research question (knowledge transfer).

Given two learning problems, can we reduce one problem to another problem? Namely, is it possible to obtain a good predictor for one learning problem (evaluation task) once we train a learner for another learning problem (learning task)?

Note that this perspective differs from traditional transfer learning [Pan and Yang, 2009], which aspires to deal with distribution shifts between training and test data. Here, we want to understand the relationship between two machine learning problems with different structures, to see whether knowledge extracted from one learning problem can help us solve another learning problem. We call this paradigm a *knowledge transfer* in this dissertation.[15]

[13] Although the machine learning community is more interested in prediction and generalization, the statistics community is more interested in how to establish a plausible mathematical model of a target natural phenomenon. Hence, interpreting model behaviors is one of the important aspects in statistical modeling. For example, a relative weight analysis [Johnson, 2000] and a dominance analysis [Azen and Budescu, 2003] have been proposed to investigate which variables heavily affect the outcomes in a regression analysis. These analysis methods provide a post-hoc interpretation in modeling behaviors for understanding the target phenomena. In comparison with statistical modeling, which is mainly devoted to a regression analysis, machine learning problems often have richer structures. Therefore, understanding the underlying mechanism of learning problems will help us design more useful cost functions.

[14] Oftentimes, classicism prefers to semantic representation by a symbolic language, whereas connectionism is based on a belief that any concept semantically corresponds to a single (neural) unit [Fodor and Pylyshyn, 1988]. Recent research on natural language processing and knowledge bases has been mainly focusing on how to extract natural language explanation from language models [Petroni et al., 2019]. Although such symbolic explanation is indispensable for our semantical understanding, we argue for the importance of knowledge extraction that enables a learner to perform well on future tasks. Note that classicism and connectionism represent knowledge (input) by a corresponding symbol and unit (output), respectively, whereas our idea based on knowledge transfer is more interested in input-output relationships and how they can be applied to the other tasks.

[15] Apart from transfer learning, this idea may remind one of meta learning. Here, we briefly discuss the difference between knowledge transfer and meta learning. In meta learning, we

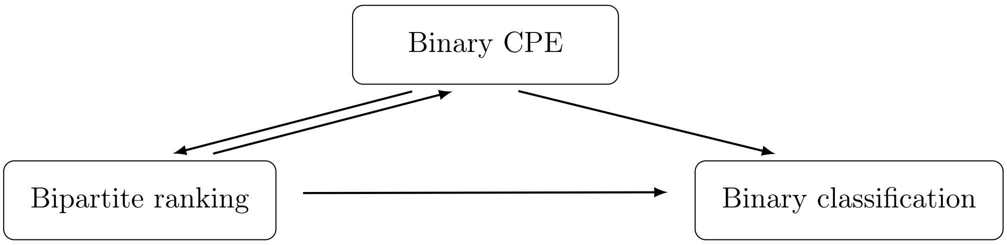


Figure 1.8: The relationship among binary classification, bipartite ranking, and binary class-posterior probability estimation. An arrow indicates that a procedure exists to transfer a model obtained from the source problem to the destination problem. The figure is based on the results of Narasimhan and Agarwal [2013]. We will explicate this relationship in Section 2.5.

Originally, our idea of knowledge transfer was inspired by a seminal work conducted by Narasimhan and Agarwal [2013], who studied the relationship among three different machine learning problems, i.e., binary classification, bipartite ranking, and binary class-posterior probability estimation (CPE). Although all three problems deal with binary outcomes, bipartite ranking seeks a ranking model to output higher scores for positive examples than negative examples (formulated by the binary rank loss), and binary CPE seeks to build a probabilistic model that approximates well a class-posterior probability (often formulated through the squared loss). Thus, one might naturally imagine that a good binary classifier can be obtained immediately after obtaining a good CPE model. In other words, a learner can obtain richer knowledge through binary CPE than binary classification. Narasimhan and Agarwal [2013] investigated such a relationship among the three learning problems and established excess risk transfer bounds, as shown in Figure 1.8.

Why does knowledge transfer matter? This is because we should design a learning criterion capable of eliciting knowledge that we need for the ultimate purpose, whereas we should not design a criterion that is too difficult to handle. This idea is illustrated through an example in Figure 1.9. Solving an excessively difficult and general learning problem can result in an insufficient prediction performance given a limited number of data and resources; hence, designing an intermediate learning criterion is an important task. This is reminiscent of the so-called Vapnik's principle [Vapnik, 2006]:

> *When solving a problem of interest, do not solve a more general problem as an intermediate step. Try to get the answer that you really need but not a more general one.* — Vladimir Vapnik

With this idea in mind, Vladimir Vapnik introduced a notion of transductive inference, which is a reasoning framework used to infer outcomes for test cases directly from training cases without eliciting a general law as is done in an inductive inference. Note that whereas transductive inference has stimulated several learning algorithms including transductive SVMs [Joachims, 2003], they mainly

consider a set of tasks, which consist of a loss function and a distribution. Oftentimes, to incorporate reinforcement learning, a transition distribution and episode length are considered to be a part of a task. A learner is herein expected to quickly adapt to a new task drawn from the underlying task distribution [Finn, 2018]. This idea appears in Thrun and Pratt [1998, p. 6]: "by identifying the 'right' properties, the hypothesis space can be diminished, yielding more accurate generalization from less data." In other words, meta learning somewhat seeks *invariant* properties of tasks from the data. By contrast, knowledge transfer aims to extract knowledge from the data, transform it, and apply to the other problem. We are rather interested in *different* characteristics of tasks.

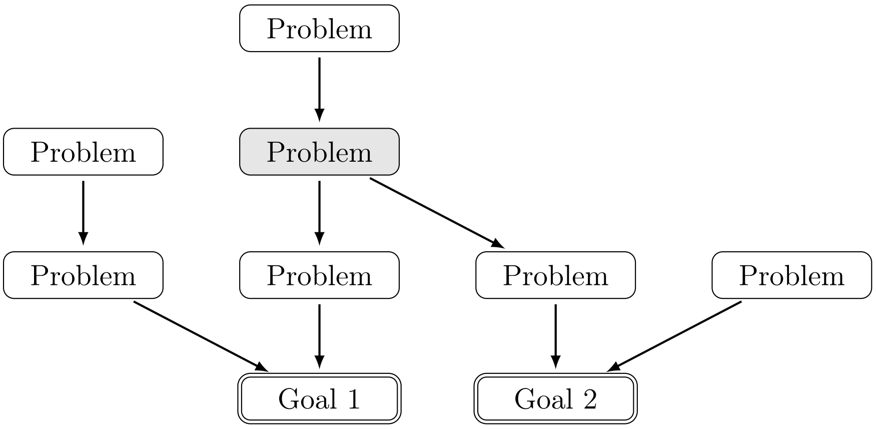


Figure 1.9: Illustration of a hierarchical relationship among learning problems. Each node corresponds to a learning problem defined by an evaluation criterion. Each arrow indicates the knowledge transfer relationships. In this example, we have two goals owing to certain real-world requirements. Under this situation, one possible strategy is to solve the problem corresponding to a common ancestral node of the two goals (the shaded node), but not to solve an excessively general one (such as the ancestral problem of the shaded one).

focused on eliciting useful information directly from training data for a test prediction of the same learning problem. By contrast, our idea here is motivated by transferring knowledge elicited from one learning problem to another learning problem with as little effort as possible.

We quote an earlier remark made in Russell [1912], which has the same spirit as Vapnik's principle:

> *We shall reach the conclusion that Socrates is mortal with a greater approach to certainty if we make our argument purely inductive than if we go by way of "all men are mortal" and then use deduction.*
>
> — Bertrand Russel

1.4 Scope and Contributions of this Dissertation

From the discussions provided thus far, the learning of reliable and transferrable knowledge is a fundamental goal in machine learning. In this dissertation, we show that this can be achieved by elucidating the relationship between the learning and evaluation criteria. Herein, we state our goal in a general form.

Scope of this dissertation (excess risk transfer).

Given an evaluation criterion, verify whether a learning criterion leads to the optimal solution in terms of the evaluation criterion, or, design an appropriate learning criterion leading to the optimal solution.

Mathematically, we adopt the idea of calibration analysis and excess risk transfer to connect the excess risk of a learning criterion to the excess risk of an evaluation criterion.

Subsequently, we explicate the contributions of this dissertation to describe how we attempt to establish the framework of excess risk transfer and relate it to reliable and transferrable machine learning.

	Predicted positive	Predicted negative
Positive	True positive ($\nearrow$)	False negative ($\searrow$)
Negative	False positive ($\searrow$)	True negative ($\nearrow$)

Figure 1.10: Confusion matrix. We expect the entries with a $\nearrow$ to be as high as possible and those with a $\searrow$ to be as low as possible.

1.4.1 Chapter 3: Design of Learning Criteria for Complex Classification Metrics

Motivation. In classification problems, class-prior distributions are often skewed such that some categories appear infrequently in the training data. This situation is common in scenarios such as information retrieval for query and document classification [Manning and Schütze, 2008]. When a class-prior distribution is heavily skewed, the classification problem is often called imbalanced [Japkowicz and Stephen, 2002].

In classical literature, the (binary) classification performance is visualized using a confusion matrix, where each row of the matrix represents the number/ratio of instances in the true classes, and each column represents the number/ratio of instances in the predicted classes [Sokolova and Lapalme, 2009]. A confusion matrix contains the four values, i.e., true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), and a learner is expected to achieve as high a TP and TN as possible while keeping the FP and FN as low as possible (see Figure 1.10). Although the four values are often visually analyzed based on a precision-recall (PR) curve and the receiver operating characteristic (ROC) curve (see Figure 1.11), dealing with the multiple metrics at the same time in the optimization is not straightforward and is not interpretable for users. For these reasons, several aggregated metrics have been proposed and used in the literature. For example, the following three performance metrics are commonly applied.

- The F$_\beta$-measure [van Rijsbergen, 1974] is defined by

  $$\frac{(1+\beta^2)\mathsf{TP}}{(1+\beta^2)\mathsf{TP}+\beta^2\mathsf{FN}+\mathsf{FP}}$$

  with a trade-off hyperparameter $\beta \geq 0$, is usually set to 1. This is the harmonic mean of the precision ($= \mathsf{TP}/(\mathsf{TP}+\mathsf{FP})$) and recall ($= \mathsf{TP}/(\mathsf{TP}+\mathsf{FN})$). This performance metric is common in many fields facing imbalanced data such as information retrieval [Manning and Schütze, 2008], natural language processing [Derczynski, 2016], and medical image analysis [Milletari et al., 2016]. The F$_\beta$-measure is also known as the Sørensen-Dice coefficient.

- The Jaccard index [Jaccard, 1901] is defined by

  $$\frac{\mathsf{TP}}{\mathsf{TP}+\mathsf{FN}+\mathsf{FP}}.$$

  The Jaccard index was originally proposed to measure the similarity between two sets A and B using $|A \cap B|/|A \cup B|$, namely, the intersection-over-union (IoU). For this reason, this index has been commonly used in semantic segmentation in the field of computer vision community [Berman et al., 2018]. When the two sets are the set of the predicted positive and true positive instances, the definition based on the set similarity reduces to the above. The Jaccard index is also known as the Tanimoto distance [Tanimoto, 1958].

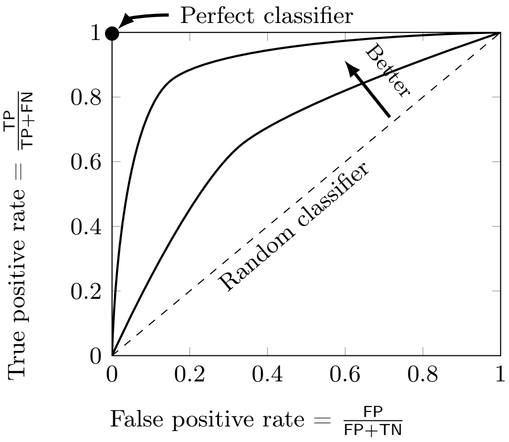


Figure 1.11: The ROC curve. For a binary classifier, a curve is plotted by repeatedly changing the classification threshold and measuring the true positive and false positive rates. The perfect classifier achieves the top-left point (false positive rate = 0 and true positive rate = 1).

- The Tversky index [Tversky, 1977] is defined by

$$\frac{\mathsf{TP}}{\mathsf{TP}+\alpha\mathsf{FN}+\beta\mathsf{FP}}$$

with trade-off parameters $\alpha, \beta \geq 0$. It reduces to an F_1-measure with $\alpha = \beta = \frac{1}{2}$ and a Jaccard index with $\alpha = \beta = 1$. Owing to its high flexibility, it has been sometimes been used in semantic segmentation [Salehi et al., 2017].

For the aforementioned wide application domains, the machine learning community has paid more attention to these complex classification performance metrics. In particular, we are interested in the following general form:

$$\frac{a_0\mathsf{TP}+b_0\mathsf{TN}+c_0\mathsf{FP}+d_0\mathsf{FN}}{a_1\mathsf{TP}+b_1\mathsf{TN}+c_1\mathsf{FP}+d_1\mathsf{FN}},$$

with the given values $a_0, b_0, c_0, d_0, a_1, b_1, c_1, d_1 \in \mathbb{R}$. This generalized form of the classification performance metric is called the *linear-fractional metric* [Koyejo et al., 2015, Bao and Sugiyama, 2020, Nordström et al., 2020]. Because the reported value of the classification performance metrics in these domains has a specific interpretation, achieving a better performance in terms of these given metrics is an important task. Although recent studies have been interested in training a classifier by maximizing the linear-fractional metrics, these approaches are based on either heuristics [Milletari et al., 2016, Berman et al., 2018] or plug-in classifiers relying on class-posterior probability estimators [Koyejo et al., 2014, Narasimhan et al., 2014, Yan et al., 2018], which is usually sample-inefficient.[16] Hence, the design of sample-efficient learning criteria with the consistency guarantee to the metrics of the classification performance remains an important open problem.

[16] Suppose that the evaluation metric is the classification accuracy. Audibert and Tsybakov [2007] then showed that the plug-in classifier cannot achieve fast rates unless a strong assumption is imposed on the underlying probability distribution, although Audibert and Tsybakov [2007] was originally intended to provide a scenario in which the plug-in classifier achieves faster rates than the ERM approaches. Although it remains an open question whether the plug-in classifier achieves the same rates as the surrogate risk minimization, we naturally speculate that the answer is negative.

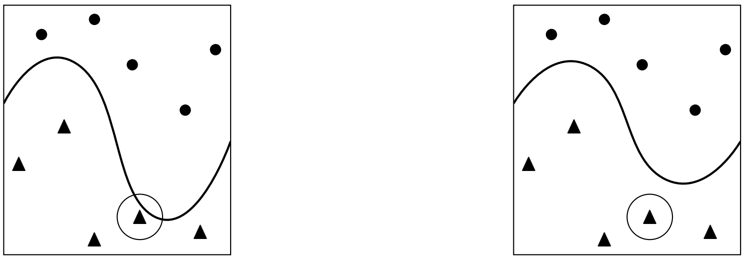
Contribution. We focus on binary classification with an evaluation criterion in the linear-fractional family. We propose surrogate utility functions as the learning criterion, which smoothens the original classification performance metric. This surrogate utility is shown to be calibrated to the given evaluation criterion, meaning that our predictor approaches the optimal performance by optimizing the proposed utility. The optimization problem of the proposed utility function is the quasiconcave maximization, and hence it is computationally feasible. Experimentally, we confirm that the proposed utility function outperforms the plug-in classifiers, particularly when the sample size is small.

1.4.2 Chapter 4: Certification of Learning Criteria for Adversarially Robust Classification

Motivation. Whereas deep learning has achieved significant success in building models with a significantly high predictive performance, it has been empirically observed that the prediction of deep neural networks can be arbitrarily manipulated by perturbing test patterns with small noises that are imperceptible to humans [Kurakin et al., 2016]. Such malicious manipulations are called adversarial attacks. Adversarial attacks are not only observed with deep neural networks but also using SVMs [Xiao et al., 2015]. Improving the robustness of machine learning models against adversarial attacks is a significantly important issue from the viewpoint of machine learning security. Figure 1.12 illustrates the idea of an adversarial example and a decision boundary vulnerable to an adversarial attack.

Goodfellow et al. [2015] introduced a simple method for crafting adversarial examples and trained a model with data augmented with adversarial examples. This can be essentially regarded as a robust optimization problem with the minimax classification risk, i.e., minimizing the classification risk under the worst-case (max) adversarial attacks, and Shaham et al. [2018] solved the robust optimization problem directly. In addition, Wong and Kolter [2018] proposed a tractable upper bound of the minimax objective. More recently, a learning procedure called randomized smoothing has received attention, which outputs a stochastic classifier built on top of a base neural network by injecting appropriate noise such that the resulting classifier is robust against adversarial attacks [Cohen et al., 2019]. Although these proposals have been reported to be effective in practice, it has yet to be determined whether the existing learning algorithms lead to the optimal solution in terms of the evaluation criterion of adversarially robust classification, i.e., robust (or worst-case) classification risk. This perspective should not be overlooked because it is necessary to assess the potential security risks quantitatively before we deploy machine learning models.

Contribution. We focus on binary classification in the presence of adversarial attacks. Using the calibration analysis formulated by Steinwart [2007], we show that no convex surrogate loss functions can lead to the optimal solution in terms of the adversarially robust classification risk when we use the linear-in-input models. Roughly speaking, convex surrogate losses cannot guarantee sufficiently large prediction margins, which is essential for adversarial robustness. This is in stark contrast to the classical result of Bartlett et al. [2006], which characterizes that convex surrogate loss functions are classification-calibrated under mild sufficient conditions. We also investigate alternative nonconvex candidates for calibrated surrogate loss functions in adversarially robust classification. This chapter provides a new insight in that not only a predictive performance but also the robustness of a classifier can be incorporated into the calibration analysis.



(a) A vulnerable decision boundary.

(b) A robust decision boundary.

Figure 1.12: Illustration of vulnerable and robust binary decision boundaries against adversarial attacks. The ball indicates the budget for an adversarial attack. An adversary is allowed to perturb the input associated with the ball such that the input crosses the decision boundary in (a).

1.4.3 Chapter 5: Reduction from Classification to Similarity Learning

Motivation. Data similarity is an important concept in machine learning. With an appropriately defined metric, clustering can be conduced to analyze the data structure [MacQueen, 1967]. Even if obtaining Euclidean feature embeddings is not straightforward in domains such as graphs, sequences, and logics, sophisticated similarity measures have been developed for such structured data [Ontañón, 2020], and we can construct feature representations from the similarity information [Wang et al., 2009, Chen et al., 2009]. Semantic similarity information has recently been popularly used in self-supervised learning based on contrastive representation learning [Jaiswal et al., 2021].

Given that similarity information is valuable in many situations, the learning of a good similarity has been an area of focus for the past two decades. One of the most common approaches is distance metric learning, which models the similarity based on a parametrized Mahalanobis distance, and makes similar inputs closer, and vice versa. Distance metric learning is usually used for downstream clustering and k-nearest neighbor classification [Xing et al., 2003, Weinberger and Saul, 2009, Kulis, 2013]. Other approaches attempt to obtain a good similarity function used for constructing feature embedding [Bellet et al., 2012]. Overall, when we learn a similarity function, we usually have other purposes in our mind such as downstream classification. However, the goodness of the similarity has seldom been evaluated in terms of the performance metrics of the downstream tasks.

Contribution. We focus on binary classification as a downstream task of similarity learning. By giving a simple formulation of similarity learning, we show that solving similarity learning leads directly to solving binary classification. Specifically, we consider the formulation of similarity learning in which a pairwise model is asked to predict whether two inputs share the underlying classes. The pairwise classification risk is then shown to be monotonically related to the clustering error, which is equivalent to the classification risk but ignores the permutation of the predicted class labels. If we do not care about the label flipping, similarity learning in our formulation is sufficient to elicit a binary decision boundary. If we need to modify the label flipping, our proposed method can fix it with an exponentially small sample complexity. Herein, we treat the pairwise classification risk as a learning criterion and the clustering error as an evaluation criterion and draw a connection between two different learning problems, i.e., similarity learning and binary classification. This relationship is reminiscent of the results of Narasimhan and Agarwal [2013]. As a result, we elucidate that a learner can

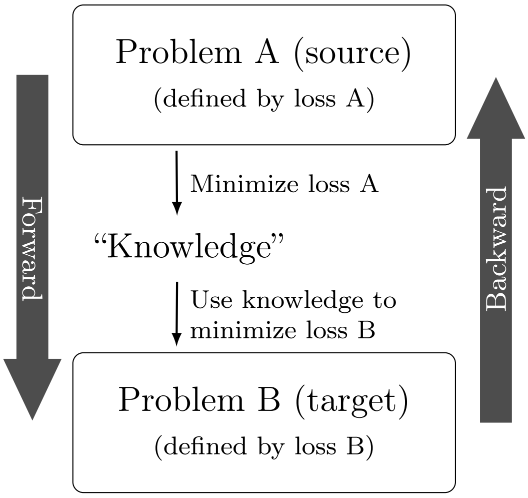


Figure 1.13: Two ways to investigate the problem reduction. In this figure, problem A is the source problem and problem B is the target problem, and we are interested in whether problem B can be solved by using the knowledge obtained when we solve problem A. In the forward way, the source problem A is fixed first, and we seek what elicitable knowledge and the target problem are. In the backward way, the target problem B is fixed first, and we seek what source problem leads to the desired target property.

essentially learn a decision boundary through similarity learning.

1.4.4 Why Contributions of This Dissertation Matter

Why do we need the contributions of this dissertation based on the two perspectives, reliable predictions and knowledge transfer, briefly summarized so far? Why do we specifically focus on the linear-fractional metrics (Chapter 3) and adversarial robustness (Chapter 4) to discuss the reliability? Why do we discuss only similarity learning (Chapter 5) to establish knowledge transfer in this dissertation? Before concluding the introduction, we provide additional explanations to answer those questions.

Our ultimate motivation to conduct research on learning theory is to *reveal the relationship between two distinct learning problems* as we stated in Section 1.3.3. As a result of elucidating the problem relationships, we expect to better understand the hierarchy of learning problems, which may lead to a more sophisticated design of learning criteria (see Figure 1.9). When one wonders which learning criterion should be used, there are two ways to approach this (see Figure 1.13 for the concept of the forward/backward approaches). The first way is the *backward* approach, where a target problem that one wants to solve or target properties that a learner should acquire are fixed first, and we seek which learning criterion can achieve them. The second way is the *forward* approach, where a source problem or learning criterion is fixed first, and we check what property a learner can eventually acquire after solving the source problem. The backward approach is mainly concerned about desired properties such as reliability, and the forward approach is concerned about whether knowledge from the source problem can be transferred to the target problem. This is why we focus on reliability and transferability.

Yet, why do we need to focus on the linear-fractional metrics and adversarial robustness among many reliability properties in machine learning? Here, we argue that the linear-fractional metrics are evaluation criteria involving the confusion matrix in a general form and that adversarial robustness is essentially related to the prediction margin. These two quantities, the confusion matrix and prediction margin, can encompass many features of classifiers, and thereby we believe that they are "canonical" quantities to be investigated.[17] Of course, there are many other choices of how to characterize the properties of a given classifier. Nonetheless, the confusion matrix and prediction margin would be interpretable quantities for users because the former is a statistically aggregated quantity and the latter is a geometrical quantity. For these reasons, we focus on them as properties characterizing the prediction reliability.[18]

Another important question is why we choose to discuss similarity learning as a single instance of knowledge transfer among many other machine learning problems. In fact, the perspective of knowledge transfer and problem reduction provides a novel framework, even though we were inspired by a prior work [Narasimhan and Agarwal, 2013], and we have infinitely many candidates on which learning problems to work. We choose similarity learning because the notion of similarity is fundamental in human perception—in Aristotle's metaphysics, humans recognize concepts in the world by utilizing similarity of five sense information, characteristics, structures, and topology to form clusters [Nagao, 2019]. Whereas similarity has got a lot of attention in recent machine learning research [Jaiswal et al., 2021], it must be valuable to reveal what similarity learning elicits not only from the practical viewpoint but also the philosophical viewpoint. As machine learning has begun from imitation of human perception, we hope to provide feedback to the understanding of human perception by analyzing similarity learning.

1.5 Organization

The organization of this dissertation is summarized as follows. Chapter 2 provides background materials for the formulation of supervised learning and learning theory. The basics of both generalization analysis and calibration analysis are briefly introduced. Chapters 3 and 4 are devoted to reliable machine learning. Specifically, we consider the design of reliable learning criteria for the linear-fractional metrics under a class imbalance in Chapter 3 and the certification of the learning criteria for adversarially robust classification in Chapter 4. Then, Chapter 5 considers the knowledge transfer; in particular, the relationship between classification and similarity learning is discussed. Lastly, we conclude this dissertation in Chapter 6. The organization of this dissertation is summarized in Figure 1.14.

[17] For example, fairness constraints can be represented by the confusion matrix [Narasimhan, 2018]. The model calibration is essentially concerned about the prediction margin [Guo et al., 2017]. These two quantities can characterize many perspectives of a given classifier.

[18] Indeed, it is a notoriously hard task to disentangle one notion in a "canonical" way. Historically, Immanuel Kant tackled to define twelve *categories* to characterize how synthetic *a priori* judgments are possible [Kant, 1781]. Each category corresponds to a condition and framework of human thought in general, and Kant claimed that the twelve categories cannot be derived from any more general concept. However, a number of philosophers criticized the validity of Kant's categories. Smith [1918] argued that the twelve categories do not match the logical framework provided by the earlier logicians and lack consensus. Yet, Kant claimed how he formulated the twelve categories based on the general framework of human thought based on the discussions provided by earlier philosophers and logicians. In any cases, it is important to clarify why one chooses a specific categorization. In our case, we believe that evaluation criteria based on the confusion matrix and prediction margin can cover a broad range of prediction characteristics and are user-friendly quantities hence we focus on these two quantities.

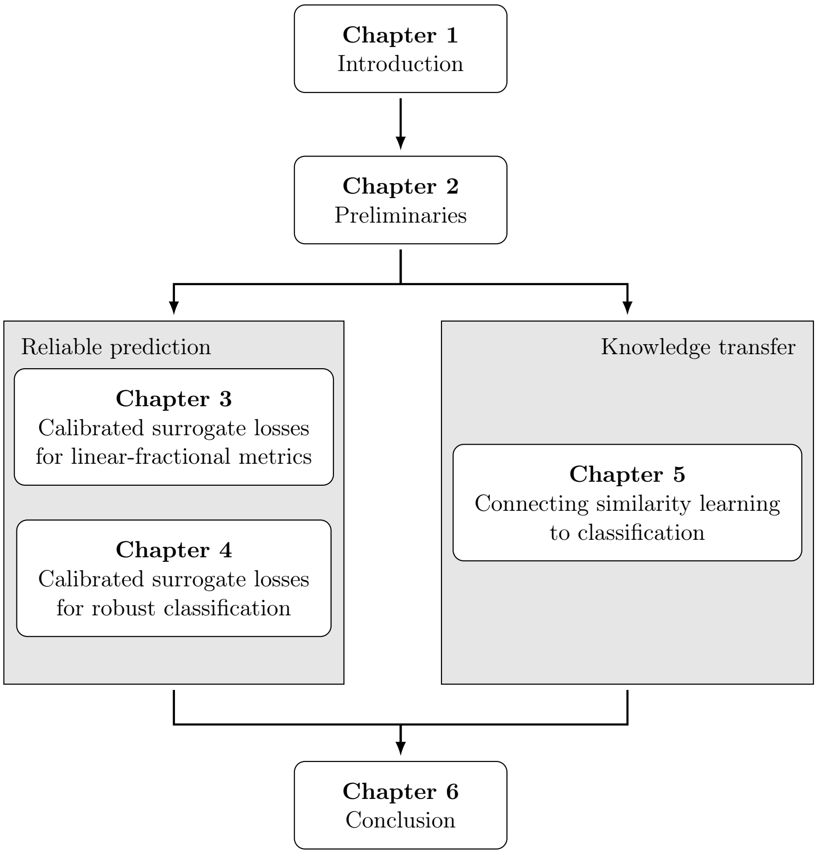

Chapter 1
Introduction
Chapter 2
Preliminaries
Reliable prediction
Chapter 3
Calibrated surrogate losses for linear-fractional metrics
Chapter 4
Calibrated surrogate losses for robust classification
Knowledge transfer
Chapter 5
Connecting similarity learning to classification
Chapter 6
Conclusion

Figure 1.14: The organization of this dissertation.

Chapter 2

Preliminaries

Science without religion is lame, religion without science is blind.
— Albert Einstein, *Science and Religion*

In this chapter, the formulation of machine learning and the background knowledge on loss functions are introduced. We mainly focus on a supervised formulation of classification in the subsequent chapters. In addition, we provide an overview of the calibration analysis conducted to study the relationship between surrogate and target risks.

2.1 Notation and Formulation of Supervised Classification

Let $\mathbb{R}$, $\mathbb{R}_{\geq 0}$, $\mathbb{N}$, and $\mathbb{Z}$ denote the real line, the non-negative real line, the set of natural numbers, and the set of integers, respectively. For $n \in \mathbb{N}$, let $[n]$ be the index set $\{1, \ldots, n\}$. Let $\mathbb{1}_{\{A\}}$ be an indicator function that takes a value of 1 if the predicate A holds and a value of 0 otherwise. The sign of a number $\alpha \in \mathbb{R}$ is denoted by $\operatorname{sgn}(\alpha)$. We adopt the convention $\operatorname{sgn}(0) = -1$. In addition, $\overline{\operatorname{sgn}} : \mathbb{R} \to \{+1, -1\}$ is defined by $\overline{\operatorname{sgn}}(\alpha) = \operatorname{sgn}(\alpha)$ for $\alpha \neq 0$ and $\overline{\operatorname{sgn}}(0) = +1$. We use bold font (such as $\mathbf{x}$) to denote vectors and sans-serif font (such as X) to denote random variables. For a vector $\mathbf{x} \in \mathbb{R}^n$, the i-th element is denoted by x_i, and $\|\mathbf{x}\|_p$ denotes the ℓ_p-norm $\sqrt[p]{|x_1|^p + \cdots + |x_n|^p}$. The norm $\|\cdot\|$ without a subscript denotes the ℓ_2-norm. For a measure μ and a measurable function f, the pushforward measure of μ by f is denoted by $f_\sharp\mu$.[1]

For random variables $\mathsf{X}_1, \ldots, \mathsf{X}_n$ drawn from a joint distribution $\mathbb{P}$, the joint expectation is denoted by $\mathbb{E}_{(\mathsf{X}_1,\ldots,\mathsf{X}_n)\sim\mathbb{P}(\mathsf{X}_1,\ldots,\mathsf{X}_n)}[\cdot]$. Alternatively, simpler expressions such as $\mathbb{E}_{(\mathsf{X}_1,\ldots,\mathsf{X}_n)}[\cdot]$, $\mathbb{E}_{\mathbb{P}}[\cdot]$, or $\mathbb{E}[\cdot]$ are used if there is no ambiguity from the context.

For a real vector space $\mathcal{Y}$ and a function $h : \mathcal{T} \to \mathbb{R}$, $\operatorname{dom}(h)$ denotes the domain of h. Let $h^\star$ denote the Fenchel-Legendre conjugate, defined by $h^\star(\mathbf{p}) \coloneqq \sup_{\mathbf{x}\in\operatorname{dom}(h)} \mathbf{p}^\top\mathbf{x} - h(\mathbf{x})$. In addition, $h^{\star\star}$ denotes the Fenchel-Legendre biconjugate, namely, the Fenchel-Legendre conjugate of $h^\star$. For more technical details of convex analysis, please refer to standard textbooks such as Rockafellar [1970].

Common notation used in this dissertation is summarized in Table 2.1.

[1] The pushforward measure $f_\sharp\mu$ is a (probability) measure obtained by transferring measure μ from the original measurable space to another space by using a measurable function f. See standard textbooks such as Halmos [1946] for details.

Table 2.1: Notation table.

Symbol	Description
$\mathbb{R}$	Real line
$\mathbb{R}_{\geq 0}$	Non-negative real line
$\mathbb{N}$	Set of natural numbers
$\mathbb{Z}$	Set of integers
$[n]$	$\{1, 2, \ldots, n\}$ for $n \in \mathbb{N}$
$\mathrm{sgn}(\alpha)$	$2 \cdot \mathbb{1}_{\{\alpha > 0\}} - 1$ for $\alpha \in \mathbb{R}$
$\overline{\mathrm{sgn}}(\alpha)$	$2 \cdot \mathbb{1}_{\{\alpha \geq 0\}} - 1$ for $\alpha \in \mathbb{R}$
$\lVert \cdot \rVert_p$	ℓ_p-norm
$\lVert \cdot \rVert$	ℓ_2-norm
$\mathcal{X}$	Feature space
$\mathcal{Y}$	Outcome space
$\mathcal{T}$	Prediction score space (e.g., $\mathcal{T} = \mathbb{R}$ in binary classification)
$\mathcal{G} \subseteq \mathcal{Y}^{\mathcal{X}}$	Hypothesis space
$\mathcal{F} \subseteq \mathcal{T}^{\mathcal{X}}$	Score function space (e.g., $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ in binary classification)
$\mathcal{F}_{\mathrm{all}} \subseteq \mathcal{T}^{\mathcal{X}}$	Set of all measurable score functions
$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$	Target loss function
$\phi : \mathcal{T} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$	Surrogate loss function
$R_\ell(g)$	Target risk function (or ℓ-risk) for $g \in \mathcal{G}$
$R_\phi(f)$	Surrogate risk function (or ϕ-risk) for $f \in \mathcal{F}$
$\widehat{R}_\phi(f)$	Empirical ϕ-risk function
R_ϕ^*	Bayes (ϕ-)risk

2.1.1 Supervised Learning

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a *feature* space and $\mathcal{Y}$ be an *outcome* space. The outcome space is typically $\mathcal{Y} = \{+1, -1\}$ for binary classification and $\mathcal{Y} = [C]$ for C-way classification, where C is an integer larger than 2. In the case of classification, an outcome is usually referred to as a *label.* In supervised learning, it is commonly assumed that a labeled example $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ is independently and identically drawn from an unknown joint probability distribution. Let $\mathbb{P}(\mathsf{X}, \mathsf{Y})$ denote the density function of this joint distribution. The $\mathcal{X}$-marginal distribution is denoted by $\mathbb{P}_{\mathsf{X}}$, and $\mathbb{P}$ is occasionally used if there is no confusion. One of the primary goals of supervised learning is to predict the best outcome for a given input feature vector from a finite number of examples drawn from $\mathbb{P}(\mathsf{X}, \mathsf{Y})$. Formally, we are given a finite number of labeled examples $\mathcal{S} := \{(\mathbf{x}_i, y_i)\}_{i \in [n]}$. To construct a prediction rule, we prepare a search space of the prediction rules, called a *hypothesis space*, $\mathcal{G} \subseteq \mathcal{Y}^{\mathcal{X}}$.[2] Then, the best model g^* across a specified hypothesis space $\mathcal{G}$ is sought by minimizing the following quantity

$$R_\ell(g) := \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim \mathbb{P}(\mathsf{X},\mathsf{Y})}[\ell(g(\mathsf{X}), \mathsf{Y})],$$

[2] The set of functions from $\mathcal{X}$ to $\mathcal{Y}$ is denoted by either $\mathcal{X} \to \mathcal{Y}$ or $\mathcal{Y}^{\mathcal{X}}$.

and g^* is defined as[3]

$$g^* \in \operatorname*{arg\,min}_{g \in \mathcal{Y}^{\mathcal{X}}:\text{measurable}} R_\ell(g).$$

Here, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is a *loss function*, which measures the "closeness" between the prediction $g(\mathsf{X})$ and the outcome Y. Loss functions are often simply called *losses*. The functional R_ℓ is called the *expected ℓ-risk function*, defining the goodness of a model. When it is clear from the context, we often simply refer to it as an ℓ-risk function. The expected risk is often referred to as *population risk* or *theoretical risk* in some studies [Mohammadi and van de Geer, 2005, Jin et al., 2018]. Note that a loss function ℓ essentially defines the desirable predictions, that is, an evaluation criterion of a supervised learning task is essentially defined by ℓ. For this reason, we occasionally call this a *target loss function* to distinguish it from surrogate loss functions that we introduce later. Similarly, we also occasionally call it a *target risk function* to clearly emphasize the difference.

Although the expected ℓ-risk function requires access to an infinite sample to be computed, the *empirical ℓ-risk function* introduced below may serve as a good approximation.

$$\widehat{R}_\ell(g) \coloneqq \frac{1}{n} \sum_{i \in [n]} \ell(g(\mathbf{x}_i), y_i).$$

In Section 2.2, we show that empirical risk serves as a good approximator of the expected risk in a certain sense.

The popular target loss functions are listed below.

- Binary 0-1 loss (binary classification): $\ell(g(\mathbf{x}), y) = \mathbb{1}_{\{yg(\mathbf{x}) \leq 0\}}$.
- Binary cost-sensitive loss (binary classification): $\ell(g(\mathbf{x}), y) = \alpha_{+1}\mathbb{1}_{\{g(\mathbf{x}) \neq +1\}} + \alpha_{-1}\mathbb{1}_{\{g(\mathbf{x}) \neq -1\}}$, where $\alpha_{\pm 1} > 0$ are misclassification costs.
- Multi-class 0-1 loss (multi-class classification): $\ell(g(\mathbf{x}), y) = \mathbb{1}_{\{g(\mathbf{x}) \neq y\}}$.
- Normalized Hamming loss (structured prediction): $\ell(\mathbf{g}(\mathbf{x}), \mathbf{y}) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\{g_t(\mathbf{x}) \neq y_t\}}$, for tuples of T binary outputs g_t and y_t.

Each loss function corresponds to some target supervised learning task. For example, the 0-1 loss penalizes predictions by a cost of 1 when prediction $g(\mathbf{x})$ does not match a given label y. Note, however, that target losses are often unpreferable in terms of optimization—the 0-1 loss has a discrete nature. To mitigate such optimization issues, surrogate loss functions are commonly introduced, which is discussed in Section 2.1.2.

Although the above definition is not confined to classification, this dissertation focuses on classification. That is, $|\mathcal{Y}| < \infty$ is assumed subsequently throughout the dissertation.

2.1.2 Surrogate Loss Functions

In the supervised learning formulation, we aimed at minimizing the expected ℓ-risk $R_\ell(g)$ in a hypothesis space $\mathcal{G}$. Because the outcome space $\mathcal{Y}$ is often discrete (such as $\{+1, -1\}$ in binary classification), the hypothesis space is usually replaced with a function space with a continuous output space $\mathcal{T}^{\mathcal{X}}$, where $\mathcal{T}$ is a complete

[3] Assume that the minimizer g^* exists.

metric space called a *prediction score space*. A score $t \in \mathcal{T}$ usually has a one-to-one relationship with a prediction in $\mathcal{Y}$. This mapping (from a prediction to a score) is called a *link function* [McCullagh and Nelder, 1989, Finocchiaro et al., 2019]. For example, in the binary classification case, a score space is usually taken as $\mathcal{T} = \mathbb{R}$, and a *score function space* $\mathcal{F} \subseteq \mathcal{T}^{\mathcal{X}}$ is introduced. When clear from the context, we also call $\mathcal{F}$ a hypothesis space. Then, a score $f(\mathbf{x})$ is transformed into a prediction $\operatorname{sgn}(f(\mathbf{x})) \in \{+1, -1\}$. In this case, $\operatorname{sgn}(\cdot)$ is the *inverse* link function. A prediction score is often referred to as a margin [Mohri et al., 2018], a report [Frongillo and Kash, 2015], or simply a score [Mohri et al., 2018].[4]

As a score function space, the following candidates are commonly used in binary classification.

- Linear-in-input models:
  $\mathcal{F} \coloneqq \left\{ \mathbf{x} \mapsto \mathbf{w}^\top \mathbf{x} \;\middle|\; \mathbf{w} \in \mathbb{R}^d \right\}$.
- Linear-in-parameter models:
  $\mathcal{F} \coloneqq \left\{ \mathbf{x} \mapsto \mathbf{w}^\top \boldsymbol{\varphi}(\mathbf{x}) \;\middle|\; \mathbf{w} \in \mathbb{R}^b \right\}$, where $\boldsymbol{\varphi} : \mathcal{X} \to \mathbb{R}^b$ are fixed basis functions.
- Kernel models:
  $\mathcal{F} \coloneqq \left\{ \mathbf{x} \mapsto \sum_{i \in [n]} w_i k(\mathbf{x}, \mathbf{x}_i) \;\middle|\; \mathbf{w} \in \mathbb{R}^n \right\}$, where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ is a fixed kernel function.
- Perceptron models:
  $\mathcal{F} \coloneqq \left\{ \mathbf{x} \mapsto \sum_{j \in [b]} w_{2,j} \varphi(\mathbf{w}_{1,j}^\top \mathbf{x} + b_1) + b_2 \right\}$, where $\varphi : \mathbb{R} \to \mathbb{R}$ is an activation function, b is the number of hidden units, $\mathbf{w}_{1,1}, \ldots, \mathbf{w}_{1,b} \in \mathbb{R}^d, \mathbf{w}_2 \in \mathbb{R}^b$ are weight parameters, and $b_1, b_2 \in \mathbb{R}$ are bias parameters.

Although a hypothesis space typically becomes more flexible in approximating a desirable target function as the number of parameters increases, there is a trade-off relationship between the flexibility and generalization performance, which will be discussed in Section 2.2.

In many cases, the optimization of the target risk is not straightforward owing to the undesirable nature of the target loss ℓ. For example, when the target loss is the binary 0-1 loss $\ell(f(\mathbf{x}), y) = \mathbb{1}_{\{yf(\mathbf{x}) \leq 0\}}$, minimizing the binary 0-1 risk is NP-hard even if the hypothesis space $\mathcal{F}$ is a linear-in-input model [Kearns et al., 1994]. Intuitively, its reason can be understood by the discrete nature of the binary 0-1 loss—the gradient descent [Boyd and Vandenberghe, 2004], one of the most commonly-used optimizers, may suffer from this loss function because the gradient of the binary 0-1 risk in f vanishes almost everywhere. Owing to this discrete nature, the target loss is replaced with a *surrogate loss function* in many situations [Zhang, 2004a, Bartlett et al., 2006].

A surrogate loss function $\phi : \mathcal{T} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ measures the "closeness" of a *prediction score* $t \in \mathcal{T}$ and an outcome $y \in \mathcal{Y}$, whereas a target loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ measures the "closeness" of a prediction $f(\mathbf{x}) \in \mathcal{Y}$ and an outcome $y \in \mathcal{Y}$. Unlike a target loss, a surrogate loss is usually chosen from the optimization perspective [Zhang, 2004a, Bartlett et al., 2006]. Common choices of surrogate losses in binary classification are shown below.

- Squared loss: $\phi(t, y) = \frac{1}{4}(1 - ty)^2$.

[4] Please be careful to distinguish a prediction score from a proper scoring rule [Buja et al., 2005]. Some proper scoring rules (proper losses) are simply called scores, such as the Brier score [Brier, 1950].

- Logistic loss: $\phi(t, y) = \ln(1 + e^{-ty})$.
- Hinge loss: $\phi(t, y) = \max\{0, 1 - ty\}$.
- Exponential loss: $\phi(t, y) = e^{-ty}$.
- Ramp loss: $\phi(t, y) = \frac{1}{2}\max\{2, \min\{0, 1 - ty\}\}$.
- Sigmoid loss: $\phi(t, y) = \frac{1}{1+e^{ty}}$.

Several surrogate losses correspond to popular machine learning algorithms. For example, the hinge loss corresponds to the support vector machine [Cortes and Vapnik, 1995], whereas the exponential loss corresponds to AdaBoost [Freund and Schapire, 1997]. Notably, many surrogate losses (such as the squared, logistic, hinge, and exponential losses) are convex in the first argument t (score), and thus optimization with such losses has a significant merit over the 0-1 loss. From the perspective of surrogate losses, our objective function is called a *surrogate risk*, or the *(expected) ϕ-risk*, given by

$$R_\phi(f) \coloneqq \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim\mathbb{P}(\mathsf{X},\mathsf{Y})}[\phi(f(\mathsf{X}), \mathsf{Y})].$$

The corresponding *empirical ϕ-risk* is given by

$$\widehat{R}_\phi(f) \coloneqq \frac{1}{n}\sum_{i\in[n]} \phi(f(\mathbf{x}_i), y_i).$$

Although surrogate losses have many nice properties, they are different from the original target loss in its functional form. Hence, the surrogate risk minimization does not necessarily lead to the target risk minimization [Lin, 2004, Zhang, 2004a, Bartlett et al., 2006]. This gap is one of the central focuses in this dissertation, and will be discussed in detail in Section 2.3.

2.2 Generalization Analysis

In Section 2.1, two qualitatively different risk functionals, expected and empirical risks, were introduced to formulate supervised learning. In this section, we discuss learning theory to bridge the gap between the expected and empirical risks.

2.2.1 Generalization Error Bounds

Assume that a (surrogate) loss function $\phi : \mathcal{T} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is fixed. The corresponding expected and empirical ϕ-risks are given as

$$R_\phi(f) = \mathbb{E}[\phi(f(\mathsf{X}), \mathsf{Y})] \quad \text{and} \quad \widehat{R}_\phi(f) = \frac{1}{n}\sum_{i\in[n]} \phi(f(\mathbf{x}_i), y_i),$$

given a sample $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i\in[n]}$. Because our ultimate goal in supervised learning is to derive a prediction rule aligning with the underlying law governed by the (unknown) distribution $\mathbb{P}$, the desideratum is the expected risk R_ϕ. Owing to the lack of the full information of $\mathbb{P}$, we need to resort to using a finite sample approximation based on the empirical risk $\widehat{R}_\phi$. Hence, there exists a gap between the two quantities, i.e., for a fixed function $f \in \mathcal{F}$,

$$R_\phi(f) - \widehat{R}_\phi(f),$$

which is called a *generalization gap* [Keskar et al., 2017]. The generalization gap measures how accurate the empirical risk can approximate the expected risk at a single point. The community of statistical learning theory has developed many different theories to evaluate the generalization gap by introducing finer complexity measures of a function class $\mathcal{F}$ [Dudley, 1967, Vapnik and Chervonenkis, 1971, Bartlett and Mendelson, 2002]. For example, Bartlett and Mendelson [2002] introduced the *Rademacher complexity*, one of the complexity measures of a function class.

Definition 2.1 (Rademacher complexity)**.** *Let the variables $\boldsymbol{\sigma} \in \{\pm 1\}^n$ be i.i.d. according to $\mathbb{P}(\sigma_i = +1) = \mathbb{P}(\sigma_i = -1) = \frac{1}{2}$.*[5] *In addition, let $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ be a function class. Then, the* Rademacher complexity *of $\mathcal{H}$ with respect to a set $\{ \mathbf{z}_1, \dots, \mathbf{z}_n \}$ distributed i.i.d. from $\mathbb{P}$ is defined as*

$$\mathfrak{R}_n(\mathcal{H}) \coloneqq \mathop{\mathbb{E}}_{\mathbf{z}_1,\dots,\mathbf{z}_n \sim \mathbb{P}} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i \in [n]} \sigma_i h(\mathbf{z}_i) \right].$$

Intuitively, the Rademacher complexity of $\mathcal{H}$ can be understood as measuring the maximum correlation between a function $h \in \mathcal{H}$ and "all possible binary labels" $\boldsymbol{\sigma}$. The larger this correlation, the more complex and flexible the function class $\mathcal{H}$ is. Based on the Rademacher complexity, the following generalization gap bound can be established.

Theorem 2.2 (Shalev-Shwartz and Ben-David [2014])**.** *Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ be a function class and $\phi : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a loss function such that $\phi(f(\mathbf{x}), y) \leq B_\phi$ for all $\mathbf{x} \in \mathcal{X}$, $y \in \mathcal{Y}$, and $f \in \mathcal{F}$. Then, for any fixed $f \in \mathcal{F}$, with probability at least $1 - \delta$ over the repeated sampling of $\mathcal{S} \sim \mathbb{P}$,*

$$R_\phi(f) - \widehat{R}_\phi(f) \leq 2\mathfrak{R}_n(\phi \circ \mathcal{F}) + B_\phi \sqrt{\frac{2 \ln \frac{2}{\delta}}{n}},$$

where $\mathfrak{R}_n(\phi \circ \mathcal{F})$ denotes the Rademacher complexity of the composed function class $\{ (\mathbf{x}, y) \mapsto \phi(f(\mathbf{x}), y) \mid f \in \mathcal{F} \}$. Furthermore, assume that $\phi(t, y)$ is Lipschitz continuous in $t \in \mathbb{R}$ when y is fixed, and let L_ϕ denote the largest Lipschitz norm of ϕ. As a result of the contraction lemma *[Ledoux and Talagrand, 1991], the following inequality holds for any fixed $f \in \mathcal{F}$ with a probability of at least $1 - \delta$.*

$$R_\phi(f) - \widehat{R}_\phi(f) \leq 2L_\phi \mathfrak{R}_n(\mathcal{F}) + B_\phi \sqrt{\frac{2 \ln \frac{2}{\delta}}{n}}.$$

The generalization error bound consists of two terms: the Rademacher complexity of $\mathcal{F}$ and the sample complexity term $B_\phi \sqrt{\frac{2\ln(2/\delta)}{n}}$. The sample complexity term obviously vanishes at the infinite sample limit $n \to \infty$. The behavior of the complexity term $\mathfrak{R}_n(\mathcal{F})$ depends on the flexibility of the hypothesis space $\mathcal{F}$ is. For example, the Rademacher complexity of the linear-in-input models $\mathcal{F} = \{ \mathbf{x} \mapsto \mathbf{w}^\top \mathbf{x} \mid \|\mathbf{w}\|_2 \leq B_{\mathbf{w}} \}$ is estimated as follows [Shalev-Shwartz and Ben-David, 2014].

$$\mathfrak{R}(\mathcal{F}) = O\left(\frac{B_{\mathbf{w}}}{\sqrt{n}} \right).$$

[5] Such random variables are called *Rademacher variables.*

By contrast, the Rademacher complexity of M-layer neural networks $\mathcal{F}$, having the parameter matrices of each layer m with a bounded Frobenius norm B_m, is estimated as follows:

$$\mathfrak{R}(\mathcal{F}) = O\left(\frac{\sqrt{M+1+\ln(d)} \cdot \prod_{m\in[M]} B_m}{\sqrt{n}}\right).$$

The detailed conditions needed for deriving the above order are described in Golowich et al. [2018]. Importantly, (i) the Rademacher complexity decreases as the sample size n increases, as does the sample complexity. However, (ii) excessively flexible models (such as a function class with a large parameter norm, or a neural network with many layers) may slow down the speed of the shrinkage. That is, a function class trades off its Rademacher complexity for the model flexibility.

Overall, by ignoring constant dependencies, the generalization analysis claims that $R_\phi(f) - \widehat{R}_\phi(f) = O_p\left(\frac{1}{\sqrt{n}}\right)$. In other words, *the generalization gap vanishes at the infinite sample limit* in probability.

Note that other complexity measures such as the Vapnik-Chervonenkis (VC) dimension [Vapnik and Chervonenkis, 1971] and Dudley's entropy integral [Dudley, 1967] have been also popularly used, although the qualitative observations on generalization gaps remain the same, i.e., the sample complexity term vanishes at the infinite sample limit, and there is a trade-off between the flexibility of a function space and its complexity.

2.2.2 Estimation Error and Approximation Error

The generalization gap estimates how close the empirical risk is to the expected risk *for a fixed function* f. In this analysis, we cannot discuss the quality of the minimizer of the empirical risk because it depends on the training sample. From an algorithmic perspective, we are interested in the quality of the minimizer $\widehat{f} := \arg\min_{f\in\mathcal{F}} \widehat{R}_\phi(f)$, which is called the *empirical risk minimizer*. In light of our goal, it is desirable for the expected risk of $\widehat{f}$ to approach the best possible expected risk. This is measured by the *excess (ϕ-)risk*:

$$R_\phi(\widehat{f}) - R_\phi^*,$$

where $R_\phi^* := \inf_{f\in\mathcal{F}_{\text{all}}} R_\phi(f)$ is the *Bayes (ϕ-)risk* and $\mathcal{F}_{\text{all}}$ is the set of all measurable functions. A classifier that attains the Bayes ϕ-risk is called the *Bayes (ϕ-)classifier*. When we simply state the Bayes classifier, it subsequently refers to the Bayes classifier with respect to the target loss function subsequently.

However, the excess risk does not involve the capacity of the function class $\mathcal{F}$. By taking account of $\mathcal{F}$, the excess risk is decomposed as follows.

$$R_\phi(\widehat{f}) - R_\phi^* = \underbrace{R_\phi(\widehat{f}) - R_\phi(f^\dagger)}_{\text{estimation error}} + \underbrace{R_\phi(f^\dagger) - R_\phi^*}_{\text{approximation error}},$$

where $f^\dagger := \arg\min_{f\in\mathcal{F}} R_\phi(f)$ is the expected risk minimizer. This decomposition consists of an estimation error, namely, how close the empirical risk minimizer is to the best possible function in $\mathcal{F}$, and an approximation error, namely, how close the Bayes classifier to $\mathcal{F}$. The estimation error can be bounded by using the generalization gap as follows.

$$\begin{aligned}R_\phi(\widehat{f}) - R_\phi(f^\dagger) &= \{R_\phi(\widehat{f}) - \widehat{R}_\phi(\widehat{f})\} + \{\widehat{R}_\phi(\widehat{f}) - \widehat{R}_\phi(f^\dagger)\} + \{\widehat{R}_\phi(f^\dagger) - R_\phi(f^\dagger)\} \\ &\leq \{R_\phi(\widehat{f}) - \widehat{R}_\phi(\widehat{f})\} + 0 + \{\widehat{R}_\phi(f^\dagger) - R_\phi(f^\dagger)\} \\ &\leq 2 \sup_{f \in \mathcal{F}} |R_\phi(f) - \widehat{R}_\phi(f)|,\end{aligned}$$

where the first inequality is due to $\widehat{R}_\phi(\widehat{f}) \leq \widehat{R}_\phi(f)$ for any $f \in \mathcal{F}$ (the definition of the empirical risk minimizer). The gap $|R_\phi(f) - \widehat{R}_\phi(f)|$ can be bounded by applying Theorem 2.2 twice (i.e., on $R_\phi(f) - \widehat{R}_\phi(f)$ and $\widehat{R}_\phi(f) - R_\phi(f)$). Therefore, the estimation error can be controlled by the generalization gap. By contrast, the approximation error remains difficult to theoretically analyze. In practice, the structural risk minimization (SRM) was proposed to nicely control the trade-off between the prediction performance and model capacity [Mohri et al., 2018]. The cross-validation is one of the widely used procedures for SRM.

2.3 Calibration Analysis

In Section 2.2, we studied the relationship between the expected and empirical risks for a fixed loss function ϕ, and showed that the empirical risk approaches the expected risk with a sufficiently large number of observations. Herein, we stress that our supervised learning formulation aims at minimizing an expected target risk, whereas a surrogate risk is introduced in light of the optimization perspective in Section 2.1.2. This section is devoted to introducing the basic important aspects required to fill in the gap between surrogate and target risks. We mainly focus on the expected surrogate and target risks; hence, the discussions in this section are free from an analysis of the sample complexity. Hence, the gap analysis between surrogate and empirical risks constitutes an axis conceptually orthogonal to the generalization analysis.

2.3.1 Target and Surrogate Risks

First, we review the important concepts in this section. For an output space $\mathcal{Y}$ and a prediction score space $\mathcal{T}$, $\psi : \mathcal{T} \to \mathcal{Y}$ denotes the inverse link function mapping a prediction score to an outcome. Let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a target loss representing a user demand, and $\phi : \mathcal{T} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a surrogate loss introduced by the user. Assuming a hypothesis space $\mathcal{F} \subseteq \mathcal{T}^\mathcal{X}$, the target risk ($\ell$-risk) and surrogate risk (ϕ-risk) are correspondingly introduced:

- Target risk (ℓ-risk): $R_\ell(f) := \mathbb{E}\,\ell(\psi \circ f(\mathsf{X}), \mathsf{Y})$.
- Surrogate risk (ϕ-risk): $R_\phi(f) := \mathbb{E}\,\phi(f(\mathsf{X}), \mathsf{Y})$.

Note that with the aid of a fixed inverse link ψ, the target risk $R_\ell(f)$ is defined over the score function space $\mathcal{F}(\subseteq \mathcal{T}^\mathcal{X})$ instead of the hypothesis space $(\subseteq \mathcal{Y}^\mathcal{X})$, unlike that in Section 2.1. When the link function ψ is clear from the context, we will often adopt this definition of the target risk. The best possible risk value is called *Bayes risk*. Formally, the Bayes ϕ-risk is defined as

$$R_\phi^* := \inf_{f \in \mathcal{F}_{\mathrm{all}}} R_\phi(f),$$

where $\mathcal{F}_{\mathrm{all}}$ is the set of all measurable functions over $\mathcal{X} \to \mathcal{T}$. The Bayes ℓ-risk R_ℓ^* is defined in the same way:

$$R_\ell^* := \inf_{f \in \mathcal{F}_{\mathrm{all}}} R_\ell(f).$$

An *excess risk* refers to the difference between a risk and its best possible value. The excess ℓ-risk and excess ϕ-risk are $R_\ell(f) - R_\ell^*$ and $R_\phi(f) - R_\phi^*$, respectively. Ideally, the surrogate loss has a property such that $R_\phi(f_i) \to R_\phi^* \implies R_\ell(f_i) \to R_\ell^*$ as $i \to \infty$, for a given sequence of prediction functions $\{f_i\}_i \subseteq \mathcal{F}$. This property can be considered a minimal necessary condition for the surrogate risk minimization. We formally call this property *ℓ-consistency*.

Definition 2.3 (ℓ-consistent loss [Steinwart, 2007])**.** *Let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a target loss function and $\phi : \mathcal{T} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a surrogate loss function. We say that ϕ is* ℓ-consistent *if for any probability distribution $\mathbb{P}$ over $\mathcal{X} \times \mathcal{Y}$ and sequence of functions $\{f_i\}_i \subseteq \mathcal{F}$,*

$$R_\phi(f_i) - R_\phi^* \to 0 \implies R_\ell(f_i) - R_\ell^* \to 0,$$

as $i \to \infty$.

Subsequently, we proceed with technical details revealing the necessary conditions on ϕ.

2.3.2 Pointwise Perspective of Risk Functionals

Directly analyzing the (full) risks R_ℓ and R_ϕ is not straightforward because they are functionals, requiring a variational calculus to achieve an optimization of these quantities. Instead, Steinwart [2007] introduced a *pointwise* analysis to ease these technical difficulties. In this subsection, we take a look at the pointwise forms of risk functionals first. For simplicity, we focus on binary classification: $\mathcal{Y} = \{+1, -1\}$ and $\mathcal{T} = \mathbb{R}$.

For a loss function ϕ, the *conditional ϕ-risk* $C_\phi : \mathcal{F} \times [0,1] \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ is defined as follows:[6]

$$C_\phi(f, \eta, \mathbf{x}) := \eta\phi(f(\mathbf{x}), +1) + (1 - \eta)\phi(-f(\mathbf{x}), -1).$$

The above is considered pointwise because C_ϕ recovers the full risk with $\eta = \mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = \mathbf{x})$:

$$\begin{aligned} R_\phi(f) &= \int \mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = \mathbf{x})\phi(f(\mathbf{x}), +1) \\ &\qquad + \mathbb{P}(\mathsf{Y} = -1 \mid \mathsf{X} = \mathbf{x})\phi(-f(\mathbf{x}), -1)\mathrm{d}\mathbb{P}(\mathsf{X} = \mathbf{x}) \\ &= \int C_\phi(f, \mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X}), \mathsf{X})\mathrm{d}\mathbb{P}(\mathsf{X}). \end{aligned}$$

In the same way, the *conditional Bayes ϕ-risk* $C_\phi^*(\eta, \mathbf{x})$ is defined as follows:

$$C_\phi^*(\eta, \mathbf{x}) := \inf_{f \in \mathcal{F}} C_\phi(f, \eta, \mathbf{x}).$$

The conditional Bayes risk can be regarded as a pointwise form of the Bayes risk at each point $\mathbf{x} \in \mathcal{X}$ with $\mathbb{P}(\mathsf{Y} = +1 \mid \mathbf{x}) = \eta$. Unlike the ϕ-risk, the Bayes ϕ-risk is not immediately representable by the conditional Bayes ϕ-risk because an expectation and an infimum are not commutable and $R_\phi^* \geq \int C_\phi^*(\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X}), \mathsf{X})\mathrm{d}\mathbb{P}(\mathsf{X})$ in general. To make them commutable, Steinwart [2007] introduced a condition called $\mathbb{P}$-*minimizability*.

[6] The conditional ϕ-risk is also called the *pointwise ϕ-risk* or *inner ϕ-risk* [Reid and Williamson, 2010].

Definition 2.4 (Minimizable loss function [Steinwart, 2007])**.** *Let $\phi : \mathcal{T} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a loss function and $\mathbb{P}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$. We state that ϕ is* $\mathbb{P}$-minimizable *if for all $\varepsilon > 0$ there exists $f_\varepsilon \in \mathcal{F}$ such that for all $\mathbf{x} \in \mathcal{X}$, we have*

$$C_\phi(f_\varepsilon, \mathbb{P}(\mathsf{Y} = +1 \mid \mathbf{x}), \mathbf{x}) < C^*_\phi(\mathbb{P}(\mathsf{Y} = +1 \mid \mathbf{x}), \mathbf{x}) + \varepsilon.$$

Steinwart [2007, Lemma 2.5] showed that a $\mathbb{P}$-minimizable loss function ϕ has a nice formula $R^*_\phi = \int C^*_\phi(\mathbb{P}(\mathsf{Y} = +1 \mid \mathbf{x}), \mathbf{x}) \mathrm{d}\mathbb{P}(\mathbf{x})$. Under this condition with $R^*_\phi < \infty$, the excess ϕ-risk can be expressed in a pointwise manner as well:

$$R_\phi(f) - R^*_\phi = \int C_\phi(f, \mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X}), \mathsf{X}) - C^*_\phi(\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X}), \mathsf{X}) \mathrm{d}\mathbb{P}(\mathsf{X}).$$

Hence, in the subsequent analysis on excess risks, we can split our analysis into

- an analysis of the conditional form of the excess risks, and
- an analysis of the minimizability of the loss function.

In Section 2.3.3, the analysis of the conditional excess risks is discussed. The latter analysis of the minimizability of the loss function is notoriously difficult. When $\mathcal{F}$ is a set of all measurable functions $\mathcal{F}_{\text{all}}$, Steinwart [2007, Theorem 3.2] showed that a loss ϕ is $\mathbb{P}$-minimizable if and only if $C^*_\phi(\mathbb{P}(\mathsf{Y} = +1 \mid \mathbf{x}), \mathbf{x}) < \infty$ for all $\mathbf{x} \in \mathcal{X}$. However, for a general function space $\mathcal{F}$, further discussion remains open and several researchers have been trying to pursue such analyses [Long and Servedio, 2013, Zhang and Agarwal, 2020].

2.3.3 Calibration Function

According to the discussion provided thus far, we are interested in the conditions that imply ℓ-consistency $R_\phi(f_i) \overset{i\to\infty}{\to} R^*_\phi \implies R_\ell(f_i) \overset{i\to\infty}{\to} R^*_\ell$. As discussed in Section 2.3.2, the conditional excess risks are useful in an analysis, provided that the regularity conditions hold for the given loss functions. Here, we specifically focus on an analysis of the surrogate consistency by utilizing the conditional excess risks.

First, we define a pointwise notion of ℓ-consistency.

Definition 2.5 (ℓ-calibrated loss [Steinwart, 2007])**.** *Let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a target loss function and $\phi : \mathcal{T} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a surrogate loss function. We state that ϕ is* ℓ-calibrated *if for any $\varepsilon > 0$, $\eta \in [0, 1]$, and $\mathbf{x} \in \mathcal{X}$, there exists $\delta > 0$ such that for all $f \in \mathcal{F}$,*

$$C_\phi(f, \eta, \mathbf{x}) - C^*_\phi(\eta, \mathbf{x}) < \delta \implies C_\ell(f, \eta, \mathbf{x}) - C^*_\ell(\eta, \mathbf{x}) < \varepsilon.$$

In light of Definition 2.3, it can be intuitively understood that ℓ-calibration is a pointwise counterpart of ℓ-consistency. If ϕ and ℓ are $\mathbb{P}$-minimizable loss functions, their excess risks can be expressed by the corresponding conditional excess risks, and thus an ℓ-calibrated loss immediately entails ℓ-consistency. The formal statement is found in Steinwart [2007, Theorem 2.8]. In addition, Steinwart [2007, Theorem 3.3] ensures that ℓ-calibration is not only sufficient but necessary for ℓ-consistency. For this reason, ℓ-calibration is an important property used to study the ℓ-consistency.

To check ℓ-calibration, the following tool is useful.

Definition 2.6 (ℓ-calibration function [Steinwart, 2007])**.** *The ℓ-calibration function of a loss function ϕ $\bar{\delta} : \mathbb{R}_{\geq 0} \times [0,1] \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ is defined as*

$$\bar{\delta}(\varepsilon, \eta, \mathbf{x}) := \inf_{f \in \mathcal{F}} \left\{ C_\phi(f, \eta, \mathbf{x}) - C_\phi^*(\eta, \mathbf{x}) \;\middle|\; C_\ell(f, \eta, \mathbf{x}) - C_\ell^*(\eta, \mathbf{x}) \geq \varepsilon \right\}.$$

Calibration functions are defined as constrained optimization problems. Because this optimization is reduced to finite-dimensional optimization in many cases, as described later, calibration functions are useful by alleviating the difficulty of the variational calculus. An ℓ-calibration function is defined as the largest δ in Definition 2.5. By definition, an ℓ-calibration function is tightly connected to ℓ-consistency through the following statement.

Proposition 2.7 (Steinwart [2007])**.** *A surrogate loss ϕ is ℓ-calibrated if and only if its ℓ-calibration function $\bar{\delta}$ satisfies $\bar{\delta}(\varepsilon, \eta, \mathbf{x}) > 0$ for all $\varepsilon > 0$, $\eta \in [0,1]$, and $\mathbf{x} \in \mathcal{X}$.*

In this way, ℓ-calibration provides a "qualitative" statement for ℓ-consistency, namely, the convergence of the excess ℓ-risk. Can we further elucidate a "quantitative" statement for ℓ-consistency, namely, the convergence rate of excess risk? For this purpose, a stronger notion than ℓ-calibration is necessary.

Definition 2.8 (Uniform ℓ-calibrated loss [Steinwart, 2007])**.** *Let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a target loss function and $\phi : \mathcal{T} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a surrogate loss function. We state that ϕ is* uniformly ℓ-calibrated *if for any $\varepsilon > 0$, there exists $\delta > 0$ such that for all $\eta \in [0,1]$, $f \in \mathcal{F}$, and $\mathbf{x} \in \mathcal{X}$,*

$$C_\phi(f, \eta, \mathbf{x}) - C_\phi^*(\eta, \mathbf{x}) < \delta \implies C_\ell(f, \eta, \mathbf{x}) - C_\ell^*(\eta, \mathbf{x}) < \varepsilon.$$

The corresponding uniform ℓ-calibration function $\delta : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ *is defined as*

$$\delta(\varepsilon) := \inf_{\eta \in [0,1]} \inf_{\mathbf{x} \in \mathcal{X}} \inf_{f \in \mathcal{F}} \left\{ C_\phi(f, \eta, \mathbf{x}) - C_\phi^*(\eta, \mathbf{x}) \;\middle|\; C_\ell(f, \eta, \mathbf{x}) - C_\ell(\eta, \mathbf{x}) \geq \varepsilon \right\}.$$

Remark that uniform and vanilla ℓ-calibrations differ in their positions of the quantifiers. Correspondingly, the calibration functions are different in how the infimum is taken. Obviously, ϕ is uniformly ℓ-calibrated if and only if $\delta(\varepsilon) > 0$ for all $\varepsilon > 0$. Steinwart [2007, Theorem 2.13] provides a connection of the excess ϕ-risk and the excess ℓ-risk, i.e., the so called *excess risk bound.*

Proposition 2.9 (Steinwart [2007])**.** *Let $\delta : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be the uniform ℓ-calibration function of a loss function ϕ. Let $\check{\delta} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be $\check{\delta}(\varepsilon) = \delta(\varepsilon)$ for $\varepsilon > 0$ and $\check{\delta}(0) = 0$. Suppose that ℓ and ϕ are $\mathbb{P}$-minimizable and $R_\ell^*, R_\phi^* < \infty$. Then, for all $f \in \mathcal{F}$,*

$$\check{\delta}^{\star\star}\left(R_\ell(f) - R_\ell^*\right) \leq R_\phi(f) - R_\phi^*,$$

where $\check{\delta}^{\star\star}$ denotes the Fenchel-Legendre biconjugate of $\check{\delta}$. In addition, $\check{\delta}^{\star\star}$ is invertible if and only if ϕ is uniformly ℓ-calibrated.

Hence, a uniformly ℓ-calibrated ϕ entails a bound

$$R_\ell(f) - R_\ell^* \leq \left(\check{\delta}^{\star\star}\right)^{-1} \left(R_\phi(f) - R_\phi^*\right),$$

providing a quantitative relationship between the surrogate and target excess risks. We occasionally refer to an excess risk bound as an *excess risk transform.*

Note that in this section we specialized the definition of conditional risks and calibration functions by Steinwart [2007] into the binary case $\mathcal{Y} = \{+1, -1\}$, although they are originally not restricted to binary classification. We confine ourselves to the binary case for simplicity of the exposition.

2.4 Classification-calibrated Loss and Proper Loss

In Section 2.3, calibration analysis was introduced to investigate the relationship between the target and surrogate excess risks. In this section, two types of surrogate loss functions are reviewed, *classification-calibrated losses* and *proper losses*, which elicit some underlying laws with respect to the distribution. Again, to avoid unnecessary technical complications, we focus on binary classification $\mathcal{Y} = \{+1, -1\}$.

2.4.1 Classification-calibrated Loss

In binary classification, the most popularly used target loss function is arguably the binary 0-1 loss $\ell(\widehat{y}, y) = \mathbb{1}_{\{\widehat{y} \neq y\}}$, for $\widehat{y} = \operatorname{sgn}(f(\mathbf{x}))$ (Section 2.1.1). The corresponding binary 0-1 risk is often called the *classification risk*. However, under the agnostic learning setting, which is a common setup in machine learning, where the Bayes risk is strictly larger than zero, minimization of the 0-1 loss is NP-hard unless RP = NP even if the hypothesis space $\mathcal{F}$ is linear-in-input models [Kearns et al., 1994].[7] For this reason, many machine learning algorithms bring in surrogate loss functions. Common examples of surrogate loss functions are listed in Section 2.1.2. One of the important questions to be asked is whether the surrogate risk minimization implies the minimization of the classification risk. Classification-calibrated loss functions are the minimal class of surrogate loss functions that imply the classification risk minimization [Bartlett et al., 2006].

The notion of classification-calibrated losses was developed in several studies [Lin, 2004, Bartlett et al., 2006]. Whereas a seminal study [Bartlett et al., 2006] introduced classification-calibrated losses in a constructive manner, we derive classification-calibrated losses by subsequently invoking calibration analysis. In this section, we focus on *margin-based loss functions* as surrogate loss functions, namely, a loss function $\phi(t, y)$ that can be written in the form $\bar{\phi}(ty)$ for some $\bar{\phi} : \mathbb{R} \to \mathbb{R}_{\geq 0}$. We consistently use a symbol $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ to denote a margin-based loss function with a slight abuse of notation. Indeed, all surrogate loss functions introduced in Section 2.1.2 are margin-based; for example, the hinge loss can be written as $\phi(m) = \max\{0, 1 - m\}$, where $m = ty$. When we focus on margin-based losses, conditional risks and the related notion can be simplified as follows. For a margin-based loss ϕ we have the following:

- Conditional ϕ-risk: $C_\phi(m, \eta) \coloneqq \eta\phi(m) + (1 - \eta)\phi(-m)$.
- Conditional Bayes ϕ-risk: $C^*_\phi(\eta) \coloneqq \inf_{m \in \mathbb{R}} C_\phi(m, \eta)$.
- Uniform ℓ-calibration function:
  $\delta(\varepsilon) \coloneqq \inf_{\eta \in [0,1]} \inf_{m \in \mathbb{R}} \left\{ C_\phi(m, \eta) - C^*_\phi(\eta) \;\middle|\; C_\ell(m, \eta) - C^*_\ell(\eta) \geq \varepsilon \right\}$.

For the binary 0-1 loss ℓ, the Bayes classifier is $f^*(\mathbf{x}) = \mathbb{P}(\mathsf{Y} = +1 \mid \mathbf{x}) - \frac{1}{2}$. The conditional ℓ-risk is

$$C_\ell(f, \eta, \mathbf{x}) = \eta \mathbb{1}_{\{\operatorname{sgn}(f(\mathbf{x})) = -1\}} + (1 - \eta)\mathbb{1}_{\{\operatorname{sgn}(f(\mathbf{x})) = +1\}}.$$

Then, the conditional Bayes ℓ-risk is $C^*_\ell(\eta, \mathbf{x}) = \min\{\eta, 1 - \eta\}$, and the conditional

[7] The randomized polynomial (RP) time is the complexity class of problems for which a probabilistic Turing machine that provides correct answers within the polynomial time with a certain probability [Arora and Barak, 2009].

excess ℓ-risk is

$$C_\ell(f,\eta,\mathbf{x}) - C_\ell^*(\eta,\mathbf{x}) = \begin{cases} 0 & \text{if } 2\eta-1>0 \text{ and } \operatorname{sgn}(f(\mathbf{x}))=+1 \\ 2\eta-1 & \text{if } 2\eta-1>0 \text{ and } \operatorname{sgn}(f(\mathbf{x}))=-1 \\ 1-2\eta & \text{if } 2\eta-1\le 0 \text{ and } \operatorname{sgn}(f(\mathbf{x}))=+1 \\ 0 & \text{if } 2\eta-1\le 0 \text{ and } \operatorname{sgn}(f(\mathbf{x}))=-1 \end{cases}$$
$$= |2\eta-1|\cdot \mathbb{1}_{\{(2\eta-1)f(\mathbf{x})\le 0\}}.$$

Hence, the uniform ℓ-calibration function is

$$\begin{aligned}\delta(\varepsilon) &= \inf_{\eta\in[0,1]} \inf_{f\in\mathcal{F}_{\text{all}}} \inf_{\mathbf{x}\in\mathcal{X}} \left\{ C_\phi(f,\eta,\mathbf{x}) - C_\phi^*(\eta,\mathbf{x}) \;\middle|\; |2\eta-1|\cdot \mathbb{1}_{\{(2\eta-1)f(\mathbf{x})\le 0\}} \ge \varepsilon \right\} \\ &= \inf_{\eta\in[0,1]} \inf_{m\in\mathbb{R}} \left\{ C_\phi(m,\eta) - C_\phi^*(\eta) \;\middle|\; |2\eta-1|\cdot \mathbb{1}_{\{(2\eta-1)m\le 0\}} \ge \varepsilon \right\}.\end{aligned}$$

If $\varepsilon > |2\eta-1|$, $\delta(\varepsilon)=\infty>0$ for all ε. By contrast, if $\varepsilon\le|2\eta-1|$, $|2\eta-1|\cdot\mathbb{1}_{\{(2\eta-1)m\le 0\}}$ holds if and only if $(2\eta-1)m\le 0$. Hence,

$$\delta(\varepsilon) = \begin{cases} \inf_{\eta\in[0,1]}\inf_{m\in\mathbb{R}:(2\eta-1)m\le 0} C_\phi(m,\eta)-C_\phi^*(\eta) & \text{if } \eta\le\frac{1-\varepsilon}{2} \text{ or } \frac{1+\varepsilon}{2}\le\eta, \\ \infty & \text{otherwise.}\end{cases}$$

Recall that the uniform ℓ-consistency is equivalent to $\delta(\varepsilon)>0$ for all $\varepsilon>0$. This is equivalent to

$$\underbrace{\inf_{m\in\mathbb{R}:(2\eta-1)m\le 0} C_\phi(m,\eta)}_{:=C_\phi^-(\eta)} - C_\phi^*(\eta) > 0 \text{ for } \eta\ne\frac{1}{2}.$$

The newly introduced quantity $C_\phi^-(\eta)$ can be interpreted as a conditional "sub-optimal" risk because the condition $(2\eta-1)m\le 0$ means the prediction m is inconsistent with the (conditional) Bayes classifier $\eta-\frac{1}{2}$ in a pointwise manner. Using $C_\phi^-(\eta)$, the uniform ℓ-calibration function can be simplified as follows:

$$\delta(\varepsilon) = C_\phi^-\left(\frac{1+\varepsilon}{2}\right) - C_\phi^*\left(\frac{1+\varepsilon}{2}\right),$$

because of the symmetry of the conditional risk $C_\phi(m,\eta)$ in $\eta=\frac{1}{2}$: $C_\phi(m,\eta)=C_\phi(-m,1-\eta)$. Based on the above discussion, Bartlett et al. [2006] introduced *classification-calibrated loss functions.*

Definition 2.10 (Classification-calibrated loss). *We say that ϕ is* classification-calibrated *if for all $\eta\ne\frac{1}{2}$,*

$$C_\phi^-(\eta) > C_\phi^*(\eta).$$

In our discussion, classification-calibrated losses naturally emerge from the uniform calibration function, whereas Bartlett et al. [2006] introduced the condition $C_\phi^-(\eta)>C_\phi^*(\eta)$ in a top-down approach. Bartlett et al. [2006] summarized the implications as follows.

Theorem 2.11 (Bartlett et al. [2006]). *For a non-negative margin-based loss function ϕ, any measurable function $f:\mathcal{X}\to\mathbb{R}$, and the probability distribution $\mathbb{P}$ over $\mathcal{X}\times\mathcal{Y}$,*

$$\check{\delta}^{\star\star}\left(R_\ell(f)-R_\ell^*\right) \le R_\phi(f)-R_\phi^*.$$

In addition, the following are equivalent.

Table 2.2: The conditional Bayes ϕ-risks and ℓ-calibration functions for some margin-based surrogate losses ϕ (excerpted from Steinwart [2007]).

Loss function	$\phi(m)$	$C^*_\phi(\eta)$	$\check{\delta}^{\star\star}(\varepsilon)$
Hinge	$\max\{0, 1-m\}$	$\lvert 2\eta - 1 \rvert$	ε
Truncated squared	$(\max\{0, 1-m\})^2$	$(2\eta-1)^2$	ε^2
Squared	$(1-m)^2$	$(2\eta-1)^2$	ε^2
Exponential	$\exp(-m)$	$1 - 2\sqrt{\eta(1-\eta)}$	$1 - \sqrt{1-\varepsilon^2}$
Sigmoid	$\frac{1}{1+\exp(m)}$	$\lvert 2\eta - 1 \rvert$	ε

1. *ϕ is classification-calibrated.*
2. *For any sequence $\{\varepsilon_i\} \subseteq [0,1]$, $\check{\delta}^{\star\star}(\varepsilon_i) \to 0$ if and only if $\varepsilon_i \to 0$.*
3. *For every sequence of measurable functions $f_i : \mathcal{X} \to \mathbb{R}$, $R_\phi(f_i) \to R^*_\phi$ implies $R_\ell(f_i) \to R^*_\ell$.*

Hence, classification-calibrated losses imply ℓ-consistency and the excess risk bounds can be obtained given C^-_ϕ and C^*_ϕ. The explicit forms of $\check{\delta}^{\star\star}$ for some examples are shown in Table 2.2. As an example of a surrogate loss that is not classification-calibrated, the perceptron loss $\phi(m) = \max\{0, -m\}$ is known [Rosenblatt, 1957].

Importantly, our discussion in this section is limited to binary classification. In other supervised learning tasks, a parallel notion of calibrated losses has been commonly introduced. Nevertheless, *calibration does not necessarily imply consistency*, unlike binary classification. This can be seen in multi-class classification [Zhang, 2004b, Tewari and Bartlett, 2007] and bipartite ranking [Gao and Zhou, 2015]. Hence, we have to be aware of the difference between calibration and consistency.

2.4.2 Proper Loss

Apart from classification-calibrated losses, there is another well-known family of loss functions called *proper losses*. Proper losses, also known as proper scoring rules, were firstly introduced for class-posterior probability estimation (CPE) of binary outcomes $\mathbb{P}(\mathsf{Y} = +1 \mid \mathbf{x})$ [Buja et al., 2005].[8] (Binary) CPE is a problem of estimating $\mathbb{P}(\mathsf{Y} = +1 \mid \mathbf{x})$ given a binary samples $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i \in [n]} \subseteq \mathcal{X} \times \mathcal{Y}$. As an insightful fact, a binary classifier can be constructed given a class probability estimate η, based on $\eta - \frac{1}{2}$. The aim of this subsection is to demonstrate the connection in that solving binary CPE leads to solving binary classification.

To measure the quality of binary class probability estimates, a loss function $\ell : [0,1] \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is introduced, i.e., $\ell(\widehat{\eta}, y)$ for $\widehat{\eta} \in [0,1]$, where y is a true class

[8] Proper scoring rules have their roots in statistics. Brier [1950] sparked debate on what is probability in meteorology. Since then, this question has received attention in the statistics community. Winkler and Murphy [1968] coined the term "proper scoring rules" for assessing the goodness. Savage [1971] discussed how to define and elicit personal probabilities in terms of the scoring rules and drew an interesting connection between the goodness and convex functions. Proper scoring rules are not limited to binary outcomes as in our discussions, and we can see some scoring rules defined in terms of the quantiles, intervals, and distributions [Gneiting and Raftery, 2007].

and $\widehat{\eta}$ is a class probability estimate. The conditional ℓ-risk is defined as

$$\begin{aligned} C_\ell(\widehat{\eta}, \eta) &\coloneqq \mathop{\mathbb{E}}_{\mathsf{Y} \sim \mathsf{Bernoulli}(\eta)}[\ell(\widehat{\eta}, \mathsf{Y})] \\ &= \eta\ell(\widehat{\eta}, 1) + (1-\eta)\ell(\widehat{\eta}, -1). \end{aligned}$$

The conditional Bayes ℓ-risk is denoted by $C_\ell^*(\eta) \coloneqq \inf_{\widehat{\eta} \in [0,1]} C_\ell(\widehat{\eta}, \eta)$. Proper losses are then defined.

Definition 2.12 (Proper loss [Buja et al., 2005]). *A loss function $\ell : [0,1] \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is said to be* proper *if $C_\ell(\widehat{\eta}, \eta)$ is minimized at $\widehat{\eta} = \eta$ for all $\eta \in [0,1]$. If the minimizer is unique, ℓ is said to be* strictly proper*.*

This definition guarantees that proper losses meet the minimal requirements for the CPE. Proper losses have two nice forms of alternative representations. One is given by Savage [1971], which has a deep connection to the Bregman divergence.

Proposition 2.13 (Savage [1971], Buja et al. [2005], Gneiting and Raftery [2007], Reid and Williamson [2009]). *The conditional Bayes risk $C_\ell^* : [0,1] \to \mathbb{R}$ for a proper loss ℓ is concave. Conversely, given a concave function $\Lambda : [0,1] \to \mathbb{R}$, there exists a proper loss ℓ satisfying $\Lambda(\eta) = C_\ell^*(\eta)$ for all $\eta \in [0,1]$, and its conditional risk satisfies*

$$C_\ell(\widehat{\eta}, \eta) = C_\phi^*(\eta) - (\widehat{\eta} - \eta)\nabla C_\phi^*(\widehat{\eta}).$$

Hence, the conditional risk of a proper loss is represented as a Bregman divergence generated from $(-C_\ell^*)$. The next representation is given by Shuford et al. [1966], in which partial proper losses $\ell(\widehat{\eta}, +1)$ and $\ell(\widehat{\eta}, -1)$ are bound into a single *weight function*.

Proposition 2.14 (Shuford et al. [1966], Buja et al. [2005], Reid and Williamson [2009]). *Suppose that $\ell(\widehat{\eta}, y)$ is differentiable in $\widehat{\eta}$. Then, ℓ is proper if and only if for all $\widehat{\eta} \in (0,1)$,*

$$-\frac{\nabla_{\widehat{\eta}}\ell(\widehat{\eta}, +1)}{1 - \widehat{\eta}} = -\frac{\nabla_{\widehat{\eta}}\ell(\widehat{\eta}, -1)}{\widehat{\eta}} = w(\widehat{\eta})$$

for some weight function *$w : (0,1) \to \mathbb{R}_{\geq 0}$ such that $\int_\varepsilon^{1-\varepsilon} w(q)\mathrm{d}q < \infty$.*

In practice, it is common to model a class probability estimator by transforming the outputs of a real-valued prediction function $f : \mathcal{X} \to \mathbb{R}$ into $[0,1]$ using a *link function* $\psi : [0,1] \to \mathbb{R}$. Given a proper loss ℓ and an invertible link ψ, a loss function over real-valued predictions is defined by $\ell_\psi(t, y) \coloneqq \ell(\psi^{-1}(t), y)$, which is called a *composite proper loss* [Reid and Williamson, 2010]. Computationally, it is useful if the composite proper loss $\ell_\psi(t, y)$ is convex in $t \in \mathbb{R}$. Buja et al. [2005] showed that a *canonical link* $\psi = -\nabla C_\phi^*$ satisfies the convexity.

Now, we are ready to review the connection between binary CPE and binary classification. Let ℓ_{01} denote the binary 0-1 loss to distinguish from a proper loss ℓ:

$$\ell_{01}(\widehat{\eta}, y) \coloneqq \frac{1}{2}\mathbb{1}_{\{y=+1\}}\mathbb{1}_{\{\widehat{\eta} > \frac{1}{2}\}} + \frac{1}{2}\mathbb{1}_{\{y=-1\}}\mathbb{1}_{\{\widehat{\eta} \leq \frac{1}{2}\}}.$$

The following theorem provides an upper bound of the conditional excess 0-1 risk by the conditional excess risk of a proper loss.

Table 2.3: Examples of proper losses. $\mathsf{H}_{\mathrm{S}}(\eta)$ denotes the binary Shannon entropy $\mathsf{H}_{\mathrm{S}}(\eta) \coloneqq -\eta \ln \eta - (1-\eta)\ln(1-\eta)$.

Loss function	$\ell(\widehat{\eta}, +1)$	$\ell(\widehat{\eta}, -1)$	Canonical link $\psi(\widehat{\eta})$	$C^*_\phi(\eta)$
Squared	$(1-\widehat{\eta})^2$	$\widehat{\eta}^2$	$2\widehat{\eta}-1$	$4\eta(1-\eta)$
Log	$-\ln \widehat{\eta}$	$-\ln(1-\widehat{\eta})$	$\ln \frac{\widehat{\eta}}{1-\widehat{\eta}}$	$\mathsf{H}_{\mathrm{S}}(\eta)$
Exponential	$\sqrt{\frac{1-\widehat{\eta}}{\widehat{\eta}}}$	$\sqrt{\frac{\widehat{\eta}}{1-\widehat{\eta}}}$	$\frac{2\widehat{\eta}-1}{\sqrt{\widehat{\eta}(1-\widehat{\eta})}}$	$2\sqrt{\eta(1-\eta)}$

Theorem 2.15 (Reid and Williamson [2009]). *Assume that a proper loss $\ell : [0,1] \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ satisfies $C^*_\ell\left(\frac{1}{2} - c\right) = C^*_\ell\left(c - \frac{1}{2}\right)$ for $c \in [0,1]$. Let $\psi : [0,1] \to \mathbb{R}$ be an invertible link function. Let $\delta(\varepsilon) \coloneqq C^*_\ell\left(\frac{1}{2}\right) - C^*_\ell\left(\frac{1}{2} + \varepsilon\right)$. Then,*

$$\delta\left(C_{\ell_{01}}(\psi^{-1}(m), \eta) - C^*_{\ell_{01}}(\eta)\right) \leq C_\ell\left(\psi^{-1}(m), \eta\right) - C^*_\ell(\eta)$$

for any $m \in \mathbb{R}$.

Note that δ is convex and increasing because of the concavity of C^*_ℓ for a proper loss ℓ. Hence, the (full) excess risk bound

$$R_{\ell_{01}}\left(\psi^{-1} \circ f\right) - R^*_{\ell_{01}} \leq \delta^{-1}\left(R^*_\ell\left(\psi^{-1} \circ f\right) - R^*_\ell\right)$$

is obtained based on Jensen's inequality, indicating that minimizing the excess ℓ-risk implies minimizing the excess 0-1 risk.

Proper losses and composite proper losses for multi-class classification have been discussed in Williamson et al. [2016].

2.5 Excess Risk Transform between Learning Tasks

Thus far, we have reviewed the framework of calibration analysis and the connection between the excess target and surrogate risks. As examples of loss functions, classification-calibrated losses and proper losses have been reviewed. Classification-calibrated losses are used purely as a "proxy" for the targeted 0-1 loss, whereas proper losses serve as a proxy but were originally designed for CPE, connecting CPE to classification. That is, surrogate losses are not necessarily merely a proxy for the true target but often represent a different learning task, eventually enabling us to connect the learning task to the true target task. In this section, we will see such a connection between two learning tasks through the lens of surrogate losses. We will review Narasimhan and Agarwal [2013], elucidating the relationship among binary classification, bipartite ranking, and binary CPE.

2.5.1 Binary Classification, Bipartite Ranking, and Binary CPE

First, each problem under consideration is summarized. For each problem, let $\mathbb{P}$ be the underlying probability distribution over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} = \{+1, -1\}$, namely, all problems are binary. Assume that sample $\mathcal{S} \coloneqq \{(\mathbf{x}_i, y_i)\}_{i \in [n]} \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ is given.

Binary classification. The goal is to learn a hypothesis $g : \mathcal{X} \to \{+1, -1\}$. The target loss is the 0-1 loss $\ell_{01}(g(\mathbf{x}), y) = \mathbb{1}_{\{g(\mathbf{x}) \neq y\}}$, and the associated 0-1 risk is denoted by $R_{01}(g) \coloneqq \mathbb{E}\left[\ell_{01}(g(\mathsf{X}), \mathsf{Y})\right]$.

Bipartite ranking [Menon and Williamson, 2016]. The goal is to learn a ranking model $f : \mathcal{X} \to \mathbb{R}$ that outputs a higher prediction score for positive data and a lower prediction score for negative data. The target loss is the *rank loss*

$$\ell_{\text{rank}}(f(\mathbf{x}), f(\mathbf{x}'), y, y') := \mathbb{1}_{\{(y-y')(f(\mathbf{x})-f(\mathbf{x})')<0\}} + \frac{1}{2}\mathbb{1}_{\{f(\mathbf{x})=f(\mathbf{x}')\}},$$

and the associated rank risk is

$$R_{\text{rank}}(f) := \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}),(\mathsf{X}',\mathsf{Y}')\sim\mathbb{P}} \left[\ell_{\text{rank}}(f(\mathsf{X}), f(\mathsf{X}'), \mathsf{Y}, \mathsf{Y}') \mid \mathsf{Y} \neq \mathsf{Y}'\right].$$

The optimal ranking function is any function that is strictly monotonically increasing with respect to $\mathbb{P}(\mathsf{Y} = +1 \mid \mathbf{x})$. The rank risk is equivalent to one minus the area under the receiver operating characteristic (ROC) curve (AUC) and hence bipartite ranking is the same problem as AUC optimization.

Binary CPE [Buja et al., 2005]. The goal is to learn a CPE model $\widehat{\eta} : \mathcal{X} \to [0, 1]$. Despite that many proper losses can be used as a target loss, herein we focus on the *squared loss*

$$\ell_{\text{CPE}}(\widehat{\eta}(\mathbf{x}), y) := \left(\widehat{\eta}(\mathbf{x}) - \frac{y+1}{2}\right)^2,$$

and the associated risk is $R_{\text{CPE}}(\widehat{\eta}) := \mathbb{E}\left[\ell_{\text{CPE}}(\widehat{\eta}(\mathsf{X}), \mathsf{Y})\right]$. The optimal CPE model is clearly $\mathbb{P}(\mathsf{Y} = +1 \mid \mathbf{x})$.

2.5.2 Reduction to Binary CPE

Once a CPE model $\widehat{\eta}$ is learned, it is ready to be used for binary classification and bipartite ranking. Indeed, $\widehat{\eta} - \frac{1}{2}$ serves as a binary classifier because we know that the Bayes classifier is of the form $\mathbb{P}(\mathsf{Y} = +1 \mid \mathbf{x}) - \frac{1}{2}$, and $\widehat{\eta}$ directly serves as a ranking function. These reductions are justified through the following excess risk bounds [Narasimhan and Agarwal, 2013]:

$$\begin{aligned} &\text{(classification} \to \text{CPE)} \quad R_{01}\left(\operatorname{sgn}\left(\widehat{\eta} - \tfrac{1}{2}\right)\right) - R_{01}^* \leq \sqrt{R_{\text{CPE}}(\widehat{\eta}) - R_{\text{CPE}}^*}, \\ &\text{(ranking} \to \text{CPE)} \quad R_{\text{rank}}(\widehat{\eta}) - R_{\text{rank}}^* \leq \frac{1}{\pi(1-\pi)}\sqrt{R_{\text{CPE}}(\widehat{\eta}) - R_{\text{CPE}}^*}, \end{aligned}$$

where $\pi := \mathbb{P}(\mathsf{Y} = +1)$. Note that the excess risk bound (classification → CPE) can be obtained as a corollary of Theorem 2.15 because ℓ_{CPE} is proper (and equivalent to the squared loss).

2.5.3 Reduction from Binary Classification to Bipartite Ranking

Recall that the Bayes classifier is $\operatorname{sgn}\left(\mathbb{P}(\mathsf{Y} = +1 \mid \mathbf{x}) - \frac{1}{2}\right)$, whereas the optimal ranking function is any strictly increasing transform of $\mathbb{P}(\mathsf{Y} = +1 \mid \mathbf{x})$. One may imagine that a good binary classifier is obtained once a good ranking function is learned by thresholding the ranking function at a certain point. This idea is formalized below.

Given a distribution $\mathbb{P}$ over $\mathcal{X} \times \mathcal{Y}$ and a ranking function $f : \mathcal{X} \to \mathbb{R}$, we define the optimal classification transform Thresh by

$$\mathsf{Thresh}(f) := \mathop{\arg\min}_{\theta \in \mathcal{F}_{\text{thresh}}} R_{01}(\theta \circ f),$$

where

$$\mathcal{F}_{\text{thresh}} := \left\{ \theta \in \{\pm 1\}^{\mathbb{R}} \;\middle|\; \theta(u) = \operatorname{sgn}(u-t), \overline{\operatorname{sgn}}(u-t) \text{ for some } t \in [-\infty, \infty] \right\}.$$

Then, the excess risk reduction is established.

Proposition 2.16 (Narasimhan and Agarwal [2013])**.** *Given any ranking function $f : \mathcal{X} \to \mathbb{R}$, assume that the pushforward $f_\sharp \mathbb{P}_{\mathsf{X}}$ is either discrete, continuous, or mixed with at most finitely many point masses. Then,*

$$R_{01}\left(\mathsf{Thresh} \circ f\right) - R_{01}^* \leq \sqrt{\pi(1-\pi)\left\{R_{\text{rank}}(f) - R_{\text{rank}}^*\right\}},$$

where $\pi := \mathbb{P}(\mathsf{Y} = +1)$.

This reduction (classification → ranking) requires access to the underlying distribution $\mathbb{P}$ when we obtain Thresh, unlike the reductions {classification, ranking} → CPE. Hence, the reduction from classification to ranking is said to be *weak* compared with a *strong* reduction such as the reductions {classification, ranking} → CPE. In practice, it is still possible to construct Thresh by splitting $\mathcal{S}$ into a training sample and validation sample for bipartite ranking and determining the threshold, respectively.

2.5.4 Reduction from Binary CPE to Bipartite Ranking

Recall that the optimal ranking function is any strictly increasing transform of $\mathbb{P}(\mathsf{Y} = +1 \mid \mathbf{x})$, which is the target CPE model. If we can find an inverse transform from a ranking function to a CPE model, then the reduction from bipartite ranking to binary CPE is possible.

Given a distribution $\mathbb{P}$ over $\mathcal{X} \times \mathcal{Y}$ and a ranking function $f : \mathcal{X} \to [a, b]$, define the optimal CPE transform Cal by

$$\mathsf{Cal}(f) := \operatorname*{arg\,min}_{\theta \in \mathcal{F}_{\text{cal}}} R_{\text{CPE}}(\theta \circ f),$$

where

$$\mathcal{F}_{\text{cal}} := \left\{ \theta \in [0,1]^{\mathbb{R}} \;\middle|\; \theta \text{ is monotonically increasing} \right\}.$$

The excess risk reduction is then established.

Proposition 2.17 (Narasimhan and Agarwal [2013])**.** *Given any ranking function $f : \mathcal{X} \to [a, b]$, assume that the pushforward $f_\sharp \mathbb{P}_{\mathsf{X}}$ is either discrete, continuous, or mixed with at most finitely many point masses without any masses at a and b. We further assume that $\eta_f : [a,b] \to [0,1]$ is square-integrable with respect to the density of the continuous part of $f_\sharp \mathbb{P}_{\mathsf{X}}$, where $\eta_f(t) := \mathbb{P}(\mathsf{Y} = +1 \mid f(\mathbf{x}) = t)$. Then,*

$$R_{\text{CPE}}\left(\mathsf{Cal} \circ f\right) - R_{\text{CPE}}^* \leq \sqrt{8\pi(1-\pi)\left\{R_{\text{rank}}(f) - R_{\text{rank}}^*\right\}},$$

where $\pi := \mathbb{P}(\mathsf{Y} = +1)$.

Remark that this reduction is also weak and hence requires access to $\mathbb{P}$ when we learn Cal. Again, it is possible to prepare a validation sample to learn Cal. In practice, isotonic regression [Kalai and Sastry, 2009] can be used for learning Cal.

Table 2.4: Relationship between binary problems.

Target		Surrogate	Reduction
classification	→	classification-calibrated loss	strong
classification	→	proper loss	strong
classification	→	CPE	strong
classification	→	ranking	weak
ranking	→	classification	N/A
ranking	→	CPE	strong
CPE	→	classification	N/A
CPE	→	ranking	weak
ranking	→	strongly proper loss	strong
linear-fractional metric	→	strongly proper loss	weak

2.5.5 Relationship between Binary Problems

In Sections 2.5.2 to 2.5.4, several examples show that a target learning task can be reduced to a different learning task through an excess risk bound; for instance, binary classification can be solved by using a good bipartite ranking function. Under this learning task reduction, we call a problem reduced from a target task a *surrogate learning task*. The reduction relationships are summarized in Table 2.4. We can also include excess risk bounds of classification-calibrated losses and proper losses (Section 2.4) into this table because all target and surrogate learning tasks are defined only through loss functions. Importantly, *every learning task has a one-to-one relationship with a loss function.*

In the existing literature, Agarwal [2014] introduced a loss function class called *strongly proper losses*, which is smaller than that of proper losses and convenient from the perspective of a proof. They also showed an excess risk bound for bipartite ranking. Kotlowski and Dembczyński [2016] dealt with a *linear-fractional metric*, i.e., target loss functions different from the 0-1 loss, which will be discussed in Chapter 3, and demonstrated that a good classifier in terms of a linear-fractional metric can be obtained by optimizing a strongly proper loss. These results can also be included in Table 2.4.

2.6 Summary

In this chapter, the basic formulation of supervised learning was introduced from the viewpoints of loss and risk functions (Section 2.1), and two conceptually orthogonal approaches of learning theory were stated, i.e., generalization analysis (Section 2.2) and calibration analysis (Section 2.3). Whereas the former studies the relationship between empirical and expected risks, the latter focuses on the gap between surrogate and target risks. In the end, generalization and calibration analyses are to be combined—we are ultimately interested in the convergence of an empirical surrogate risk to the expected target risk, which can be analyzed by investigating the gap between the empirical and expected surrogate risks (generalization analysis), as well as the gap between the expected surrogate and target risks (calibration analysis). To illustrate the calibration analysis, we showed surrogates from two different perspectives: surrogate loss functions designed for a specific target problem (Section 2.4) and connections between different learning tasks (Section 2.5). Although these two perspectives may look distinct at a glance, we stress that there is no need to distinguish them because a learning task is tied to a loss function and can be treated as a surrogate. Finally, in Table 2.4, we summarized that many learning problems can be reduced to/from each other through an excess risk transfer.

Chapter 3

Calibrated Surrogate Losses for Linear-fractional Metrics

To handle class-imbalanced cases such as information retrieval and image segmentation, complex classification performance metrics such as the F_β-measure and Jaccard index are often used. These performance metrics are not decomposable, that is, they cannot be expressed in a per-example manner, which hinders a straightforward application of the M-estimation widely used in supervised learning. In this chapter, we consider *linear-fractional metrics*, which are a family of classification performance metrics that encompasses many standard metrics such as the F_β-measure and Jaccard index, and propose methods for directly maximizing the performances under such metrics. A clue tackling their direct optimization is a *calibrated surrogate utility*, which is a tractable lower bound of the true utility function representing a given metric. We characterize sufficient conditions that make the surrogate maximization coincide with the maximization of the true utility. Simulation results on benchmark datasets validate the effectiveness of our calibrated surrogate maximization, particularly if the sample sizes are extremely small.

3.1 Introduction

Binary classification, one of the main focuses in machine learning, is a problem of predicting the binary responses for the input covariates. Classifiers are usually evaluated based on the *classification accuracy*, which is the expected proportion of correct predictions. Because the accuracy cannot evaluate the classifiers appropriately under a class imbalance [Menon et al., 2013] or in the presence of label noises [Menon et al., 2015], alternative performance metrics have been employed such as the F_β-measure [van Rijsbergen, 1974, Joachims, 2005, Nan et al., 2012, Koyejo et al., 2014], Jaccard index [Koyejo et al., 2014, Berman et al., 2018], and balanced error rate (BER) [Brodersen et al., 2010, Menon et al., 2013, 2015, Charoenphakdee et al., 2019]. Once a performance metric is given, it is a natural strategy to optimize the performance of classifiers directly under the given performance metric. However, alternative performance metrics generally have difficulties in terms of direct optimization, because they are non-decomposable, for which a per-example loss decomposition is unavailable. In other words, the M-estimation procedure [van de Geer, 2000] cannot be applied, which makes it difficult to optimize the non-decomposable metrics.

One of the earliest studies tackling the non-traditional metrics [Koyejo et al., 2014] generalized the performance metrics into the linear-fractional metrics, which are the linear-fractional form of entries in a confusion matrix, and encompasses the BER, F_β-measure, Jaccard index, Gower-Legendre index [Gower and Legendre,

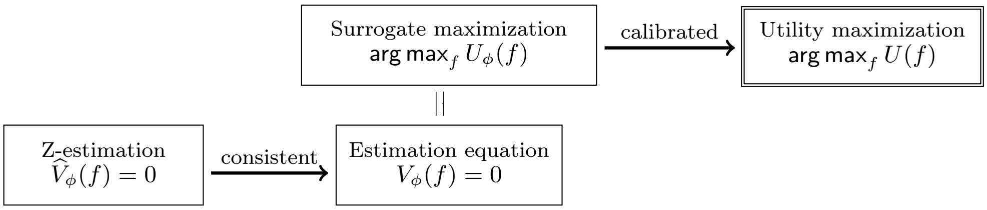


Figure 3.1: Overview of this chapter. Intuitively, we can obtain the utility maximizer by solving $\widehat{V}_\phi(f) = 0$.

1986, Natarajan et al., 2016], and weighted accuracy [Koyejo et al., 2014]. In addition, Koyejo et al. [2014] formulated the optimization problem in two ways. One is a plug-in rule [Koyejo et al., 2014, Narasimhan et al., 2014, Yan et al., 2018] to estimate the class-posterior probability and its optimal threshold, and the other is an iterative weighted empirical risk minimization approach [Koyejo et al., 2014, Parambath et al., 2014] to find a better cost through which the minimizer of the cost-sensitive risk [Scott, 2012] achieves higher utilities. Although they provide statistically consistent estimators, the former suffers from a high sample complexity owing to the class-posterior probability estimation, whereas the latter is computationally demanding because of iterative classifier training.

Without sacrificing the statistical consistency, our goal is to seek computationally more efficient procedures to directly optimize the linear-fractional metrics. Specifically, we provide a novel calibrated surrogate utility, which is a tractable lower bound of the true utility representing the metric of interest. The surrogate maximization is formulated as a combination of concave and quasiconcave programs, which can be efficiently optimized. We then derive sufficient conditions of the surrogate calibration, under which the surrogate maximization implies the maximization of the true utility. In addition, we show the consistency of the empirical estimation procedure based on the theory of a Z-estimation [van der Vaart, 2000]. An overview of our proposed method is illustrated in Figure 3.1.

3.1.1 Contributions of this Chapter

The contributions of this chapter are summarized below.[1]

1. *Surrogate calibration* (Section 3.4): We propose a tractable lower bound of the linear-fractional metrics with calibration conditions, guaranteeing that the surrogate maximization implies the maximization of the true utility. This approach is model-agnostic, differing from many previous approaches [Koyejo et al., 2014, Narasimhan et al., 2014, 2015, Yan et al., 2018], in the sense that our classifier is not restricted to any specific plug-in forms.

2. *Efficient gradient-based optimization* (Sections 3.3.2 and 3.3.3): The surrogate utility has affinity with gradient-based optimization owing to its non-vanishing gradient and unbiased estimator of the gradient direction. Although the linear-fractional objective does not admit concavity in general, our proposed algorithm is a two-step approach combining concave and quasiconcave programs and hence is computationally efficient.

[1] In this chapter, we refer to the convergence of a surrogate optimizer to the target metric optimizer as surrogate calibration, whereas the convergence of the empirical optimizer to the theoretical optimizer is called consistency, to distinguish two notion clearly.

3. *Consistency analysis* (Section 3.5): The estimator obtained through the surrogate maximization with a finite sample is shown to be consistent with the maximizer of the expected utility.

3.2 Preliminaries

Throughout this chapter, we focus on binary classification. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a feature space and $\mathcal{Y} = \{\pm 1\}$ be the label space. We assume that a sample $\mathcal{S} := \{(\mathbf{x}_i, y_i)\}_{i=1}^n \subseteq \mathcal{X} \times \mathcal{Y}$ independently follows the joint distribution $\mathbb{P}$ with a density p. We often split $\mathcal{S}$ into two independent samples $\mathcal{S}_0 = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ and $\mathcal{S}_1 = \{(\mathbf{x}_i, y_i)\}_{i=m+1}^n$. Usually, $m = \lfloor \frac{n}{2} \rfloor$. A classifier is given as a function $f : \mathcal{X} \to \mathbb{R}$, where $\operatorname{sgn}(f(\cdot))$ determines the predictions. Here, we adopt the convention $\operatorname{sgn}(0) = -1$. Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ be a hypothesis space of classifiers. In addition, let $\pi := p(\mathsf{Y} = +1)$ and $\eta(\mathbf{x}) := p(\mathsf{Y} = +1 \mid \mathsf{X} = \mathbf{x})$ be the class-prior/-posterior probabilities of $Y = +1$, respectively. The 0-1 loss is denoted by $\ell(m) := \mathbb{1}_{\{m \le 0\}}$, whereas $\phi : \mathbb{R} \to \mathbb{R}_{\ge 0}$ denotes a margin-based surrogate loss. For a set $\mathcal{A} \subseteq \mathcal{F}$, $\mathcal{A}^{\complement}$ denotes the complement of $\mathcal{A}$, namely, $\mathcal{A}^{\complement} := \mathcal{F} \setminus \mathcal{A}$.

The following four quantities are focal targets in binary classification: the true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN).

Definition 3.1 (Confusion matrix)**.** *Given a classifier $f \in \mathcal{F}$ and a distribution $\mathbb{P}$, its confusion matrix is defined as $C(f, \mathbb{P}) := [\mathsf{TP}, \mathsf{FN}; \mathsf{FP}, \mathsf{TN}]$, where*

$$\begin{aligned}
\mathsf{TP}(f, \mathbb{P}) &:= \mathbb{P}(\mathsf{Y} = +1, \operatorname{sgn}(f(\mathsf{X})) = +1), \\
\mathsf{FN}(f, \mathbb{P}) &:= \mathbb{P}(\mathsf{Y} = +1, \operatorname{sgn}(f(\mathsf{X})) = -1), \\
\mathsf{FP}(f, \mathbb{P}) &:= \mathbb{P}(\mathsf{Y} = -1, \operatorname{sgn}(f(\mathsf{X})) = +1), \\
\mathsf{TN}(f, \mathbb{P}) &:= \mathbb{P}(\mathsf{Y} = -1, \operatorname{sgn}(f(\mathsf{X})) = -1).
\end{aligned}$$

Here, FN and TP, and TN and FP, can be transformed into the other, i.e., $\mathsf{FN}(f, \mathbb{P}) = \pi - \mathsf{TP}(f, \mathbb{P})$ and $\mathsf{TN}(f, \mathbb{P}) = (1 - \pi) - \mathsf{FP}(f, \mathbb{P})$. These can be expressed with ℓ and η, such as $\mathsf{TP}(f, \mathbb{P}) = \mathbb{E}[\ell(-f(\mathsf{X}))\eta(\mathsf{X})]$. The goal of binary classification is to obtain a classifier that "maximizes" TP and TN, while keeping FP and FN as "low" as possible. Classifiers are evaluated based on performance metrics that have a trade off with those four quantities. Performance metrics need to be chosen based on user preference on the confusion matrix [Sokolova and Lapalme, 2009, Menon et al., 2015]. In this chapter, we focus on the following family of utilities representing the linear-fractional metrics.

Definition 3.2 (Linear-fractional metrics)**.** *A* linear-fractional metrics $U : \mathcal{F} \to [0, 1]$ *is defined as*[2]

$$U(f) := \frac{\mathbb{E}[W_0(f(\mathsf{X}), \eta(\mathsf{X}))]}{\mathbb{E}[W_1(f(\mathsf{X}), \eta(\mathsf{X}))]}, \tag{3.1}$$

where $W_0, W_1 : \mathbb{R} \times [0, 1] \to \mathbb{R}$ are class-conditional score functions given as

$$W_k(\xi, q) := a_{k,+1}\ell(-\xi)q + a_{k,-1}\ell(-\xi)(1 - q) + b_k,$$

[2] As mentioned by Dembczyński et al. [2017], there is a dichotomy in the definition of the performance metrics, i.e., the population utility (PU) and expected test utility (ETU). The PU is a functional transform of the expected confusion matrix, whereas the ETU evaluates the utility over a fixed-size test set. Here, the PU is adopted for the definition of the linear-fractional utilities because we are partly interested in statistical consistency, namely, the behavior of the empirical utility optimizer for sufficiently large n. See Dembczyński et al. [2017] for the precise definitions of the PU and ETU and further discussions.

Table 3.1: Examples of the linear-fractional performance metrics. $\beta > 0$ is a trade-off parameter for the F_β-measure, whereas $\alpha \geq 0$ is for the Gower-Legendre index.

Metric	F_β-measure [van Rijsbergen, 1974]	Jaccard index [Jaccard, 1901]	Gower-Legendre index [Gower and Legendre, 1986]
Definition	$\frac{(1+\beta^2)\mathsf{TP}}{(1+\beta^2)\mathsf{TP}+\beta^2\mathsf{FN}+\mathsf{FP}}$	$\frac{\mathsf{TP}}{\mathsf{TP}+\mathsf{FN}+\mathsf{FP}}$	$\frac{\mathsf{TP}+\mathsf{TN}}{\mathsf{TP}+\alpha(\mathsf{FP}+\mathsf{FN})+\mathsf{TN}}$
$(a_{0,+1}, a_{0,-1})$	$(1+\beta^2, 0)$	$(1, 0)$	$(1, -1)$
b_0	0	0	$1-\pi$
$(a_{1,+1}, a_{1,-1})$	$(1, 1)$	$(0, 1)$	$(1-\alpha, \alpha-1)$
b_1	$\beta^2\pi$	π	$1+(\alpha-1)\pi$

and $a_{0,+1} > 0, a_{0,-1} \leq 0, b_0 \in \mathbb{R}, a_{1,+1} \geq 0, a_{1,-1} \geq 0, b_1 \in \mathbb{R}$ *are constants such that* $0 \leq U(f) \leq 1$ *for any* f*.*

The class-conditional score functions correspond to a linear-transformation of TP and FP: $\mathbb{E}[W_k(f(\mathsf{X}), \eta(\mathsf{X}))] = a_{k,+1}\mathsf{TP}(f, \mathbb{P}) + a_{k,-1}\mathsf{FP}(f, \mathbb{P}) + b_k$. Examples of U are shown in Table 3.1. Given a utility function U, our goal is to obtain a classifier $f^\dagger$ that maximizes U.

$$f^\dagger = \arg\max_{f \in \mathcal{F}} U(f). \tag{3.2}$$

3.2.1 Traditional Supervised Classification

Here, we briefly review a traditional procedure for supervised classification [Vapnik, 1998]. Our aim is to obtain a classifier with high accuracy, which corresponds to minimizing the classification risk $R(f) := \mathbb{E}[\ell(\mathsf{Y}f(\mathsf{X}))]$. Because optimizing the 0-1 loss ℓ is a computationally infeasible problem [Ben-David et al., 2003, Feldman et al., 2012], it is a common practice to instead minimize a surrogate risk $R_\phi(f) := \mathbb{E}[\phi(\mathsf{Y}f(\mathsf{X}))]$, where $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is a margin-based surrogate loss. If ϕ is a classification-calibrated loss [Bartlett et al., 2006], it is known that minimizing R_ϕ corresponds to minimizing R. Eventually, what we actually minimize is the empirical (surrogate) risk $\widehat{R}_\phi(f) := \frac{1}{n}\sum_{i=1}^n \phi(y_i f(\mathbf{x}_i))$. The empirical risk $\widehat{R}_\phi(f)$ is an unbiased estimator of the true risk $R_\phi(f)$ for a fixed $f \in \mathcal{F}$, and the uniform law of large numbers guarantees that $\widehat{R}_\phi(f)$ converges to $R_\phi(f)$ for any $f \in \mathcal{F}$ in probability [Vapnik, 1998, van de Geer, 2000, Mohri et al., 2018]. This strategy to minimize $\widehat{R}_\phi$ is called empirical risk minimization (ERM).

Traditional ERM is devoted to maximizing the accuracy, which is not necessarily suitable when another metric is used for evaluation. Our aim is to give an alternative procedure to maximize U directly as in Equation (3.2). In the next section, we introduce a tractable counterpart of the true utility U because U contains the 0-1 loss ℓ and is intractable, similar to R_ϕ above.

3.3 Surrogate Utility and Optimization

The true utility in Equation (3.1) consists of the 0-1 loss ℓ, which is difficult to optimize. In this section, we introduce a surrogate utility to make the optimization problem in Equation (3.2) easier.

3.3.1 Lower Bounding True Utility

Assume that we are given a surrogate loss $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$. We hope that the surrogate utility should be a lower bound of the true utility U, and that TP/FP should become larger/smaller as a result of the optimization, respectively. We realize these ideas by constructing surrogate class-conditional score functions $W_{0,\phi}$ and $W_{1,\phi}$ as follows:

$$\begin{aligned} W_{0,\phi}(\xi, q) &\coloneqq a_{0,+1}(1-\phi(\xi))q + a_{0,-1}\phi(-\xi)(1-q) + b_0, \\ W_{1,\phi}(\xi, q) &\coloneqq a_{1,+1}(1+\phi(\xi))q + a_{1,-1}\phi(-\xi)(1-q) + b_1. \end{aligned} \tag{3.3}$$

When clear from the context, we often abbreviate $\mathbb{E}[W_{k,\phi}(f(\mathsf{X}), \eta(\mathsf{X}))]$ as $\mathbb{E}[W_{k,\phi}]$. Given the surrogate class-conditional scores, we define the surrogate utility as follows:

$$U_\phi(f) \coloneqq \frac{\mathbb{E}_X[W_{0,\phi}(f(\mathsf{X}), \eta(\mathsf{X}))]}{\mathbb{E}_X[W_{1,\phi}(f(\mathsf{X}), \eta(\mathsf{X}))]} = \frac{\mathbb{E}[W_{0,\phi}]}{\mathbb{E}[W_{1,\phi}]}. \tag{3.4}$$

To construct U_ϕ, the 0-1 losses appearing in the true utility U are replaced with the surrogate loss ϕ. The surrogate class-conditional scores in Equation (3.3) are designed such that the surrogate utility in Equation (3.4) bounds U from below.

Lemma 3.3. *For all f and a surrogate loss $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ such that $\phi(m) \geq \ell(m)$ for all $m \in \mathbb{R}$, $U_\phi(f) \leq U(f)$.*

Proof. Fix $\xi \in \mathbb{R}$ and $q \in [0,1]$. Because $\ell(-\xi) = 1 - \ell(\xi)$, $a_{0,+1}\ell(-\xi) = a_{0,+1}(1 - \ell(\xi)) \geq a_{0,+1}(1-\phi(\xi))$ ($\because a_{0,+1} \geq 0$). Together with $a_{0,-1}\ell(-\xi) \geq a_{0,-1}\phi(-\xi)$ ($\because a_{0,-1} \leq 0$), we confirm $W_0(\xi, q) \geq W_{0,\phi}(\xi, q)$. It can be confirmed that $W_1(\xi, q) \leq W_{1,\phi}(\xi, q)$ as well. Hence, $U(f) \geq U_\phi(f)$ is easy to see. □

Owing to this property, maximizing U_ϕ is at least maximizing a lower bound of U. We will discuss the goodness of this lower bound in Section 3.4; however, we can immediately see $U_\phi(f)(\leq U(f)) \leq 1$ for any f. In the rest of this chapter, we assume that U_ϕ is Fréchet differentiable.

3.3.2 Tractability of Surrogate Utility

The surrogate utility U_ϕ comes to have a non-vanishing gradient by using a smooth ϕ, and is guaranteed to be a lower bound of U. In this subsection, we discuss how it can be efficiently maximized.

Let us consider an empirical estimator of U_ϕ:

$$\widehat{U}_\phi(f) = \frac{\frac{1}{m}\sum_{i=1}^{m} \widetilde{W}_{0,\phi}(f(\mathbf{x}_i), y_i)}{\frac{1}{n-m}\sum_{i=m+1}^{n} \widetilde{W}_{1,\phi}(f(\mathbf{x}_i), y_i)}, \tag{3.5}$$

where

$$\widetilde{W}_{0,\phi}(\xi, y) \coloneqq \begin{cases} a_{0,+1}(1-\phi(\xi)) + b_0 & \text{if } y = +1, \\ a_{0,-1}\phi(-\xi) + b_0 & \text{if } y = -1, \end{cases}$$

$$\widetilde{W}_{1,\phi}(\xi, y) \coloneqq \begin{cases} a_{1,+1}(1+\phi(\xi)) + b_1 & \text{if } y = +1, \\ a_{1,-1}\phi(-\xi) + b_1 & \text{if } y = -1. \end{cases}$$

A global maximizer of $\widehat{U}_\phi$ could be efficiently obtained if $\widehat{U}_\phi$ was concave. However, this is difficult to achieve in our case regardless of the choice of ϕ owing to its fractional form. Nonetheless, we may formulate our optimization problem as a *quasiconcave* program under a certain condition. First, we introduce the notion of quasiconcavity.

Definition 3.4 (Quasiconcavity [Boyd and Vandenberghe, 2004])**.** *A function $h : A \to \mathbb{R}$ is said to be quasiconcave if the super-level set $\{x \in A \mid h(x) \geq \alpha\}$ is a convex set for $\forall \alpha \in \mathbb{R}$.*

A quasiconcave function is a generalization of a concave function and has the unimodality despite not necessarily being concave, which ensures the uniqueness of the solution. We then show that the surrogate utility $\widehat{U}_\phi$ is quasiconcave in a subset of the domain. Let

$$\widehat{U}^{\mathrm{n}}_\phi(f) \coloneqq \frac{1}{m}\sum_{i=1}^{m} \widetilde{W}_{0,\phi}(f(\mathbf{x}_i), y_i)$$

be the numerator of $\widehat{U}_\phi$.

Lemma 3.5. *Let $\bar{\mathcal{F}} \coloneqq \left\{ f \in \mathcal{F} \,\middle|\, \widehat{U}^{\mathrm{n}}_\phi(f) \geq 0 \right\}$. If ϕ is convex, $\widehat{U}_\phi$ in Equation (3.5) is quasiconcave over $\bar{\mathcal{F}}$ and $\widehat{U}^{\mathrm{n}}_\phi$ is concave over $\mathcal{F}$.*

Proof. Define an α-super-level set of $\widehat{U}_\phi$ restricted in $\bar{\mathcal{F}}$ as

$$\mathcal{A}_\alpha \coloneqq \left\{ f \in \bar{\mathcal{F}} \,\middle|\, \widehat{U}_\phi(f) \geq \alpha \right\}.$$

It is sufficient to show that $\mathcal{A}_\alpha$ is a convex set for any $\alpha \geq 0$ owing to $f \in \bar{\mathcal{F}}$.

Fix any $\alpha \geq 0$. Then,

$$\begin{aligned} \widehat{U}_\phi(f) \geq \alpha &\iff \frac{\frac{1}{m}\sum_{i=1}^{m} \widetilde{W}_{0,\phi}(f(\mathbf{x}_i), y_i)}{\frac{1}{n-m}\sum_{j=m+1}^{n} \widetilde{W}_{1,\phi}(f(\mathbf{x}_j), y_j)} \geq \alpha \\ &\iff \underbrace{\frac{1}{m}\sum_{i=1}^{m} \widetilde{W}_{0,\phi}(f(\mathbf{x}_i), y_i) - \alpha\frac{1}{n-m}\sum_{j=m+1}^{n} \widetilde{W}_{1,\phi}(f(\mathbf{x}_j), y_j)}_{(*)} \geq 0. \end{aligned}$$

Here, $\frac{1}{m}\sum_{i=1}^{m} \widetilde{W}_{0,\phi}(f(\mathbf{x}_i), y_i)$ is concave in f because it is a non-negative sum of concave functions. Note that $\widetilde{W}_{0,\phi}(f(\mathbf{x}_i), y_i)$ is concave in f for any $(\mathbf{x}_i, y_i)$ owing to the definition of $\widetilde{W}_{0,\phi}$ in Equation (3.3) and the assumption ϕ is convex. Similarly, $\frac{1}{n-m}\sum_{j=m+1}^{n} \widetilde{W}_{1,\phi}(f(\mathbf{x}_j), y_j)$ is convex as well. Thus, $(*)$ is concave in f, which means that $\mathcal{A}_\alpha$ is a convex set because any super-level set of a concave function is convex.

Hence, we confirm that $\mathcal{A}_\alpha$ is convex for any $\alpha \geq 0$. □

From Lemma 3.5, we observe the following two important facts: First, within the range of $f \notin \bar{\mathcal{F}}$, our objective $\widehat{U}_\phi$ is generally neither concave nor quasiconcave, but its numerator $\widehat{U}^{\mathrm{n}}_\phi$ is concave. Second, $\widehat{U}_\phi$ is quasiconcave over $\bar{\mathcal{F}}$. These observations motivate us to employ Algorithm 3.1, which first increases the numerator $\widehat{U}^{\mathrm{n}}_\phi$ only to make it positive, and then maximizes the fractional form $\widehat{U}_\phi$. Because the former is a concave program and the latter is a quasiconcave program within $\bar{\mathcal{F}}$, the entire optimization can be conducted in a computationally efficient manner. For quasiconcave optimization, a normalized gradient ascent (NGA) [Hazan et al., 2015] is applied, which is guaranteed to find a global solution to the quasiconcave objectives. The behavior of Algorithm 3.1 is illustrated in Figure 3.2.

Algorithm 3.1: Hybrid Optimization Algorithm

```
Input  : $\phi$ convex loss, $\boldsymbol{\theta}$ initial classifier parameter
while $\widehat{U}^{\mathrm{n}}_{\phi}(f_{\boldsymbol{\theta}}) \leq 0$ do
    $\mathrm{g}^{\mathrm{n}} \longleftarrow \nabla_{\boldsymbol{\theta}} \widehat{U}^{\mathrm{n}}_{\phi}(f_{\boldsymbol{\theta}})$
    $\boldsymbol{\theta} \longleftarrow$ gradient_based_update$(\boldsymbol{\theta}, \mathrm{g}^{\mathrm{n}})$
end
while stopping criterion is not satisfied do
    $\mathrm{g} \longleftarrow \nabla_{\boldsymbol{\theta}} \widehat{U}_{\phi}(f_{\boldsymbol{\theta}})$, $\widehat{\mathrm{g}} = \mathrm{g}/\|\mathrm{g}\|$
    $\boldsymbol{\theta} \longleftarrow$ gradient_based_update$(\boldsymbol{\theta}, \widehat{\mathrm{g}})$
end
Output: maximizer $f_{\boldsymbol{\theta}}$
```

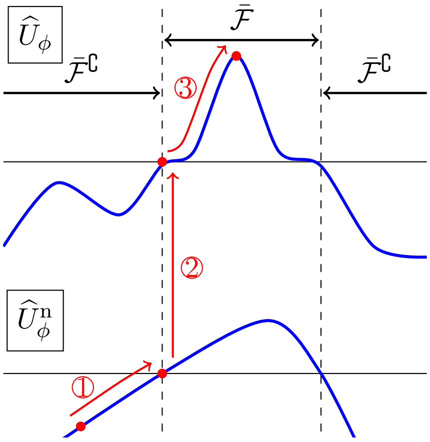

$\widehat{U}_\phi$
$\bar{\mathcal{F}}$
$\bar{\mathcal{F}}^\complement$
$\bar{\mathcal{F}}^\complement$
③
②
$\widehat{U}^{\mathrm{n}}_\phi$
①

Figure 3.2: Illustration of our hybrid optimization approach in Algorithm 3.1. ① maximize the numerator $\widehat{U}^{\mathrm{n}}_\phi$ (concave), ② once $\widehat{U}^{\mathrm{n}}_\phi(f) \geq 0$, optimize the fraction $\widehat{U}_\phi$, ③ maximize the fraction $\widehat{U}_\phi$ (quasiconcave only in $\bar{\mathcal{F}}$).

3.3.3 Gradient Direction Estimator

The empirical estimator $\widehat{U}_\phi$ in Equation (3.5) is generally biased owing to its fractional form. Nevertheless, its gradient $\nabla_f \widehat{U}_\phi$ is unbiased to the expected gradient $\nabla_f U_\phi$ up to a positive scalar multiple. Hence, we may safely use $\nabla_f \widehat{U}_\phi$ as the update direction in the NGA.

We state this idea formally below. Under the interchangeability of the expectation and derivative, the gradient of the expected utility U_ϕ is expressed as

$$\begin{aligned}\nabla_f U_\phi(f) &= \underbrace{\frac{1}{(\mathbb{E}[W_{1,\phi}])^2}}_{\text{positive scalar}} \underbrace{\mathbb{E}[W_{1,\phi}]\,\mathbb{E}[\nabla W_{0,\phi}] - \mathbb{E}[W_{0,\phi}]\,\mathbb{E}[\nabla W_{1,\phi}]}_{\text{gradient direction } (:= V_\phi(f))} \\ &= c V_\phi(f), \qquad \text{where } c = (\mathbb{E}[W_{1,\phi}])^{-2} > 0,\end{aligned}$$

from which we can see that its gradient direction is parallel to V_ϕ. In addition, V_ϕ can be unbiasedly estimated.

Lemma 3.6. *Denote $\widetilde{W}_{0,\phi}(f(\mathbf{x}_i), y_i) = \widetilde{W}_{0,\phi}(z_i)$ for simplicity. Define*

$$\widehat{V}_\phi(f) := \frac{1}{m(n-m)} \sum_{i=1}^{m} \sum_{j=m+1}^{n} \left\{ \widetilde{W}_{1,\phi}(z_j) \nabla_f \widetilde{W}_{0,\phi}(z_i) - \widetilde{W}_{0,\phi}(z_i) \nabla_f \widetilde{W}_{1,\phi}(z_j) \right\}. \tag{3.6}$$

Algorithm 3.2: Normalized Gradient Ascent

```
Input  : $\boldsymbol{\theta}$ initial classifier parameter, $\gamma$ learning rate
while stopping criterion is not satisfied do
    $\mathbf{g} \longleftarrow \widehat{V}_\phi(f_{\boldsymbol{\theta}})$, $\widehat{\mathbf{g}} = \mathbf{g}/\|\mathbf{g}\|$
    $\boldsymbol{\theta} \longleftarrow \boldsymbol{\theta} + \gamma\widehat{\mathbf{g}}$
end
Output: learned classifier parameter $\boldsymbol{\theta}$
```

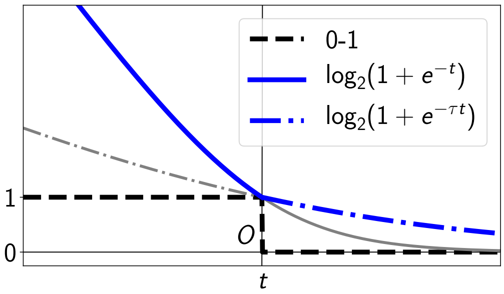


Figure 3.3: An example of τ-discrepant loss with $\tau > 0$: $\phi(m) = \log_2(1+e^{-m})$ for $m \leq 0$ and $\phi(m) = \log_2(1+e^{-\tau m})$ for $m > 0$.

We then have $V_\phi(f) = \mathbb{E}_{\mathcal{S}}[\widehat{V}_\phi(f)]$, where the expectation is taken over the sample $\mathcal{S}$.

Lemma 3.6 can be confirmed through simple algebra, noting that two samples $\mathcal{S}_0$ and $\mathcal{S}_1$ are independent and identically drawn from $\mathbb{P}$. Because solving $\nabla\widehat{U}_\phi(f) = 0$ is identical to solving $\widehat{V}_\phi(f) = 0$, gradient updates using $\nabla\widehat{U}_\phi$ are aligned to the maximization of U_ϕ. Hence, optimization procedures that only require gradients such as a gradient ascent and quasi-Newton methods [Boyd and Vandenberghe, 2004], e.g., the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [Fletcher, 2013], can be applied to maximize U_ϕ. Note that Algorithm 3.2 can be regarded as an extension of the traditional gradient ascent using $\widehat{V}_\phi$. We plug either Algorithm 3.2 or BFGS using the normalized gradient into the second half of Algorithm 3.1.

3.4 Calibration Analysis: Bridging Surrogate Utility and True Utility

In Section 3.3, we formulated the tractable surrogate utility. Given the surrogate utility U_ϕ, a natural question arises in the same way as the classification calibration in binary classification [Zhang, 2004a, Bartlett et al., 2006]: *Does maximizing the surrogate utility U_ϕ imply maximizing the true utility U?* In this section, to connect the maximization of U_ϕ and the maximization of U, we study sufficient conditions on the surrogate loss ϕ.

First, we define the notion of *U-calibration.*

Definition 3.7 (U-calibration)**.** *The surrogate utility U_ϕ is said to be U-calibrated if for any sequence of measurable functions $\{f_k\}_{k\in\mathbb{N}}$ and any distribution $\mathbb{P}$, it holds that $U_\phi(f_k) \to U_\phi^* \implies U(f_k) \to U^\dagger$ when $k \to \infty$, where $U_\phi^* := \sup_f U_\phi(f)$ and $U^\dagger := \sup_f U(f)$ are the suprema taken over all measurable functions.*

This definition is motivated by the calibration used in other learning problems such as binary classification [Bartlett et al., 2006, Theorem 3], multi-class classification [Zhang, 2004b, Theorem 3], structured prediction [Osokin et al., 2017, Theorem 2], and AUC optimization [Gao and Zhou, 2015, Definition 1]. If a surrogate utility is U-calibrated, we can safely optimize the surrogate utility instead of the true utility U. Note that U-calibration is a concept used to reduce the surrogate maximization to the maximization of U *within all measurable functions*. The approximation error of U_ϕ is not the target of our analysis, as in Bartlett et al. [2006].

Next, we give a property of loss functions that is needed to guarantee U-calibration.

Definition 3.8 (τ-discrepant loss)**.** *For a fixed $\tau > 0$, a convex margin-based loss function $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is said to be* τ-discrepant *if ϕ satisfies $\lim_{m \searrow 0} \phi'(m) \geq \tau \lim_{m \nearrow 0} \phi'(m)$.*

Intuitively, τ-discrepancy means that the gradient of ϕ around the origin is steeper in the negative domain than in the positive domain (see Figure 3.3). The value τ controls the *steepness* of the TP/FP surrogates appearing in the surrogate utility U_ϕ. Note that $\phi(\xi)$ and $\phi(-\xi)$ appearing in Equations (3.3) and (3.4) correspond to TP and FP, respectively, based on their constructions.

Below, we first provide an overview of the calibration analysis of the linear-fractional metrics. Specific linear-fractional metrics, i.e., the F_β-measure and Jaccard index, are then analyzed.

3.4.1 Overview of Calibration Analysis

Unlike the calibration analysis reviewed in Section 2.3, the analysis of the linear-fractional metrics are not straightforward. The main obstacle arises from Equation (3.1) involving a fractional form of expectations, whereas a traditional calibration analysis deals with the risk functional form $R(f) = \mathbb{E}[\ell(\mathsf{Y}f(\mathsf{X}))]$. Hence, this subsection reviews the basic idea for overcoming this hardness, and more details can be found in Section 3.8.

First, the Bayes-optimal set of classifiers is defined.

Definition 3.9 (Bayes-optimal set)**.** *Given a linear-fractional metric U, the* Bayes-optimal *set $\mathcal{B} \subseteq \mathbb{R}^{\mathcal{X}}$ is a set of functions that achieve the supremum of U, that is, $\mathcal{B} := \left\{ f \,\middle|\, U(f) = U^\dagger = \sup_{f'} U(f') \right\}$, where the supremum is taken over all measurable functions.*

Classifiers in $\mathcal{B}$ are Bayes classifiers. Note that they are not necessarily unique. In this chapter, we assume that $\mathcal{B} \neq \emptyset$, namely, maximizers of U exist.

Assumption 3.10. *For the target utility U, the associated Bayes-optimal set $\mathcal{B}$ is not empty.*

The Bayes-optimal set $\mathcal{B}$ is characterized as follows.

Proposition 3.11. *Given a linear-fractional metric U in Equation (3.1), the Bayes-optimal set $\mathcal{B}$ for U is*

$$\mathcal{B} = \{ f \mid f(\mathbf{x})\{(\Delta a_0 - \Delta a_1 U(f))\eta(\mathbf{x}) - (a_{1,-1}U(f) - a_{0,-1})\} > 0 \; \forall \mathbf{x} \in \mathcal{X} \},$$

where $\Delta a_0 := a_{0,+1} - a_{0,-1}$ and $\Delta a_1 := a_{1,+1} - a_{1,-1}$.

Proof. The maximization problem of U over all measurable functions can be restated as follows:

$$\max_{\lambda \in \Lambda} \bar{U}(\lambda) \text{ where } \bar{U}(\lambda) := \frac{\mathbb{E}[a_{0,+1}\lambda(\mathsf{X})\eta(\mathsf{X}) + a_{0,-1}\lambda(\mathsf{X})(1-\eta(\mathsf{X})) + b_0]}{\mathbb{E}[a_{1,+1}\lambda(\mathsf{X})\eta(\mathsf{X}) + a_{1,-1}\lambda(\mathsf{X})(1-\eta(\mathsf{X})) + b_1]},$$

where $\Lambda \coloneqq \{\, \mathbf{x} \mapsto \ell(-f(\mathbf{x})) \mid f\text{: measurable} \,\}$. First, the Fréchet derivative of $\bar{U}$ evaluated at x is obtained as follows.

$$
\begin{aligned}
[\nabla_\lambda \bar{U}(\lambda)]_{\mathbf{x}} &= \frac{(\Delta a_0 \eta(\mathbf{x}) + a_{0,-1})\,\mathbb{E}[W_1] - (\Delta a_1 \eta(\mathbf{x}) + a_{1,-1})\,\mathbb{E}[W_0]}{\mathbb{E}[W_1]^2} p(\mathbf{x}) \\
&= \frac{p(\mathbf{x})}{\mathbb{E}[W_1]} \left\{ \left(\Delta a_0 - \Delta a_1 \frac{\mathbb{E}[W_0]}{\mathbb{E}[W_1]} \right) \eta(\mathbf{x}) - \left(a_{1,-1} \frac{\mathbb{E}[W_0]}{\mathbb{E}[W_1]} - a_{0,-1} \right) \right\} \\
&= \frac{p(\mathbf{x})}{\mathbb{E}[W_1]} \left\{ (\Delta a_0 - \Delta a_1 \bar{U}(\lambda)) \eta(\mathbf{x}) - (a_{1,-1} \bar{U}(\lambda) - a_{0,-1}) \right\}.
\end{aligned}
$$

Let $f^\dagger$ be a function that maximizes U, and $\lambda^\dagger \coloneqq \ell(-f^\dagger)$. Such $f^\dagger$ can be chosen under Assumption 3.10. Then, $\lambda^\dagger$ maximizes $\bar{U}$, which satisfies [Koyejo et al., 2014, lemma 12]

$$
\int_{\mathcal{X}} [\nabla_\lambda \bar{U}(\lambda^\dagger)]_{\mathbf{x}} \lambda^\dagger(\mathbf{x}) \mathrm{d}\mathbf{x} \geq \int_{\mathcal{X}} [\nabla_\lambda \bar{U}(\lambda^\dagger)]_{\mathbf{x}} \lambda(\mathbf{x}) \mathrm{d}\mathbf{x} \quad \forall \lambda \in \Lambda.
$$

Thus, the necessary condition for local optimality is

$$
\operatorname{sgn}(\lambda^\dagger(\mathbf{x})) = \operatorname{sgn}([\nabla_\lambda \bar{U}(\lambda^\dagger)]_{\mathbf{x}})
$$

for all $\mathbf{x} \in \mathcal{X}$.[3] Since $\operatorname{sgn}(\lambda^\dagger(\mathbf{x})) = \operatorname{sgn}(\ell(-f^\dagger(\mathbf{x}))) = \operatorname{sgn}(f^\dagger(\mathbf{x}))$, the above condition is $\operatorname{sgn}(f^\dagger(\mathbf{x})) = \operatorname{sgn}([\nabla_\lambda \bar{U}(\lambda^\dagger)]_{\mathbf{x}})$ for all $\mathbf{x} \in \mathcal{X}$, which is equivalent to the condition $f^\dagger(\mathbf{x})\{(\Delta a_0 - \Delta a_1 U(f^\dagger))\eta(\mathbf{x}) - (a_{1,-1} U(f^\dagger) - a_{0,-1})\} > 0$ for all $\mathbf{x} \in \mathcal{X}$. This concludes the proof. Note that $p(\mathbf{x}) / \mathbb{E}[W_1]$ is a positive value, and $\bar{U}(\lambda^\dagger) = U(f^\dagger)$. □

Next, we state a proposition proving the surrogate calibration of a surrogate utility. This proposition follows the latter half of Gao and Zhou [2015, Theorem 2].

Proposition 3.12. *Fix a true utility U and a surrogate utility U_ϕ, and let $\mathcal{B}$ be the Bayes-optimal set corresponding to the utility U. In addition, assume that*

$$
\sup_{f \notin \mathcal{B}} U_\phi(f) < \sup_f U_\phi(f). \tag{3.7}
$$

The surrogate utility U_ϕ is then U-calibrated.

Proof. Recall that $U_\phi^* \coloneqq \sup_f U_\phi(f)$, and let

$$
\delta \coloneqq U_\phi^* - \sup_{f \notin \mathcal{B}} U_\phi(f) > 0,
$$

and $\{f_k\}_{k \in \mathbb{N}}$ be any sequence of measurable functions such that $U_\phi(f_k) \xrightarrow{k \to \infty} U_\phi^*$. Then, for any $\varepsilon > 0$, there exists $k_0 \in \mathbb{N}$ such that $U_\phi^* - U_\phi(f_k) < \varepsilon$ for $k \geq k_0$. Here, we set $\varepsilon = \frac{\delta}{2}$: $U_\phi^* - U_\phi(f_k) < \frac{\delta}{2}$ for $k \geq k_0$. If we assume that $f_k \notin \mathcal{B}$,

$$
U_\phi^* - U_\phi(f_k) = \underbrace{U_\phi^* - \sup_{f' \notin \mathcal{B}} U_\phi(f')}_{=\delta} + \underbrace{\sup_{f' \notin \mathcal{B}} U_\phi(f') - U_\phi(f_k)}_{\geq 0} \geq \delta,
$$

which contradicts $U_\phi^* - U_\phi(f_k) < \frac{\delta}{2}$. Thus, it holds that $f_k \in \mathcal{B}$ for $k \geq k_0$, that is, $U(f_k) = U^\dagger$, which indicates U-calibration. □

[3] This can be confirmed in a similar manner as the proof in Yan et al. [2018, Theorem 3.1].

Thus, we follow the strategy below to prove the U-calibration.

1. Characterize the Bayes-optimality condition based on Proposition 3.11.
2. Assume the converse of Equation (3.7), namely, the existence of $f^* \notin \mathcal{B}$ satisfying the stationary condition $U_\phi(f^*) = \sup_f U_\phi(f)$.
3. Show that the Bayes-optimality condition and the stationary condition cannot be satisfied simultaneously (proof by contradiction).

Throughout the proofs, we assume the following regularity condition.

Assumption 3.13. *For the true utility U and the underlying distribution $\mathbb{P}$, the critical set*

$$\mathcal{C}^\dagger := \left\{ \mathbf{x} \in \mathcal{X} \;\middle|\; (\Delta a_0 - \Delta a_1 U(f^\dagger))\eta(\mathbf{x}) - (a_{1,-1}U(f^\dagger) - a_{0,-1}) = 0, f^\dagger \in \mathcal{B} \right\}$$

is a p-null set, namely, $p(\mathcal{C}^\dagger) = 0$.

This holds for any *η-continuous* distribution [Yan et al., 2018, Assumption 2].

3.4.2 Calibration Analysis of F_β-measure

The F_β-measure has been widely used, particularly in the field of information retrieval where relevant items are rare [Manning and Schütze, 2008]. Because it is defined as the weighted harmonic mean of the precision and recall (see Table 3.1), its optimization is generally difficult. Although many previous studies have tried to directly optimize this in the context of a class-posterior probability estimation [Nan et al., 2012, Koyejo et al., 2014, Yan et al., 2018] or iterative cost-sensitive learning [Koyejo et al., 2014, Parambath et al., 2014], we show that a calibrated surrogate utility exists that can also be used in the direct optimization.

For the F_β-measure $\frac{(1+\beta^2)\mathsf{TP}}{(1+\beta^2)\mathsf{TP}+\beta^2\mathsf{FN}+\mathsf{FP}} = \frac{(1+\beta^2)\mathsf{TP}}{\mathsf{TP}+\mathsf{FP}+\beta^2\pi}$, define the true utility U^{F_β} and the surrogate utility $U_\phi^{\mathsf{F}_\beta}$ as

$$\begin{aligned} U^{\mathsf{F}_\beta}(f) &= \frac{\mathbb{E}\left[(1+\beta^2)\ell(-f(\mathsf{X}))\eta(\mathsf{X})\right]}{\mathbb{E}\left[\ell(-f(\mathsf{X}))\eta(\mathsf{X}) + \ell(-f(\mathsf{X}))(1-\eta(\mathsf{X})) + \beta^2\pi\right]}, \\ U_\phi^{\mathsf{F}_\beta}(f) &= \frac{\mathbb{E}\left[(1+\beta^2)(1-\phi(f(\mathsf{X})))\eta(\mathsf{X})\right]}{\mathbb{E}\left[(1+\phi(f(\mathsf{X})))\eta(\mathsf{X}) + \phi(-f(\mathsf{X}))(1-\eta(\mathsf{X})) + \beta^2\pi\right]}. \end{aligned}$$

For $U_\phi^{\mathsf{F}_\beta}$, we have the following F_β-calibration guarantee. Denote $(U_\phi^{\mathsf{F}_\beta})^* := \sup_f U_\phi^{\mathsf{F}_\beta}(f)$.

Theorem 3.14 (F_β-calibration)**.** *Assuming that a surrogate loss $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is convex, non-increasing, and differentiable almost everywhere, and that $(U_\phi^{\mathsf{F}_\beta})^* \geq \frac{(1+\beta^2)\tau}{\beta^2-\tau}$ and ϕ is τ-discrepant for a certain constant $\tau \in (0, \beta^2)$, $U_\phi^{\mathsf{F}_\beta}$ is then F_β-calibrated.*

An example of the τ-discrepant surrogate loss is shown in Figure 3.3. Here, τ is a discrepancy hyperparameter. From the fact that $(U_\phi^{\mathsf{F}_\beta})^* \leq 1$, τ ranges over $(0, \frac{\beta^2}{2+\beta^2}]$. We may determine τ through cross-validation, or fix it at $\tau = \frac{\beta^2}{2+\beta^2}$ by assuming $(U_\phi^{\mathsf{F}_\beta})^* \approx 1$. The proof of Theorem 3.14 is provided in Section 3.8.

Remark 3.1. *Under the assumption of Theorem 3.14, $(U_\phi^{\mathsf{F}_\beta})^* \geq \frac{(1+\beta^2)\tau}{\beta^2-\tau}$ is assumed. Herein, we briefly demonstrate that the optimal surrogate utility $(U_\phi^{\mathsf{F}_\beta})^*$ is non-negative, ensuring that we can choose $\tau \in (0, \beta^2)$. In the case of the F_β-measure,*

$$\begin{aligned} W_{0,\phi}(\xi, q) &= (1+\beta^2)(1-\phi(\xi))q, \\ W_{1,\phi}(\xi, q) &= (1+\phi(\xi))q + \phi(-\xi)(1-q) + \beta^2\pi, \\ U_\phi(f) &= \frac{\mathbb{E}[W_{0,\phi}(f(\mathsf{X}), \eta(\mathsf{X}))]}{\mathbb{E}[W_{1,\phi}(f(\mathsf{X}), \eta(\mathsf{X}))]}, \end{aligned}$$

and letting f^* and $\check{f}$ be the suprema of U_ϕ and $\mathbb{E}[W_0(f(\mathsf{X}), \eta(\mathsf{X}))]$ in f within all measurable functions, respectively, then,

$$\begin{aligned} (U_\phi^{\mathsf{F}_\beta})^* = U_\phi^{\mathsf{F}_\beta}(f^*) \geq U_\phi^{\mathsf{F}_\beta}(\check{f}) &= \frac{\sup_{f'} \mathbb{E}[W_{0,\phi}(f'(\mathsf{X}), \eta(\mathsf{X}))]}{\mathbb{E}[W_{1,\phi}(\check{f}(\mathsf{X}), \eta(\mathsf{X}))]} \\ &\overset{(a)}{=} \frac{\mathbb{E}[H_{0,\phi}(\eta(\mathsf{X}))]}{\mathbb{E}[W_{1,\phi}(\check{f}(\mathsf{X}), \eta(\mathsf{X}))]} \\ &= \frac{(1+\beta^2)\pi}{\mathbb{E}[W_{1,\phi}(\check{f}(\mathsf{X}), \eta(\mathsf{X}))]} \\ &\geq 0, \end{aligned}$$

where $H_{0,\phi}(q) \coloneqq \sup_{\xi\in\mathbb{R}} W_{0,\phi}(\xi, q)$. The equality (a) holds by assuming some regularity conditions such as the $\mathbb{P}$-minimizability (Section 2.3). Hence, we confirm $(U_\phi^{\mathsf{F}_\beta})^* \geq 0$.

3.4.3 Calibration Analysis of Jaccard Index

The Jaccard index, also referred to as the *intersection over union (IoU)*, is a metric of similarity between two sets: For two sets A and B, it is defined as $\frac{|A\cap B|}{|A\cup B|} \in [0, 1]$ [Jaccard, 1901]. The Jaccard index between the sets of examples predicted as positives and labeled as positives becomes $\frac{\mathsf{TP}}{\mathsf{TP}+\mathsf{FN}+\mathsf{FP}}$, as shown in Table 3.1. This measure is not only used for measuring the performance of binary classification [Koyejo et al., 2014, Narasimhan et al., 2015], but also for semantic segmentation [Everingham et al., 2010, Csurka et al., 2013, Ahmed et al., 2015, Berman et al., 2018].

For the Jaccard index $\frac{\mathsf{TP}}{\mathsf{TP}+\mathsf{FN}+\mathsf{FP}} = \frac{\mathsf{TP}}{\mathsf{FP}+\pi}$, define the true utility U^{Jac} and the surrogate utility U_ϕ^{Jac} as

$$\begin{aligned} U^{\mathsf{Jac}}(f) &= \frac{\mathbb{E}[\ell(-f(\mathsf{X}))\eta(\mathsf{X})]}{\mathbb{E}[\ell(-f(\mathsf{X}))(1-\eta(\mathsf{X}))+\pi]}, \\ U_\phi^{\mathsf{Jac}}(f) &= \frac{\mathbb{E}[(1-\phi(f(\mathsf{X})))\eta(\mathsf{X})]}{\mathbb{E}[\phi(-f(\mathsf{X}))(1-\eta(\mathsf{X}))+\pi]}. \end{aligned}$$

Denote $(U_\phi^{\mathsf{Jac}})^* \coloneqq \sup_f U_\phi^{\mathsf{Jac}}(f)$. As for U_ϕ^{Jac}, we have the following Jaccard-calibration.

Theorem 3.15 (Jaccard-calibration)**.** *Assume that a surrogate loss $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ convex, non-increasing, and differentiable almost everywhere, and that $(U_\phi^{\mathsf{Jac}})^* \geq \tau$ and ϕ is τ-discrepant for a certain constant $\tau \in (0, 1)$. Then, U_ϕ^{Jac} is Jaccard-calibrated.*

Theorem 3.15 also relies on the τ-discrepancy as in Theorem 3.14. Thus, the loss shown in Figure 3.3 can also be used in the Jaccard case with a certain range of τ. In the same manner as the F_β-measure, a hyperparameter τ ranges over $(0, 1)$, which we may either determine through cross-validation or fix to a certain value.

As is the case for the F_β-measure, $(U_\phi^{\mathsf{Jac}})^* \geq 0$ can be guaranteed, and $\tau \in (0, 1)$ such that $\tau \leq (U_\phi^{\mathsf{Jac}})^*$ can be chosen. The proof of Theorem 3.15 is provided in Section 3.8.

3.5 Consistency Analysis: Bridging Finite Sample and Asymptotics

In this section, we analyze the statistical properties of the estimator $\widehat{V}_\phi$ in Equation (3.6). To simplify our analysis, the linear-in-input model $f_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\theta}^\top \mathbf{x}$ is considered throughout this section, where $\boldsymbol{\theta} \in \Theta$ is a classifier parameter and $\Theta \subseteq \mathbb{R}^d$ is a compact parameter space. The maximization procedure introduced above can be naturally seen as *Z-estimation* [van der Vaart, 2000], which is an estimation procedure used to solve an estimation equation. In our case, the maximization of U_ϕ is reduced to a Z-estimation problem to solve the system $\widehat{V}_\phi(f) = 0$. The first lemma shows that the derivative estimator $\widehat{V}_\phi$ admits a uniform convergence. For its proof, please refer to Section 3.8.3.

Lemma 3.16 (Uniform convergence)**.** *For simplicity, assume that $m = n/2$. For $k = 0, 1$, let $c_k := \sup_{\xi \in \mathbb{R}, y \in \mathcal{Y}} |W_{k,\phi}(\xi, y)| < +\infty$. Assume that $W_k(\cdot, y)$ for $y \in \mathcal{Y}$ are ρ_k-Lipschitz continuous for some $0 < \rho_k < \infty$, and that $\|\mathbf{x}\| < c_\mathcal{X}$ $(\forall \mathbf{x} \in \mathcal{X})$ and $\|\boldsymbol{\theta}\| < c_\Theta$ $(\forall \boldsymbol{\theta} \in \Theta)$ for some $0 < c_\mathcal{X}, c_\Theta < \infty$. Then,*

$$\sup_{\boldsymbol{\theta} \in \Theta} \left\| \widehat{V}_\phi(f_{\boldsymbol{\theta}}) - V_\phi(f_{\boldsymbol{\theta}}) \right\| = O_p(n^{-\frac{1}{2}}), \tag{3.8}$$

where O_p denotes the order in probability.

The proof of Lemma 3.16 is provided in Section 3.8.3. The Lipschitz continuity and smoothness assumptions in Lemma 3.16 can be satisfied if the surrogate loss ϕ satisfies a certain Lipschitzness and smoothness. Note that Lemma 3.16 still holds for τ-discrepant surrogates because we allow surrogates to have different smoothness parameters for both positive and negative domains. In addition, Lemma 3.16 is the basis for showing the consistency. Let $\boldsymbol{\theta}^* := \arg\max_{\boldsymbol{\theta} \in \Theta} U_\phi(f_{\boldsymbol{\theta}})$ and $\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta} \in \Theta} \widehat{U}_\phi(f_{\boldsymbol{\theta}})$. Under the identifiability described below, $f_{\boldsymbol{\theta}^*}$ and $f_{\widehat{\boldsymbol{\theta}}_n}$ are the roots of V_ϕ and $\widehat{V}_\phi$, respectively. We can then show the consistency of $\widehat{\boldsymbol{\theta}}_n$.

Theorem 3.17 (Consistency)**.** *Assume that $\boldsymbol{\theta}^*$ is* identifiable*, that is,*

$$\inf \{ \ \| V_\phi(f_{\boldsymbol{\theta}})\| \,|\, \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \geq \varepsilon \} > \|V_\phi(f_{\boldsymbol{\theta}^*})\| = 0$$

for all $\varepsilon > 0$, and that Equation (3.8) holds for $\widehat{V}_\phi$. Then, $\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}^*$.

Given a uniform convergence (Lemma 3.16) and an identifiability assumption, Theorem 3.17 is an immediate result of van der Vaart [2000, Theorem 5.9]. Note that the identifiability assumes that V_ϕ has a unique zero $f_{\boldsymbol{\theta}^*}$, which is also usual in an M-estimation, i.e., the global optimizer is identifiable. In addition, because Algorithm 3.1 is a combination of concave and quasiconcave programs, the identifiability is reasonably assumed.

Table 3.2: Comparison of related work.

Method	Consistency	Avoids to estimate η	Efficient optimization
ours	✓	✓	✓
(i)	✗	✓	✓
(ii)	✓	✗	✓
(iii)	✓	✓	✗

3.6 Related Work

In this section, we summarize the existing lines of research conducted on the optimization of the generalized performance metrics, elucidating the advantages of our approach.

(i) Surrogate optimization: One of the earliest attempts to optimize non-decomposable performance metrics dates back to Joachims [2005], in which a structured SVM was formulated as a surrogate objective. However, Dembczyński et al. [2013] showed that this surrogate is inconsistent, which means that the surrogate maximization does not necessarily imply the maximization of the true metric. In addition, Kar et al. [2014] demonstrated a sublinear regret for the structural surrogate developed by Joachims [2005] in an online setting. Later, Yu and Blaschko [2015], Eban et al. [2017], Berman et al. [2018], and Zhao et al. [2019] attempted different surrogates, although their calibration has yet to be studied.[4] As a concurrent approach, Fathony and Kolter [2020] provided a surrogate objective for generalized performance metrics based on the framework of adversarial prediction [Asif et al., 2015, Fathony et al., 2018] and studied its consistency. Their definition of a performance metric is based on the expected test utility (ETU) [Dembczyński et al., 2017]. We believe that the population utility is more appropriate than the ETU for a consistent analysis.[5] There have been several recent works dealing with scenarios where not only the performance metrics are complex but also they are constrained. Whereas the constrained optimization problems are more intricate to be handled by surrogate optimization, Kumar et al. [2021] proposed a clever approach to make them unconstrained by the implicit function theorem. Nevertheless, surrogate consistency has yet to be known for this approach.

As a final remark, one may be reminded of the multi-label F-measure calibrated surrogate loss proposed in Zhang et al. [2020]. We stress that optimizing the F-measure has more difficulty in binary classification than multi-label learning if we follow the metric definition based on the population utility (PU) [Dem-

[4] In particular, Yu and Blaschko [2015] proposed the Lovász hinge loss as a convex surrogate loss for submodular loss functions, to which the Jaccard index belongs. However, Finocchiaro et al. [2019] showed that the Lovász hinge loss is not calibrated unless the original target loss function is modular, where the Lovász hinge loss reduces to the weighted Hamming loss. Since surrogate losses proposed in Eban et al. [2017], Berman et al. [2018], and Zhao et al. [2019] are more complicated, their calibration results have yet to be known. We remark that the Lovász hinge loss relies on the ETU framework [Dembczyński et al., 2017], namely, the loss function takes a set of binary predictions and target labels as inputs.

[5] Under the metric definition based on the ETU, a classifier takes the form of $\mathcal{X}^k \to \mathcal{Y}^k$, where k is the fixed size of the test set. In other words, the binary classification problem is transformed into a structured prediction problem. Because we are interested in a pure binary classifier in the form of $\mathcal{X} \to \mathcal{Y}$, our analysis is based on the PU metric definition. The proof of our calibration analysis initially invokes the stationary condition of the optimal classifier (see Section 3.4.1), which is similar to the proof technique used in adversarial prediction [Fathony et al., 2018].

bczyński et al., 2017]. The multi-label F-measure has similar nature to the ETU definition of the binary F-measure and is relatively easy to handle in surrogate optimization since the problem is transformed into structured prediction. In any cases, our constructions and proofs of calibrated surrogates in the PU framework is notable.

(ii) Plug-in rule: Instead of a surrogate optimization, in Dembczyński et al. [2013], the authors mentioned that a plug-in rule is consistent, where η and a threshold parameter are estimated independently. We can estimate η by minimizing strictly proper losses [Reid and Williamson, 2009]. The plug-in rule has been investigated under many different settings [Nan et al., 2012, Dembczyński et al., 2013, Koyejo et al., 2014, Narasimhan et al., 2014, Busa-Fekete et al., 2015, Yan et al., 2018]. Liu et al. [2018] proposed an online algorithm to estimate η and the threshold on-the-fly to optimize the F-measure. However, as one of the weaknesses of the plug-in rule, it requires an accurate estimate of η, which is less sample-efficient than the usual classification with convex surrogates [Audibert and Tsybakov, 2007].[6] Moreover, estimation of class-posterior probability η is known to be biased when the class distribution is imbalanced [King and Zeng, 2001, Menon et al., 2012, Bao and Sugiyama, 2021], which makes the choice of the threshold parameter very sensitive.

(iii) Cost-sensitive risk minimization: By contrast, Parambath et al. [2014] is a pioneering study focusing on the *pseudo-linearity* of the metrics, which reduces their maximization to an alternative optimization with respect to a classifier and its sublevel. This can be formulated as an iterative cost-sensitive risk minimization [Koyejo et al., 2014, Narasimhan et al., 2015, 2016, Sanyal et al., 2018]. Bascol et al. [2019] derived a tighter upper bound on the F-measure and eventually obtains a scheme to reduce the number of candidates of the cost parameter. Although these methods achieve consistency, they need to train classifiers numerous times, which may lead to high computational costs, particularly for complex hypothesis spaces. Many recent studies on constrained performance metric optimization have extended the cost-sensitive approach to handle metric constraints such as fairness constraints. Narasimhan [2018], Tavker et al. [2020] formulated the constrained problem as an optimization problem over the confusion matrix. At each update, the cost is updated by optimizing the performance metric, and the confusion matrix is updated by the Frank-Wolfe algorithm [Jaggi, 2013]. Narasimhan et al. [2019] got rid of the oracle access to the cost-sensitive risk minimizers assumed in Narasimhan [2018], instead by formulating as a three-player game.

Remark 3.2. *Although our proposed methods can be considered to belong to the family (i), one of the crucial differences is the fact that we have calibration guarantee. We do not need to estimate the class-posterior probability as in (ii), or train classifiers many times as in (iii). This comparison is summarized in Table 3.2. Some recent works have not been able to be categorized as such. For example, Jiang et al. [2020] considered black-box optimization of the performance metrics. They proposed an approach to iteratively estimate the gradient of the black-box metric, update the confusion matrix, and project the confusion matrix onto the parameter space.*

[6] Audibert and Tsybakov [2007] showed that plug-in classifiers achieve fast rates of the excess classification risk only when a strong density assumption is imposed on the underlying distribution. Without this relatively strong assumption, plug-in classifiers cannot achieve as fast rates as the ERM minimizers. Even though we do not know whether plug-in classifiers converge as fast as the surrogate utility maximizers in the case of the linear-fractional metrics, we speculate that the latter behaves well.

3.7 Experiments

In this section, we investigate the empirical performances of the surrogate optimizations (Algorithm 3.1 using NGA and normalized BFGS). Datasets that we use throughout this section are obtained from the *UCI Machine Learning Repository* [Lichman, 2013] and the *LIBSVM* [Chang and Lin, 2011]. For those having independent training data, validation data, and test data, all such data are merged into a single dataset. We randomly split the original dataset at a ratio of 8 to 2, and the former data are used for training and the latter are used for an evaluation. Each feature value is scaled between zero and one.

Subsequently, we describe the implementation details of the proposed and baseline methods.

Proposed Methods. The linear-in-input model $f_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\theta}^\top \mathbf{x}$ was used for the hypothesis space $\mathcal{F}$. As the initializer of $\boldsymbol{\theta}$, the ERM minimizer trained using an SVM was applied. For both NGA and BFGS, gradient updates were iterated 300 times. NGA and normalized BFGS are referred to below as U-GD and U-BFGS below, respectively. The surrogate loss shown in Figure 3.3 was used, i.e., $\phi(m) = \log_2(1 + e^{-m})$ when $m \le 0$ and $\phi(m) = \log_2(1 + e^{-\tau m})$ for $m > 0$, where τ was set to 0.33 in the F_1-measure case and 0.75 in the Jaccard index case.[7] The training set was divided into a ratio of 4 to 1 and the latter dataset was used for validation. We used a common learning rate in Algorithm 3.1, which was chosen from $\left\{ 10^1, 10^{-1}, 10^{-3}, 10^{-5} \right\}$ through cross-validation.

Baseline 1 (ERM). The first baseline is a usual empirical risk minimization, which optimizes not the metric of interest but the accuracy. The hinge loss and ℓ_2-regularization are employed with the regularization parameter 10^{-2}.

Baseline 2 (W-ERM). A weighted empirical risk minimization, or cost-sensitive empirical risk minimization, is often used to optimize the non-linear performance metrics [Koyejo et al., 2014, Narasimhan et al., 2014, Parambath et al., 2014]. Herein, we apply a simple approach: A cost parameter is found from a given cost parameter space, which provides the maximum validation performance of a classifier trained through the cost-sensitive empirical risk minimization [Scott, 2012]. The training dataset is split into a ratio of 4 to 1 at random, and the latter dataset is saved for a validation of a regularization parameter. The former dataset is further split into ratio of 9 to 1 at random, where the former 90% is used for training the base classifier, and the latter 10% is used for the validation. As the base cost-sensitive learner, we use the hinge loss minimizer with ℓ_2-regularization (a regularization parameter is chosen from $\left\{ 10^{-1}, 10^{-3}, 10^{-5} \right\}$ through cross-validation). The cost parameter is chosen from the range $[10^{-3}, 1 - 10^{-3}]$ evenly split into 20 ranges, that is, $\left\{ 10^{-3} + \frac{1 - 2 \cdot 10^{-3}}{20} i \;\middle|\; i = 1, \ldots, 20 \right\}$.

Baseline 3 (Plug-in). A plug-in estimator is one of the other common methods used to optimize the non-linear performance metrics [Koyejo et al., 2014, Yan et al., 2018], which is a two-step method, i.e., the class posterior probability $\widehat{\eta}(\mathbf{x}) = \mathbb{P}(\mathsf{Y} = +1 | \mathsf{X} = \mathbf{x})$ is first estimated, and the optimal threshold $\widehat{\delta}$ is then determined. The classifier is constructed as $\mathbf{x} \mapsto \mathrm{sgn}(\widehat{\eta}(\mathbf{x}) - \widehat{\delta})$. The training

[7] The discrepancy parameter τ should be chosen within $(0, \frac{1}{3})$ and $(0, 1)$ for the F_1-measure and Jaccard index, respectively. Here, we fix them to the slightly smaller values than the upper limits of their ranges.

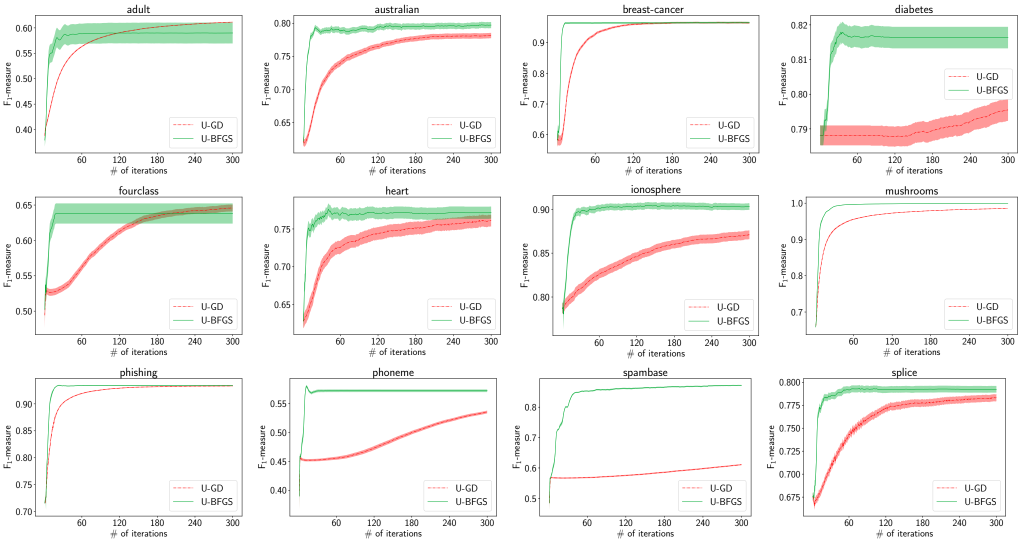


Figure 3.4: Convergence comparison of the F$_1$-measure (vertical axes). Standard errors of 50 trials are shown as shaded areas.

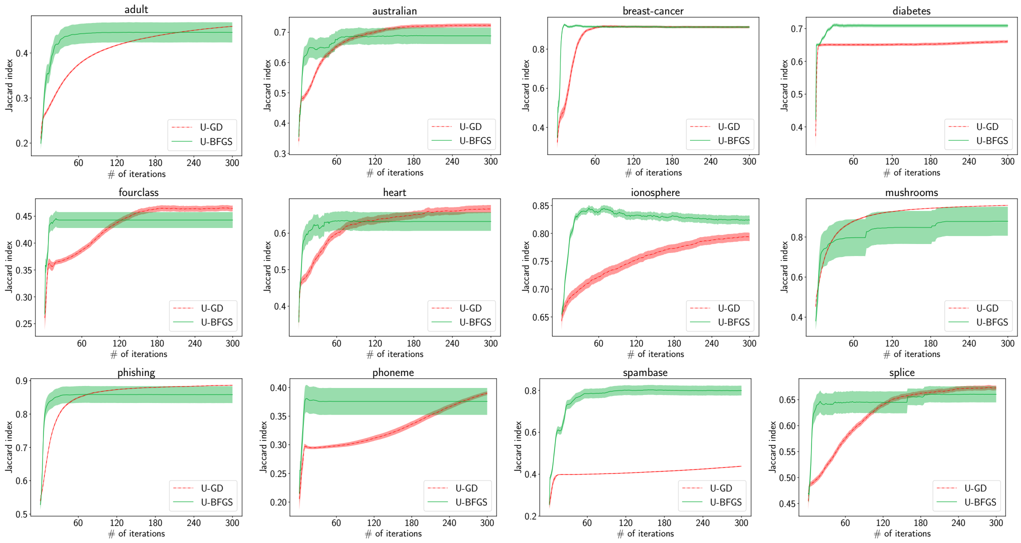


Figure 3.5: Convergence comparison of the Jaccard index (vertical axes). Standard errors of 50 trials are shown as shaded areas.

dataset is split into a ratio of 4 to 1 at random, and the latter dataset is saved for validation of a regularization parameter. The former dataset is further split into a ratio of 9 to 1 at random, and the resulting datasets are independently used for the first and second steps. To estimate $\mathbb{P}(\mathsf{Y} = +1|\mathbf{x})$ (the first step), the logistic regression is used [Reid and Williamson, 2009], with ℓ_2-regularization (a regularization parameter is chosen from $\left\{ 10^{-1}, 10^{-3}, 10^{-5} \right\}$ through cross-validation). To determine $\widehat{\delta}$, we pick a threshold with the highest validation metric from $\left\{ 10^{-3} + \frac{1-2\cdot10^{-3}}{20} i \;\middle|\; i = 1, \dots, 20 \right\}$.

3.7.1 Convergence Comparison

We compare the convergence behaviors of U-GD and U-BFGS. During this experiment, we ran them 300 iterations from randomly initialized parameters drawn from $\mathcal{N}(0_d, I_d)$. The results in terms of the F$_1$-measure and Jaccard index are shown in Figures 3.4 and 3.5, respectively. The vertical axes show test metric

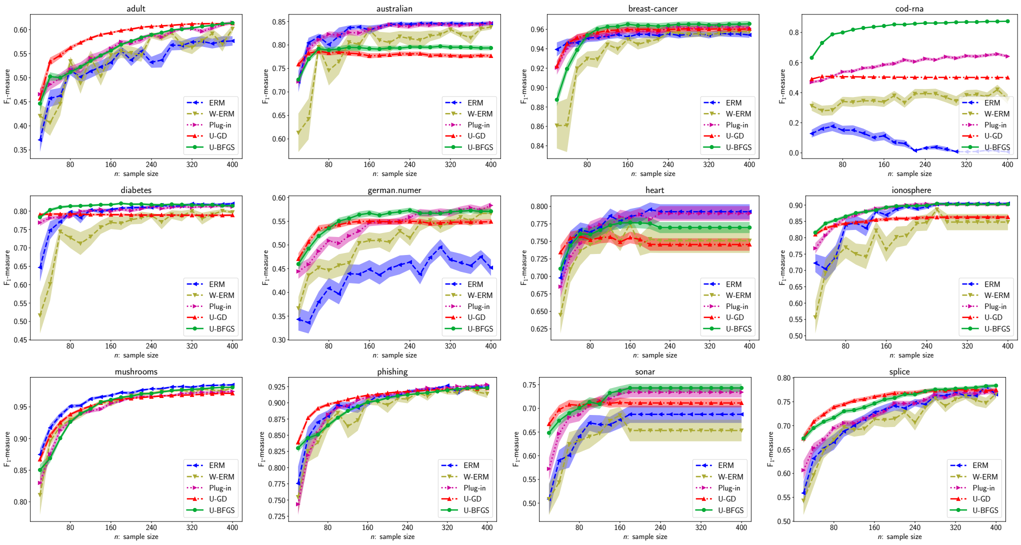


Figure 3.6: The relationship between the test F$_1$-measure (vertical axes) and sample size (horizontal axes). Standard errors of 50 trials are shown as shaded areas.

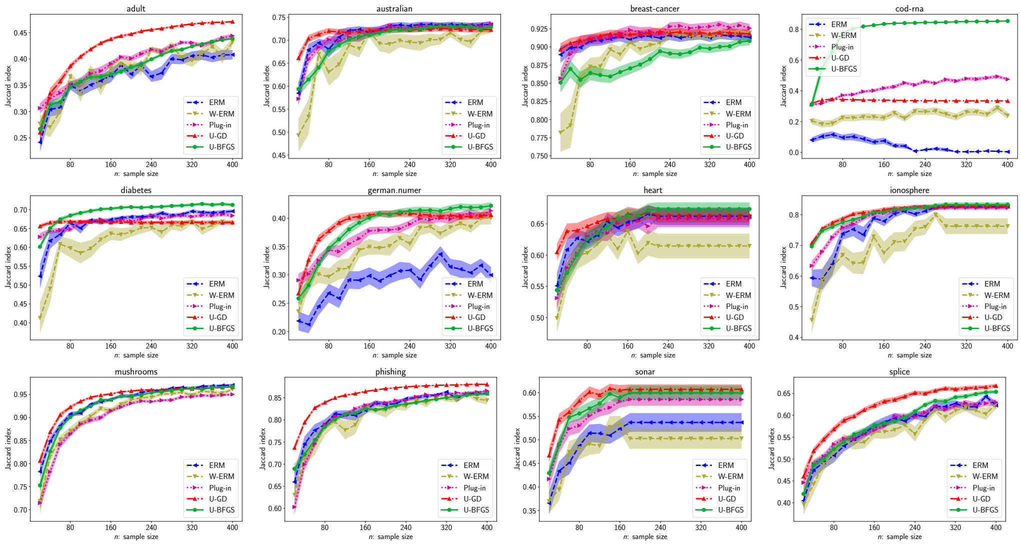


Figure 3.7: The relationship between the test Jaccard (vertical axes) and sample size (horizontal axes). Standard errors of 50 trials are shown as shaded areas.

values, where the higher the values the better. Note that both the F$_1$-measure and Jaccard index ranges from zero to one. The horizontal axes show the number of iterations.

Overall, U-BFGS shows a faster convergence than U-GD in terms of the number of iterations, which constitutes a trade-off in that the former converges within fewer steps, whereas the latter can update the solution faster during each step. In almost all cases, U-BFGS converges within 30 iterations, except for german.numer and mushrooms in the Jaccard case. Moreover, it usually achieves a higher performance than U-GD. U-GD convergences require at approximately 100 iterations at a minimum (mushrooms and phishing in F$_1$ case), and occasionally does not converge even within 300 iterations, such with as heart and ionosphere for the F$_1$ and Jaccard cases.

Table 3.3: Results of the F_1-measure: 50 trials were conducted for each pairing of a method and dataset. Standard errors (multiplied by 10^4) are shown in the parentheses. Bold font indicates outperforming methods, which were chosen through a one-sided t-test with a significant level of 5%.

(F_1-measure)	Proposed		Baselines		
Dataset	U-GD	U-BFGS	ERM	W-ERM	Plug-in
adult	0.617 (101)	0.660 (11)	0.639 (51)	0.676 (18)	**0.681 (9)**
australian	**0.843 (41)**	**0.844 (45)**	0.820 (123)	0.814 (116)	0.827 (51)
breast-cancer	**0.963 (31)**	**0.960 (32)**	0.950 (37)	0.948 (44)	0.953 (40)
cod-rna	0.802 (231)	0.594 (4)	0.927 (7)	0.927 (6)	**0.930 (2)**
diabetes	**0.834 (32)**	**0.828 (31)**	0.817 (50)	0.821 (40)	0.820 (42)
german.numer	0.561 (102)	**0.580 (74)**	0.492 (188)	0.560 (107)	**0.589 (73)**
heart	**0.796 (101)**	**0.802 (99)**	**0.792 (80)**	0.764 (151)	0.764 (137)
ionosphere	**0.908 (49)**	**0.901 (43)**	0.883 (104)	0.842 (217)	**0.897 (54)**
mushrooms	1.000 (1)	0.997 (7)	**1.000 (1)**	1.000 (2)	0.999 (4)
phishing	0.937 (29)	**0.943 (7)**	**0.944 (8)**	0.940 (12)	**0.944 (8)**
phoneme	**0.648 (27)**	0.559 (22)	0.530 (201)	0.616 (135)	0.633 (35)
skin_nonskin	0.870 (3)	0.856 (4)	0.854 (7)	**0.877 (8)**	0.838 (5)
sonar	**0.735 (95)**	**0.740 (91)**	0.706 (121)	0.655 (189)	**0.721 (113)**
spambase	0.876 (27)	0.756 (61)	0.887 (42)	0.881 (58)	**0.903 (18)**
splice	0.785 (49)	**0.799 (46)**	0.785 (55)	0.771 (67)	**0.801 (45)**
w8a	0.297 (80)	0.284 (96)	0.735 (35)	**0.742 (29)**	**0.745 (26)**

Table 3.4: Results of the Jaccard index: 50 trials were conducted for each pairing of a method and dataset. Standard errors (multiplied by 10^4) are shown in the parentheses. Bold font indicates outperforming methods, which were chosen through a one-sided t-test with a significant level of 5%.

(Jaccard index)	Proposed		Baselines		
Dataset	U-GD	U-BFGS	ERM	W-ERM	Plug-in
adult	0.499 (44)	0.498 (11)	0.471 (51)	0.510 (20)	**0.516 (10)**
australian	**0.735 (63)**	**0.733 (59)**	0.702 (144)	0.693 (143)	0.707 (76)
breast-cancer	**0.921 (54)**	**0.918 (55)**	0.905 (66)	0.903 (78)	**0.913 (69)**
cod-rna	0.854 (3)	0.785 (8)	0.864 (11)	0.865 (9)	**0.869 (3)**
diabetes	**0.714 (44)**	0.702 (50)	0.692 (70)	0.698 (56)	0.695 (60)
german.numer	**0.433 (64)**	**0.429 (69)**	0.335 (153)	0.391 (98)	**0.418 (71)**
heart	**0.665 (135)**	**0.675 (135)**	**0.664 (102)**	0.629 (178)	0.626 (163)
ionosphere	**0.826 (76)**	**0.829 (65)**	0.796 (134)	0.749 (245)	**0.815 (87)**
mushrooms	0.999 (2)	0.995 (4)	**1.000 (1)**	0.999 (4)	0.997 (7)
phishing	0.883 (43)	**0.893 (11)**	**0.894 (14)**	0.888 (22)	**0.894 (15)**
phoneme	0.435 (51)	0.436 (24)	0.371 (160)	**0.450 (104)**	**0.461 (34)**
skin_nonskin	0.744 (5)	0.751 (5)	0.746 (10)	**0.780 (13)**	0.722 (7)
sonar	**0.600 (125)**	**0.600 (111)**	0.552 (147)	0.495 (202)	**0.572 (134)**
spambase	**0.827 (22)**	0.708 (22)	0.798 (67)	0.790 (86)	**0.824 (31)**
splice	**0.670 (60)**	**0.672 (56)**	0.646 (71)	0.629 (84)	**0.672 (57)**
w8a	0.496 (151)	0.452 (28)	0.580 (44)	**0.590 (35)**	**0.595 (33)**

3.7.2 Performance Comparison with Benchmark Data

We compared the proposed methods with the baselines. The results of the F_1-measure and Jaccard index are summarized in Tables 3.3 and 3.4, respectively. For each dataset, we first picked the method with the highest test performance as an outperforming method within that dataset, and then conducted a one-sided t-test with a significant level of 5%. They are also regarded as outperforming methods if the performance differences are insignificant as a result of hypothesis tests. Outperforming methods are indicated in bold.

As general tendencies, we observed that U-BFGS and Plug-in work well for both the F_1-measure and Jaccard index. As for the F_1-measure, their performances are competitive, whereas U-BFGS is better as for the Jaccard index. In practice, both U-BFGS and Plug-in are worth applying.

For the other methods, ERM does not work good as we expected because it does not optimize the metrics of interests, i.e., the F_1-measure and Jaccard index, at all. In addition, W-ERM does not work as well as Plug-in even though both of them are known to be consistent with the linear-fractional utilities. We may need to apply a finer split of the threshold search space, or try a binary-search-type algorithm provided by a recent study [Yan et al., 2018]. U-GD does not work as well as U-BFGS contrary to our expectation. We may need more iterations to make U-GD converge, as we can see in Figures 3.4 and 3.5.

3.7.3 Sample Complexity

To confirm our hypothesis that Plug-in requires large sample sizes for a probability estimation, we empirically studied the relationship between the sample size and performance. We randomly subsampled each original dataset to reduce the sample sizes to $\{20, 40, \dots, 400\}$, and trained all methods on the reduced samples. Herein, Figures 3.6 and 3.7 show the sample complexity results. Although learning is unstable for small samples (e.g., heart and w8a), we can observe clear differences in certain cases, such as cod-rna, diabetes, german.numer, ionosphere, sonar, and splice in terms of the F$_1$-measure; and australian, cod-rna, diabetes, ionosphere, phishing, sonar, and spambase in terms of the Jaccard index, where either U-GD or U-BFGS works better than Plug-in even if sample sizes are quite small at approximately 20 to 40. In addition, Plug-in seldom works significantly better than the gradient-based methods in sample sizes ranging from approximately 100 to 400 as investigated in this section. This is contrary to the behavior shown in Tables 3.3 and 3.4, where the full-size datasets are used to train the classifiers. Hence, we claim that the gradient-based methods are good options when the sample sizes are extremely small.

3.8 Proofs

3.8.1 Proof of Theorem 3.14

We simply let U denote the F$_\beta$-measure and U_ϕ denote the surrogate utility such as

$$U_\phi(f) = \frac{\int_{\mathcal{X}}\{(1+\beta^2)(1-\phi(f(\mathbf{x})))\eta(\mathbf{x})\}p(\mathbf{x})\mathrm{d}\mathbf{x}}{\int_{\mathcal{X}}\{(1+\phi(f(\mathbf{x})))\eta(\mathbf{x})+\phi(-f(\mathbf{x}))(1-\eta(\mathbf{x}))+\beta^2\pi\}p(\mathbf{x})\mathrm{d}\mathbf{x}}.$$

Define

$$\begin{aligned} W_{0,\phi}(\xi, q) &:= (1+\beta^2)(1-\phi(\xi))q, \\ W_{1,\phi}(\xi, q) &:= (1+\phi(\xi))q + \phi(-\xi)(1-q) + \beta^2\pi. \end{aligned}$$

Then, $U_\phi(f) = \mathbb{E}[W_{0,\phi}]/\mathbb{E}[W_{1,\phi}]$. From Proposition 3.11, the Bayes-optimal set $\mathcal{B}$ for the F$_\beta$-measure is

$$\mathcal{B} := \left\{ f \;\middle|\; f(\mathbf{x})((1+\beta^2)\eta(\mathbf{x}) - U(f)) > 0 \quad \forall \mathbf{x} \in \mathcal{X} \right\}.$$

By Proposition 3.12, it is sufficient to show Equation (3.7). We prove this by contradiction. Assume that

$$\sup_{f \notin \mathcal{B}} U_\phi(f) = \sup_f U_\phi(f).$$

This implies that there exists an optimal function $f^* \notin \mathcal{B}$ that achieves $U_\phi(f^*) = \sup_f U_\phi(f) = U_\phi^*$, that is, $U_\phi(f^*) = U_\phi^*$ and $f^*(\bar{\mathbf{x}})((1+\beta^2)\eta(\bar{\mathbf{x}}) - U(f^*)) \le 0$ for some $\bar{\mathbf{x}} \in \mathcal{X}$.

Let us describe the *stationary condition* of f^*. We introduce a function δ_f:

$$\delta_f(\mathbf{x}) := \begin{cases} 1 & \text{if } \mathbf{x} = \bar{\mathbf{x}}, \\ 0 & \text{if } \mathbf{x} \neq \bar{\mathbf{x}}. \end{cases}$$

Let $G(\gamma) := U_\phi(f^* + \gamma\delta_f)$. Because Gâteaux derivative of U_ϕ at f^* must be zero in any direction, we claim that $G'(0) = 0$, where $G'(0)$ corresponds to Gâteaux derivative of U_ϕ at f^* in the direction of δ_f. Here, $G'(0)$ is computed as

$$\begin{aligned} G'(0) = \frac{p(\bar{\mathbf{x}})}{\mathbb{E}[W_{1,\phi}(f^*(\mathsf{X}), \eta(\mathsf{X}))]} \Big\{ & -(1+\beta^2)\phi'(f^*(\bar{\mathbf{x}}))\eta(\bar{\mathbf{x}}) - \phi'(f^*(\bar{\mathbf{x}}))U_\phi^*\eta(\bar{\mathbf{x}}) \\ & + \phi'(-f^*(\bar{\mathbf{x}}))U_\phi^*(1-\eta(\bar{\mathbf{x}})) \Big\}. \end{aligned}$$

Thus, the stationary condition $G'(0) = 0$ is equivalent to

$$\underbrace{\left\{(1+\beta^2)\phi'(f^*(\bar{\mathbf{x}})) + (\phi'(f^*(\bar{\mathbf{x}})) + \phi'(-f^*(\bar{\mathbf{x}})))U_\phi^*\right\}}_{\neq 0\ (\because\ \phi' < 0 \text{ and } U_\phi^* \geq 0)} \eta(\bar{\mathbf{x}}) = \phi'(-f^*(\bar{\mathbf{x}}))U_\phi^*,$$

which is equivalent to

$$\eta(\bar{\mathbf{x}}) = \frac{\phi'(-f^*(\bar{\mathbf{x}}))U_\phi^*}{(1+\beta^2)\phi'(f^*(\bar{\mathbf{x}})) + (\phi'(f^*(\bar{\mathbf{x}})) + \phi'(-f^*(\bar{\mathbf{x}})))U_\phi^*}. \tag{3.9}$$

From now on, we divide the cases to consider the Bayes-optimality condition $f^* \notin \mathcal{B}$, namely, $f^*(\bar{\mathbf{x}})((1+\beta^2)\eta(\bar{\mathbf{x}}) - U(f^*)) \leq 0$. By Assumption 3.13, the case $(1+\beta^2)\eta(\bar{\mathbf{x}}) - U(f^*) = 0$ is excluded.

1) If $f^*(\bar{\mathbf{x}}) > 0$ and $\eta(\bar{\mathbf{x}}) < \frac{1}{1+\beta^2}U(f^*)$, we show $\eta(\bar{\mathbf{x}}) \geq \frac{U(f^*)}{1+\beta^2}$, leading to the contradiction. In addition, we write the difference $\eta(\bar{\mathbf{x}}) - \frac{U(f^*)}{1+\beta^2} = \frac{D_\mathrm{n}}{D_\mathrm{d}}$, where

$$\begin{aligned} D_\mathrm{n} &= (1+\beta^2)\phi'(-f^*(\bar{\mathbf{x}}))U_\phi^* - (1+\beta^2)\phi'(f^*(\bar{\mathbf{x}}))U(f^*) \\ &\quad - (\phi'(f^*(\bar{\mathbf{x}})) + \phi'(-f^*(\bar{\mathbf{x}}))U_\phi^* U(f^*), \\ D_\mathrm{d} &= (1+\beta^2)\Big\{(1+\beta^2)\phi'(f^*(\bar{\mathbf{x}})) + (\phi'(f^*(\bar{\mathbf{x}})) + \phi'(-f^*(\bar{\mathbf{x}})))U_\phi^*\Big\}. \end{aligned}$$

The denominator D_d is always negative because of $\phi' < 0$. The numerator D_n can be simplified as

$$\begin{aligned} & \frac{D_\mathrm{n}}{U(f^*)\{(1+\beta^2) + U_\phi^*\}} \\ & = \underbrace{\frac{U_\phi^*}{(1+\beta^2) + U_\phi^*}}_{\geq \tau/\beta^2} \underbrace{\frac{(1+\beta^2) - U(f^*)}{U(f^*)}}_{\geq \beta^2} \phi'(-f^*(\bar{\mathbf{x}})) - \phi'(f^*(\bar{\mathbf{x}})) \\ & \leq \tau\phi'(-f^*(\bar{\mathbf{x}})) - \phi'(f^*(\bar{\mathbf{x}})) \\ & \leq 0, \end{aligned}$$

where the first inequality holds because $\frac{U_\phi^*}{(1+\beta^2)+U_\phi^*} \geq \frac{\tau}{\beta^2}$ when $\frac{(1+\beta^2)\tau}{\beta^2-\tau} \leq U_\phi^* \leq 1$ (see Figure 3.8) and $\frac{(1+\beta^2)-U(f^*)}{U(f^*)} \geq \beta^2$ when $0 \leq U(f^*) \leq 1$ (see Figure 3.9). Note that $\phi'(-f^*(\bar{\mathbf{x}})) < 0$. The second inequality holds because of the assumption that $\lim_{m\searrow 0}\phi'(m) \geq \tau \lim_{m\nearrow 0}\phi'(m)$ and ϕ is convex, which implies $\tau\phi'(-m) - \phi'(m) \leq 0$ for every $m > 0$.

Thus, $\eta(\bar{\mathbf{x}}) \geq \frac{U(f^*)}{1+\beta^2}$ holds (contradiction).

2) If $f^*(\bar{\mathbf{x}}) \leq 0$ and $\eta(\bar{\mathbf{x}}) > \frac{1}{1+\beta^2}U(f^*)$, in addition the previous case, we begin from a stationary condition (3.9). If $\phi'(-f^*(\bar{\mathbf{x}})) < 0$,

$$\begin{aligned}
\eta(\bar{\mathbf{x}}) &= \frac{\phi'(-f^*(\bar{\mathbf{x}}))U_\phi^*}{(1+\beta^2)\phi'(f^*(\bar{\mathbf{x}})) + (\phi'(f^*(\bar{\mathbf{x}})) + \phi'(-f^*(\bar{\mathbf{x}})))U_\phi^*} \\
&= \frac{U_\phi^*}{(1+\beta^2)\frac{\phi'(f^*(\bar{\mathbf{x}}))}{\phi'(-f^*(\bar{\mathbf{x}}))} + \left(\frac{\phi'(f^*(\bar{\mathbf{x}}))}{\phi'(-f^*(\bar{\mathbf{x}}))} + 1\right)U_\phi^*} \\
&\leq \frac{1}{1+\beta^2}\frac{U_\phi^*}{1+U_\phi^*} \\
&\leq \frac{1}{1+\beta^2}U(f^*) \\
&< \eta(\bar{\mathbf{x}}), \qquad \text{(contradiction)}
\end{aligned}$$

where the first inequality holds because $\frac{\phi'(-m)}{\phi'(m)} \geq 1$ for every $m \geq 0$ and $f^*(\bar{\mathbf{x}}) \leq 0$, and the second inequality holds because $U_\phi(f) \leq U(f)$ $(\forall f)$ implies $\frac{U_\phi^*}{1+U_\phi^*} \leq U(f^*)$ when $\frac{(1+\beta^2)\tau}{\beta^2-\tau} \leq U_\phi^* \leq 1$ (see Figure 3.10).

If $\phi'(-f^*(\bar{\mathbf{x}})) = 0$, it is easy to see the contradiction.

Combining the above cases, it follows that

$$\sup_{f \notin \mathcal{B}} U_\phi(f) < \sup_f U_\phi(f).$$

□

3.8.2 Proof of Theorem 3.15

We simply let U denote the Jaccard index and U_ϕ denote the surrogate utility such that

$$U_\phi(f) = \frac{\int_{\mathcal{X}}(1-\phi(f(\mathbf{x})))\eta(\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x}}{\int_{\mathcal{X}}\{\phi(-f(\mathbf{x}))(1-\eta(\mathbf{x})) + \pi\}p(\mathbf{x})\mathrm{d}\mathbf{x}}.$$

In addition, we let $\mathcal{B}$ denote the Bayes-optimal set such that

$$\mathcal{B} := \{\, f \mid f(\mathbf{x})\{(1+U(f))\eta(\mathbf{x}) - U(f)\} > 0 \quad \forall \mathbf{x} \in \mathcal{X} \,\},$$

as characterized by Proposition 3.11. We follow the same proof technique, i.e., proof by contradiction, that we used in the proof of Theorem 3.14. Assume that

$$\sup_{f \notin \mathcal{B}} U_\phi(f) = \sup_f U_\phi(f),$$

which implies that there exists an optimal function $f^* \notin \mathcal{B}$ that achieves $U_\phi(f^*) = \sup_f U_\phi(f) := U_\phi^*$, that is, $U_\phi(f^*) = U_\phi^*$ and $f^*(\bar{\mathbf{x}})\{(1+U(f^*))\eta(\bar{\mathbf{x}}) - U(f^*)\} \leq 0$ for some $\bar{\mathbf{x}} \in \mathcal{X}$. By Assumption 3.13, the case $(1+U(f^*))\eta(\bar{\mathbf{x}}) - U(f^*) = 0$ is excluded.

The stationary condition of U_ϕ around f^* can be stated along with Equation (3.9) in Theorem 3.14:

$$\eta(\bar{\mathbf{x}}) = \frac{\phi'(-f^*(\bar{\mathbf{x}}))U_\phi^*}{\phi'(f^*(\bar{\mathbf{x}})) + \phi'(-f^*(\bar{\mathbf{x}}))U_\phi^*}. \tag{3.10}$$

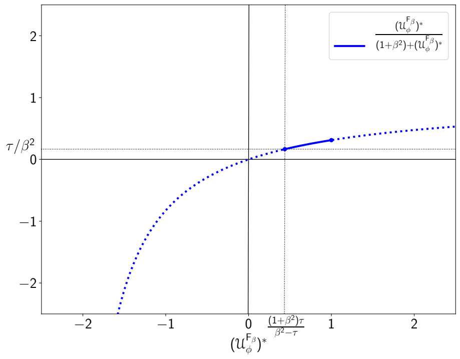


Figure 3.8: The range of $\frac{U_\phi^*}{(1+\beta^2)+U_\phi^*}$ in $\frac{(1+\beta^2)\tau}{\beta^2-\tau} \le U_\phi^* \le 1$.

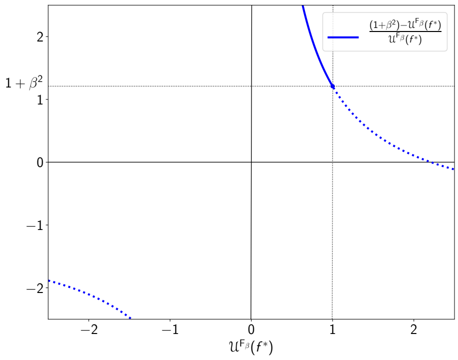


Figure 3.9: The range of $\frac{(1+\beta^2)-U(f^*)}{U(f^*)}$ in $0 < U(f^*) \le 1$.

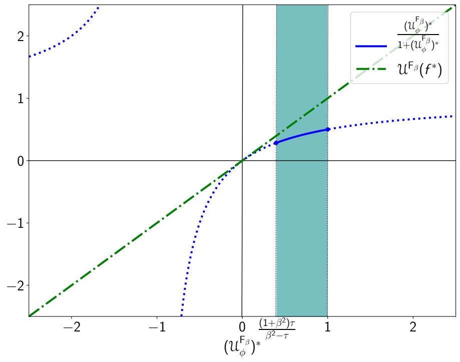


Figure 3.10: If $\frac{(1+\beta^2)\tau}{\beta^2-\tau} \le U_\phi^* \le 1$, then $\frac{U_\phi^*}{1+U_\phi^2} \le U(f^*)$.

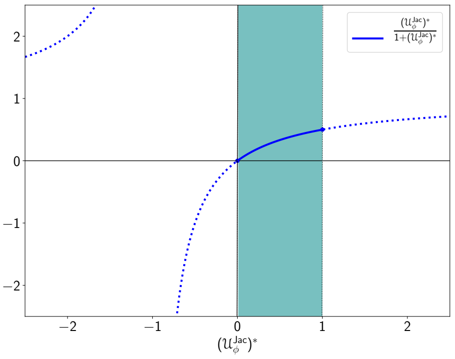


Figure 3.11: $\frac{U_\phi^*}{1+U_\phi^*}$ is monotonically increasing if $0 \le U_\phi^* \le 1$.

1) If $f^*(\bar{\mathbf{x}}) > 0$ and $\eta(\bar{\mathbf{x}}) < \frac{U(f^*)}{1+U(f^*)}$, we show $\eta(\bar{\mathbf{x}}) \ge \frac{U(f^*)}{1+U(f^*)}$. First, take the difference between the left- and the right-hand sides.

$$\eta(\bar{\mathbf{x}}) - \frac{U(f^*)}{1+U(f^*)} = \frac{\phi'(-f^*(\bar{\mathbf{x}}))U_\phi^* - \phi'(f^*(\bar{\mathbf{x}}))U(f^*)}{(\phi'(f^*(\bar{\mathbf{x}})) + \phi'(-f^*(\bar{\mathbf{x}}))U_\phi^*)(1+U(f^*))},$$

where the denominator is always negative. Next, we show that the numerator is always negative. If $\phi'(-f^*(\bar{\mathbf{x}})) < 0$,

$$\begin{aligned}
&\phi'(-f^*(\bar{\mathbf{x}}))U_\phi^* - \phi'(f^*(\bar{\mathbf{x}}))U(f^*) \\
&\quad = \phi'(-f^*(\bar{\mathbf{x}}))\left(U_\phi^* - \frac{\phi'(f^*(\bar{\mathbf{x}}))}{\phi'(-f^*(\bar{\mathbf{x}}))}U(f^*)\right) \\
&\quad \le \phi'(-f^*(\bar{\mathbf{x}}))\left(U_\phi^* - \frac{\phi'(f^*(\bar{\mathbf{x}}))}{\phi'(-f^*(\bar{\mathbf{x}}))}\right) && (\because U(f^*) \le 1) \\
&\quad \le \phi'(-f^*(\bar{\mathbf{x}}))(U_\phi^* - \tau) \\
&\quad \le 0, && (\because U_\phi^* \ge \tau)
\end{aligned}$$

where the assumption $\lim_{m\searrow 0}\phi'(m) \ge \tau \lim_{m\nearrow 0}\phi'(m)$ (for $m > 0$) and convexity of ϕ imply the second inequality. Thus, $\eta(\bar{\mathbf{x}}) \ge \frac{U(f^*)}{1+U(f^*)}$ holds, which is contradiction.

If $\phi'(-f^*(\bar{\mathbf{x}})) = 0$, then $\phi'(f^*(\bar{\mathbf{x}})) = 0$ from the assumption $\lim_{m\searrow 0}\phi'(m) \ge \tau \lim_{m\nearrow 0}\phi'(m)$, which immediately results in a contradiction.

2) If $f^*(\bar{\mathbf{x}}) \le 0$ and $\eta(\bar{\mathbf{x}}) > \frac{U(f^*)}{1+U(f^*)}$, we begin from the stationary condition Equation (3.10). If $\phi'(-f^*(\bar{\mathbf{x}})) < 0$,

$$\eta(\bar{\mathbf{x}}) = \frac{U_\phi^*}{\frac{\phi'(f^*(\bar{\mathbf{x}}))}{\phi'(-f^*(\bar{\mathbf{x}}))} + U_\phi^*} \le \frac{U_\phi^*}{1+U_\phi^*} \le \frac{U(f^*)}{1+U(f^*)},$$

where the first inequality follows from $\frac{\phi'(m)}{\phi'(-m)} \ge 1$ for $m \le 0$, and the second inequality follows because $U_\phi(f) \le U(f)$ $(\forall f)$ and the function $x \mapsto \frac{x}{1+x}$ $(0 \le x \le 1)$ is monotonically increasing (see Figure 3.11). This contradicts $\eta(\bar{\mathbf{x}}) > \frac{U(f^*)}{1+U(f^*)}$.

It is easy to see such a contradiction in the case of $\phi'(-f^*(\bar{\mathbf{x}})) = 0$.

Combining the above cases, it follows that

$$\sup_{f \notin \mathcal{B}} U_\phi(f) < \sup_f U(f).$$

□

3.8.3 Proof of Lemma 3.16

First, we need to carefully analyze our *non-smooth* surrogate loss to handle the Rademacher complexity.

Definition 3.18 (Rademacher complexity). *Let $\mathcal{S} := \{ z_1, \dots, z_n \} \subseteq \mathcal{Z}$ be a sample with size n. In addition, let $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{Z}}$ be a subset of measurable functions, and $\boldsymbol{\sigma} := (\sigma_1, \dots, \sigma_n)$ be the Rademacher variables. Then, the Rademacher complexity of $\mathcal{G}$ of the sample size n is defined as*

$$\mathfrak{R}_n(\mathcal{G}) := \mathop{\mathbb{E}}_{\mathcal{S}} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i g(z_i) \right| \right].$$

We usually analyze the Rademacher complexity of the composite function class $\phi \circ \mathcal{F} := \{(\mathbf{x}, y) \mapsto \phi(yf(\mathbf{x})) \mid f \in \mathcal{F}\}$ by applying the Ledoux-Talagrand's contraction inequality [Ledoux and Talagrand, 1991] when the surrogate ϕ is Lipschitz continuous: $\mathfrak{R}_n(\phi \circ \mathcal{F}) \le 2\rho_\phi \mathfrak{R}_n(\mathcal{F})$, where ρ_ϕ is the Lipschitz norm of ϕ. By contrast, we need to deal with the case of the uniform convergence of the gradients, which requires a smoothness of the surrogate, whereas the τ-discrepant loss is a non-smooth surrogates. Thus, we need an alternative analysis.

Lemma 3.19. *Assume that ϕ is τ-discrepant and can be decomposed as $\phi(m) = \phi_{+1}(m)\mathbb{1}_{\{m>0\}} + \phi_{-1}(m)\mathbb{1}_{\{m \le 0\}}$. For $k = 0, 1$, define*

$$\widetilde{W}'_{k,\phi} \circ \mathcal{F} := \left\{ (\mathbf{x}, y) \mapsto \widetilde{W}'_{k,\phi}(f(\mathbf{x}), y) \;\middle|\; f \in \mathcal{F} \right\}.$$

Then,

$$\mathfrak{R}_n(\widetilde{W}'_{k,\phi} \circ \mathcal{F}) \le 2(\gamma_{+1} + \gamma_{-1}) \mathfrak{R}_n(\mathcal{F}).$$

Proof. First, we provide a proof for $k = 0$. Note that $\widetilde{W}'_{0,\phi}(f(\mathbf{x}), y) = -y\phi'(yf(\mathbf{x}))$.

$$\begin{aligned}
\mathfrak{R}_n(\widetilde{W}'_{0,\phi} \circ \mathcal{F}) &= \mathop{\mathbb{E}}_{\mathcal{S},\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i(-y_i\phi'(y_i f(\mathbf{x}_i)))\right|\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{S},\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\phi'(y_i f(\mathbf{x}_i))\right|\right] \\
&\le \underbrace{\mathop{\mathbb{E}}_{\mathcal{S},\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\phi'_{-1}(y_i f(\mathbf{x}_i))\mathbb{1}_{\{y_i f(\mathbf{x}_i)\le 0\}}\right|\right]}_{\text{(A)}} \\
&\quad + \underbrace{\mathop{\mathbb{E}}_{\mathcal{S},\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\phi'_{+1}(y_i f(\mathbf{x}_i))\mathbb{1}_{\{y_i f(\mathbf{x}_i)> 0\}}\right|\right]}_{\text{(B)}},
\end{aligned}$$

where the triangular inequality is invoked at the last inequality. For (A), let $\psi_{-1}(m) := \phi'_{-1}(m)\frac{m}{|m|}$ if $m \neq 0$, and $\psi_{-1}(0) := 0$. Because $\psi'_{-1}(m) = \phi''_{-1}(m)\frac{m}{|m|}$, the Lipschitz norm of ψ_{-1} can be computed as

$$\sup_{f\in\mathcal{F},(\mathbf{x},y)\in\mathcal{X}\times\mathcal{Y}} |\psi'_{-1}(f(\mathbf{x}))| = \sup_{f,\mathbf{x},y}|\phi''_{-1}(yf(\mathbf{x}))| \cdot \sup_{f,\mathbf{x},y}\left|\frac{yf(\mathbf{x})}{|yf(\mathbf{x})|}\right| = \gamma_{-1}.$$

Note that the Lipschitz norm of ϕ'_{-1} is γ_{-1} because ϕ_{-1} is γ_{-1}-smooth. We then further bound (A) by using the fact $\mathbb{1}_{\{y_i f(\mathbf{x}_i)\le 0\}} = \frac{1-\frac{y_i f(\mathbf{x}_i)}{|y_i f(\mathbf{x}_i|}}{2}$.

$$\begin{aligned}
\text{(A)} &= \mathop{\mathbb{E}}_{\mathcal{S},\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\phi'_{-1}(y_i f(\mathbf{x}_i))\frac{1-\frac{y_i f(\mathbf{x}_i)}{|y_i f(\mathbf{x}_i)|}}{2}\right|\right] \\
&\le \frac{1}{2}\mathop{\mathbb{E}}_{\mathcal{S},\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\phi'_{-1}(y_i f(\mathbf{x}_i))\right|\right] + \frac{1}{2}\mathop{\mathbb{E}}_{\mathcal{S},\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\psi_{-1}(y_i f(\mathbf{x}_i))\right|\right] \\
&= \frac{1}{2}\mathfrak{R}_n(\phi'_{-1}\circ\mathcal{F}) + \frac{1}{2}\mathfrak{R}_n(\psi_{-1}\circ\mathcal{F}) \\
&\le \frac{1}{2}\cdot 2\gamma_{-1}\mathfrak{R}_n(\mathcal{F}) + \frac{1}{2}\cdot 2\gamma_{-1}\mathfrak{R}_n(\mathcal{F}) \\
&= 2\gamma_{-1}\mathfrak{R}_n(\mathcal{F}),
\end{aligned}$$

where the second inequality is the result of Ledoux-Talagrand's contraction inequality [Ledoux and Talagrand, 1991, Theorem 4.12]. Note that both ϕ'_{-1} and ψ_{-1} are γ_{-1}-Lipschitz. The bound (B) $\le 2\gamma_{+1}\mathfrak{R}_n(\mathcal{F})$ can be proven as well. We omit the proof for $k = 1$, which follows the same proof strategy. □

Now, we move on to the proof of Lemma 3.16.

Proof of Lemma 3.16. We write $V_\phi(f_{\boldsymbol{\theta}})$ as $V_\phi(\boldsymbol{\theta})$. If we explicit note for which sample we take the empirical average in $\widehat{V}_\phi(\boldsymbol{\theta})$, let us write $\widehat{V}_\phi(\boldsymbol{\theta};\mathcal{S})$. Let $E(\mathcal{S}) := \sup_{\boldsymbol{\theta}\in\Theta}\|\widehat{V}_\phi(\boldsymbol{\theta};\mathcal{S}) - V_\phi(\boldsymbol{\theta})\|$. For simplicity, we write $z_i := (\mathbf{x}_i, y_i)$ and $\widetilde{W}_{0,\phi}(\boldsymbol{\theta}; z_i) := \widetilde{W}_{0,\phi}(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i)$. First, we observe $E(\mathcal{S})$ admits the bounded difference property [McDiarmid, 1989].

Denote $\mathcal{S} := \{ z_i \}_{i \in [n]}$ and $\mathcal{S}' := \{ z_1, \dots, z_k', \dots, z_n \}$. If $1 \le k \le m$,

$$\begin{aligned}
&\sup_{\substack{\mathcal{S} \subseteq \mathcal{X} \times \mathcal{Y} \\ z_k' \in \mathcal{X} \times \mathcal{Y}}} |E(\mathcal{S}) - E(\mathcal{S}')| \\
&\le \sup_{\mathcal{S}, z_k', \boldsymbol{\theta}} \| \widehat{V}_\phi(\boldsymbol{\theta}; \mathcal{S}) - \widehat{V}_\phi(\boldsymbol{\theta}; \mathcal{S}') \| \qquad (\because \text{triangular inequality}) \\
&= \frac{1}{m(n-m)} \sup_{\mathcal{S}, z_k', \boldsymbol{\theta}} \left\| \left\{ \nabla \widetilde{W}_{0,\phi}(\boldsymbol{\theta}; z_k) - \nabla \widetilde{W}_{0,\phi}(\boldsymbol{\theta}; z_k') \right\} \sum_{j=m+1}^{n} \widetilde{W}_{1,\phi}(\boldsymbol{\theta}; z_j) \right. \\
&\qquad\qquad \left. - \left\{ \widetilde{W}_{0,\phi}(\boldsymbol{\theta}; z_k) - \widetilde{W}_{0,\phi}(\boldsymbol{\theta}; z_k') \right\} \sum_{j=m+1}^{n} \nabla \widetilde{W}_{1,\phi}(\boldsymbol{\theta}; z_j) \right\| \\
&\le \frac{1}{m(n-m)} \sup_{\mathcal{S}, z_k', \boldsymbol{\theta}} \left\{ \left(\| \nabla \widetilde{W}_{0,\phi}(\boldsymbol{\theta}; z_k) \| + \| \nabla \widetilde{W}_{0,\phi}(\boldsymbol{\theta}; z_k') \| \right) \sum_{j=m+1}^{n} | \widetilde{W}_{1,\phi}(\boldsymbol{\theta}; z_j) | \right. \\
&\qquad\qquad \left. + \left(| \widetilde{W}_{0,\phi}(\boldsymbol{\theta}; z_k) | + | \widetilde{W}_{0,\phi}(\boldsymbol{\theta}; z_k') | \right) \sum_{j=m+1}^{n} \| \nabla \widetilde{W}_{1,\phi}(\boldsymbol{\theta}; z_j) \| \right\} \\
&\le \frac{2\rho_0 c_{\mathcal{X}} \cdot (n-m) c_1 + 2 c_0 \cdot (n-m) \rho_1 c_{\mathcal{X}}}{m(n-m)} \\
&= \frac{4 c_{\mathcal{X}} (\rho_1 c_0 + \rho_0 c_1)}{n},
\end{aligned}$$

where the second inequality also holds owing to the triangular inequality, and the last inequality follows from the fact that $\widetilde{W}_{0,\phi}$ and $\widetilde{W}_{1,\phi}$ are ρ_0-/ρ_1-Lipschitz and bounded by c_0 and c_1, respectively. The same holds for the case $m+1 \le k \le n$.

Thus, E is the bounded difference with a constant $(4c_{\mathcal{X}}(\rho_1 c_0 + \rho_0 c_1))/n$ for each index, and we can obtain the following inequality by McDiarmid's inequality [McDiarmid, 1989], i.e., with a probability of at least $1-\delta$,

$$E(\mathcal{S}) - \mathop{\mathbb{E}}_{\mathcal{S}}[E(\mathcal{S})] \le \sqrt{\frac{8 c_{\mathcal{X}}^2 (\rho_1 c_0 + \rho_0 c_1)^2 \log \frac{2}{\delta}}{n}}.$$

Next, we bound $\mathbb{E}_{\mathcal{S}}[E(\mathcal{S})]$ based on the *symmetrization device* [Ledoux and Talagrand, 1991, Lemma 6.3].

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{S}}[E(\mathcal{S})] \\
&\le \underbrace{\mathop{\mathbb{E}}_{\mathcal{S}} \sup_{\boldsymbol{\theta}} \left\| \frac{1}{m(n-m)} \sum_{i=1}^{m} \sum_{j=m+1}^{n} \widetilde{W}_{1,\phi}(\boldsymbol{\theta}; z_j) \nabla \widetilde{W}_{0,\phi}(\boldsymbol{\theta}; z_i) - \mathbb{E}[W_{1,\phi} \nabla W_{0,\phi}] \right\|}_{\text{(A)}} \\
&\quad + \underbrace{\mathop{\mathbb{E}}_{\mathcal{S}} \sup_{\boldsymbol{\theta}} \left\| \frac{1}{m(n-m)} \sum_{i=1}^{m} \sum_{j=m+1}^{n} \widetilde{W}_{0,\phi}(\boldsymbol{\theta}; z_i) \nabla \widetilde{W}_{1,\phi}(\boldsymbol{\theta}; z_j) - \mathbb{E}[W_{0,\phi} \nabla W_{1,\phi}] \right\|}_{\text{(B)}},
\end{aligned} \tag{3.11}$$

where the second line is the result of the triangular inequality, and

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{S}}[(\mathrm{A})] \\
&= \mathop{\mathbb{E}}_{\mathcal{S}} \sup_{\boldsymbol{\theta}} \Bigg\| \frac{1}{m(n-m)} \sum_{i=1}^{m} \sum_{j=m+1}^{n} \widetilde{W}_{1,\phi}(\boldsymbol{\theta}; z_j) \left(\nabla \widetilde{W}_{0,\phi}(\boldsymbol{\theta}; z_i) - \mathbb{E}[\nabla W_{0,\phi}] \right) \\
&\qquad\qquad + \frac{1}{m(n-m)} \sum_{i=1}^{m} \sum_{j=m+1}^{n} \mathbb{E}[\nabla W_{0,\phi}] \left(\widetilde{W}_{1,\phi}(\boldsymbol{\theta}; z_j) - \mathbb{E}[W_{1,\phi}] \right) \Bigg\| \\
&\le \mathop{\mathbb{E}}_{\mathcal{S}} \sup_{\boldsymbol{\theta}} \Bigg\{ \frac{1}{m(n-m)} \sum_{j=m+1}^{n} |\widetilde{W}_{1,\phi}(\boldsymbol{\theta}; z_j)| \cdot \left\| \sum_{i=1}^{m} \nabla \widetilde{W}_{0,\phi}(\boldsymbol{\theta}; z_i) - \mathbb{E}[\nabla W_{0,\phi}] \right\| \\
&\qquad\qquad + \frac{1}{m(n-m)} \sum_{i=1}^{m} \| \mathbb{E}[\nabla W_{0,\phi}] \| \cdot \left| \sum_{j=m+1}^{n} \widetilde{W}_{1,\phi}(\boldsymbol{\theta}; z_j) - \mathbb{E}[W_{1,\phi}] \right| \Bigg\} \\
&= c_1 \underbrace{\mathop{\mathbb{E}}_{\mathcal{S}} \left[\sup_{\boldsymbol{\theta}} \left\| \frac{1}{m} \sum_{i=1}^{m} \nabla \widetilde{W}_{0,\phi}(\boldsymbol{\theta}; z_i) - \mathbb{E}[\nabla W_{0,\phi}] \right\| \right]}_{\text{(A')}} \\
&\quad + \rho_0 c_{\mathcal{X}} \underbrace{\mathop{\mathbb{E}}_{\mathcal{S}} \left[\sup_{\boldsymbol{\theta}} \left| \frac{1}{n-m} \sum_{j=m+1}^{n} \widetilde{W}_{1,\phi}(\boldsymbol{\theta}; z_j) - \mathbb{E}[W_{1,\phi}] \right| \right]}_{\text{(A")}},
\end{aligned}
$$

where the first inequality is the triangular inequality. Now we introduce the Rademacher random variables $\boldsymbol{\sigma}_{1:n} \coloneqq (\sigma_1, \ldots, \sigma_n)$.

- For (A'),

$$
\begin{aligned}
\text{(A')} &\le \mathop{\mathbb{E}}_{\mathcal{S}, \boldsymbol{\sigma}_{1:m}} \left[\sup_{\boldsymbol{\theta}} \sum_{l=1}^{d} \left| \frac{1}{n} \sum_{i=1}^{m} \nabla_{\theta_l} \widetilde{W}_{0,\phi}(\theta; z_i) - \mathbb{E}[\nabla_{\theta_l} W_{0,\phi}] \right| \right] \\
&\le \sum_{l=1}^{d} \mathop{\mathbb{E}}_{\mathcal{S}, \boldsymbol{\sigma}_{1:m}} \left[\sup_{\boldsymbol{\theta}} \left| \frac{2}{m} \sum_{i=1}^{m} \sigma_i \nabla_{\theta_l} \widetilde{W}_{0,\phi}(\boldsymbol{\theta}; z_i) \right| \right] \\
&= \sum_{l=1}^{d} 2 \mathop{\mathbb{E}}_{\mathcal{S}, \boldsymbol{\sigma}_{1:m}} \left[\sup_{\boldsymbol{\theta}} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i \widetilde{W}'_{0,\phi}(\boldsymbol{\theta}; z_i) \cdot x_l \right| \right] \\
&\le \sum_{l=1}^{d} 2 \mathop{\mathbb{E}}_{\mathcal{S}, \boldsymbol{\sigma}_{1:m}} \left[\sup_{\boldsymbol{\theta}} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i \widetilde{W}'_{0,\phi}(\boldsymbol{\theta}; z_i) \right| \cdot c_{\mathcal{X}} \right] \\
&\le 4 d c_{\mathcal{X}} (\gamma_{+1} + \gamma_{-1}) \mathfrak{R}_m(\mathcal{F}) \\
&= 4 d c_{\mathcal{X}} (\gamma_{+1} + \gamma_{-1}) \mathfrak{R}_{n/2}(\mathcal{F}),
\end{aligned}
$$

  where the first inequality is owing to $\|\cdot\|_2 \le \|\cdot\|_1$, the second inequality is owing to the symmetrization device, the third line is owing to $\nabla_{\boldsymbol{\theta}} \widetilde{W}_{0,\phi}(\boldsymbol{\theta}^\top \mathbf{x}), y) = -y\phi'(y\boldsymbol{\theta}^\top \mathbf{x}) \cdot \mathbf{x}$, and the last inequality uses Lemma 3.19.

- For (A"),

$$
\begin{aligned}
\text{(A")} &\le \mathop{\mathbb{E}}_{\mathcal{S}, \boldsymbol{\sigma}_{m:n-m}} \left[\sup_{\boldsymbol{\theta}} \left| \frac{2}{n-m} \sum_{j=m+1}^{n} \sigma_j \widetilde{W}_{1,\phi}(\boldsymbol{\theta}; z_i) \right| \right] \\
&\le 4\rho_1 \mathfrak{R}_{n-m}(\mathcal{F}) = 4\rho_1 \mathfrak{R}_{n/2}(\mathcal{F}),
\end{aligned}
$$

where the first inequality is owing to the symmetrization device and the second inequality uses Ledoux-Talagrand's contraction inequality [Ledoux and Talagrand, 1991, Theorem 4.12], together with the fact that $\widetilde{W}_{1,\phi}$ is ρ_1-Lipschitz continuous.

Thus, Equation (3.11) can be bounded as follows.

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{S}}[E(\mathcal{S})] \\
&\le c_1(\text{A'}) + \rho_0 c_{\mathcal{X}}(\text{A"}) + \mathop{\mathbb{E}}_{\mathcal{S}}[(\text{B})] \\
&\le 4dc_{\mathcal{X}}c_1(\gamma_{+1} + \gamma_{-1})\mathfrak{R}_{n/2}(\mathcal{F}) + 4\rho_0\rho_1 c_{\mathcal{X}}\mathfrak{R}_{n/2}(\mathcal{F}) \\
&\qquad + \underbrace{4dc_{\mathcal{X}}c_0(\gamma_{+1} + \gamma_{-1})\mathfrak{R}_{n/2}(\mathcal{F}) + 4\rho_1\rho_0 c_{\mathcal{X}}\mathfrak{R}_{n/2}(\mathcal{F})}_{\text{can be proven in the same manner as (A)}} \\
&= (4c_{\mathcal{X}}c_0 d\gamma + 4c_{\mathcal{X}}c_1 d\gamma + 8\rho_0\rho_1 c_{\mathcal{X}})\mathfrak{R}_{n/2}(\mathcal{F}) \qquad (\gamma \coloneqq \gamma_{+1} + \gamma_{-1}) \\
&\le (4c_{\mathcal{X}}c_0 d\gamma + 4c_{\mathcal{X}}c_1 d\gamma + 8\rho_0\rho_1 c_{\mathcal{X}})\frac{\sqrt{2}c_{\mathcal{X}}c_\Theta}{\sqrt{n}},
\end{aligned}$$

where the last inequality comes from Mohri et al. [2018, Theorem 4.3].

Finally, we obtain the desired uniform bound: with a probability of at least $1 - \delta$,

$$\begin{aligned}
\sup_{\boldsymbol{\theta}\in\Theta} \left\| \widehat{V}_\phi(\boldsymbol{\theta}; \mathcal{S}) - V_\phi(\boldsymbol{\theta}) \right\| &= E(\mathcal{S}) \\
&\le \mathop{\mathbb{E}}_{\mathcal{S}}[E(\mathcal{S})] + \frac{\sqrt{8}c_{\mathcal{X}}(\rho_1 c_0 + \rho_0 c_1)\sqrt{\log\frac{2}{\delta}}}{\sqrt{n}} \\
&\le \frac{(4c_{\mathcal{X}}c_0 d\gamma + 4c_{\mathcal{X}}c_1 d\gamma + 8\rho_0\rho_1 c_{\mathcal{X}}) + \sqrt{8}c_{\mathcal{X}}(\rho_1 c_0 + \rho_0 c_1)\sqrt{\log\frac{2}{\delta}}}{\sqrt{n}}.
\end{aligned}$$

□

3.9 Conclusion

In this chapter, we provided a new insight into a calibrated surrogate for the linear-fractional metrics. Sufficient conditions for a surrogate calibration were given, which to the best of our knowledge are the first calibration results for the linear-fractional metrics. A surrogate maximization can be conducted through the combination of concave and quasiconcave programs, and its performance is validated based on simulations.

Chapter 4

Calibrated Surrogate Losses for Robust Classification

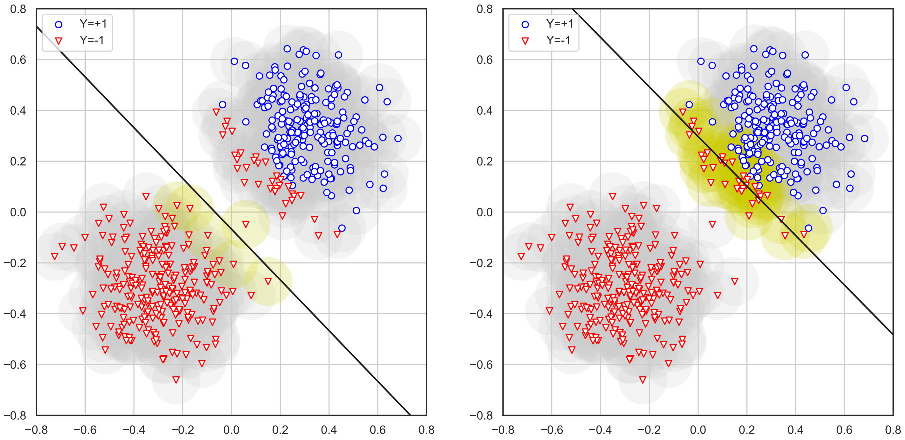
Adversarially robust classification seeks a classifier that is insensitive to adversarial perturbations of the test patterns. This problem is often formulated through a minimax objective, where the target loss is the worst-case value of the 0-1 loss subject to a bound placed on the size of the perturbation. In an effort to make the optimization more tractable, recent studies have proposed convex surrogates for the adversarial 0-1 loss. A primary question is that of consistency, that is, whether a minimization of the surrogate risk implies a minimization of the adversarial 0-1 risk. In this chapter, we analyze this question through the lens of calibration, which is a pointwise notion of consistency. We show that no convex surrogate loss is calibrated with respect to the adversarial 0-1 loss when restricted to the class of linear models. We further introduce a class of nonconvex losses and offer necessary and sufficient conditions for losses in this class to be calibrated. We also show that if the underlying distribution satisfies Massart's noise condition, convex losses can also be calibrated within an adversarial setting.

4.1 Introduction

In conventional machine learning, training and testing instances are assumed to follow the same probability distribution. In *adversarially robust* machine learning, test instances may be perturbed by an adversary before being presented to the predictor. Recent studies have shown that seemingly insignificant adversarial perturbations can lead to significant performance degradations of otherwise highly accurate classifiers [Goodfellow et al., 2015]. This has led to the development of a number of methods for learning predictors with decreased sensitivity to adversarial perturbations [Xu et al., 2009, Xu and Mannor, 2012, Goodfellow et al., 2015, Cisse et al., 2017, Wong and Kolter, 2018, Raghunathan et al., 2018a, Tsuzuku et al., 2018].

Adversarially robust classification is typically formulated as an empirical risk minimization with an *adversarial 0-1 loss*, which is the maximum of the usual 0-1 loss over a set of possible perturbations of the test instance. This minimax optimization problem is nonconvex, and a recent study, reviewed in Section 4.4, has proposed several convex surrogate losses. However, it is still unknown whether minimizing these convex surrogates leads to a minimization of the adversarial 0-1 loss.

In this chapter, we examine the question of which surrogate losses are calibrated with respect to (w.r.t.) the adversarial 0-1 loss. The term "calibration", defined accurately below, means that for each possible input $\mathbf{x}$, minimization of the excess surrogate risk (over a specified class of decision functions) implies a



(a) Ramp loss ($\beta = 0.5$)

(b) Hinge loss ($\beta = 0.5$)

Figure 4.1: The best linear classifier under each loss. The shift parameter β for a surrogate loss is defined in Section 4.8. The ℓ_2-balls associated with each instance indicate adversarial perturbations with a radius of 0.1. The yellow balls indicate instances vulnerable to perturbations, in that they are within 0.1 of the decision boundary. In this example, 1.2% of instances are vulnerable under the ramp loss, whereas 10.4% of the instances are vulnerable under the hinge loss.

minimization of the excess target risk. Calibration thus ensures pointwise consistency, and this notion has been repeatedly used to prove the *consistency* of algorithms based on the surrogate losses. Employing the calibration function perspective of Steinwart [2007], we show that no convex surrogate loss is calibrated w.r.t. the adversarial 0-1 loss for general distributions when restricted to the class of linear models (Section 4.6). Intuitively, this is because convex losses prefer predictions close to the decision boundary on average when $\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X}) \approx \frac{1}{2}$, whereas predictions that are too close to the decision boundary should be penalized in adversarially robust classification. We also provide necessary and sufficient conditions for a certain class of nonconvex losses to be calibrated w.r.t. the adversarial 0-1 loss (Section 4.7). These calibrated losses attain robustness by penalizing predictions that are too close to the decision boundary. Finally, we show that under a certain type of low-noise condition [Massart and Nédélec, 2006], convex losses can be calibrated (Section 4.9).

Our analysis depends on the fact that the adversarially robust 0-1 loss equals the horizontally shifted (non-robust) 0-1 loss when restricted to linear models (Proposition 4.1). In summary, we argue against the use of convex losses in adversarially robust classification (with linear models), and calibrated nonconvex losses serve as good alternatives.

Our results demonstrate that adversarial robustness requires different surrogates than other notions of robustness. For example, symmetric losses such as the sigmoid and ramp losses are robust to label noise [Ghosh et al., 2015], but not calibrated w.r.t. the adversarial 0-1 loss. Figure 4.1 illustrates the results of learning a linear classifier w.r.t. a *shifted* ramp loss, which is calibrated w.r.t. the adversarial 0-1 loss, and a shifted hinge loss, which is not (these losses are discussed in detail later). Whereas the hinge loss yields a classifier with smaller misclassification rate w.r.t. the conventional 0-1 loss, this classifier is quite sensitive to small perturbations of the test instances. The classifier learned by the ramp loss, by contrast, makes fewer errors when subjected to adversarial perturbations.

4.1.1 Organization of This Chapter

The rest of this chapter is organized as follows. Section 4.3 formalizes the notations and the problem. Related studies on robust learning and calibration analysis are reviewed in Section 4.4. Technical details of the calibration analysis are reviewed in Section 4.5. Section 4.6 describes the nonexistence of convex calibrated surrogate losses, whereas Section 4.7 presents general calibration conditions for a certain class of nonconvex losses. Section 4.8 applies our theory to several convex and nonconvex losses for the calibrated nonconvex losses. Calibration analysis under low-noise conditions is shown in Section 4.9. Section 4.10 shows simulation results verifying that calibrated losses achieve an excess target risk tending toward zero under a robust 0-1 loss. Finally, some concluding remarks are given in Section 4.12.

4.2 Notation and Preliminaries

4.2.1 Basic Notation

Let $\mathcal{B}_p(r) := \left\{ \mathbf{v} \in \mathbb{R}^d \;\middle|\; \|\mathbf{v}\|_p \le r \right\}$ be a d-dimensional closed ℓ_p-ball with radius r, and $\mathcal{B}_p^\circ(r) := \left\{ \mathbf{v} \in \mathbb{R}^d \;\middle|\; \|\mathbf{v}\|_p < r \right\}$ be an open ℓ_p-ball. We define the infimum over an empty set as $+\infty$. For a function $h : S \to \mathbb{R}$, we write $h^{\star\star} : S \to \mathbb{R}$ for the Fenchel-Legendre biconjugate of h, characterized by $\mathrm{epi}(h^{\star\star}) = \overline{\mathrm{co}}\,\mathrm{epi}(h)$, where $\overline{\mathrm{co}}\,S$ is the closure of the convex hull of the set S, and $\mathrm{epi}(h)$ is the epigraph of the function h: $\mathrm{epi}(h) := \{ (\mathbf{x}, t) \mid x \in S, h(\mathbf{x}) \le t \}$.

4.2.2 Convex and Quasiconvex Analysis

This subsection summarizes the basic tools used for convex and quasiconvex analyses.

Quasiconvex function. A function $h : S \to \mathbb{R}$ on a (finite-dimensional) vector space S is said to be *quasiconvex* if for all $\mathbf{x}, \mathbf{y} \in S$ and $\lambda \in [0, 1]$, $h(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \le \max\{ h(\mathbf{x}), h(\mathbf{y}) \}$. A function h is said to be quasiconcave if $-h$ is quasiconvex: For all $\mathbf{x}, \mathbf{y} \in S$ and $\lambda \in [0, 1]$, $h(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \ge \min\{ h(\mathbf{x}), h(\mathbf{y}) \}$. Intuitively, quasiconvexity relaxes the convexity in that a function still preserves the "unimodality" although it loses a definite curvature. There is an equivalent definition (herein, we only show this for quasiconcavity) such that h is quasiconcave if every superlevel set $\{ \mathbf{x} \mid h(\mathbf{x}) \ge t \}$ for $t \in \mathbb{R}$ is a convex set [Boyd and Vandenberghe, 2004].

Subderivative. To analyze the convexity and quasiconvexity, a subderivative is a useful tool. We adopt the Clarke definition of subderivative [Clarke, 1990, Aussel et al., 1994]. Let S^* be the dual space of S and $\langle \cdot, \cdot \rangle$ be the dual pairing.[1] The (Clarke) subderivative of a lower semicontinuous function h is the operator $\partial h : S \to S^*$ defined for each $\mathbf{x} \in S$ such that

$$\partial h(x) := \{ \mathbf{x}_* \in S^* \mid \langle \mathbf{x}_*, \mathbf{x} \rangle \le h^\circ(\mathbf{x}; \mathbf{v}) \quad \forall \mathbf{v} \in S \},$$

[1] For two vector spaces U and V over the same field F and a bilinear map $\langle \cdot, \cdot, \rangle : U \times V \to F$, we state that a triple $(U, V, \langle \cdot, \cdot, \rangle)$ is a dual pair if there exists $\mathbf{v} \in V$ such that $\langle \mathbf{u}, \mathbf{v} \rangle \ne 0$ for all $\mathbf{u} \in U$ and there exists $\mathbf{u} \in U$ such that $\langle \mathbf{u}, \mathbf{v} \rangle \ne 0$ for all $\mathbf{v} \in V$. Here, V is called a dual space of V, and $\langle \cdot, \cdot \rangle$ is called a dual pairing.

where $h^\circ(\mathbf{x};\mathbf{v})$ is the Rockafellar directional derivative (see Clarke [1990] and Aussel et al. [1994] for the formal definition). When h is locally Lipschitz at $\mathbf{x} \in S$, Clarke [1990] states that this is equivalent to

$$\partial h(\mathbf{x}) = \operatorname{co}\{\, \lim \nabla f(\mathbf{x}_i) \mid \mathbf{x}_i \to \mathbf{x}, \mathbf{x}_i \notin \Upsilon \cup \Omega_h \,\},$$

where co is the convex hull, Υ is any set of measure zero, and Ω_h is the set of points where h is non-differentiable.

Properties of subderivative. Several basic properties of subderivatives are shown in Clarke [1990, Section 2.3] such as

- (scalar multiples) $\partial(th)(\mathbf{x}) = t\partial h(\mathbf{x}) \coloneqq \{\, t\mathbf{x}_* \mid \mathbf{x}_* \in \partial h(\mathbf{x}) \,\}$,
- (finite sums) $\partial\left(\sum h_i\right)(x) \subseteq \sum \partial h_i(\mathbf{x}) \coloneqq \{\, \sum \mathbf{x}_{i,*} \mid \mathbf{x}_{i,*} \in \partial h_i(\mathbf{x}) \,\}$,
- $0 \in \partial h(\mathbf{x})$ if h attains a local extrema at $\mathbf{x}$.

When h is locally Lipschitz, it clearly holds that $\partial h(\mathbf{x}) = \{h'(\mathbf{x})\}$ if h is differentiable at $\mathbf{x}$.

Operator monotonicity. Convex smooth functions have monotonically non-decreasing derivatives. This can be extended to non-smooth functions through the subderivatives. Let $h : S \to \mathbb{R}$ be a lower semicontinuous function. Then h is convex if and only if $\partial h : S \to S^*$ is a *monotone* operator [Aussel et al., 1994], that is, $\langle \mathbf{y}_* - \mathbf{x}_*, \mathbf{y} - \mathbf{x} \rangle \geq 0$ for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom}(h)$ and $\mathbf{x}_* \in \partial h(\mathbf{x}), \mathbf{y}_* \in \partial h(\mathbf{y})$. In addition, h is quasiconvex if and only if ∂h is a *quasimonotone* operator [Aussel et al., 1994], that is, $\langle \mathbf{x}_*, \mathbf{y} - \mathbf{x} \rangle > 0 \implies \langle \mathbf{y}_*, \mathbf{y} - \mathbf{x} \rangle \geq 0$ for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom}(h)$ and $\mathbf{x}_* \in \partial h(\mathbf{x}), \mathbf{y}_* \in \partial h(\mathbf{y})$.

4.2.3 Formulation of Loss and Risk

Let $\mathcal{X} \coloneqq \mathcal{B}_2(1)$ be the feature space, $\mathcal{Y} \coloneqq \{\pm 1\}$ be the binary label space, and $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ be a function class. We consider *symmetric* $\mathcal{F}$, that is, $-f \in \mathcal{F}$ for all $f \in \mathcal{F}$. We write $\mathcal{F}_{\text{all}} \subseteq \mathbb{R}^{\mathcal{X}}$ for the space of all measurable functions. Let $\ell : \mathcal{Y} \times \mathcal{X} \times \mathcal{F} \to \mathbb{R}_{\geq 0}$ be a loss function.[2] Then, we write

$$R_\ell(f) \coloneqq \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})}[\ell(\mathsf{Y},\mathsf{X},f)]$$

for the *ℓ-risk* of $f \in \mathcal{F}$, where $(\mathsf{X},\mathsf{Y}) \in \mathcal{X} \times \mathcal{Y}$ are random variables jointly distributed following the underlying distribution $\mathbb{P}(\mathsf{X},\mathsf{Y})$. Subsequently, $\mathbb{P}(\mathsf{X})$ and $\mathbb{P}(\mathsf{Y}|\mathsf{X})$ denote the $\mathcal{X}$-marginal and the posterior distributions, respectively. If ℓ can be represented by $\ell(y,\mathbf{x},f) = \phi(yf(\mathbf{x}))$ with some $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ for any $y \in \mathcal{Y}$, $\mathbf{x} \in \mathcal{X}$, and $f \in \mathcal{F}$, ϕ is called a *margin-based* loss function. We define the *ϕ-risk* of $f \in \mathcal{F}$ for a margin-based loss ϕ by

$$R_\phi(f) \coloneqq \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})}[\phi(\mathsf{Y}f(\mathsf{X}))] = \mathop{\mathbb{E}}_{\mathsf{X}}\mathop{\mathbb{E}}_{\mathsf{Y}|\mathsf{X}}[\phi(\mathsf{Y}f(\mathsf{X}))], \tag{4.1}$$

where $\mathbb{E}_{\mathsf{X}}$ and $\mathbb{E}_{\mathsf{Y}|\mathsf{X}}$ indicate the expectation over $\mathbb{P}(\mathsf{X})$ and $\mathbb{P}(\mathsf{Y}|\mathsf{X})$, respectively. We can rewrite Equation (4.1) as $R_\phi(f) = \mathbb{E}_{\mathsf{X}}[C_\phi(f, \mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X}), \mathsf{X})]$ with

$$C_\phi(f,\eta,\mathbf{x}) \coloneqq \eta\phi(f(\mathbf{x})) + (1-\eta)\phi(-f(\mathbf{x})).$$

[2] Although the loss function ℓ introduced in Section 2.1.2 is defined over $(f(\mathbf{x}), y) \in \mathbb{R} \times \mathcal{Y}$, the loss function introduced in this chapter is defined over $(y, \mathbf{x}, f) \in \mathcal{Y} \times \mathcal{X} \times \mathcal{F}$. This is because the norm of the feature $\mathbf{x}$ has an influence on the loss value in adversarial robust classification.

We call $C_\phi(f, \eta, \mathbf{x})$ the *class-conditional ϕ-risk*, or ϕ-CCR. The minimal ϕ-risk (over a function class $\mathcal{F}$)

$$R^*_{\phi,\mathcal{F}} \coloneqq \inf_{f \in \mathcal{F}} R_\phi(f)$$

is called the *Bayes (ϕ, $\mathcal{F}$)-risk*, and the minimal ϕ-CCR on $\mathcal{F}$ at $\mathbf{x}$ is denoted by

$$C^*_{\phi,\mathcal{F}}(\eta, \mathbf{x}) \coloneqq \inf_{f \in \mathcal{F}} C_\phi(f, \eta, \mathbf{x}).$$

We refer to $R_\phi(f) - R^*_{\phi,\mathcal{F}}$ as the $(\phi, \mathcal{F})$-excess risk. We occasionally use the abbreviation

$$\Delta C_{\phi,\mathcal{F}}(f, \eta, \mathbf{x}) \coloneqq C_\phi(f, \eta, \mathbf{x}) - C^*_{\phi,\mathcal{F}}(\eta, \mathbf{x})$$

to denote the excess $(\phi, \mathcal{F})$-CCR at $\mathbf{x}$. For non-margin-based loss function ℓ, we define the ℓ-CCR $C_{\ell,\mathcal{F}}(f, \eta, \mathbf{x})$, the minimal ℓ-CCR $C^*_{\ell,\mathcal{F}}(\eta, \mathbf{x})$, and $\Delta C_{\ell,\mathcal{F}}(f, \eta, \mathbf{x})$ in the same manner.

4.3 Surrogate Losses for Adversarial Robust Classification

In supervised binary classification, a learner is asked to output a predictor $f : \mathcal{X} \to \mathbb{R}$ that minimizes the classification error $\mathbb{P}\{\mathsf{Y} f(\mathsf{X}) \le 0\}$, where $\mathbb{P}$ is the unknown underlying distribution. This can be equivalently interpreted as the minimization of the risk $\mathbb{E}_{(\mathsf{X},\mathsf{Y})}[\ell_{01}(\mathsf{Y}, \mathsf{X}, f)]$ w.r.t. f, where

$$\ell_{01}(y, \mathbf{x}, f) \coloneqq \begin{cases} 1 & \text{if } y \neq \operatorname{sgn}(f(\mathbf{x})), \\ 0 & \text{otherwise} \end{cases}$$

is the 0-1 loss. Here, we adopt the convention $\operatorname{sgn}(0) \coloneqq +1$. By contrast, an adversarially robust learner is asked to output a predictor f that minimizes the 0-1 loss while being tolerant to small perturbations to the input data points. Following existing studies [Xu et al., 2009, Tsuzuku et al., 2018, Bubeck et al., 2019], we consider ℓ_2-ball perturbations and define the goal as the minimization of

$$\mathbb{P}\{\exists \Delta_\mathbf{x} \in \mathcal{B}_2(\gamma) \text{ s.t. } \mathsf{X} + \Delta_\mathbf{x} \in \mathcal{X} \text{ and } \mathsf{Y} f(\mathsf{X} + \Delta_\mathbf{x}) \le 0\},$$

where $\Delta_\mathbf{x}$ is a perturbation vector and $\gamma \in (0, 1)$ is a pre-defined perturbation budget. Equivalently, the goal of adversarially robust classification is to minimize $\mathbb{E}_{(\mathsf{X},\mathsf{Y})}[\ell_\gamma(\mathsf{Y}, \mathsf{X}, f)]$ w.r.t. f, where

$$\ell_\gamma(y, \mathbf{x}, f) \coloneqq \begin{cases} 1 & \text{if } \exists \Delta_\mathbf{x} \in \mathcal{B}_2(\gamma) \text{ s.t. } \mathbf{x} + \Delta_\mathbf{x} \in \mathcal{X} \text{ and } y f(\mathbf{x} + \Delta_\mathbf{x}) \le 0, \\ 0 & \text{otherwise.} \end{cases}$$

We call this loss function ℓ_γ *adversarially robust 0-1 loss*, or *robust 0-1 loss* for short.

The robust 0-1 loss is a margin-based loss when restricted to the class of linear models $\mathcal{F}_\text{lin} \coloneqq \left\{ \mathbf{x} \mapsto \boldsymbol{\theta}^\top \mathbf{x} \;\middle|\; \boldsymbol{\theta} \in \mathbb{R}^d, \|\boldsymbol{\theta}\|_2 = 1 \right\} \subseteq \mathbb{R}^\mathcal{X}$. Note that $\mathcal{F}_\text{lin}$ is symmetric.

Proposition 4.1. *For any $\mathbf{x} \in \mathcal{X}$, $y \in \mathcal{Y}$, and $f \in \mathcal{F}_\text{lin}$, we have $\ell_\gamma(y, \mathbf{x}, f) = \mathbb{1}_{\{y f(\mathbf{x}) \le \gamma\}}$.*

Proof. Fix $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ and $f \in \mathcal{F}_{\text{lin}}$ associated with parameter $\boldsymbol{\theta} \in \mathbb{R}^d$. Because we can prove the case $y = -1$ in the same manner, assume $y = +1$ below without loss of generality.

We will check the existence of $\Delta_{\mathbf{x}} \in \mathcal{B}_2(\gamma)$ such that $\boldsymbol{\theta}^\top(\mathbf{x} + \Delta_{\mathbf{x}}) \le 0$ and $\mathbf{x} + \Delta_{\mathbf{x}} \in \mathcal{X}$, depending on the value $\boldsymbol{\theta}^\top\mathbf{x}$. If $\boldsymbol{\theta}^\top\mathbf{x} \le 0$, the trivial choice $\Delta_{\mathbf{x}} = \mathbf{0}$ satisfies $\boldsymbol{\theta}^\top(\mathbf{x} + \Delta_{\mathbf{x}}) \le 0$.

If $0 < \boldsymbol{\theta}^\top\mathbf{x} \le \gamma$, the choice $\Delta_{\mathbf{x}} := -(\boldsymbol{\theta}^\top\mathbf{x})\boldsymbol{\theta}$ satisfies them, i.e., $\|\Delta_{\mathbf{x}}\|_2 = \boldsymbol{\theta}^\top\mathbf{x} \le \gamma$ implies $\Delta_{\mathbf{x}} \in \mathcal{B}_2(\gamma)$, $\boldsymbol{\theta}^\top(\mathbf{x} + \Delta_{\mathbf{x}}) = \boldsymbol{\theta}^\top\mathbf{x} - \boldsymbol{\theta}^\top\mathbf{x} = 0$, and $\|\mathbf{x} + \Delta_{\mathbf{x}}\|_2^2 = \|\mathbf{x}\|^2 - \left(\boldsymbol{\theta}^\top\mathbf{x}\right)^2 \le \|\mathbf{x}\|^2 \le 1$ implies $\mathbf{x} + \Delta_{\mathbf{x}} \in \mathcal{X}$.

If $\boldsymbol{\theta}^\top\mathbf{x} > \gamma$, we can check $\boldsymbol{\theta}^\top(\mathbf{x}+\Delta_{\mathbf{x}}) > 0$ for any $\Delta_{\mathbf{x}} \in \mathcal{B}_2(\gamma)$. We consider the convex optimization problem $\min_{\Delta_{\mathbf{x}} \in \mathcal{B}_2(\gamma)} \boldsymbol{\theta}^\top(x + \Delta_{\mathbf{x}})$. Consider the Lagrangian

$$\mathcal{L}(\Delta_{\mathbf{x}}, \mu) := \boldsymbol{\theta}^\top(\mathbf{x} + \Delta_{\mathbf{x}}) + \mu(\|\Delta_{\mathbf{x}}\|_2 - \gamma),$$

where $\mu \in \mathbb{R}$ is a KKT multiplier. Its KKT conditions are

$$\begin{cases} -\boldsymbol{\theta} = \mu \frac{\Delta_{\mathbf{x}}}{\|\Delta_{\mathbf{x}}\|_2}, \\ \|\Delta_{\mathbf{x}}\|_2 \le \gamma, \\ \mu \ge 0, \\ \mu(\|\Delta_{\mathbf{x}}\|_2 - \gamma) = 0. \end{cases}$$

The objective is minimized when the constraint $\|\Delta_{\mathbf{x}}\|_2 \le \gamma$ is activated, where the multiplier $\mu > 0$ and $\Delta_{\mathbf{x}} = -\frac{\gamma}{\mu}\boldsymbol{\theta}$, meaning that $\Delta_{\mathbf{x}}$ is parallel to $\boldsymbol{\theta}$ in the opposite direction. Hence, $\Delta_{\mathbf{x}} = -\gamma\boldsymbol{\theta}$ is the minimizer. We have $\boldsymbol{\theta}^\top(\mathbf{x}+\Delta_{\mathbf{x}}) = \boldsymbol{\theta}^\top\mathbf{x} - \gamma > 0$ with this minimizer $\Delta_{\mathbf{x}}$.

By combining the three cases, we have $\ell_\gamma(+1, \mathbf{x}, f) = \mathbb{1}_{\{\boldsymbol{\theta}^\top\mathbf{x} \le \gamma\}}$. □

Subsequently, when considering $\mathcal{F}_{\text{lin}}$, we work with the loss function

$$\phi_\gamma(\alpha) := \mathbb{1}_{\{\alpha \le \gamma\}}$$

and call ϕ_γ the *γ-robust 0-1 loss*. We will study calibrated surrogates w.r.t. ϕ_γ instead of ℓ_γ, and both are equivalent under the restricted function class $\mathcal{F}_{\text{lin}}$.

In many machine learning problems, there are often dichotomies between optimization (learning) and evaluation. For instance, binary classification is evaluated by the 0-1 loss, whereas common learning methods such as the support vector machine and logistic regression minimize surrogates to the 0-1 loss. This dichotomy arises because minimizing the 0-1 loss directly is known to be NP-hard [Feldman et al., 2012]. Many studies has investigated surrogates ϕ satisfying

$$R_\phi(f_i) - R^*_{\phi,\mathcal{F}} \to 0 \Longrightarrow R_\ell(f_i) - R^*_{\ell,\mathcal{F}} \to 0, \tag{4.2}$$

for all probability distributions and the sequence of $\{f_i\}_{i\in\mathbb{N}} \subseteq \mathcal{F}$. When Equation (4.2) is satisfied, the surrogate ϕ is said to be $(\ell, \mathcal{F})$-*consistent.*

In this chapter, we study a pointwise form of consistency, known as *calibration*, which can be viewed as consistency of the excess $(\phi, \mathcal{F})$-CCR $C_{\phi,\mathcal{F}}(f, \eta, \mathbf{x}) - C^*_{\phi,\mathcal{F}}(\eta, \mathbf{x})$ at each $\mathbf{x} \in \mathcal{X}$ (formally defined in Section 4.5). Because CCRs are defined in a pointwise manner, calibration analysis is easier than analyzing the consistency directly, and has been used to prove the consistency in a number of learning settings, as we will see in Section 4.4. For example, calibration analysis has been conducted in standard binary classification [Bartlett et al., 2006], where calibration is necessary [Steinwart, 2007, Theorem 3.3] and sufficient [Steinwart, 2007, Theorem 2.8] for consistency when $\mathcal{F} = \mathcal{F}_{\text{all}}$. When $\mathcal{F} \neq \mathcal{F}_{\text{all}}$, calibration may not be sufficient for consistency, although it remains an important first step to analyze and understand the consistency in standard classification [Long and Servedio, 2013, Zhang and Agarwal, 2020]. This motivated the study of calibration in the context of adversarially robust classification.

4.4 Related Work

From the viewpoint of robust optimization [Ben-Tal et al., 2009, Bertsimas et al., 2011], adversarially robust binary classification can be formulated as[3]

$$\min_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})} \left[\max_{\widetilde{\mathsf{X}} \in \mathcal{U}(\mathsf{X})} \ell(\mathsf{Y}, \widetilde{\mathsf{X}}, f) \right], \tag{4.3}$$

where ℓ is a loss function and $\mathcal{U}(\mathbf{x})$ is a user-specified uncertainty set. The optimization problem of adversarially robust classification $\min_{f \in \mathcal{F}} R_{\ell_\gamma}(f)$ can be regarded as a special case, $\ell = \ell_{01}$ and $\mathcal{U}(\mathbf{x}) = \mathbf{x} + \mathcal{B}_2(\gamma)$.

Because the minimax problem (4.3) is generally nonconvex, it is traditionally tackled by minimizing a convex upper bound. Lanckriet et al. [2002] and Shivaswamy et al. [2006] chose $\mathcal{U}(\mathbf{x}) = \{ \mathbf{x} \sim (\bar{\mathbf{x}}, \Sigma_{\mathbf{x}}) \}$ as an uncertainty set, where $\mathbf{x} \sim (\bar{\mathbf{x}}, \Sigma_{\mathbf{x}})$ means that $\mathbf{x}$ is drawn from a distribution that has a prespecified mean $\bar{\mathbf{x}}$, covariance $\Sigma_{\mathbf{x}}$, and arbitrarily higher moments. Lanckriet et al. [2002] and Shivaswamy et al. [2006] convexified Equation (4.3) and obtained a second-order cone program. In addition, Xu et al. [2009] studied the relationship between robustness and regularization, and showed that Equation (4.3) with the hinge loss and $\mathcal{U}(\mathbf{x}) = \mathbf{x} + \mathcal{B}_2(\gamma)$ is equivalent to ℓ_2-regularized SVM. Recently, Wong and Kolter [2018], Madry et al. [2018], Raghunathan et al. [2018a], Raghunathan et al. [2018b], and Khim and Loh [2019] examined Equation (4.3) using the softmax cross entropy loss and $\mathcal{U}(\mathbf{x}) = \mathbf{x} + \mathcal{B}_\infty^d(\gamma)$ when $\mathcal{F}$ is a set of deep nets, and provided convex upper bounds of the worst-case loss in Equation (4.3). However, no approach other than Cranko et al. [2019] considered whether the surrogate objectives minimize the robust 0-1 excess risk. In addition, Cranko et al. [2019] showed that no canonical proper loss [Reid and Williamson, 2010] can minimize the robust 0-1 loss. Because canonical proper losses are convex, this result aligns with our results. Dan et al. [2020] proposed the plug-in classifier based on the linear discriminant analysis and showed that this procedure is consistent to the robust 0-1 loss but under the assumption that data are normally distributed. We show more general results through calibration analysis for $\mathcal{U}(\mathbf{x}) = \mathbf{x} + \mathcal{B}_2(\gamma)$.

There are several other approaches to the robust classification such as minimizing the Taylor approximation of the worst-case loss in Equation (4.3) [Goodfellow et al., 2015, Gu and Rigazio, 2015, Shaham et al., 2018], regularization on the Lipschitz norm of models [Cisse et al., 2017, Hein and Andriushchenko, 2017, Tsuzuku et al., 2018], and an injection of random noises to model parameters called randomized smoothing [Lecuyer et al., 2019, Cohen et al., 2019, Pinot et al., 2019, Salman et al., 2019]. It is not known whether these methods imply a minimization of the robust 0-1 excess risk.

Other forms of robustness have also been considered in the literature. A number of existing studied have considered the worst-case test distribution. This

[3] This formulation of adversarial robustness is often called *loss-based robustness* [Seshia et al., 2018]. For other types of adversarial robustness, one may consider local robustness and global robustness (also known as certification [Cohen et al., 2019]), which are interested in how a classifier behaves around a fixed or each data point. Loss-based robustness is a stricter notion because it involves the distributional information as well. Please refer to Seshia et al. [2018], Dreossi et al. [2019] for more details.

line includes divergence-based methods [Namkoong and Duchi, 2016, 2017, Hu et al., 2018, Sinha et al., 2018], domain adaptation [Mansour et al., 2009, Ben-David et al., 2010, Germain et al., 2013, Kuroki et al., 2019, Zhang et al., 2019b], and methods based on constraints on the feature moments [Farnia and Tse, 2016, Fathony et al., 2016].

In addition to adversarial robustness, it is worth mentioning outlier and label-noise robustness. It is known that convex losses are vulnerable to outliers, thus a truncation making the losses nonconvex is useful [Huber, 2011]. In a machine learning context, Masnadi-Shirazi and Vasconcelos [2009] and Holland [2019] designed nonconvex losses robust to outliers. By contrast, label-noise robustness, particularly a random classification noise model, has been extensively studied [Angluin and Laird, 1988], where training labels are flipped with a fixed probability. Long and Servedio [2010] showed that there is no convex loss that is robust to label noises. Later, Ghosh et al. [2015], van Rooyen et al. [2015], and Charoenphakdee et al. [2019] discovered that a certain class of nonconvex losses is a good alternative for label-noise robustness. In both outlier and label-noise robustness, nonconvex loss functions play an important role, as we can see in adversarial robustness.

Calibration analysis has been formalized in Lin [2004], Zhang [2004a], Bartlett et al. [2006], and Steinwart [2007], and employed to analyze not only binary classification, but also complicated problems such as multi-class classification [Zhang, 2004b, Tewari and Bartlett, 2007, Long and Servedio, 2013, Ávila Pires and Szepesvári, 2016, Ramaswamy and Agarwal, 2016], multi-label classification [Gao and Zhou, 2011, Dembczynski et al., 2012], cost-sensitive learning [Scott, 2011, 2012, Ávila Pires et al., 2013], ranking [Duchi et al., 2010, Ravikumar et al., 2011, Ramaswamy et al., 2013], structured prediction [Hazan et al., 2010, Ramaswamy and Agarwal, 2012, Osokin et al., 2017, Blondel, 2019], AUC optimization [Gao and Zhou, 2015], and optimization of non-decomposable metrics [Bao and Sugiyama, 2020]. In addition, Zhang [2004a], Ravikumar et al. [2011], and Gao and Zhou [2015] determined *ad hoc* derivations of excess risk bounds, whereas Bartlett et al. [2006], Steinwart [2007], Scott [2012], Ávila Pires et al. [2013], Ávila Pires and Szepesvári [2016], Osokin et al. [2017], and Blondel [2019] used more systematic approaches. For adversarially robust classification, Zhang et al. [2019a, Theorem 3.1] applied the classical result of calibration analysis on convex losses to upper bound the robust classification risk, resulting in a term requiring numerical approximation in practice.

Finally, Awasthi et al. [2021a] contributed calibration analysis of adversarially robust classification by showing that realizability assumptions are sufficient for calibrated losses to imply consistency. They showed that no *continuous* margin-based losses are calibrated and that some nonconvex and minimax-type losses are consistent w.r.t. the robust 0-1 loss. Awasthi et al. [2021b] independently corrected our main results and extended them to more general function classes beyond $\mathcal{F}_{\text{lin}}$.

4.5 Calibration Analysis

Calibration analysis is a tool used to study the relationship between surrogate losses and target losses. This section is devoted to describing the calibration function introduced in Steinwart [2007] and specializing it to the current study.[4]

[4] We import toolsets from Steinwart [2007] for two reasons: (i) Steinwart [2007] formalized calibration analysis that is *dependent* on the user-specified function classes, which is useful for

Note that the subsequent definitions are more general than those introduced in Section 2.3 in that a function class $\mathcal{F}$ is fixed. This is useful for our subsequent analysis restricted to a linear model.

Definition 4.2. *For a loss $\psi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ and a function class $\mathcal{F}$, we say a loss $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is* calibrated w.r.t. $(\psi, \mathcal{F})$*, or* $(\psi, \mathcal{F})$-calibrated*, if for any $\varepsilon > 0$, $\eta \in [0, 1]$, and $\mathbf{x} \in \mathcal{X}$, there exists $\delta > 0$ such that for all $f \in \mathcal{F}$, we have*

$$C_\phi(f, \eta, \mathbf{x}) < C^*_{\phi,\mathcal{F}}(\eta, \mathbf{x}) + \delta \implies C_\psi(f, \eta, \mathbf{x}) < C^*_{\psi,\mathcal{F}}(\eta, \mathbf{x}) + \varepsilon. \tag{4.4}$$

If ϕ is $(\psi, \mathcal{F})$-calibrated, the condition (4.2) holds for any probability distribution on $\mathcal{X} \times \mathcal{Y}$ satisfying the regularity conditions [Steinwart, 2007, Theorem 2.8].[5]

Next, we introduce the *calibration function* [Steinwart, 2007, Lemma 2.9].

Definition 4.3. *For a margin-based loss ψ and ϕ, and a function class $\mathcal{F}$, the* calibration function of ϕ w.r.t. $(\psi, \mathcal{F})$*, or simply* calibration function *if the context is clear, is defined as*

$$\bar{\delta}(\varepsilon, \eta, \mathbf{x}) = \inf_{f \in \mathcal{F}} \left\{ C_\phi(f, \eta, \mathbf{x}) - C^*_{\phi,\mathcal{F}}(\eta, \mathbf{x}) \;\middle|\; C_\psi(f, \eta, \mathbf{x}) - C^*_{\psi,\mathcal{F}}(\eta, \mathbf{x}) \geq \varepsilon \right\}. \tag{4.5}$$

Note that $\bar{\delta}(\varepsilon, \eta, \mathbf{x})$ is nondecreasing for $\varepsilon > 0$. The calibration function $\bar{\delta}(\varepsilon, \eta, \mathbf{x})$ is the maximal δ satisfying the CCR condition (4.4). Steinwart [2007] established the following important result confirming whether a surrogate is $(\psi, \mathcal{F})$-calibrated.

Proposition 4.4 (Steinwart [2007])**.** *A surrogate loss ϕ is $(\psi, \mathcal{F})$-calibrated if and only if its calibration function $\bar{\delta}$ satisfies $\bar{\delta}(\varepsilon, \eta, \mathbf{x}) > 0$ for all $\varepsilon > 0$, $\eta \in [0, 1]$, and $\mathbf{x} \in \mathcal{X}$.*

To see the relationship between $(\psi, \mathcal{F})$-excess risk and $(\phi, \mathcal{F})$-excess risk, a stronger notion of calibrated losses than Definition 4.2 is necessary.

Definition 4.5. *For a loss $\psi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ and a function class $\mathcal{F}$, we state that a loss $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is* uniformly $(\psi, \mathcal{F})$-calibrated*, if for any $\varepsilon > 0$, there exists $\delta > 0$ such that for all $\eta \in [0, 1]$, $f \in \mathcal{F}$, and $\mathbf{x} \in \mathcal{X}$, we have*

$$C_\phi(f, \eta, \mathbf{x}) < C^*_{\phi,\mathcal{F}}(\eta, \mathbf{x}) + \delta \implies C_\psi(f, \eta, \mathbf{x}) < C^*_{\psi,\mathcal{F}}(\eta, \mathbf{x}) + \varepsilon.$$

The corresponding uniform calibration function *is defined as*

$$\delta(\varepsilon) = \inf_{\eta \in [0,1]} \inf_{\mathbf{x} \in \mathcal{X}} \bar{\delta}(\varepsilon, \eta, \mathbf{x}).$$

Note that Definition 4.5 is slightly but substantially different from Definition 4.2 in that the order of quantifiers of δ and $(\eta, \mathbf{x})$ is reversed. With this notion, we can connect the surrogate excess risk to the target excess risk, as shown in the following statement.

our analysis on $\mathcal{F}_{\text{lin}}$. (ii) Steinwart [2007] gave a general form of the calibration function (4.5), whereas most of literature has focused on specific target losses.

[5] To imply $(\psi, \mathcal{F})$-consistency (4.2), the two loss functions ϕ and ψ are required to be $\mathbb{P}$-*minimizable* for the underlying distribution $\mathbb{P}$, roughly indicating that their CCRs can be made arbitrarily small by a function in $\mathcal{F}$. This ensures that $R^*_{\phi,\mathcal{F}} = \mathbb{E}_{\mathsf{X}}[C^*_{\phi,\mathcal{F}}(\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X}), \mathsf{X})]$. The precise statements and more details regarding $\mathbb{P}$-minimizability can be found in Section 2.3.

Proposition 4.6 (Theorem 2.13 in Steinwart [2007])**.** *Let $\delta : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be the uniform calibration function of ϕ w.r.t. $(\psi, \mathcal{F})$. Define $\check{\delta} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ as $\check{\delta}(\varepsilon) = \delta(\varepsilon)$ if $\varepsilon > 0$ and $\check{\delta}(0) = 0$. Suppose that ϕ and ψ are $\mathbb{P}$-minimizable and $R^*_{\phi,\mathcal{F}}, R^*_{\psi,\mathcal{F}} < \infty$. Then, for all $f \in \mathcal{F}$, we have*

$$\check{\delta}^{\star\star}\left(R_\psi(f) - R^*_{\psi,\mathcal{F}}\right) \leq R_\phi(f) - R^*_{\phi,\mathcal{F}}, \tag{4.6}$$

where $\check{\delta}^{\star\star}$ denotes the Fenchel-Legendre biconjugate of $\check{\delta}$.

The relationship in Equation (4.6) is called an excess risk transform. The excess risk transform is invertible iff ϕ is uniformly $(\psi, \mathcal{F})$-calibrated [Steinwart, 2007, Remark 2.14]. In this case, we obtain the excess risk bound $R_\psi(f) - R^*_{\psi,\mathcal{F}} \leq (\check{\delta}^{\star\star})^{-1}(R_\phi(f) - R^*_{\phi,\mathcal{F}})$.[6] In the end, the calibration function can be used in two ways: Proposition 4.4 enables us to check if a surrogate loss is calibrated, and Proposition 4.6 gives us a quantitative relationship between the surrogate excess risk and the target excess risk. Such an analysis has been carried out in a number of learning problems, as we mentioned in Section 4.4.

We review an important result regarding convex surrogates for the non-robust 0-1 loss ℓ_{01}.

Proposition 4.7 (Theorem 6 in Bartlett et al. [2006])**.** *Let ϕ be a convex surrogate loss. Then, ϕ is uniformly calibrated w.r.t. $(\ell_{01}, \mathcal{F}_{\mathrm{all}})$ if and only if it is differentiable at 0 and $\phi'(0) < 0$.*

As a result of Proposition 4.7, we know that many surrogate losses used in practice such as the hinge loss, logistic loss, and squared loss are uniformly calibrated w.r.t. $(\ell_{01}, \mathcal{F}_{\mathrm{all}})$. One of our objectives in this paper is to establish a general class of loss functions that are calibrated w.r.t. the adversarial 0-1 loss, in analogy to Proposition 4.7.

Before proceeding to our main results, we present two lemmas that facilitate our analysis. All proofs are deferred to Section 4.11.

Lemma 4.8. *Let $\widetilde{\mathcal{X}}_\rho \coloneqq \mathcal{X} \setminus \mathcal{B}_2^\circ(\gamma + \rho)$ and ϕ be a continuous surrogate loss. Denote*

$$\delta_\rho(\varepsilon) = \inf_{\substack{\eta \in [0,1], \\ \mathbf{x} \in \widetilde{\mathcal{X}}_\rho, \\ f \in \mathcal{F}_{\mathrm{lin}}}} \left\{ C_\phi(f, \eta, \mathbf{x}) - C^*_{\phi,\mathcal{F}_{\mathrm{lin}}}(\eta, \mathbf{x}) \;\middle|\; C_{\phi_\gamma}(f, \eta, \mathbf{x}) - C^*_{\phi_\gamma,\mathcal{F}_{\mathrm{lin}}}(\eta, \mathbf{x}) \geq \varepsilon \right\}.$$

Then, ϕ is $(\phi_\gamma, \mathcal{F}_{\mathrm{lin}})$-calibrated if and only if $\delta_\rho(\varepsilon) > 0$ for all $\varepsilon > 0$ and $\rho \in (0, 1 - \gamma)$.

Proof. By Proposition 4.4, we need to show the following conditions are equivalent.

(i) For all $\varepsilon > 0$, $\eta \in [0, 1]$, and $\mathbf{x} \in \mathcal{X}$, $\bar{\delta}(\varepsilon, \eta, \mathbf{x}) > 0$.

(ii) For all $\varepsilon > 0$ and $\rho \in (0, 1 - \gamma)$, $\delta_\rho(\varepsilon) > 0$.

We have $\Delta C_{\phi_\gamma,\mathcal{F}_{\mathrm{lin}}}(f, \eta, \mathbf{x}) = 0$ for x with $\|\mathbf{x}\|_2 \leq \gamma$, from Equation (4.9) in the proof of Lemma 4.9 (below). This means that the constraint $\Delta C_{\phi_\gamma,\mathcal{F}_{\mathrm{lin}}}(f, \eta, \mathbf{x}) \geq \varepsilon$

[6] In addition, it is known that a non-vacuous and distribution-independent excess risk transform is available only if a surrogate is uniformly calibrated provided that the biconjugate of the calibration function is invertible [Steinwart, 2007, Theorem 2.17]. Hence, a uniform calibration is necessary to obtain an excess risk bound.

in $\bar{\delta}$ would never be satisfied for $\varepsilon > 0$, where the infimum value of $\bar{\delta}(\varepsilon, \eta, \mathbf{x}) = \infty$ for all $\varepsilon > 0$, $\eta \in [0, 1]$. Note that

$$\begin{aligned}\delta_\rho(\varepsilon) &= \inf_{\eta\in[0,1]} \inf_{\|\mathbf{x}\|_2\geq\gamma+\rho} \inf_{f\in\mathcal{F}_{\mathrm{lin}}} \left\{ \Delta C_{\phi,\mathcal{F}_{\mathrm{lin}}}(f, \eta, \mathbf{x}) \;\middle|\; \Delta C_{\phi_\gamma,\mathcal{F}_{\mathrm{lin}}}(f, \eta, \mathbf{x}) \geq \varepsilon \right\} \\ &= \inf_{\eta\in[0,1]} \inf_{\|\mathbf{x}\|_2\geq\gamma+\rho} \bar{\delta}(\varepsilon, \eta, \mathbf{x}).\end{aligned}$$

For (i) $\Rightarrow$ (ii), let $\widetilde{\mathcal{X}}_\rho := \mathcal{X} \setminus \mathcal{B}_2^\circ(\gamma + \rho) = \{ x \in \mathcal{X} \mid \|\mathbf{x}\|_2 \geq \gamma + \rho \}$. For a fixed $\varepsilon > 0$, the extreme value theorem states that $\delta_\rho(\varepsilon) = \bar{\delta}(\varepsilon, \eta_\varepsilon, \mathbf{x}_\varepsilon)$ for some $(\eta_\varepsilon, \mathbf{x}_\varepsilon) \in [0, 1] \times \widetilde{\mathcal{X}}_\rho$, by noting that $\bar{\delta}(\varepsilon, \cdot, \cdot) : [0, 1] \times \widetilde{\mathcal{X}}_\rho \to \mathbb{R}_{\geq 0}$ is continuous and its domain $[0, 1] \times \widetilde{\mathcal{X}}_\rho$ is compact. Indeed, $\bar{\delta}(\varepsilon, \cdot, \cdot)$ is continuous because it is the infimum function of a continuous function over a compact set (see Equation (4.7) in Lemma 4.9). Eventually, we have $\delta_\rho(\varepsilon) \geq \bar{\delta}(\varepsilon, \eta_\varepsilon, \mathbf{x}_\varepsilon) > 0$ by using (i).

Subsequently, we check (ii) $\Rightarrow$ (i). The condition (ii) implies that $\bar{\delta}(\varepsilon, \eta, \mathbf{x}) \geq \delta_\rho(\varepsilon) > 0$ for all $\varepsilon > 0$, $\eta \in [0, 1]$, and $\mathbf{x} \in \mathcal{X}$ with $\|\mathbf{x}\|_2 > \gamma$. Together with $\bar{\delta}(\varepsilon, \eta, \mathbf{x}) = \infty$ for all $\varepsilon > 0$, $\eta \in [0, 1]$, and $\mathbf{x} \in \mathcal{X}$ with $\|\mathbf{x}\|_2 \leq \gamma$, (i) is assured. □

The calibration function with the restricted domain δ_ρ is easier to work with in the subsequent analyses.

Finally, we characterize the calibration function of an arbitrary surrogate loss ϕ w.r.t. $(\phi_\gamma, \mathcal{F}_{\mathrm{lin}})$.

Lemma 4.9. *Let ϕ be a surrogate loss. Then, the $(\phi_\gamma, \mathcal{F}_{\mathrm{lin}})$-calibration function is*

$$\begin{aligned}&\bar{\delta}(\varepsilon, \eta, \mathbf{x}) \\ &= \begin{cases} \infty & \text{if } \varepsilon > \max\{ \eta, 1 - \eta \}, \\ \inf\limits_{f\in\mathcal{F}_{\mathrm{lin}}:|f(\mathbf{x})|\leq\gamma} \Delta C_{\phi,\mathcal{F}_{\mathrm{lin}}}(f, \eta, \mathbf{x}) & \text{if } |2\eta - 1| < \varepsilon \leq \max\{ \eta, 1 - \eta \}, \\ \inf\limits_{\substack{f\in\mathcal{F}_{\mathrm{lin}}: \\ (2\eta-1)f(\mathbf{x})\leq 0 \text{ or } |f(\mathbf{x})|\leq\gamma}} \Delta C_{\phi,\mathcal{F}_{\mathrm{lin}}}(f, \eta, \mathbf{x}) & \text{if } \varepsilon \leq |2\eta - 1|, \end{cases}\end{aligned} \tag{4.7}$$

when $\|\mathbf{x}\| > \gamma$, and $\bar{\delta}(\varepsilon, \eta, \mathbf{x}) = \infty$ when $\|\mathbf{x}\| \leq \gamma$.

Proof. We first simplify the constraint in the calibration function (4.5). The ϕ_γ-CCR for $f \in \mathcal{F}_{\mathrm{lin}}$ at $\mathbf{x}$ is

$$\begin{aligned}C_{\phi_\gamma}(f, \eta, \mathbf{x}) &= \eta \mathbb{1}_{\{f(\mathbf{x})\leq\gamma\}} + (1 - \eta)\mathbb{1}_{\{f(\mathbf{x})\geq-\gamma\}} \\ &= \begin{cases} 1 & \text{if } |f(\mathbf{x})| \leq \gamma, \\ 1 - \eta & \text{if } \gamma < f(\mathbf{x}), \\ \eta & \text{if } f(\mathbf{x}) < -\gamma. \end{cases}\end{aligned} \tag{4.8}$$

To compute the minimal $(\phi_\gamma,\mathcal{F}_{\mathrm{lin}})$-CCR, we divide it into two cases.

- If $\underline{\|\mathbf{x}\|_2 \leq \gamma}$, $C_{\phi_\gamma}(f, \eta, \mathbf{x}) = 1$ for any $f \in \mathcal{F}_{\mathrm{lin}}$ because $|f(\mathbf{x})| \leq \gamma$. Thus, we have $C^*_{\phi_\gamma,\mathcal{F}_{\mathrm{lin}}}(\eta, \mathbf{x}) = 1$ and $\Delta C_{\phi_\gamma,\mathcal{F}_{\mathrm{lin}}}(f, \eta, \mathbf{x}) = 0$.
- If $\underline{\|\mathbf{x}\|_2 > \gamma}$, there exists $f \in \mathcal{F}_{\mathrm{lin}}$ such that $C_{\phi_\gamma}(f, \eta, \mathbf{x}) = \min\{ \eta, 1 - \eta \}$. Thus, we have $C^*_{\phi_\gamma,\mathcal{F}_{\mathrm{lin}}}(\eta, \mathbf{x}) = \min\{ \eta, 1 - \eta \}$.

This implies that

$$\Delta C_{\phi_\gamma,\mathcal{F}_{\mathrm{lin}}}(f,\eta,\mathbf{x}) = \begin{cases} \max\{\,\eta, 1-\eta\,\} & \text{if } |f(\mathbf{x})| \le \gamma, \\ |2\eta-1| \cdot \mathbb{1}_{\{(2\eta-1)f(\mathbf{x})\le 0\}} & \text{if } \gamma < |f(\mathbf{x})|. \end{cases}$$

Note that the latter case is obtained in the same manner as Bartlett et al. [2006, Proof of Theorem 3]. To summarize, we obtain the expression of $\Delta C_{\phi_\gamma,\mathcal{F}_{\mathrm{lin}}}$ as

$$\Delta C_{\phi_\gamma,\mathcal{F}_{\mathrm{lin}}}(f,\eta,\mathbf{x}) = \begin{cases} 0 & \text{if } \|\mathbf{x}\|_2 \le \gamma, \\ \max\{\eta, 1-\eta\} & \text{if } \|\mathbf{x}\|_2 > \gamma \text{ and } |f(\mathbf{x})| \le \gamma, \\ |2\eta-1| \cdot \mathbb{1}_{\{(2\eta-1)f(\mathbf{x})\le 0\}} & \text{if } \|\mathbf{x}\|_2 > \gamma \text{ and } \gamma < |f(\mathbf{x})|. \end{cases} \tag{4.9}$$

Next, we simplify the infimum on f

$$\inf_{f\in\mathcal{F}_{\mathrm{lin}}} \left\{ \Delta C_{\phi,\mathcal{F}_{\mathrm{lin}}}(f,\eta,\mathbf{x}) \;\middle|\; \Delta C_{\phi_\gamma,\mathcal{F}_{\mathrm{lin}}}(f,\eta,\mathbf{x}) \ge \varepsilon \right\} = \bar{\delta}(\varepsilon,\eta,\mathbf{x})$$

in Equation (4.5), for a fixed $\eta \in [0,1]$ and $\mathbf{x} \in \mathcal{X}$.

- If $\|\mathbf{x}\|_2 \le \gamma$ or $\varepsilon > \max\{\eta, 1-\eta\}$, no $f \in \mathcal{F}_{\mathrm{lin}}$ achieves $\Delta C_{\phi_\gamma,\mathcal{F}_{\mathrm{lin}}}(f,\eta,\mathbf{x}) \ge \varepsilon$, meaning that $\bar{\delta}(\varepsilon,\eta,\mathbf{x}) = \infty$.

- If $\|\mathbf{x}\|_2 > \gamma$ and $|2\eta-1| < \varepsilon \le \max\{\,\eta, 1-\eta\,\}$, $\Delta C_{\phi_\gamma,\mathcal{F}_{\mathrm{lin}}}(f,\eta,\mathbf{x}) \ge \varepsilon$ is achieved when $|f(\mathbf{x})| \le \gamma$. Hence,

  $$\bar{\delta}(\varepsilon,\eta,\mathbf{x}) = \inf_{f\in\mathcal{F}_{\mathrm{lin}}} \{\, \Delta C_{\phi,\mathcal{F}_{\mathrm{lin}}}(f,\eta,\mathbf{x}) \mid |f(\mathbf{x})| \le \gamma \,\}.$$

  Note that $|2\eta-1| \le \max\{\,\eta, 1-\eta\,\} = \frac{1+|2\eta-1|}{2}$ for all $\eta \in [0,1]$.

- If $\|\mathbf{x}\|_2 > \gamma$ and $\varepsilon \le |2\eta-1|$, $\Delta C_{\phi,\mathcal{F}_{\mathrm{lin}}}(f,\eta,\mathbf{x}) \ge \varepsilon$ is achieved if either $|f(\mathbf{x})| \le \gamma$ or $(2\eta-1)f(\mathbf{x}) \le 0$ holds. Hence,

  $$\bar{\delta}(\varepsilon,\eta,\mathbf{x}) = \inf_{f\in\mathcal{F}_{\mathrm{lin}}} \{\, \Delta C_{\phi,\mathcal{F}_{\mathrm{lin}}}(f,\eta,\mathbf{x}) \mid |f(\mathbf{x})| \le \gamma \text{ or } (2\eta-1)f(\mathbf{x}) \le 0 \,\}.$$

These verify the statement of this lemma. □

Lemmas 4.8 and 4.9 are used in the proofs and examples below.

4.6 Convex Surrogates are Not (ϕ_γ, $\mathcal{F}_{\mathrm{lin}}$)-calibrated

Our first result concerns calibration of convex surrogate losses w.r.t. the γ-robust 0-1 loss.

Theorem 4.10. *For any margin-based surrogate loss $\phi : \mathbb{R} \to \mathbb{R}_{\ge 0}$, if ϕ is convex, then ϕ is not calibrated w.r.t. (ϕ_γ,$\mathcal{F}_{\mathrm{lin}}$).*

Proof. (Sketch) In a non-robust setup, Bartlett et al. [2006] showed that a surrogate loss is calibrated w.r.t. (ℓ_{01},$\mathcal{F}_{\mathrm{all}}$) iff $\inf_{(2\eta-1)f(\mathbf{x})\le 0} C_\phi(f,\eta,\mathbf{x})$ (the minimum ϕ-risk over the "wrong" predictions) is larger than $\inf_{f(\mathbf{x})\in\mathbb{R}} C_\phi(f,\eta,\mathbf{x})$ (the minimum ϕ-risk over all predictions) for $\eta \ne \frac{1}{2}$. This means that the wrong predictions

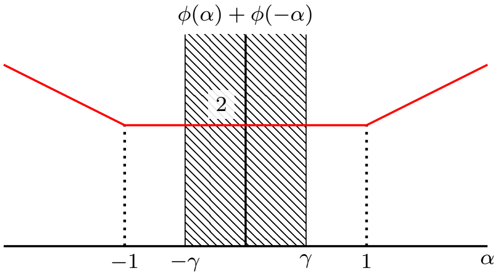


Figure 4.2: $\phi(\alpha)+\phi(-\alpha) = 2C_\phi\left(f, \frac{1}{2}, \mathbf{x}\right)$ is illustrated with $\alpha = f(\mathbf{x})$, where ϕ is the hinge loss and $\gamma = 0.5$. $\phi(\alpha)+\phi(-\alpha)$ has the same minimizers in both $|\alpha| \leq \gamma$ and $|\alpha| \leq 1$.

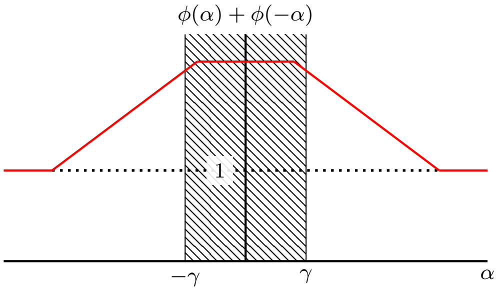


Figure 4.3: $\phi(\alpha)+\phi(-\alpha) = 2C_\phi\left(f, \frac{1}{2}, \mathbf{x}\right)$ is illustrated with $\alpha = f(\mathbf{x})$, where ϕ is the ramp loss with $\beta = 0.6$ (defined in Section 4.8) and $\gamma = 0.5$. The condition $\phi(\gamma)+\phi(-\gamma) > \phi(\alpha)+\phi(-\alpha)$ for $\alpha \in (\gamma, 1]$ reflects the idea that predictions falling into the shaded area ($|\alpha| \leq \gamma$) must be penalized more than the others.

must be penalized more. In our robust setup, we must penalize not only the wrong predictions but also predictions that fall within the γ-margin, i.e.,

$$\inf_{f \in \mathcal{F}_{\text{lin}} : |f(\mathbf{x})| \leq \gamma} C_\phi(f, \eta, \mathbf{x}) > \inf_{f \in \mathcal{F}_{\text{lin}}} C_\phi(f, \eta, \mathbf{x}), \tag{4.10}$$

which is an immediate corollary of Proposition 4.4 and Lemma 4.9 and stated in part 3 of Lemma 4.16 in Section 4.11. Equation (4.10) becomes harder to satisfy as a data point becomes more uncertain ($\eta \to \frac{1}{2}$). In the limit, we have $\inf_{|\alpha| \leq \gamma} \phi(\alpha) + \phi(-\alpha) > \inf_{\alpha \in \mathbb{R}} \phi(\alpha) + \phi(-\alpha)$, meaning that the even part of ϕ should take larger values in $|\alpha| \leq \gamma$ than in the rest of α. However, $\phi(\alpha)+\phi(-\alpha)$ attains the infimum at $\alpha = 0$ because $\phi(\alpha)+\phi(-\alpha)$ is convex and even as long as ϕ is convex. Therefore, Equation (4.10) would never be satisfied through a convex surrogate ϕ. This idea is illustrated in Figure 4.2. □

Hence, many popular surrogate losses such as the hinge, logistic, and squared error losses are not calibrated w.r.t. (ϕ_γ,$\mathcal{F}_{\text{lin}}$). We defer all proofs to Section 4.11.

Note that the definition of calibration makes no assumptions regarding the conditional distribution $\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = \mathbf{x})$. If we additionally adopt the low noise assumption [Massart and Nédélec, 2006], then it is possible for a convex loss to be calibrated w.r.t. (ϕ_γ,$\mathcal{F}_{\text{lin}}$). We will discuss the details of this later in Section 4.9.

4.7 Calibration Conditions for Nonconvex Surrogates

As described in Section 4.6, convex surrogate losses that are calibrated w.r.t. (ϕ_γ,$\mathcal{F}_{\text{lin}}$) do not exist. This has motivated a search for nonconvex surrogate losses. Nonconvex surrogates are used for outlier robustness [Collobert et al., 2006, Masnadi-Shirazi and Vasconcelos, 2009, Holland, 2019] or label-noise robustness [Ghosh et al., 2015, van Rooyen et al., 2015, Charoenphakdee et al., 2019]. Bounded monotone surrogates such as the ramp loss and the sigmoid loss are simple and common choices for such purposes. In this section, we also look for good surrogates from bounded monotone losses.

The following assumption will be adopted.

Assumption 4.11. *For a margin-based loss function $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$, $\phi(-\alpha) > \phi(\alpha)$ for $\alpha \in (\gamma, 1]$, and its ϕ-CCR $C_\phi(\cdot, \eta)$ is quasiconcave for all $\eta \in [0, 1]$.*

The assumption $\phi(-\alpha) > \phi(\alpha)$ for $\alpha \in (\gamma, 1]$ is naturally satisfied by surrogates strictly decreasing in $[-\alpha_0, \alpha_0]$ with sufficiently large $\alpha_0 > 0$.

Next, we state our main positive result, the proof of which is included in Section 4.11.

Theorem 4.12. *Let $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a surrogate loss. Assume that ϕ is bounded, continuous, nonincreasing, and satisfies Assumption 4.11. Let $\mathcal{F} = \mathcal{F}_{\mathrm{lin}}$. Then,*

1. *ϕ is $(\ell_{01}, \mathcal{F}_{\mathrm{lin}})$-calibrated.*
2. *ϕ is $(\phi_\gamma, \mathcal{F}_{\mathrm{lin}})$-calibrated if and only if $\phi(\gamma) + \phi(-\gamma) > \phi(\alpha) + \phi(-\alpha)$ for all $\alpha \in (\gamma, 1]$.*

Proof. (Sketch of 2) As in the proof sketch of Theorem 4.10, Equation (4.10) is needed for $(\phi_\gamma, \mathcal{F}_{\mathrm{lin}})$-calibration, and thus $\phi(\alpha) + \phi(-\alpha)$ should take larger values in $|\alpha| \leq \gamma$ than in the rest of α. Quasiconcavity of $\phi(\alpha) + \phi(-\alpha)$ naturally implies this property with a non-strict inequality, and the condition $\phi(\gamma) + \phi(-\gamma) > \phi(\alpha) + \phi(-\alpha)$ (for all $\alpha > \gamma$) ensures a strict inequality. Figure 4.3 illustrates this idea with the ramp loss. □

To the best our knowledge, this is the first characterization of losses calibrated to ϕ_γ.

Remark 4.1. *For all $\alpha > \gamma$, $\phi(\gamma) + \phi(-\gamma) \geq \phi(\alpha) + \phi(-\alpha)$ always holds when ϕ is bounded, continuous, nonincreasing, and satisfies Assumption 4.11 (see part 3 of Lemma 4.17 in Section 4.11). The strict inequality $\phi(\gamma) + \phi(-\gamma) > \phi(\alpha) + \phi(-\alpha)$ is then necessary and sufficient for $(\phi_\gamma, \mathcal{F}_{\mathrm{lin}})$-calibration.*

Remark 4.2. *Charoenphakdee et al. [2019] shows that the ramp loss and the sigmoid loss are $(\ell_{01}, \mathcal{F}_{\mathrm{all}})$-calibrated. Note that these two losses are bounded, continuous, nonincreasing, and satisfy Assumption 4.11, hence $(\ell_{01}, \mathcal{F}_{\mathrm{lin}})$-calibrated.*

4.8 Examples

Several examples of loss functions are shown in Figure 4.4. For each base surrogate ϕ, we consider the shifted surrogate $\phi_\beta(\alpha) := \phi(\alpha - \beta)$ with the horizontal shift parameter β. The ramp, sigmoid, and modified squared losses are examples of nonconvex losses satisfying Assumption 4.11 when $\beta \geq 0$, whereas the hinge, logistic, and squared losses are examples of convex losses. We show $(\phi_\gamma, \mathcal{F}_{\mathrm{lin}})$-calibration functions in this subsection.[7] As a result, we will see that the ramp, sigmoid, and modified squared losses are calibrated with appropriate shift parameters. Detailed derivations of the calibration functions and the proofs of quasiconcavity are described in Section 4.11.7.

[7] In this section, we call $\delta_\rho(\varepsilon)$ defined in Lemma 4.8 as a calibration function instead of $\bar{\delta}(\varepsilon, \eta, \mathbf{x})$ with a slight abuse of terminology.

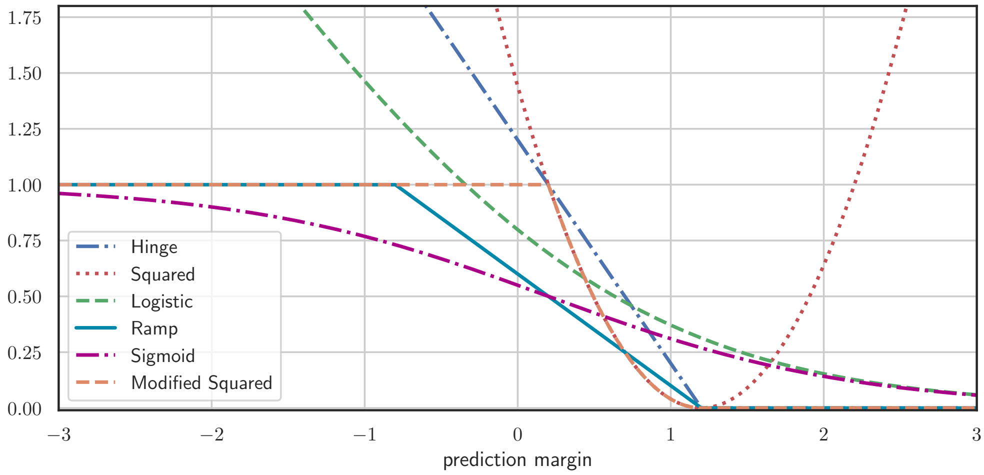
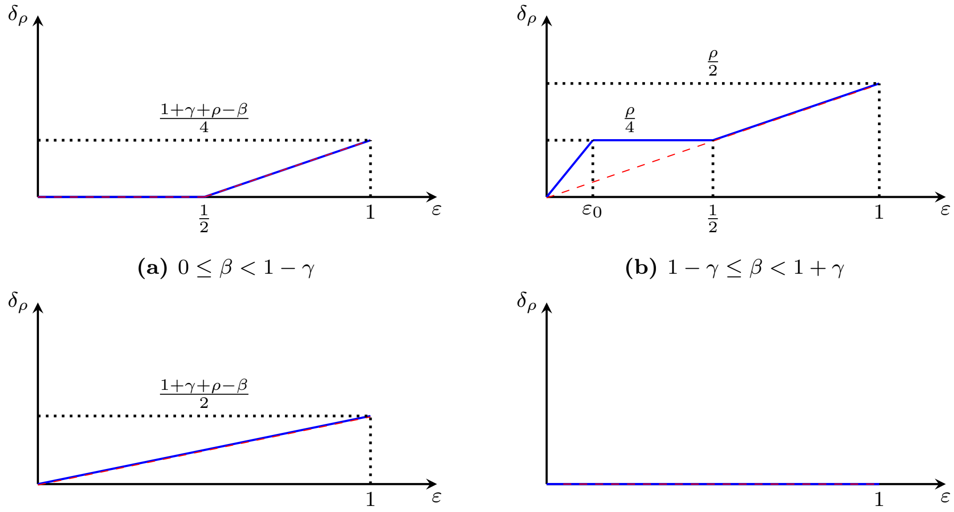


Figure 4.4: Surrogate losses. They are different from the traditional losses through a horizontal translation of $+\beta$ ($\beta = 0.2$ here).



(a) $0 \le \beta < 1 - \gamma$

(b) $1 - \gamma \le \beta < 1 + \gamma$

(c) $1 + \gamma \le \beta < 1 + \gamma + \rho$

(d) $1 + \gamma + \rho \le \beta$

Figure 4.5: Calibration function of the ramp loss (plotted in the real blue lines). $\varepsilon_0 := \frac{\rho}{4(1+\gamma+\rho-\beta)}$. The dashed red line is $\check{\delta}_\rho^{\star\star}$.

4.8.1 Ramp Loss

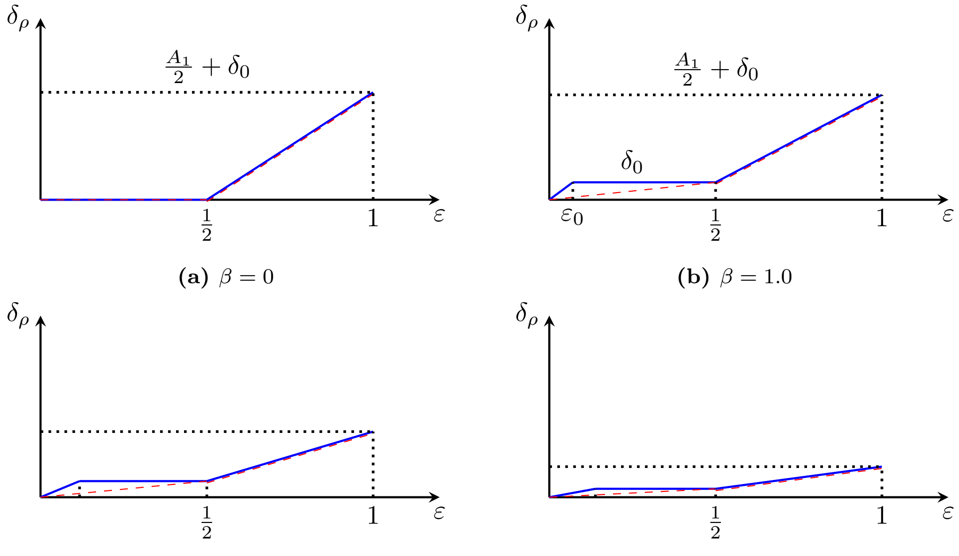
The ramp loss is

$$\phi(\alpha) = \min\left\{1, \max\left\{0, \frac{1-\alpha}{2}\right\}\right\}.$$

We consider the shifted ramp loss

$$\phi_\beta(\alpha) = \phi(\alpha - \beta) = \min\left\{1, \max\left\{0, \frac{1-\alpha+\beta}{2}\right\}\right\}.$$

The $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-calibration function and its Fenchel-Legendre biconjugate of the ramp loss are plotted in Figure 4.5. We can see that the ramp loss is calibrated w.r.t. $(\phi_\gamma,\mathcal{F}_{\text{lin}})$ when $1 - \gamma < \beta < 1 + \gamma$. Since the ramp loss satisfies Assumption 4.11 when $\beta \ge 0$, we also observe that the ramp loss is not calibrated when $\beta = 0$ because it is a symmetric loss [Charoenphakdee et al., 2019], that



(a) $\beta = 0$ **(b)** $\beta = 1.0$

(c) $\beta = 2.0$ **(d)** $\beta = 3.0$

Figure 4.6: Calibration function of the sigmoid loss. $A_0 := \phi_\beta(-\gamma - \rho) - \phi_\beta(\gamma + \rho)$, $A_1 := \phi_\beta(\gamma) - \phi_\beta(-\gamma) - \phi_\beta(\gamma+\rho) + \phi_\beta(-\gamma-\rho)$, $\delta_0 := (\phi_\beta(\gamma) + \phi_\beta(-\gamma) - \phi_\beta(\gamma+\rho) - \phi_\beta(-\gamma-\rho))/2$, and $\varepsilon_0 := \frac{\delta_0}{A_0}$. The dashed line is $\breve{\delta}_\rho^{\star\star}$.

is, $\phi_0(\alpha) + \phi_0(-\alpha) = 1$ for all $\alpha \in \mathbb{R}$, which does not satisfy the condition $\phi_0(\gamma) + \phi_0(-\gamma) > \phi_0(\alpha) + \phi_0(-\alpha)$ for all $\alpha \in (\gamma, 1]$ in Theorem 4.12.

4.8.2 Sigmoid Loss

The sigmoid loss is

$$\phi(\alpha) = \frac{1}{1 + e^{\alpha}}.$$

We consider the shifted sigmoid loss

$$\phi_\beta(\alpha) = \frac{1}{1 + e^{\alpha - \beta}}$$

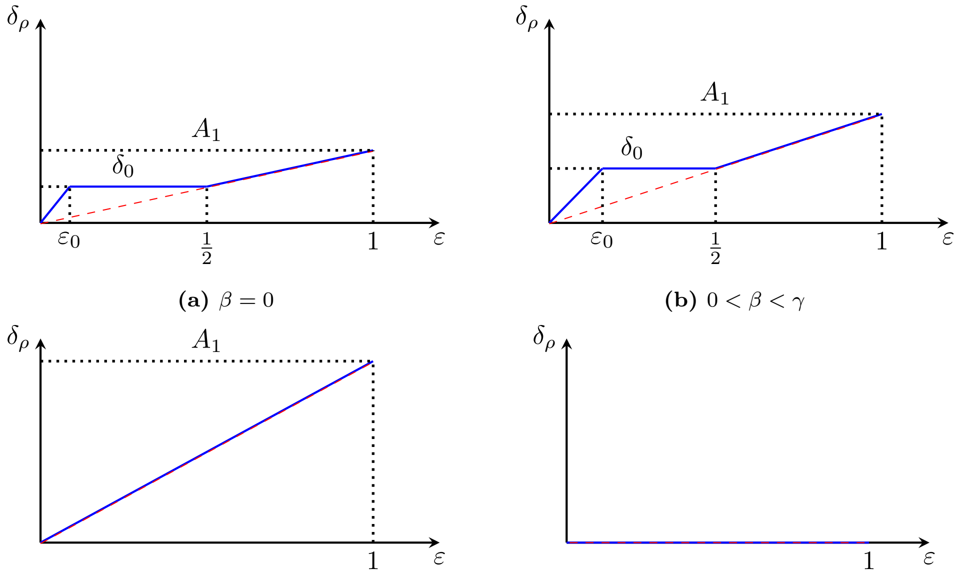
for $\beta > 0$. The $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-calibration function is plotted in Figure 4.6. Thus, the sigmoid loss is $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-calibrated when $\delta_0 > 0$, which is equivalent to $\beta > 0$. Again, we observe that the sigmoid loss with $\beta = 0$ is not calibrated in the same way as the ramp loss because it is symmetric.

4.8.3 Modified Squared Loss

We make a bounded monotone surrogate

$$\phi(\alpha) = \text{clip}_{[0,1]}\left(\max\{\,0, 1 - \alpha\,\}^2\right)$$

by modifying the squared loss, where $\text{clip}_{[a,b]}(\cdot)$ clips values outside the interval $[a, b]$, and consider the shifted version $\phi_\beta(\alpha) := \phi(\alpha - \beta)$. The $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-calibration function and its Fenchel-Legendre biconjugate are plotted in Figure 4.7. We can deduce that the modified squared loss is calibrated w.r.t. $(\phi_\gamma,\mathcal{F}_{\text{lin}})$ for all $0 \leq \beta \leq \gamma$. In contrast to the proceeding examples, the modified squared loss is not symmetric.



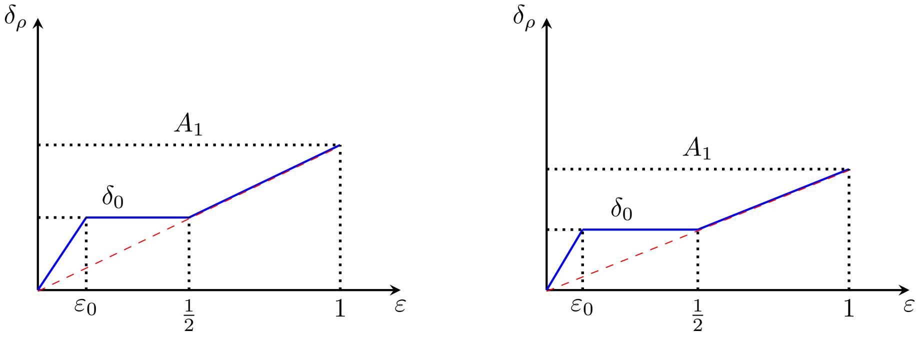
(a) $\beta = 0$

(b) $0 < \beta < \gamma$

(c) $\gamma \le \beta < \gamma + \rho$

(d) $\gamma + \rho \le \beta$

Figure 4.7: Calibration function of the modified squared loss. The dashed line is $\check{\delta}_\rho^{\star\star}$. $A_0 \coloneqq (\gamma + \rho - \beta)(2 + \beta - \gamma - \rho)$, $A_1 \coloneqq \rho(2 + 2\beta - 2\gamma - \rho)$, $\delta_0 \coloneqq \frac{A_1}{2}$, and $\varepsilon_0 \coloneqq \frac{\delta_0}{A_0}$.



(a) $\beta = -0.1$, $\gamma = 0.2$

(b) $\beta = -0.2$, $\gamma = 0.2$

Figure 4.8: Calibration function of the modified squared loss when $\beta < 0$. A_0, A_1, δ_0, ε_0 are the same as in the caption of Figure 4.7.

Moreover, the modified squared loss is $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-calibrated even if ϕ_β for $\beta < 0$ does not satisfy Assumption 4.11.[8] We plot two examples in Figure 4.8. As seen in the proof sketch of Theorem 4.12, it is crucial that $\phi_\beta(\alpha) + \phi_\beta(-\alpha)$ takes higher values in $|\alpha| \le \gamma$ than in $|\alpha| > \gamma$. When $\gamma \le \frac{2}{5}$, the modified squared loss with $-1 - \gamma + \sqrt{1 + 2\gamma^2} < \beta < 0$ satisfies this property (see Figure 4.9).

4.8.4 Hinge Loss and Squared Losses

Here we consider the shifted hinge loss

$$\phi_\beta(\alpha) = \max\{0, 1 - \alpha + \beta\},$$

and the shifted squared loss

$$\phi_\beta(\alpha) = (1 - \alpha + \beta)^2$$

[8] Indeed, its CCR is not necessarily quasiconcave. See Figure 4.17 in Section 4.11.7.

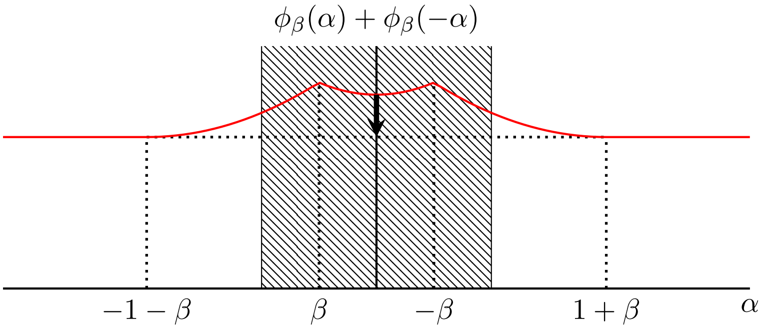
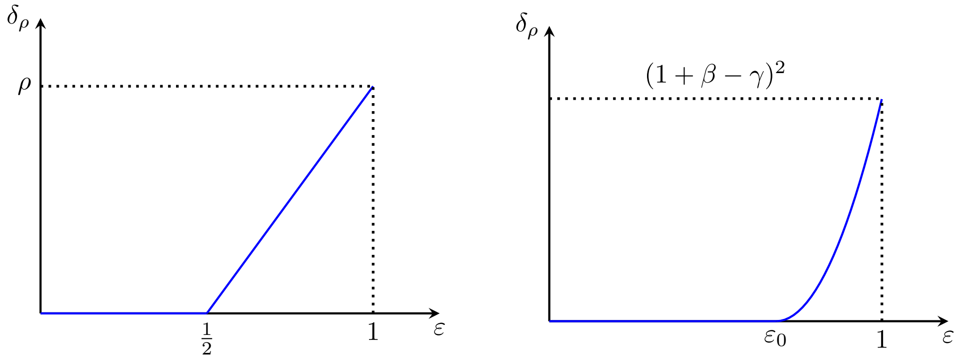


Figure 4.9: Illustration of $\phi_\beta(\alpha) + \phi_\beta(-\alpha)$ for the modified squared loss when $\gamma \leq 0.4$ and $-1-\gamma+\sqrt{1+2\gamma^2} < \beta < 0$. Here, $\beta = -0.2$ and $\gamma = 0.4$.



(a) Hinge loss.

(b) Squared loss.

Figure 4.10: Calibration functions of the hinge and squared losses. $\varepsilon_0 := \frac{1+\beta+\gamma}{2(1+\beta)}$.

as examples of convex losses. Their $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-calibration functions are shown in Figure 4.10, which tell us that the hinge and squared losses are not $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-calibrated. This result aligns with Theorem 4.10.

4.9 Calibrated Losses under Low-noise Condition

In Sections 4.6 and 4.7, we have seen that convex ϕ would not be $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-calibrated whereas some nonconvex ϕ can be calibrated. In this section, we will see that convex losses can be $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-calibrated under a certain assumption on the conditional distribution.

Assumption 4.13. *Let $\xi \in (0,1)$. The conditional distribution satisfies*

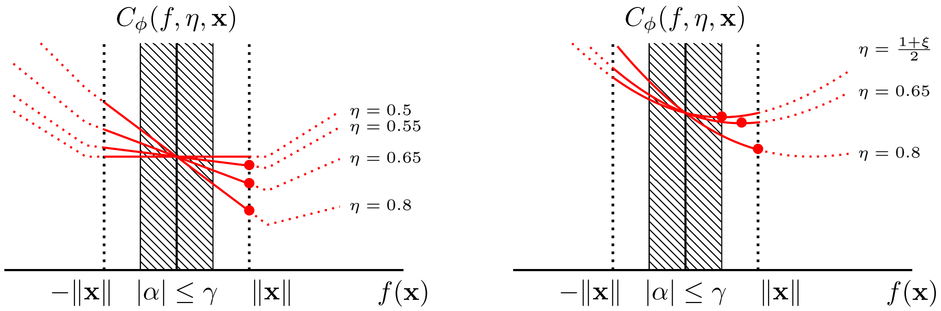
$$|2\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = \mathbf{x}) - 1| \geq \xi$$

almost surely.

This assumption is commonly known as Massart's noise condition and we sometimes refer to it as the ξ-Massart condition [Massart and Nédélec, 2006]. With the Massart condition, we further introduce a modified version of $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-calibrated losses and the calibration function.

Definition 4.14. *For the robust 0-1 loss ϕ_γ and a function class $\mathcal{F}$, we state that a loss $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is* $(\phi_\gamma, \mathcal{F})$-calibrated under the ξ-Massart condition *if for any $\varepsilon > 0$, there exists $\delta > 0$ such that for all $\eta \in [0,1]$ with $|2\eta - 1| \geq \xi$, $\mathbf{x} \in \mathcal{X}$, and $f \in \mathcal{F}$, we have*

$$C_\phi(f, \eta, \mathbf{x}) < C^*_{\phi,\mathcal{F}}(\eta, \mathbf{x}) + \delta \implies C_{\phi_\gamma}(f, \eta, \mathbf{x}) < C^*_{\phi_\gamma,\mathcal{F}}(\eta, \mathbf{x}) + \varepsilon.$$



(a) Hinge loss ($\beta = 0$).

(b) Logistic loss.

Figure 4.11: ϕ-CCR for the hinge and logistic losses with different η. The dots are minimizers of each line.

The corresponding $(\phi_\gamma, \mathcal{F})$-calibration function is defined as

$$\delta_\xi^{\text{Massart}}(\varepsilon) = \inf_{\substack{\eta \in [0,1] \\ |2\eta - 1| \geq \xi}} \inf_{x \in \mathcal{X}} \inf_{f \in \mathcal{F}} \left\{ \Delta C_\phi(f, \eta, \mathbf{x}) \;\middle|\; \Delta C_{\phi_\gamma}(f, \eta, \mathbf{x}) \geq \varepsilon \right\}.$$

As shown in the case of $(\phi_\gamma, \mathcal{F})$-calibration (Proposition 4.4), to check $(\phi_\gamma, \mathcal{F})$-calibration under ξ-Massart condition, it is necessary and sufficient to check $\delta_\xi^{\text{Massart}}(\varepsilon) > 0$ for all $\varepsilon > 0$.

We can then obtain a positive result for convex losses under the Massart condition.

Theorem 4.15. *Under the ξ-Massart condition,*

- *the shifted hinge loss $\phi(\alpha) = [1 - \alpha + \beta]_+$ with any shift $\beta \geq 0$ is $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-calibrated for any $\xi > 0$, and*
- *the logistic loss $\phi(\alpha) = \log(1 + e^{-\alpha})$ is $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-calibrated for $\xi > \tanh\left(\frac{\gamma}{2}\right)$.*

Proof. (Sketch) As we can see in the proof sketches of Theorems 4.10 and 4.12, the suboptimal predictions should be penalized more strictly than the optimal predictions. Under the ξ-Massart condition, let us focus on predictions $f(\mathbf{x})$ for $\eta \geq \frac{1+\xi}{2}$. Because $f(\mathbf{x})$ spans $[-\|\mathbf{x}\|, \|\mathbf{x}\|]$ when $f \in \mathcal{F}_{\text{lin}}$ for a fixed x, the suboptimal predictions are obtained by the infimum of ϕ-CCR in $f(\mathbf{x}) \in [-\|\mathbf{x}\|, \gamma]$ ($[-\|\mathbf{x}\|, 0]$ indicates incorrect predictions and $(0, \gamma]$ indicates non-robust predictions), whereas the optimal predictions are obtained by the infimum in $f(\mathbf{x}) \in [-\|\mathbf{x}\|, \|\mathbf{x}\|]$. Now, take a look at Figure 4.11. Figure 4.11a tells us that the optimal minimizers of the hinge loss is always $f(\mathbf{x}) = \|\mathbf{x}\|$ unless $\eta = \frac{1}{2}$. By contrast, we can see that the optimal minimizers of the logistic loss satisfy $f(\mathbf{x}) > \gamma$ if $\eta > \frac{1+\xi}{2}$. They are strictly less penalized than the calibrated suboptimal minimizers. □

Theorem 4.15 shows that surrogate losses can be $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-calibrated under the Massart condition even if they are not calibrated for all distributions.

Remark 4.3. *Awasthi et al. [2021a, Theorem 25] provides a sufficient condition for $(\ell_\gamma, \mathcal{F}_{\text{lin}})$-consistency to hold for $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-calibrated surrogate loss. Their condition assumes $R^*_{\ell_\gamma, \mathcal{F}_{\text{lin}}} = 0$. Because $R_{\ell_{01}, \mathcal{F}_{\text{lin}}} \leq R_{\ell_\gamma}$, this assumption immediately implies $R^*_{\ell_{01}} = 0$, which is equivalent to Assumption 4.13 with $\xi = 1$. Hence, convex losses lead to $(\ell_\gamma, \mathcal{F}_{\text{lin}})$-consistency under the assumptions of Awasthi et al. [2021a, Theorem 25].*

4.10 Simulation

4.10.1 Learning Curve on Synthetic Data

We use two synthetic datasets.

- **Twonorm.** Positive data are generated from $\mathcal{N}([0.3\ 0.3]^\top, 0.1^2 I_2)$, and negative data are generated from $\mathcal{N}(-[0.3\ 0.3]^\top, 0.1^2 I_2)$. The class ratio is 0.5. All data points lie in the ℓ_2 unit ball with high probability. The classifier $\boldsymbol{\theta} = [1/\sqrt{2}\ 1/\sqrt{2}]^\top$ achieves $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-Bayes risk.
- **Advnorm.** First, clean positive data are generated from

$$\mathcal{N}\left(\begin{bmatrix}0.3\\0.3\end{bmatrix}, 0.142^2 I_2\right),$$

and clean negative data are generated from

$$\mathcal{N}\left(-\begin{bmatrix}0.3\\0.3\end{bmatrix}, 0.142^2 I_2\right).$$

Then, the labels of $(\mathbf{x}, y = +1)$ with $0 < x_1 + x_2 < 0.25$ are flipped to $y = -1$. All data points lie within the ℓ_2 unit ball with high probability. The classifier $\boldsymbol{\theta} = [1/\sqrt{2}\ 1/\sqrt{2}]^\top$ achieves $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-Bayes risk. This dataset is the same that used in the illustration of Figure 4.1.

For each dataset, we generate 500 training and 500 test points.

Linear models $f(\mathbf{x}) = \boldsymbol{\theta}^\top \mathbf{x} + \theta_0$ are used, where $\boldsymbol{\theta}$ and θ_0 are learnable parameters. As surrogate losses, we use the ramp, sigmoid, logistic, and hinge losses. Batch gradient descent with the fixed step size 0.01 is used in the optimization, and 3,000 steps are applied for each trial. After every parameter update, the parameters are normalized to ensure $\|[\boldsymbol{\theta}\ \theta_0]^\top\|_2 = 1$.

Bayes risk computation. The robust 0-1 loss is used as the target loss. The Bayes risk for each surrogate loss and the robust 0-1 loss is numerically computed, which is used to compute the excess risk. To compute the Bayes $(\phi, \mathcal{F}_{\text{lin}})$-risk for a loss ϕ, we substitute the Bayes $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-classifier $f_\gamma^* \in \mathcal{F}_{\text{lin}}$ into

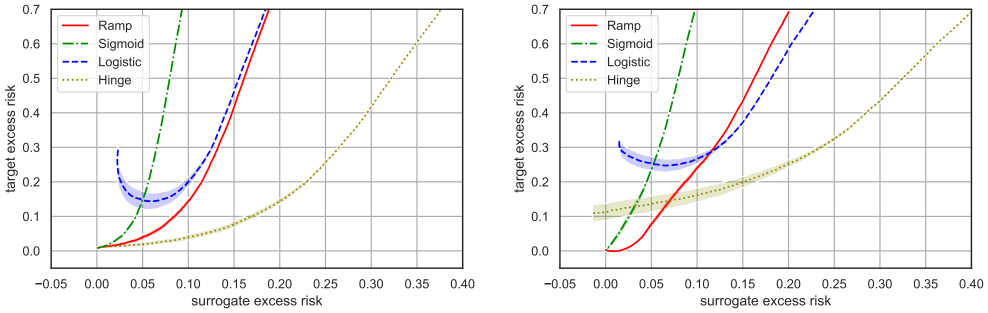
$$\begin{aligned}
&R_\phi(f_\gamma^*) \\
&= \mathbb{E}[\phi(\mathsf{Y} f_\gamma^*(\mathsf{X}))] \\
&= \int_{\mathsf{X}} \left\{ \phi(f_\gamma^*(\mathbf{x}))\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = \mathbf{x}) + \phi(-f_\gamma^*(\mathbf{x}))\mathbb{P}(\mathsf{Y} = -1 \mid \mathsf{X} = \mathbf{x}) \right\} \mathrm{d}\mathbb{P}(\mathsf{X} = \mathbf{x})
\end{aligned}$$

and apply numerical integration. The partitioning quadrature method was used with a grid size of 0.05. The Bayes $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-classifier is $f_\gamma^*(\mathbf{x}) = (x_1 + x_2)/\sqrt{2}$ for both twonorm and advnorm datasets.

To apply a numerical integration, $\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = \mathbf{x})$ needs to be estimated. Note that $\mathbb{P}(\mathsf{X})$ can be estimated given $\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = \mathbf{x})$. For the advnorm dataset, we estimate $\mathbb{P}(\mathsf{X} = \mathbf{x} \mid \mathsf{Y} = +1)$ with the kernel density estimator and then compute $\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = \mathbf{x})$ and $\mathbb{P}(\mathsf{X})$. Subsequently, we focus on the twonorm dataset and derive the closed-form expression of $\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = \mathbf{x})$. Let q_+ and q_- be the probability density functions of $\mathcal{N}([0.3\ 0.3]^\top, 0.1^2 I_2)$ and

Table 4.1: Approximated Bayes risks. For advnorm, we used a kernel density estimator with an RBF kernel (bandwidth: 0.25) to estimate $\mathbb{P}(\mathsf{X} = \mathbf{x} \mid \mathsf{Y} = +1)$.

Loss	twonorm	advnorm
Robust 0-1	0.012	0.067
Ramp	0.389	0.550
Sigmoid	0.445	0.525
Hinge	0.778	1.100
Logistic	0.590	0.750



(a) Twonorm dataset ($\gamma = 0.1, \beta = 0.2$)

(b) Advnorm dataset ($\gamma = 0.1, \beta = 0.5$)

Figure 4.12: Optimization trajectories are shown. The horizontal (vertical) axis shows surrogate excess risk (excess risk of the robust 0-1 loss) on the test data.

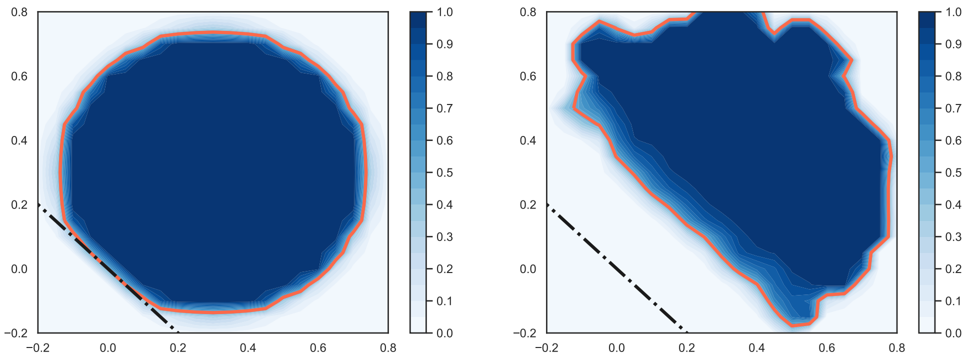
$\mathcal{N}([-0.3 \;\; -0.3], 0.1^2 I_2)$, respectively. Then,

$$
\begin{aligned}
&\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = \mathbf{x}) \\
&\quad = \frac{\mathbb{P}(\mathsf{Y} = +1)\mathbb{P}(\mathsf{X} = \mathbf{x} \mid \mathsf{Y} = +1)}{\mathbb{P}(\mathsf{Y} = +1)\mathbb{P}(\mathsf{X} = \mathbf{x} \mid \mathsf{Y} = +1) + \mathbb{P}(\mathsf{Y} = -1)\mathbb{P}(\mathsf{X} = \mathbf{x} \mid \mathsf{Y} = -1)} \\
&\quad = \frac{\frac{1}{2}q_+(\mathbf{x})}{\frac{1}{2}q_+(\mathbf{x}) + \frac{1}{2}q_-(\mathbf{x})}.
\end{aligned}
$$

The approximated Bayes risks are listed in Table 4.1.

Results. The surrogate and target excess risks are shown in Figure 4.12. A total of 20 trials are run for each data realization. As we can see from Figure 4.12, for both twonorm and advnorm, the optimization trajectories of the calibrated surrogates (the ramp and sigmoid) have target excess risks tending toward zero, whereas the logistic loss fails. This observation agrees with our theoretical findings in Theorems 4.10 and 4.12 for the logistic loss. As for the hinge loss, we observe that it achieves a near-optimal target excess risk on twonorm. This distribution does not satisfy Massart's condition for any $\xi > 0$, which suggests there might be a more general condition that guarantees calibration for certain convex losses. For advnorm, which does not satisfy Massart's condition, hinge fails to converge to a zero target excess risk, most likely because $\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = \mathbf{x})$ changes more smoothly around $\frac{1}{2}$ for advnorm compared to twonorm (see Figure 4.13).

Note again that even if a surrogate loss ϕ is $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-calibrated, it does not immediately imply $(\ell_\gamma, \mathcal{F}_{\text{lin}})$-consistency as pointed out by Awasthi et al. [2021a]. Nonetheless, nonconvex calibrated surrogate losses are useful in practice as illustrated above, and the hinge loss may also perform reasonably when there is not too much noise near the decision boundary.



(a) Twonorm dataset

(b) Advnorm dataset

Figure 4.13: The estimated posterior distributions $\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X})$ are plotted with the same scale. The black dashed line indicates the Bayes $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-classifier and the red solid line indicates the contour line for $\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X}) = 0.5$. As can be seen, the posterior value changes more gradually around 0.5 in Figure 4.13b.

Table 4.2: The simulation results of the γ-adversarially robust 0-1 loss with $\gamma = 0.1$ and $\beta = 0.5$. A total of 50 trials were conducted for each pair of a method and dataset. Standard errors (multiplied by 10^4) are shown in parentheses. Bold font indicates outperforming methods, which were chosen by a one-sided t-test with a significance level of 5%.

	Ramp	Sigmoid	Hinge	Logistic
0 vs 1	0.034 (3)	**0.017 (2)**	0.087 (12)	0.321 (19)
0 vs 2	**0.111 (7)**	0.133 (10)	**0.109 (8)**	0.281 (19)
0 vs 3	**0.107 (7)**	0.126 (8)	0.120 (9)	0.307 (18)
0 vs 4	**0.069 (6)**	0.093 (12)	0.072 (7)	0.269 (21)
0 vs 5	**0.233 (21)**	0.340 (25)	**0.233 (21)**	0.269 (16)
0 vs 6	**0.129 (8)**	0.167 (13)	**0.127 (8)**	0.287 (22)
0 vs 7	**0.067 (6)**	0.073 (6)	0.090 (9)	0.302 (18)
0 vs 8	**0.096 (7)**	0.123 (12)	0.100 (9)	0.263 (20)
0 vs 9	**0.082 (6)**	0.101 (8)	0.092 (8)	0.279 (22)

4.10.2 Benchmark Data

We compare the ramp, sigmoid, hinge, and logistic losses on MNIST. Simulation details are as follows.

- Dataset: MNIST extracted with two digits (7,000 instances for each digit).
- Preprocessing: Reduced to 2-dimensions with a principal component analysis.
- Train-test split: 14,000 instances are randomly split into training and test data with a ratio of 4 to 1.
- Model: Linear models $f(\mathbf{x}) = \boldsymbol{\theta}^\top \mathbf{x} + \theta_0$ ($\boldsymbol{\theta}$ and θ_0 are learnable parameters)
- Surrogate loss: The ramp, sigmoid, hinge, and logistic losses with shift $\beta = +0.5$.
- Target loss: The γ-adversarially robust 0-1 loss with $\gamma = 0.1$.
- Optimization: Batch gradient descent with 1,000 iterations.

The results are shown in Tables 4.2 and 4.3, where we can see that nonconvex losses, particularly the ramp loss, outperform convex losses in terms of the robust 0-1 loss.

Table 4.3: The simulation results of the 0-1 loss with $\beta = 0.5$. A total of 50 trials were conducted for each pair of a method and dataset. Standard errors (multiplied by 10^4) are shown in parentheses. Bold font indicates outperforming methods, which were chosen by a one-sided t-test with a significance level of 5%.

	Ramp	Sigmoid	Hinge	Logistic
0 vs 1	0.012 (2)	**0.005 (1)**	0.038 (7)	0.228 (18)
0 vs 2	**0.050 (5)**	0.059 (7)	0.058 (7)	0.206 (18)
0 vs 3	**0.047 (4)**	0.054 (6)	0.064 (8)	0.229 (15)
0 vs 4	**0.028 (4)**	**0.029 (4)**	0.032 (6)	0.184 (18)
0 vs 5	**0.117 (11)**	0.185 (20)	**0.117 (11)**	0.193 (15)
0 vs 6	**0.060 (5)**	0.080 (8)	0.063 (6)	0.206 (18)
0 vs 7	**0.027 (3)**	**0.027 (4)**	0.045 (6)	0.214 (18)
0 vs 8	**0.050 (6)**	0.054 (6)	0.054 (7)	0.186 (18)
0 vs 9	**0.040 (4)**	0.044 (5)	0.046 (6)	0.192 (20)

4.11 Proofs

4.11.1 Useful Lemmas

The following lemmas are useful in the remaining proofs in Section 4.11. The proofs are provided in Sections 4.11.5 and 4.11.6.

Lemma 4.16. *Let $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a margin-based loss function, and $\mathcal{F} = \mathcal{F}_{\text{lin}}$.*

1. *For all $f \in \mathcal{F}$ and $\mathbf{x} \in \mathcal{X}$, $C_\phi(f, \eta, \mathbf{x})$ and $\Delta C_{\phi,\mathcal{F}}(f, \eta, \mathbf{x})$ are symmetric about $\eta = \frac{1}{2}$, i.e., $C_\phi(f, \eta, \mathbf{x}) = C_\phi(-f, 1-\eta, \mathbf{x})$ and $\Delta C_{\phi,\mathcal{F}}(f, \eta, \mathbf{x}) = \Delta C_{\phi,\mathcal{F}}(-f, 1-\eta, \mathbf{x})$ for all $\eta \in [0, 1]$.*

2. *Fix $\mathbf{x} \in \mathcal{X}$. When $\eta = \frac{1}{2}$, we have*
$$\inf_{f \in \mathcal{F}: |f(\mathbf{x})| \leq \gamma} \Delta C_{\phi,\mathcal{F}}\left(f, \tfrac{1}{2}, \mathbf{x}\right) = \inf_{f \in \mathcal{F}: 0 \leq f(\mathbf{x}) \leq \gamma} \Delta C_{\phi,\mathcal{F}}\left(f, \tfrac{1}{2}, \mathbf{x}\right).$$

3. *A surrogate loss ϕ is calibrated w.r.t. (ϕ_γ,$\mathcal{F}$) if and only if*
$$\inf_{f \in \mathcal{F}: |f(\mathbf{x})| \leq \gamma} C_\phi\left(f, \tfrac{1}{2}, \mathbf{x}\right) > \inf_{f \in \mathcal{F}} C_\phi\left(f, \tfrac{1}{2}, \mathbf{x}\right), \text{ and}$$
$$\inf_{f \in \mathcal{F}: f(\mathbf{x}) \leq \gamma} C_\phi(f, \eta, \mathbf{x}) > \inf_{f \in \mathcal{F}} C_\phi(f, \eta, \mathbf{x}),$$
for all $\eta \in \left(\frac{1}{2}, 1\right]$ and $\mathbf{x} \in \mathcal{X}$ such that $\|\mathbf{x}\|_2 > \gamma$.

4. *A surrogate loss ϕ is calibrated w.r.t. (ℓ_{01},$\mathcal{F}$) if and only if*
$$\inf_{f \in \mathcal{F}: f(\mathbf{x}) \leq 0} C_\phi(f, \eta, \mathbf{x}) > \inf_{f \in \mathcal{F}} C_\phi(f, \eta, \mathbf{x}),$$
for all $\eta \in \left(\frac{1}{2}, 1\right]$ and $\mathbf{x} \in \mathcal{X} \setminus \{\mathbf{0}\}$.

Lemma 4.17. *Let $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a margin-based loss function. In addition, let $\bar{C}_\phi(\alpha, \eta) := \eta\phi(\alpha) + (1-\eta)\phi(-\alpha)$. If ϕ is bounded, continuous, non-increasing, and satisfies Assumption 4.11, then*

1. *for all $\eta \in \left(\frac{1}{2}, 1\right]$, $\bar{C}_\phi(\alpha, \eta)$ is nonincreasing in α when $\alpha \geq 0$.*

2. *for all $\eta \in \left(\frac{1}{2}, 1\right]$ and $\alpha > 0$, $\bar{C}_\phi(-\alpha, \eta) > \bar{C}_\phi(\alpha, \eta)$ if $\phi(-\alpha) > \phi(\alpha)$.*

3. *$\phi(\alpha) + \phi(-\alpha)$ is nonincreasing in α when $\alpha \geq 0$.*

4. *for $l, u \in \mathbb{R}$ ($l \leq u$), $\inf_{\alpha \in [l,u]} \bar{C}_\phi(\alpha, \eta) = \min\{\bar{C}_\phi(l, \eta), \bar{C}_\phi(u, \eta)\}$ for all $\eta \in [0, 1]$.*

4.11.2 Proof of Theorem 4.10

Part 3 of Lemma 4.16 states that ϕ is calibrated w.r.t. $(\phi_\gamma,\mathcal{F}_{\mathrm{lin}})$ if and only if

$$\inf_{f\in\mathcal{F}_{\mathrm{lin}}:0\le f(\mathbf{x})\le\gamma} C_\phi\left(f,\tfrac{1}{2},\mathbf{x}\right) > \inf_{f\in\mathcal{F}_{\mathrm{lin}}:f(\mathbf{x})\ge 0} C_\phi\left(f,\tfrac{1}{2},\mathbf{x}\right) \quad \text{and}$$
$$\inf_{f\in\mathcal{F}_{\mathrm{lin}}:f(\mathbf{x})\le\gamma} C_\phi(f,\eta,\mathbf{x}) > \inf_{f\in\mathcal{F}_{\mathrm{lin}}:f(\mathbf{x})\ge 0} C_\phi(f,\eta,\mathbf{x}) \quad \text{for any } \eta\in\left(\tfrac{1}{2},1\right],$$

for all $\mathbf{x}\in\mathcal{X}$ such that $\|\mathbf{x}\|_2>\gamma$. To show ϕ is not calibrated w.r.t. $(\phi_\gamma,\mathcal{F}_{\mathrm{lin}})$, it is sufficient to show the existence of $\mathbf{x}\in\mathcal{X}$ such that $\|\mathbf{x}\|_2>\gamma$ and

$$\inf_{f\in\mathcal{F}_{\mathrm{lin}}:0\le f(\mathbf{x})\le\gamma} C_\phi\left(f,\tfrac{1}{2},\mathbf{x}\right) = \inf_{f\in\mathcal{F}_{\mathrm{lin}}:f(\mathbf{x})\ge 0} C_\phi\left(f,\tfrac{1}{2},\mathbf{x}\right),$$

which is equivalent to

$$\inf_{f\in\mathcal{F}_{\mathrm{lin}}:0\le f(\mathbf{x})\le\gamma} \phi(f(\mathbf{x}))+\phi(-f(\mathbf{x})) = \inf_{f\in\mathcal{F}_{\mathrm{lin}}:f(\mathbf{x})\ge 0} \phi(f(\mathbf{x}))+\phi(-f(\mathbf{x})). \tag{4.11}$$

Because $\bar{\phi}(\alpha)\coloneqq\phi(\alpha)+\phi(-\alpha)$ is a convex even function, we have $\bar{\phi}(0)\le\bar{\phi}(\alpha)$ for all $\alpha\in\mathbb{R}$. To see this, assume that there exists $\alpha_*\in\mathbb{R}$ such that $\alpha_*\ne 0$ and $\bar{\phi}(0)>\bar{\phi}(\alpha_*)$. We then also have $\bar{\phi}(-\alpha_*)<\bar{\phi}(0)$ because $\bar{\phi}$ is even. It follows that $\frac{1}{2}\{\bar{\phi}(-\alpha_*)+\bar{\phi}(\alpha_*)\}<\bar{\phi}(0)$. However, we have $\frac{1}{2}\{\bar{\phi}(-\alpha_*)+\bar{\phi}(\alpha_*)\}\ge\bar{\phi}\left(\frac{-\alpha_*+\alpha_*}{2}\right)=\bar{\phi}(0)$ because of the convexity of $\bar{\phi}$. Hence, we can see that $\bar{\phi}(0)\le\bar{\phi}(\alpha)$ for all $\alpha\in\mathbb{R}$. This means that $\inf_{0\le\alpha\le\gamma}\bar{\phi}(\alpha)=\inf_{\alpha\in\mathcal{A}:0\le\alpha}\bar{\phi}(\alpha)=\bar{\phi}(0)$, where $\mathcal{A}$ is any subset of the real line containing 0. Note that for $\mathbf{x}\in\mathcal{X}$ such that $\|\mathbf{x}\|_2>\gamma$, $f(\mathbf{x})$ ranges $[-\|\mathbf{x}\|_2,\|\mathbf{x}\|_2]\supseteq[-\gamma,\gamma]$ with $f\in\mathcal{F}_{\mathrm{lin}}$. Therefore, for any choice of a fixed $\mathbf{x}\in\mathcal{X}$ such that $\|\mathbf{x}\|_2>\gamma$, we have $\inf_{f\in\mathcal{F}_{\mathrm{lin}}:0\le f(\mathbf{x})\le\gamma}\bar{\phi}(f(\mathbf{x}))=\inf_{0\le\alpha\le\gamma}\bar{\phi}(\alpha)$ and $\inf_{f\in\mathcal{F}_{\mathrm{lin}}:f(\mathbf{x})\ge 0}\bar{\phi}(f(\mathbf{x}))=\inf_{0\le\alpha\le\|\mathbf{x}\|_2}\bar{\phi}(\alpha)$. This implies that the sufficient condition (4.11) for the nonexistence of convex surrogate losses holds. □

4.11.3 Proof of Theorem 4.12

Let $\bar{C}_\phi(\alpha,\eta)\coloneqq\eta\phi(\alpha)+(1-\eta)\phi(-\alpha)$.

Part 1. Based on part 4 of Lemma 4.16, $(\ell_{01},\mathcal{F}_{\mathrm{lin}})$-calibration is equivalent to

$$\inf_{f\in\mathcal{F}_{\mathrm{lin}}:f(\mathbf{x})\le 0} C_\phi(f,\eta,\mathbf{x}) > \inf_{f\in\mathcal{F}_{\mathrm{lin}}} C_\phi(f,\eta,\mathbf{x}) \quad \text{for all } \eta\in\left(\tfrac{1}{2},1\right] \text{ and } \mathbf{x}\in\mathcal{X}\setminus\{\mathbf{0}\}. \tag{4.12}$$

Fix an arbitrary η such that $\frac{1}{2}<\eta\le 1$ and $\mathbf{x}\in\mathcal{X}\setminus\{\mathbf{0}\}$. We can observe through part 4 of Lemma 4.17 that

$$\begin{aligned}
\inf_{\substack{f\in\mathcal{F}_{\mathrm{lin}}:\\ f(\mathbf{x})\le 0}} C_\phi(f,\eta,\mathbf{x}) &= \inf_{\alpha\in[-\|\mathbf{x}\|_2,0]} \bar{C}_\phi(\alpha,\eta) \\
&= \min\left\{\, \bar{C}_\phi(-\|\mathbf{x}\|_2,\eta), \bar{C}_\phi(0,\eta) \,\right\} \quad \text{(part 4 of Lemma 4.17)} \\
&= \min\left\{\, \eta\overline{B}+(1-\eta)\underline{B}, \phi(0) \,\right\},
\end{aligned}$$

and

$$\begin{aligned}
\inf_{f\in\mathcal{F}_{\mathrm{lin}}} C_\phi(f,\eta,\mathbf{x}) &= \inf_{\alpha\in[-\|\mathbf{x}\|_2,\|\mathbf{x}\|_2]} \bar{C}_\phi(\alpha,\eta) \\
&= \min\left\{\, \bar{C}_\phi(-\|\mathbf{x}\|_2,\eta), \bar{C}_\phi(\|\mathbf{x}\|_2,\eta) \,\right\} \quad \text{(part 4 of Lemma 4.17)} \\
&= \bar{C}_\phi(\|\mathbf{x}\|_2,\eta) \quad \text{(part 2 of Lemma 4.17)} \\
&= \eta\underline{B}+(1-\eta)\overline{B},
\end{aligned}$$

where $\underline{B} \coloneqq \phi(\|\mathbf{x}\|_2)$ and $\overline{B} \coloneqq \phi(-\|\mathbf{x}\|_2)$. Note that $\overline{B} > \underline{B}$ because $\|\mathbf{x}\|_2 > 0$. Here,

$$\begin{aligned}
\bar{C}_\phi(-\|\mathbf{x}\|_2, \eta) - \bar{C}_\phi(\|\mathbf{x}\|_2, \eta) &= (\overline{B} - \underline{B})(2\eta - 1) > 0, \\
\bar{C}_\phi(0, \eta) - \bar{C}_\phi(\|\mathbf{x}\|_2, \eta) &= \phi(0) - \overline{B} + \eta(\overline{B} - \underline{B}) \\
&\geq \frac{\overline{B} + \underline{B}}{2} - \overline{B} + \eta(\overline{B} - \underline{B}) \\
&> \frac{\overline{B} + \underline{B}}{2} - \overline{B} + \frac{\overline{B} - \underline{B}}{2} \qquad (\overline{B} > \underline{B} \text{ and } \eta > \tfrac{1}{2}) \\
&= 0,
\end{aligned}$$

where the first inequality is shown through quasiconcavity of $\alpha \mapsto \phi(\alpha) + \phi(-\alpha)$. Indeed, by letting $F(\alpha) \coloneqq \phi(\alpha) + \phi(-\alpha)$,

$$\begin{aligned}
2\phi(0) = F(0) &= F\left(\frac{\|\mathbf{x}\|_2}{2} + \frac{-\|\mathbf{x}\|_2}{2}\right) \\
&\geq \min\{\, F(\,\|\,\mathbf{x}\|_2), F(-\|\mathbf{x}\|_2)\,\} \\
&= F(\|\mathbf{x}\|_2) \\
&= \overline{B} + \underline{B}.
\end{aligned}$$

We then have

$$\begin{aligned}
&\inf_{f \in \mathcal{F}_{\mathrm{lin}} : f(\mathbf{x}) \leq 0} C_\phi(f, \eta, \mathbf{x}) - \inf_{f \in \mathcal{F}_{\mathrm{lin}}} C_\phi(f, \eta, \mathbf{x}) \\
&\qquad = \min\{\, \bar{C}_\phi(-\,\|\,\mathbf{x}\|_2, \eta) - \bar{C}_\phi(\|\mathbf{x}\|_2, \eta), \bar{C}_\phi(0, \eta) - \bar{C}_\phi(\|\mathbf{x}\|_2, \eta)\,\} \\
&\qquad > 0.
\end{aligned}$$

This verifies Equation (4.12). □

Part 2. Here, ϕ is calibrated w.r.t. (ϕ_γ,$\mathcal{F}_{\mathrm{lin}}$) if and only if

$$\begin{aligned}
&\text{(i)} \quad \inf_{f \in \mathcal{F}_{\mathrm{lin}} : |f(\mathbf{x})| \leq \gamma} C_\phi\left(f, \tfrac{1}{2}, \mathbf{x}\right) > \inf_{f \in \mathcal{F}_{\mathrm{lin}}} C_\phi\left(f, \tfrac{1}{2}, \mathbf{x}\right), \quad \text{and} \\
&\text{(ii)} \quad \inf_{f \in \mathcal{F}_{\mathrm{lin}} : f(\mathbf{x}) \leq \gamma} C_\phi(f, \eta, \mathbf{x}) > \inf_{f \in \mathcal{F}_{\mathrm{lin}}} C_\phi(f, \eta, \mathbf{x}) \quad \text{for all } \eta \in \left(\tfrac{1}{2}, 1\right]
\end{aligned} \tag{4.13}$$

for any $\mathbf{x} \in \mathcal{X}$ such that $\|\mathbf{x}\|_2 > \gamma$, based on part 3 of Lemma 4.16. Now, assuming (i) and (ii), we show $\phi(\gamma) + \phi(-\gamma) > \phi(\alpha) + \phi(-\alpha)$ for any $\alpha \in (\gamma, 1]$. For an arbitrary $\alpha \in (\gamma, 1]$, choose an x such that $\|\mathbf{x}\|_2 = \alpha$, and thus $\{\, f(\mathbf{x}) \mid f \in \mathcal{F}_{\mathrm{lin}} \,\}$ ranges within $[-\alpha, \alpha]$.

$$\begin{aligned}
\phi(\gamma) + \phi(-\gamma) &= \inf_{0 \leq \alpha' \leq \gamma} \phi(\alpha') + \phi(-\alpha') && \text{(part 3 of Lemma 4.17)} \\
&= \inf_{f \in \mathcal{F}_{\mathrm{lin}} : 0 \leq f(\mathbf{x}) \leq \gamma} \phi(f(\mathbf{x})) + \phi(-f(\mathbf{x})) \\
&= \inf_{f \in \mathcal{F}_{\mathrm{lin}} : |f(\mathbf{x})| \leq \gamma} \phi(f(\mathbf{x})) + \phi(-f(\mathbf{x})) && (\phi(\alpha) + \phi(-\alpha) \text{ is even}) \\
&> \inf_{f \in \mathcal{F}_{\mathrm{lin}}} \phi(f(\mathbf{x})) + \phi(-f(\mathbf{x})) && \text{((i) is used)} \\
&= \inf_{-\|\mathbf{x}\|_2 \leq \alpha' \leq \|\mathbf{x}\|_2} \phi(\alpha') + \phi(-\alpha') \\
&= \inf_{0 \leq \alpha' \leq \alpha} \phi(\alpha') + \phi(-\alpha') && (\phi(\alpha) + \phi(-\alpha) \text{ is even}) \\
&= \phi(\alpha) + \phi(-\alpha). && \text{(part 3 of Lemma 4.17)}
\end{aligned}$$

Conversely, assume $\phi(\gamma)+\phi(-\gamma) > \phi(\alpha)+\phi(-\alpha)$ for any $\alpha \in (\gamma, 1]$. We will show (i) and (ii) in Equation (4.13). Fix an $\mathbf{x} \in \mathcal{X}$ such that $\|\mathbf{x}\|_2 > \gamma$ arbitrarily, and thus $\{\, f(\mathbf{x}) \mid f \in \mathcal{F}_{\text{lin}} \,\}$ ranges within $[-\|\mathbf{x}\|_2, \|\mathbf{x}\|_2] \supseteq [-\gamma, \gamma]$. Because $\phi(\alpha)+\phi(-\alpha)$ is nonincreasing in $\alpha \geq 0$ (part 3 of Lemma 4.17), we have

$$\begin{aligned}
2 \inf_{\substack{f \in \mathcal{F}_{\text{lin}}: \\ |f(\mathbf{x})| \leq \gamma}} C_\phi\left(f, \tfrac{1}{2}, \mathbf{x}\right) &= \inf_{|\alpha| \leq \gamma} \phi(\alpha)+\phi(-\alpha) && (f(\mathbf{x}) \in [-\|\mathbf{x}\|_2, \|\mathbf{x}\|_2]) \\
&= \inf_{0 \leq \alpha \leq \gamma} \phi(\alpha)+\phi(-\alpha) && (\phi(\alpha)+\phi(-\alpha) \text{ is even}) \\
&= \phi(\gamma)+\phi(-\gamma) && (\text{part 3 of Lemma 4.17}) \\
&> \phi(\|\mathbf{x}\|_2)+\phi(-\|\mathbf{x}\|_2) && (\text{by assumption}) \\
&= \inf_{0 \leq \alpha \leq \|\mathbf{x}\|_2} \phi(\alpha)+\phi(-\alpha), && (\text{part 3 of Lemma 4.17}) \\
&= \inf_{-\|\mathbf{x}\|_2 \leq \alpha \leq \|\mathbf{x}\|_2} \phi(\alpha)+\phi(-\alpha) && (\phi(\alpha)+\phi(-\alpha) \text{ is even}) \\
&= 2 \inf_{f \in \mathcal{F}_{\text{lin}}} C_\phi\left(f, \tfrac{1}{2}, \mathbf{x}\right),
\end{aligned}$$

which is equivalent to (i). For (ii), fix an η such that $\frac{1}{2} < \eta \leq 1$. We first observe through parts 2 and 4 of Lemma 4.17 that

$$\begin{aligned}
\inf_{-\|\mathbf{x}\|_2 \leq \alpha \leq \gamma} \bar{C}_\phi(\alpha, \eta) &= \min\left\{\, \bar{C}_\phi(-\|\mathbf{x}\|_2, \eta), \bar{C}_\phi(\gamma, \eta) \,\right\}, \\
\inf_{-\|\mathbf{x}\|_2 \leq \alpha \leq \|\mathbf{x}\|_2} \bar{C}_\phi(\alpha, \eta) &= \min\left\{\, \bar{C}_\phi(-\|\mathbf{x}\|_2, \eta), \bar{C}_\phi(\|\mathbf{x}\|_2, \eta) \,\right\} = \bar{C}_\phi(\|\mathbf{x}\|_2, \eta).
\end{aligned}$$

Here, we have

$$\begin{aligned}
\bar{C}_\phi(\|\mathbf{x}\|_2, \eta) &= (\underline{B} - \overline{B})\eta + \overline{B}, \\
\bar{C}_\phi(\gamma, \eta) &= (\phi(\gamma) - \phi(-\gamma))\eta + \phi(-\gamma),
\end{aligned}$$

where $\underline{B} \coloneqq \phi(\|\mathbf{x}\|_2)$ and $\overline{B} \coloneqq \phi(-\|\mathbf{x}\|_2)$. Then, for all $\eta \in \left(\frac{1}{2}, 1\right]$,

$$\begin{aligned}
\bar{C}_\phi(\gamma, \eta) - \bar{C}_\phi(\|\mathbf{x}\|_2, \eta) &= (\phi(\gamma) - \phi(-\gamma) + \overline{B} - \underline{B})\eta + (\phi(-\gamma) - \overline{B}) \\
&\geq (\phi(\gamma) - \phi(-\gamma) + \overline{B} - \underline{B})\frac{1}{2} + \phi(-\gamma) - \overline{B} \\
&= \frac{\{\phi(\gamma)+\phi(-\gamma)\} - \{\phi(\|\mathbf{x}\|_2)+\phi(-\|\mathbf{x}\|_2)\}}{2} \\
&> 0,
\end{aligned}$$

where the first inequality holds because $(\phi(\gamma) - \phi(-\gamma) + \overline{B} - \underline{B}) > 0$ and $\eta > \frac{1}{2}$, and the second inequality holds because of the assumption $\phi(\gamma)+\phi(-\gamma) > \phi(\alpha)+\phi(-\alpha)$ for any $\alpha \in (\gamma, 1]$. In addition, we have $\bar{C}_\phi(-\|\mathbf{x}\|_2, \eta) > \bar{C}_\phi(\|\mathbf{x}\|_2, \eta)$ for $\eta > \frac{1}{2}$ through part 2 of Lemma 4.17. Therefore,

$$\begin{aligned}
&\inf_{f \in \mathcal{F}_{\text{lin}}: f(\mathbf{x}) \leq \gamma} C_\phi(f, \eta, \mathbf{x}) - \inf_{f \in \mathcal{F}_{\text{lin}}} C_\phi(f, \eta, \mathbf{x}) \\
&\qquad = \inf_{-\|\mathbf{x}\|_2 \leq \alpha \leq \gamma} \bar{C}_\phi(\alpha, \eta) - \inf_{-\|\mathbf{x}\|_2 \leq \alpha \leq \|\mathbf{x}\|_2} \bar{C}_\phi(\alpha, \eta) \\
&\qquad = \min\left\{\, \bar{C}_\phi(-\|\mathbf{x}\|_2, \eta) - \bar{C}_\phi(\|\mathbf{x}\|_2, \eta), \bar{C}_\phi(\gamma, \eta) - \bar{C}_\phi(\|\mathbf{x}\|_2, \eta) \,\right\} \\
&\qquad > 0
\end{aligned}$$

holds for all η such that $\frac{1}{2} < \eta \leq 1$, which verifies (ii). □

4.11.4 Proof of Theorem 4.15

First, we derive the equivalent condition for a $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-calibration under a ξ-Massart condition. Let us introduce

$$\delta_{\rho,\xi}^{\text{Massart}}(\varepsilon) := \inf_{\substack{\eta\in[0,1] \\ |2\eta-1|\ge\xi}} \inf_{x\in\widetilde{\mathcal{X}}_\rho} \inf_{f\in\mathcal{F}} \left\{ \Delta C_\phi(f,\eta,\mathbf{x}) \;\middle|\; \Delta C_{\phi_\gamma}(f,\eta,\mathbf{x}) \ge \varepsilon \right\},$$

where $\widetilde{\mathcal{X}}_\rho := \mathcal{X} \setminus \mathcal{B}_2^\circ(\gamma+\rho)$. A loss function ϕ is $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-calibrated if and only if $\delta_\xi^{\text{Massart}}(\varepsilon) > 0$ for all $\varepsilon > 0$. It is easy to see that ϕ is $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-calibrated if and only if $\delta_{\rho,\xi}^{\text{Massart}}(\varepsilon) > 0$ for all $\varepsilon > 0$ and $\rho \in (0, 1-\gamma)$ by the same argument as the proof of Lemma 4.8. By following the same argument as the proof of Lemma 4.16 (part 3), we claim that a surrogate ϕ is calibrated if and only if

$$\inf_{f\in\mathcal{F}_{\text{lin}}:f(\mathbf{x})\le\gamma} C_\phi(f,\eta,\mathbf{x}) > \inf_{f\in\mathcal{F}_{\text{lin}}} C_\phi(f,\eta,\mathbf{x})$$

for all η and $\mathbf{x}$ such that $\eta \ge \frac{1+\xi}{2}$ and $\|\mathbf{x}\|_2 > \gamma$.[9] Denote $\bar{C}_\phi(f(\mathbf{x}),\eta) = C_\phi(f,\eta,\mathbf{x}) = \eta\phi(f(\mathbf{x})) + (1-\eta)\phi(-f(\mathbf{x}))$. This is then equivalent to

$$H(\eta,\mathbf{x}) := \inf_{\alpha\in[-\|\mathbf{x}\|_2,\gamma]} \bar{C}_\phi(\alpha,\eta) - \inf_{\alpha\in[-\|\mathbf{x}\|_2,\|\mathbf{x}\|_2]} \bar{C}_\phi(\alpha,\eta) > 0 \quad \text{for all } \eta \ge \tfrac{1+\xi}{2},$$

by noting that $f(\mathbf{x})$ spans $[-\|\mathbf{x}\|_2, \|\mathbf{x}\|_2]$ for $f \in \mathcal{F}_{\text{lin}}$.

Next, we will check each loss function.

Shifted hinge loss. Because

$$\bar{C}_\phi(\alpha,\eta) = \begin{cases} -\eta\alpha + \eta(1+\beta) & \text{if } \alpha < -(1+\beta), \\ (1-2\eta)\alpha + (1+\beta) & \text{if } -(1+\beta) \le \alpha < 1+\beta, \\ (1-\eta)\alpha + (1-\eta)(1+\beta) & \text{if } 1+\beta \le \alpha, \end{cases}$$

we have

$$\inf_{\alpha\in[-\|\mathbf{x}\|_2,\gamma]} \bar{C}_\phi(\alpha,\eta) = \bar{C}_\phi(\gamma,\eta), \qquad \inf_{\alpha\in[-\|\mathbf{x}\|_2,\|\mathbf{x}\|_2]} \bar{C}_\phi(\alpha,\eta) = \bar{C}_\phi(\|\mathbf{x}\|_2,\eta).$$

Hence, $H(\eta,\mathbf{x}) = (1-2\eta)(\gamma - \|\mathbf{x}\|_2) \ge \xi(\|\mathbf{x}\|_2 - \gamma) > 0$, implying that the hinge loss is $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-calibrated under any $\xi > 0$. □

Logistic loss. The minimizer of

$$\bar{C}_\phi(\alpha,\eta) = \eta\log(1+e^{-\alpha}) + (1-\eta)\log(1+e^{\alpha})$$

in $\alpha \in \mathbb{R}$ is $\alpha^*(\eta) = \ln\left(\frac{\eta}{1-\eta}\right)$. When $\eta \ge \frac{1+\xi}{2}$ with $\xi > \tanh\left(\frac{\gamma}{2}\right)$, we have $\alpha^*(\eta) > \gamma$. Because $\bar{C}_\phi(\alpha,\eta)$ is convex in α, it is decreasing for $\alpha \le \alpha^*(\eta)$. Hence,

- when $\alpha^*(\eta) \le \|\mathbf{x}\|_2$, $H(\eta,\mathbf{x}) = \bar{C}_\phi(\gamma,\eta) - \bar{C}_\phi(\alpha^*(\eta),\eta) > 0$, and
- when $\gamma < \|\mathbf{x}\|_2 < \alpha^*(\eta)$, $H(\eta,\mathbf{x}) = \bar{C}_\phi(\gamma,\eta) - \bar{C}_\phi(\|\mathbf{x}\|_2,\eta) > 0$ since $\gamma < \|\mathbf{x}\|_2$.

Therefore, the logistic loss is $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-calibrated under $\xi > \tanh\left(\frac{\gamma}{2}\right)$. □

[9] The proof of this argument is a routine given the proof of Lemma 4.16 (part 3), which is omitted.

4.11.5 Proof of Lemma 4.16

Parts 1 and 2 are obvious from the definition of the class-conditional ϕ-risk.

Part 3. Let $\bar{\delta} : \mathbb{R}_{\geq 0} \times [0,1] \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ be the $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-calibration function, whose expression is given in Lemma 4.9: for $\mathbf{x} \in \mathcal{X}$ such that $\|\mathbf{x}\|_2 > \gamma$,

$$\bar{\delta}(\varepsilon, \eta, \mathbf{x}) = \begin{cases} \infty & \text{if } \varepsilon > \max\{\eta, 1-\eta\}, \\ \inf_{f \in \mathcal{F}_{\text{lin}} : |f(\mathbf{x})| \leq \gamma} \Delta C_{\phi, \mathcal{F}_{\text{lin}}}(f, \eta, \mathbf{x}) & \text{if } |2\eta - 1| < \varepsilon \leq \max\{\eta, 1-\eta\}, \\ \inf_{\substack{f \in \mathcal{F}_{\text{lin}} : \\ |f(\mathbf{x})| \leq \gamma \text{ or } (2\eta-1) f(\mathbf{x}) \leq 0}} \Delta C_{\phi, \mathcal{F}_{\text{lin}}}(f, \eta, \mathbf{x}) & \text{if } \varepsilon \leq |2\eta - 1|, \end{cases}$$

and $\bar{\delta}(\varepsilon, \eta, \mathbf{x}) = \infty$ for $\|\mathbf{x}\|_2 \leq \gamma$. Proposition 4.4 states that ϕ is $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-calibrated if and only if $\bar{\delta}(\varepsilon, \eta, \mathbf{x}) > 0$ for all $\varepsilon > 0$, $\eta \in [0,1]$, and $\mathbf{x} \in \mathcal{X}$ with $\|\mathbf{x}\|_2 > \gamma$. We subsequently fix $\|\mathbf{x}\|_2 > \gamma$ and simplify the third expression

$$\inf_f \{ \Delta C_{\phi, \mathcal{F}_{\text{lin}}}(f, \eta, \mathbf{x}) \mid f \in \mathcal{F}_{\text{lin}}, |f(\mathbf{x})| \leq \gamma \text{ or } (2\eta - 1) f(\mathbf{x}) \leq 0 \}$$

first. Using part 1 of Lemma 4.16 and the symmetry of $\mathcal{F}_{\text{lin}}$, for $\eta \leq \frac{1}{2}$ we have

$$\begin{aligned} & \inf_{\substack{f \in \mathcal{F}_{\text{lin}} : \\ |f(\mathbf{x})| \leq \gamma \text{ or } (2\eta-1) f(\mathbf{x}) \leq 0}} C_\phi(f, \eta, \mathbf{x}) \\ &= \inf_{f \in \mathcal{F}_{\text{lin}} : f(\mathbf{x}) \geq -\gamma} C_\phi(f, \eta, \mathbf{x}) \\ &= \inf_{f \in \mathcal{F}_{\text{lin}} : f(\mathbf{x}) \geq -\gamma} C_\phi(-f, 1-\eta, x) && \text{(part 1 of Lemma 4.16)} \\ &= \inf_{f \in \mathcal{F}_{\text{lin}} : f(\mathbf{x}) \leq \gamma} C_\phi(f, 1-\eta, x), && \text{(replace } -f \text{ with } f) \end{aligned}$$

and for $\eta \geq \frac{1}{2}$ we have

$$\inf_{f \in \mathcal{F}_{\text{lin}} : |f(\mathbf{x})| \leq \gamma \text{ or } (2\eta-1) f(\mathbf{x}) \leq 0} C_\phi(f, \eta, \mathbf{x}) = \inf_{f \in \mathcal{F}_{\text{lin}} : f(\mathbf{x}) \leq \gamma} C_\phi(f, \eta, \mathbf{x}).$$

By combining these two, we see that $\inf_{f \in \mathcal{F}_{\text{lin}} : f(\mathbf{x}) \leq \gamma} \Delta C_{\phi, \mathcal{F}_{\text{lin}}}(f, \eta, \mathbf{x}) > 0$ holds for all $\eta \geq \frac{1}{2}$ if and only if

$$\inf_{f \in \mathcal{F}_{\text{lin}} : |f(\mathbf{x})| \leq \gamma \text{ or } (2\eta-1) f(\mathbf{x}) \leq 0} \Delta C_{\phi, \mathcal{F}_{\text{lin}}}(f, \eta, \mathbf{x}) > 0$$

holds for all $\eta \in [0,1]$. Hence,

$$\inf_{f \in \mathcal{F}_{\text{lin}} : |f(\mathbf{x})| \leq \gamma \text{ or } (2\eta-1) f(\mathbf{x}) \leq 0} \Delta C_{\phi, \mathcal{F}_{\text{lin}}}(f, \eta, \mathbf{x}) > 0$$

for $\varepsilon > 0$ and $\eta \in [0,1]$ such that $\varepsilon \leq |2\eta - 1|$ if and only if

$$\inf_{f \in \mathcal{F}_{\text{lin}} : f(\mathbf{x}) \leq \gamma} \Delta C_{\phi, \mathcal{F}_{\text{lin}}}(f, \eta, \mathbf{x}) > 0$$

for $\varepsilon > 0$ and $\eta \in [\frac{1}{2}, 1]$ such that $\varepsilon \leq 2\eta - 1$.

Note that the second expression $\inf_f \{ \Delta C_{\phi, \mathcal{F}_{\text{lin}}}(f, \eta, \mathbf{x}) \mid f \in \mathcal{F}_{\text{lin}}, |f(\mathbf{x})| \leq \gamma \}$ can be simplified in the same way. Therefore, $\bar{\delta}(\varepsilon, \eta, \mathbf{x}) > 0$ for all $\varepsilon > 0$, $\eta \in [0,1]$, and $\mathbf{x} \in \mathcal{X}$ if and only if

$$\begin{cases} \inf_{f \in \mathcal{F}_{\text{lin}} : |f(\mathbf{x})| \leq \gamma} C_\phi(f, \eta, \mathbf{x}) > \inf_{f \in \mathcal{F}_{\text{lin}}} C_\phi(f, \eta, \mathbf{x}) & \forall \eta \geq \frac{1}{2} \ (2\eta - 1 < \varepsilon \leq \eta), \\ \inf_{f \in \mathcal{F}_{\text{lin}} : f(\mathbf{x}) \leq \gamma} C_\phi(f, \eta, \mathbf{x}) > \inf_{f \in \mathcal{F}_{\text{lin}}} C_\phi(f, \eta, \mathbf{x}) & \forall \eta \geq \frac{1}{2} \ (\varepsilon \leq 2\eta - 1), \end{cases}$$

for all $\varepsilon > 0$, which is equivalent to

$$\begin{cases} \inf_{f\in\mathcal{F}_{\mathrm{lin}}:|f(\mathbf{x})|\le\gamma} C_\phi(f,\eta,\mathbf{x}) > \inf_{f\in\mathcal{F}_{\mathrm{lin}}} C_\phi(f,\eta,\mathbf{x}) & \forall\eta\ge\frac{1}{2}\ (\varepsilon\le\eta<\frac{1+\varepsilon}{2}), \\ \inf_{f\in\mathcal{F}_{\mathrm{lin}}:f(\mathbf{x})\le\gamma} C_\phi(f,\eta,\mathbf{x}) > \inf_{f\in\mathcal{F}_{\mathrm{lin}}} C_\phi(f,\eta,\mathbf{x}) & \forall\eta\ge\frac{1}{2}\ (\frac{1+\varepsilon}{2}\le\eta\le 1), \end{cases}$$

for all $\varepsilon > 0$.

We immediately observe that

$$\bigcup_{\varepsilon>0}\left\{\eta\ge\frac{1}{2}\,\middle|\,\varepsilon\le\eta<\frac{1+\varepsilon}{2}\right\} = \left\{\frac{1}{2}\le\eta\le 1\right\}, \text{ and}$$
$$\bigcup_{\varepsilon>0}\left\{\eta\ge\frac{1}{2}\,\middle|\,\frac{1+\varepsilon}{2}\le\eta\le 1\right\} = \left\{\frac{1}{2}<\eta\le 1\right\}.$$

Therefore, we reduce the above conditions to

$$\begin{cases} \inf_{f\in\mathcal{F}_{\mathrm{lin}}:|f(\mathbf{x})|\le\gamma} C_\phi(f,\eta,\mathbf{x}) > \inf_{f\in\mathcal{F}_{\mathrm{lin}}} C_\phi(f,\eta,\mathbf{x}) & \text{if } \frac{1}{2}\le\eta\le 1, \\ \inf_{f\in\mathcal{F}_{\mathrm{lin}}:f(\mathbf{x})\le\gamma} C_\phi(f,\eta,\mathbf{x}) > \inf_{f\in\mathcal{F}_{\mathrm{lin}}} C_\phi(f,\eta,\mathbf{x}) & \text{if } \frac{1}{2}<\eta\le 1. \end{cases}$$

Note that $\inf_{|f(\mathbf{x})|\le\gamma} C_\phi(f,\eta,\mathbf{x}) \ge \inf_{f\in\mathcal{F}_{\mathrm{lin}}:f(\mathbf{x})\le\gamma} C_\phi(f,\eta,\mathbf{x})$ (note that inequality is not strict) always holds for all values of η. Because the first case is included in the second case except when $\eta=\frac{1}{2}$, this is equivalent to

$$\inf_{f\in\mathcal{F}_{\mathrm{lin}}:|f(\mathbf{x})|\le\gamma} C_\phi\left(f,\tfrac{1}{2},\mathbf{x}\right) > \inf_{f\in\mathcal{F}_{\mathrm{lin}}} C_\phi\left(f,\tfrac{1}{2},\mathbf{x}\right), \text{ and}$$
$$\inf_{f\in\mathcal{F}_{\mathrm{lin}}:f(\mathbf{x})\le\gamma} C_\phi(f,\eta,\mathbf{x}) > \inf_{f\in\mathcal{F}_{\mathrm{lin}}} C_\phi(f,\eta,\mathbf{x}) \text{ for } \eta\in\left(\tfrac{1}{2},1\right].$$

□

Part 4. First, we obtain the calibration function $\bar{\delta}$ w.r.t. $(\ell_{01},\mathcal{F}_{\mathrm{lin}})$ following the same strategy as Lemma 4.9. The ℓ_{01}-CCR for $f\in\mathcal{F}_{\mathrm{lin}}$ at $\mathbf{x}$ is

$$\begin{aligned} C_{\ell_{01}}(f,\eta,\mathbf{x}) &= \eta\mathbb{1}_{\{\mathrm{sgn}(f(\mathbf{x}))=+1\}} + (1-\eta)\mathbb{1}_{\{\mathrm{sgn}(f(\mathbf{x}))=-1\}} \\ &= \begin{cases} \eta & \text{if } f(\mathbf{x})\ge 0, \\ 1-\eta & \text{if } f(\mathbf{x})<0. \end{cases} \end{aligned}$$

To compute $\Delta C_{\ell_{01},\mathcal{F}_{\mathrm{lin}}}(f,\eta,\mathbf{x})$, note that given x, $f(\mathbf{x})$ ranges $[-\|\mathbf{x}\|_2,\|\mathbf{x}\|_2]$ for $f\in\mathcal{F}_{\mathrm{lin}}$. When $\|\mathbf{x}\|_2=0$, $C_{\ell_{01}}(f,\eta,\mathbf{x})=\eta$ for all $f\in\mathcal{F}_{\mathrm{lin}}$, which is equivalent to $\Delta C_{\ell_{01},\mathcal{F}_{\mathrm{lin}}}(f,\eta,\mathbf{x})=0$. When $\|\mathbf{x}\|_2>0$, we have $\Delta C_{\ell_{01},\mathcal{F}_{\mathrm{lin}}}(f,\eta,\mathbf{x}) = |2\eta-1|\cdot\mathbb{1}_{\{(2\eta-1)f(\mathbf{x})\le 0\}}$ in the same way as in Steinwart [2007, Lemma 4.1]. Hence,

$$\Delta C_{\ell_{01},\mathcal{F}_{\mathrm{lin}}}(f,\eta,\mathbf{x}) = \begin{cases} 0 & \text{if } \|\mathbf{x}\|_2=0, \\ |2\eta-1|\cdot\mathbb{1}_{\{(2\eta-1)f(\mathbf{x})\le 0\}} & \text{if } \|\mathbf{x}\|_2>0. \end{cases}$$

Now, if $\|\mathbf{x}\|_2=0$ or $\varepsilon>|2\eta-1|$, we always have $\Delta C_{\ell_{01},\mathcal{F}_{\mathrm{lin}}}(f,\eta,\mathbf{x})<\varepsilon$, and hence $\bar{\delta}(\varepsilon,\eta,\mathbf{x})=\infty$. If $\|\mathbf{x}\|_2>0$ and $\varepsilon\le|2\eta-1|$, we have $\Delta C_{\ell_{01},\mathcal{F}_{\mathrm{lin}}}(f,\eta,\mathbf{x})<\varepsilon$ if and only if $(2\eta-1)f(\mathbf{x})>0$. Therefore, we have

$$\bar{\delta}(\varepsilon,\eta,\mathbf{x}) = \begin{cases} \infty & \text{if } \|\mathbf{x}\|_2=0 \text{ or } \varepsilon>|2\eta-1|, \\ \inf_{\substack{f\in\mathcal{F}_{\mathrm{lin}}: \\ (2\eta-1)f(\mathbf{x})\le 0}} \Delta C_{\phi,\mathcal{F}_{\mathrm{lin}}}(f,\eta,\mathbf{x}) & \text{if } \|\mathbf{x}\|_2>0 \text{ and } \varepsilon\le|2\eta-1|. \end{cases}$$

Next, based on Proposition 4.4, ϕ is $(\ell_{01}, \mathcal{F}_{\text{lin}})$-calibrated if and only if $\bar{\delta}(\varepsilon, \eta, \mathbf{x}) > 0$ for all $\varepsilon > 0$, $\eta \in [0, 1]$, and $\mathbf{x} \in \mathcal{X}$. In the same way as in part 3 of Lemma 4.16, this is equivalent to

$$\inf_{f \in \mathcal{F}_{\text{lin}} : f(\mathbf{x}) \leq 0} C_\phi(f, \eta, \mathbf{x}) > \inf_{f \in \mathcal{F}_{\text{lin}}} C_\phi(f, \eta, \mathbf{x}) \quad \text{for all } \eta \geq \tfrac{1}{2} \text{ such that } \tfrac{1+\varepsilon}{2} \leq \eta \leq 1,$$

for all $\varepsilon > 0$ and $\mathbf{x} \in \mathcal{X} \setminus \{\mathbf{0}\}$ when using part 1 of Lemma 4.16 and the symmetry of $\mathcal{F}_{\text{lin}}$. This is equivalent to the lemma statement. □

4.11.6 Proof of Lemma 4.17

Denote the following: $\bar{\phi}(\alpha) \coloneqq \phi(\alpha) + \phi(-\alpha)$.

Part 1. Fix an $\eta \in \left(\frac{1}{2}, 1\right]$ and $\alpha_1, \alpha_2 \geq 0$ such that $\alpha_1 < \alpha_2$. Based on the fact that ϕ is nonincreasing, we have

$$\begin{aligned}\phi(\alpha_1) - \phi(-\alpha_1) - \phi(\alpha_2) + \phi(-\alpha_2) &= (\phi(\alpha_1) - \phi(\alpha_2)) + (\phi(-\alpha_2) - \phi(-\alpha_1)) \\ &\geq 0.\end{aligned}$$

Then,

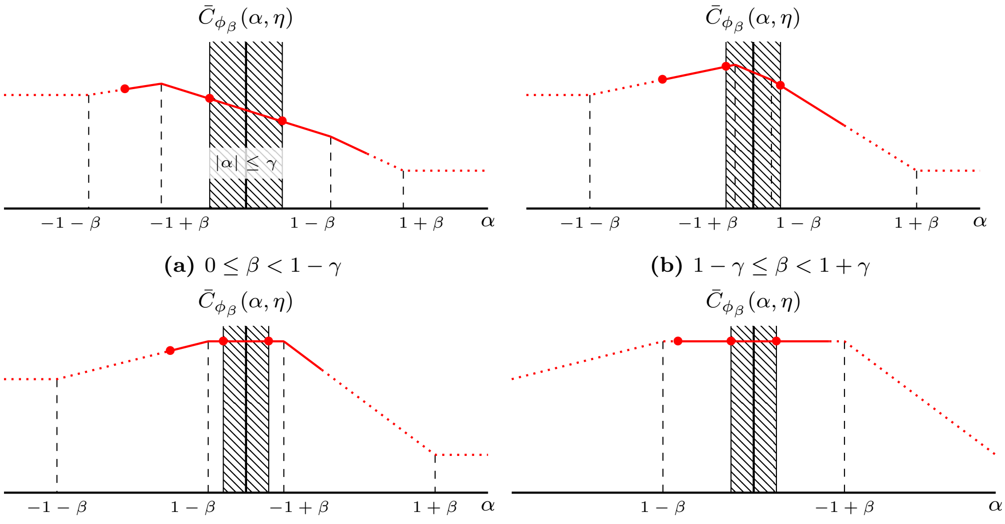
$$\begin{aligned}&\bar{C}_\phi(\alpha_1, \eta) - \bar{C}_\phi(\alpha_2, \eta) \\ &\qquad = (\phi(\alpha_1) - \phi(-\alpha_1) - \phi(\alpha_2) + \phi(-\alpha_2))\eta + \phi(-\alpha_1) - \phi(-\alpha_2) \\ &\qquad \geq (\phi(\alpha_1) - \phi(-\alpha_1) - \phi(\alpha_2) + \phi(-\alpha_2))\frac{1}{2} + \phi(-\alpha_1) - \phi(-\alpha_2) \\ &\qquad = \frac{\phi(\alpha_1) + \phi(-\alpha_1) - \phi(\alpha_2) - \phi(-\alpha_2)}{2} \\ &\qquad \geq 0,\end{aligned}$$

where the last inequality holds because $\phi(\alpha) + \phi(-\alpha)$ is nonincreasing when $\alpha \geq 0$ by part 3. Therefore, $\bar{C}_\phi(\alpha, \eta)$ is nonincreasing in $\alpha \geq 0$.

Part 2. Fix an $\eta \in \left(\frac{1}{2}, 1\right]$. Then,

$$\bar{C}_\phi(-\alpha, \eta) - \bar{C}_\phi(\alpha, \eta) = (2\eta - 1)(\phi(-\alpha) - \phi(\alpha)) > 0.$$

Part 3. Here, $\bar{\phi}$ is an even function, and thus is symmetric in $\alpha = 0$. In addition, $\bar{\phi}$ is continuous because of the continuity of ϕ. Every quasiconcave continuous function is nondecreasing, or nonincreasing, or there is global maxima in its domain [Boyd and Vandenberghe, 2004]. If $\bar{\phi}$ is either nondecreasing or nonincreasing in α, it is a constant function in α and clearly nonincreasing in $\alpha \geq 0$. If $\bar{\phi}$ has the global maxima, i.e., there is a point $\alpha_* \in \operatorname{dom}(\bar{\phi})$ such that $\bar{\phi}$ is nondecreasing for $\alpha \leq \alpha_*$ and nonincreasing for $\alpha \geq \alpha_*$, it is still nonincreasing in $\alpha \geq 0$. This is clear when $\alpha_* \leq 0$. When $\alpha_* > 0$, $\bar{\phi}$ may only be a constant function in $\alpha \in [0, \alpha_*]$; otherwise, we have a point $\widetilde{\alpha} \in [0, \alpha_*)$ such that $\bar{\phi}(\widetilde{\alpha}) < \bar{\phi}(\alpha_*)$, and hence $\bar{\phi}(\alpha_*) = \bar{\phi}(-\alpha_*)$ $(\coloneqq \bar{\phi}_*)$ by the symmetry and $\bar{\phi}_0 \coloneqq \bar{\phi}(\widetilde{\alpha}) < \bar{\phi}_*$, which means there is no convex superlevel sets for $\bar{\phi}$ within the range $(\bar{\phi}_0, \bar{\phi}_*)$. For example, choose $t \in (\bar{\phi}_0, \bar{\phi}_*)$ and consider the t-superlevel set of $\bar{\phi}$. If the t-superlevel set is convex, it must contain every point in $[-\alpha_*, \alpha_*]$ because $t < \bar{\phi}_* = \bar{\phi}(-\alpha_*) = \bar{\phi}(\alpha_*)$. However, the t-superlevel set will not contain $\widetilde{\alpha} \in [-\alpha_*, \alpha_*]$ because $t > \bar{\phi}_0 = \bar{\phi}(\widetilde{\alpha})$. This contradicts the quasiconcavity of $\bar{\phi}$. In any case, $\bar{\phi}$ is nonincreasing in $\alpha \geq 0$.



(a) $0 \le \beta < 1-\gamma$ **(b)** $1-\gamma \le \beta < 1+\gamma$

(c) $1+\gamma \le \beta < 2$ **(d)** $2 \le \beta$

Figure 4.14: Class-conditional risk for the ramp loss.

Part 4. This is an immediate consequence of the quasiconcavity and continuity of $\bar{C}_\phi(\alpha,\eta)$ (Assumption 4.11). □

4.11.7 Derivation of Calibration Functions

In this subsection, we derive closed-forms of $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-calibration functions for several surrogate losses ϕ by minimizing $\bar{\delta}(\varepsilon,\eta,\mathbf{x})$ in (4.7) w.r.t. $\eta \in [0,1]$ and $\mathbf{x} \in \widetilde{\mathcal{X}}_\rho$, or in other words, $\|\mathbf{x}\|_2 \ge \gamma + \rho$. To simplify the notation, let $\bar{C}_\phi(\alpha,\eta) := \eta\phi(\alpha) + (1-\eta)\phi(-\alpha)$.

Ramp Loss

The ramp loss is $\phi(\alpha) = \min\left\{ 1, \max\left\{ 0, \frac{1-\alpha}{2} \right\} \right\}$. We consider the shifted ramp loss $\phi_\beta(\alpha) = \phi(\alpha - \beta)$ as follows:

$$\phi_\beta(\alpha) = \begin{cases} 1 & \text{if } \alpha \le -1+\beta, \\ \frac{1-\alpha+\beta}{2} & \text{if } -1+\beta < \alpha \le 1+\beta, \\ 0 & \text{if } 1+\beta < \alpha. \end{cases}$$

The ϕ_β-CCR is plotted in Figure 4.14. We can confirm that $\bar{C}_{\phi_\beta}$ is quasiconcave with each $\beta \ge 0$.

Minimal inner risk. By part 4 of Lemma 4.17, it is easy to check

$$\begin{aligned} C^*_{\phi_\beta,\mathcal{F}_{\text{lin}}}(\eta,\mathbf{x}) &= \inf_{f\in\mathcal{F}_{\text{lin}}} C_{\phi_\beta}(f,\eta,\mathbf{x}) = \inf_{\alpha\in[-\|\mathbf{x}\|_2,\|\mathbf{x}\|_2]} \bar{C}_{\phi_\beta}(\alpha,\eta) \\ &= \min\left\{ \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta), \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta) \right\}. \end{aligned}$$

Calibration function. We analyze ϕ_β-CCR $C_{\phi_\beta}(f,\eta,\mathbf{x}) = \eta\phi_\beta(f(\mathbf{x})) + (1-\eta)\phi_\beta(-f(\mathbf{x})) = \bar{C}_{\phi_\beta}(f(\mathbf{x}),\eta)$, and restrict $\eta > \frac{1}{2}$ by virtue of the symmetry of C_{ϕ_β} (part 1 in Lemma 4.16). It is easy to see that $C^*_{\phi_\beta,\mathcal{F}_{\text{lin}}}(\eta,\mathbf{x}) = \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta)$. Subsequently, we divide the cases depending on the relationship among $\bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta)$, $\bar{C}_{\phi_\beta}(\gamma,\eta)$, and $\bar{C}_{\phi_\beta}(-\gamma,\eta)$.

(A) When $0 \le \beta < 1-\gamma$,

$$\begin{aligned}
\bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta) &= \begin{cases} \eta + \frac{1-\|\mathbf{x}\|_2+\beta}{2}(1-\eta) & \text{if } \gamma+\rho \le \|\mathbf{x}\|_2 < 1-\beta, \\ \frac{1+\|\mathbf{x}\|_2+\beta}{2}\eta + \frac{1-\|\mathbf{x}\|_2+\beta}{2}(1-\eta) & \text{otherwise,} \end{cases} \\
\bar{C}_{\phi_\beta}(\gamma,\eta) &= \frac{1-\gamma+\beta}{2}\eta + \frac{1+\gamma+\beta}{2}(1-\eta), \\
\bar{C}_{\phi_\beta}(-\gamma,\eta) &= \frac{1+\gamma+\beta}{2}\eta + \frac{1-\gamma+\beta}{2}(1-\eta),
\end{aligned}$$

from which it follows that $\bar{C}_{\phi_\beta}(-\gamma,\eta) - \bar{C}_{\phi_\beta}(\gamma,\eta) = \frac{\gamma}{2}(2\eta-1) > 0$, that is, $\bar{C}_{\phi_\beta}(-\gamma,\eta) > \bar{C}_{\phi_\beta}(\gamma,\eta)$ for all $\eta > \frac{1}{2}$. In addition, because when $\gamma+\rho \le \|\mathbf{x}\|_2 < 1-\beta$,

$$\bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta) - \bar{C}_{\phi_\beta}(\gamma,\eta) = (\gamma+\|\mathbf{x}\|_2)\left(\eta - \frac{1}{2}\right),$$

we have $C_{\phi_\beta}(\gamma,\eta) < C_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta)$ for β and $\|\mathbf{x}\|_2$. Based on part 4 in Lemma 4.17, it follows that

$$\begin{aligned}
\inf_{f\in\mathcal{F}_{\text{lin}}:|f(\mathbf{x})|\le\gamma} C_{\phi_\beta}(f,\eta,\mathbf{x}) &= \inf_{|\alpha|\le\gamma} \bar{C}_{\phi_\beta}(\alpha,\eta) = \bar{C}_{\phi_\beta}(\gamma,\eta), \\
\inf_{f\in\mathcal{F}_{\text{lin}}:f(\mathbf{x})\le\gamma} C_{\phi_\beta}(f,\eta,\mathbf{x}) &= \inf_{\alpha\in[-\|\mathbf{x}\|_2,\gamma]} \bar{C}_{\phi_\beta}(\alpha,\eta) \\
&= \min\{\bar{C}_{\phi_\beta}(\gamma,\eta), \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta)\}.
\end{aligned}$$

Thus, based on Lemma 4.9,

$$\begin{aligned}
&\bar{\delta}(\varepsilon,\eta,\mathbf{x}) \\
&= \begin{cases} \infty & \text{if } \eta < \varepsilon, \\ \bar{C}_{\phi_\beta}(\gamma,\eta) - \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta) & \text{if } \varepsilon \le \eta \text{ and } \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta) \ge \bar{C}_{\phi_\beta}(\gamma,\eta), \\ \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta) - \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta) & \text{if } \varepsilon \le \eta \text{ and } \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta) < \bar{C}_{\phi_\beta}(\gamma,\eta), \end{cases} \\
&= \begin{cases} \infty & \text{if } \eta < \varepsilon, \\ \bar{C}_{\phi_\beta}(\gamma,\eta) - \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta) & \text{if } \varepsilon \le \eta \text{ and } \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta) \ge \bar{C}_{\phi_\beta}(\gamma,\eta), \\ \frac{1+\|\mathbf{x}\|_2-\beta}{2}\left(\eta-\frac{1}{2}\right) & \text{if } \varepsilon \le \eta \text{ and } \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta) < \bar{C}_{\phi_\beta}(\gamma,\eta), \end{cases}
\end{aligned}$$

where the last identity holds because $\|\mathbf{x}\|_2 \ge 1-\beta$ when $\bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta) < \bar{C}_{\phi_\beta}(\gamma,\eta)$. Because $\bar{C}_{\phi_\beta}(\alpha,\eta)$ is nonincreasing in $\alpha \ge 0$ (part 1 of Lemma 4.17), we know that $C^*_{\phi_\beta}(\eta,\mathbf{x}) = \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta)$ is maximized at $\|\mathbf{x}\|_2 = \gamma+\rho$, which implies

$$\begin{aligned}
\delta_\rho(\varepsilon) &= \inf_{\eta\in\left(\frac{1}{2},1\right]} \inf_{x\in\mathcal{X}:\|\mathbf{x}\|_2\ge\gamma+\rho} \bar{\delta}(\varepsilon,\eta,\mathbf{x}) \\
&= \inf_{\eta\in\left(\frac{1}{2},1\right]:\varepsilon\le\eta} \frac{1+\gamma+\rho-\beta}{2}\left(\eta-\frac{1}{2}\right) \\
&= \frac{1+\gamma+\rho-\beta}{2}\left[\varepsilon-\frac{1}{2}\right]_+.
\end{aligned}$$

(B) When $1-\gamma \le \beta < 1+\gamma$,

$$\begin{aligned}
\bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta) &= \eta + \frac{1-\|\mathbf{x}\|_2+\beta}{2}(1-\eta), \\
\bar{C}_{\phi_\beta}(\gamma,\eta) &= \frac{1-\gamma+\beta}{2}\eta + (1-\eta), \\
\bar{C}_{\phi_\beta}(-\gamma,\eta) &= \eta + \frac{1-\gamma+\beta}{2}(1-\eta),
\end{aligned}$$

from which it follows that $\bar{C}_{\phi_\beta}(-\gamma,\eta) - \bar{C}_{\phi_\beta}(\gamma,\eta) = \frac{1+\gamma-\beta}{2}(2\eta-1) > 0$, that is, $\bar{C}_{\phi_\beta}(-\gamma,\eta) > \bar{C}_{\phi_\beta}(\gamma,\eta)$ for all $\eta > \frac{1}{2}$. In addition, because

$$\bar{C}_{\phi_\beta}(\gamma,\eta) - \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta) = -\frac{2-2\beta+\gamma+\|\mathbf{x}\|_2}{2}\left(\eta - \eta_0(\mathbf{x})\right),$$

where

$$\eta_0(\mathbf{x}) := \frac{1}{1+\frac{1+\gamma-\beta}{1+\|\mathbf{x}\|_2-\beta}},$$

and $2-2\beta+\gamma+\|\mathbf{x}\|_2 > 0$, we have $\bar{C}_{\phi_\beta}(\gamma,\eta) > \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta)$ if $\frac{1}{2} < \eta < \eta_0(\mathbf{x})$, and $\bar{C}_{\phi_\beta}(\gamma,\eta) \le \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta)$ if $\eta_0(\mathbf{x}) \le \eta$. Note that $\frac{1}{2} < \eta_0(\mathbf{x}) < 1$.

- If $\frac{1}{2} < \eta < \eta_0(\mathbf{x})$: By part 4 in Lemma 4.17, it follows that

  $$\inf_{f\in\mathcal{F}_{\mathrm{lin}}:|f(\mathbf{x})|\le\gamma} C_{\phi_\beta}(f,\eta,\mathbf{x}) = \inf_{|\alpha|\le\gamma} \bar{C}_{\phi_\beta}(\alpha,x) = \bar{C}_{\phi_\beta}(\gamma,\eta),$$
  $$\inf_{f\in\mathcal{F}_{\mathrm{lin}}:f(\mathbf{x})\le\gamma} C_{\phi_\beta}(f,\eta,\mathbf{x}) = \inf_{\alpha\in[-\|\mathbf{x}\|_2,\gamma]} \bar{C}_{\phi_\beta}(\alpha,x) = \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta).$$

  Thus, by Lemma 4.9,

  $$\bar{\delta}(\varepsilon,\eta,\mathbf{x}) = \begin{cases} \infty & \text{if } \eta < \varepsilon, \\ \bar{C}_{\phi_\beta}(\gamma,\eta) - \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta) = \frac{\|\mathbf{x}\|_2-\gamma}{2}\eta & \text{if } \varepsilon \le \eta < \frac{1+\varepsilon}{2}, \\ \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta) - \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta) & \\ \quad = (1+\|\mathbf{x}\|_2-\beta)\left(\eta-\frac{1}{2}\right) & \text{if } \frac{1+\varepsilon}{2} \le \eta, \end{cases}$$

  and

  $$\begin{aligned}\inf_{\eta\in\left(\frac{1}{2},\eta_0(\mathbf{x})\right)} \bar{\delta}(\varepsilon,\eta,\mathbf{x}) &= \min\left\{\frac{\|\mathbf{x}\|_2-\gamma}{2}\max\left\{\varepsilon,\tfrac{1}{2}\right\}, (1+\|\mathbf{x}\|_2-\beta)\varepsilon\right\} \\ &= \begin{cases} (1+\|\mathbf{x}\|_2-\beta)\varepsilon & \text{if } 0 < \varepsilon \le \varepsilon_0(\mathbf{x}), \\ \frac{\|\mathbf{x}\|_2-\gamma}{4} & \text{if } \varepsilon_0(\mathbf{x}) < \varepsilon \le \frac{1}{2}, \\ \frac{\|\mathbf{x}\|_2-\gamma}{2}\varepsilon & \text{if } \frac{1}{2} < \varepsilon, \end{cases}\end{aligned}$$

  where $\varepsilon_0(\mathbf{x}) := \frac{\|\mathbf{x}\|_2-\gamma}{4(1+\|\mathbf{x}\|_2-\beta)}$. Finally, by taking the infimum over $\mathbf{x} \in \widetilde{\mathcal{X}}_\rho$,

  $$\inf_{x\in\widetilde{\mathcal{X}}_\rho}\inf_{\eta\in\left(\frac{1}{2},\eta_0(\mathbf{x})\right)} \bar{\delta}(\varepsilon,\eta,\mathbf{x}) = \begin{cases} (1+\gamma+\rho-\beta)\varepsilon & \text{if } 0 < \varepsilon \le \varepsilon_0, \\ \frac{\rho}{4} & \text{if } \varepsilon_0 < \varepsilon \le \frac{1}{2}, \\ \frac{\rho}{2}\varepsilon & \text{if } \frac{1}{2} < \varepsilon, \end{cases}$$

  where $\varepsilon_0 := \frac{\rho}{4(1+\gamma+\rho-\beta)}$.

- If $\eta_0(\mathbf{x}) \le \eta \le 1$, by part 4 in Lemma 4.17, it follows that

  $$\inf_{f\in\mathcal{F}_{\mathrm{lin}}:|f(\mathbf{x})|\le\gamma} C_{\phi_\beta}(f,\eta,\mathbf{x}) = \inf_{|\alpha|\le\gamma} \bar{C}_{\phi_\beta}(\alpha,\eta) = \bar{C}_{\phi_\beta}(\gamma,\eta),$$
  $$\inf_{f\in\mathcal{F}_{\mathrm{lin}}:f(\mathbf{x})\le\gamma} C_{\phi_\beta}(f,\eta,\mathbf{x}) = \inf_{\alpha\in[-\|\mathbf{x}\|_2,\gamma]} \bar{C}_{\phi_\beta}(\alpha,\eta) = \bar{C}_{\phi_\beta}(\gamma,\eta).$$

  Thus,

  $$\bar{\delta}(\varepsilon,\eta,\mathbf{x}) = \begin{cases} \infty & \text{if } \eta < \varepsilon, \\ \bar{C}_{\phi_\beta}(\gamma,\eta) - \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta) = \frac{\|\mathbf{x}\|_2-\gamma}{2}\eta & \text{if } \varepsilon \le \eta, \end{cases}$$

and

$$\inf_{\eta\in[\eta_0(\mathbf{x}),1]} \bar{\delta}(\varepsilon,\eta,\mathbf{x}) = \frac{\|\mathbf{x}\|_2-\gamma}{2}\max\{\,\eta_0(\mathbf{x}),\varepsilon\,\}\,.$$

By taking the infimum over $\mathbf{x}\in\widetilde{\mathcal{X}}_\rho$,

$$\inf_{x\in\widetilde{\mathcal{X}}_\rho}\inf_{\eta\in[\eta_0(\mathbf{x}),1]} \bar{\delta}(\varepsilon,\eta,\mathbf{x}) = \begin{cases} \frac{\rho}{2}\varepsilon_1 & \text{if } 0<\varepsilon\le\varepsilon_1,\\ \frac{\rho}{2}\varepsilon & \text{if } \varepsilon_1<\varepsilon,\end{cases}$$

where $\varepsilon_1 := \eta_0(\mathbf{x}_{\gamma+\rho}) = \frac{1+\gamma+\rho-\beta}{2+2\gamma-2\beta+\rho}$ with $\|\mathbf{x}_{\gamma+\rho}\|_2=\gamma+\rho$.

Finally, we obtain the calibration function by combining the above cases as follows:

$$\delta_\rho(\varepsilon) = \begin{cases} (1+\gamma+\rho-\beta)\varepsilon & \text{if } 0<\varepsilon\le\varepsilon_0,\\ \frac{\rho}{4} & \text{if } \varepsilon_0<\varepsilon\le\frac{1}{2},\\ \frac{\rho}{2}\varepsilon & \text{if } \frac{1}{2}<\varepsilon.\end{cases}$$

(C) When $1+\gamma\le\beta<2$, it is easy to see that

$$\begin{aligned}
\inf_{f\in\mathcal{F}_{\text{lin}}:|f(\mathbf{x})|\le\gamma} C_{\phi_\beta}(f,\eta,\mathbf{x}) &= \inf_{|\alpha|\le\gamma}\bar{C}_{\phi_\beta}(\alpha,\eta)=1,\\
\inf_{f\in\mathcal{F}_{\text{lin}}:f(\mathbf{x})\le\gamma} C_{\phi_\beta}(\alpha,\eta) &= \inf_{\alpha\in[-\|\mathbf{x}\|_2,\gamma]}\bar{C}_{\phi_\beta}(\alpha,\eta)=\bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta)\\
&= \begin{cases} 1 & \text{if } \gamma<\|\mathbf{x}\|_2\le -1+\beta,\\ \eta+\frac{1-\|\mathbf{x}\|_2+\beta}{2}(1-\eta) & \text{if } -1+\beta<\|\mathbf{x}\|_2\le 1,\end{cases}\\
C^*_{\phi_\beta,\mathcal{F}_{\text{lin}}}(\eta,\mathbf{x}) &= \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta)\\
&= \begin{cases} 1 & \text{if } \gamma<\|\mathbf{x}\|_2\le -1+\beta,\\ \frac{1-\|\mathbf{x}\|_2+\beta}{2}\eta+(1-\eta) & \text{if } -1+\beta<\|\mathbf{x}\|_2\le 1.\end{cases}
\end{aligned}$$

Hence, by part 4 in Lemma 4.17, it follows that

$$\begin{aligned}
\bar{\delta}(\varepsilon,\eta,\mathbf{x}) &= \begin{cases} \infty & \text{if } \eta<\varepsilon,\\ 1-\bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta) & \text{if } \varepsilon\le\eta<\frac{1+\varepsilon}{2},\\ \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta)-\bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta) & \text{if } \frac{1+\varepsilon}{2}\le\eta.\end{cases}\\
&= \begin{cases} \infty & \text{if } \eta<\varepsilon,\\ 0 & \text{if } \varepsilon\le\eta<\frac{1+\varepsilon}{2} \text{ and } \gamma<\|\mathbf{x}\|_2\le-1+\beta,\\ \frac{1+\|\mathbf{x}\|_2-\beta}{2}\eta & \text{if } \varepsilon\le\eta<\frac{1+\varepsilon}{2} \text{ and } -1+\beta<\|\mathbf{x}\|_2\le 1,\\ 0 & \text{if } \frac{1+\varepsilon}{2}\le\eta \text{ and } \gamma<\|\mathbf{x}\|_2\le-1+\beta,\\ (1+\|\mathbf{x}\|_2-\beta)\left(\eta-\frac{1}{2}\right) & \text{if } \frac{1+\varepsilon}{2}\le\eta \text{ and } -1+\beta<\|\mathbf{x}\|_2\le 1.\end{cases}
\end{aligned}$$

Thus, Lemma 4.9 implies $\delta_\rho(\varepsilon)=\inf_{\eta\in(\frac{1}{2},1]}\inf_{\|\mathbf{x}\|_2\in[\gamma+\rho,1]}\bar{\delta}(\varepsilon,\eta,\mathbf{x})=0$ when $\gamma+\rho\le-1+\beta \iff 1+\gamma+\rho\le\beta$, by setting $\|\mathbf{x}\|_2\le-1+\beta$ and arbitrary η. When $1+\gamma+\rho>\beta$, $\delta_\rho(\varepsilon)=\frac{1+\gamma+\rho-\beta}{2}\varepsilon$ by setting $\|\mathbf{x}\|_2=\gamma+\rho$.

(D) When $2\le\beta$, $\bar{C}_{\phi_\beta}(\alpha,\eta)=1$ for all $\eta\in[0,1]$ and $\alpha\in[-1,1]$. Hence, $\Delta C_{\phi_\beta,\mathcal{F}_{\text{lin}}}(f,\eta,\mathbf{x})=0$ and $\delta_\rho(\varepsilon)=0$.

To summarize, the $(\phi_\gamma,\mathcal{F}_{\text{lin}})$-calibration function and its Fenchel-Legendre biconjugate of the ramp loss is as follows:

- If $0\le\beta<1-\gamma$, $\delta_\rho(\varepsilon)=\delta_\rho^{\star\star}(\varepsilon)=\frac{1+\gamma+\rho-\beta}{2}\left[\varepsilon-\frac{1}{2}\right]_+$.

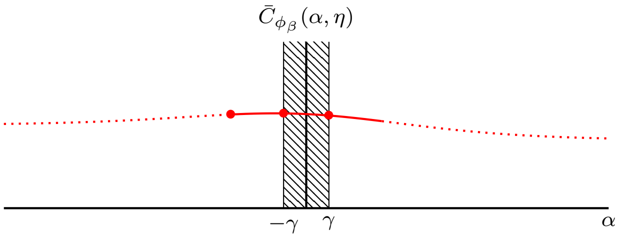


Figure 4.15: Class-conditional risk of the sigmoid loss.

- If $1-\gamma \le \beta < 1+\gamma$,

$$\delta_\rho(\varepsilon) = \begin{cases} (1+\gamma+\rho-\beta)\varepsilon & \text{if } 0 < \varepsilon \le \varepsilon_0, \\ \frac{\rho}{4} & \text{if } \varepsilon_0 < \varepsilon \le \frac{1}{2}, \\ \frac{\rho}{2}\varepsilon & \text{if } \frac{1}{2} < \varepsilon, \end{cases} \quad \text{and} \quad \delta_\rho^{\star\star}(\varepsilon) = \frac{\rho}{2}\varepsilon.$$

- If $1+\gamma \le \beta < 1+\gamma+\rho$, $\delta_\rho(\varepsilon) = \delta_\rho^{\star\star}(\varepsilon) = \frac{1+\gamma+\rho-\beta}{2}\varepsilon$.
- If $1+\gamma+\rho \le \beta$, $\delta_\rho(\varepsilon) = \delta_\rho^{\star\star}(\varepsilon) = 0$.

We can see that the ramp loss is $(\phi_\rho, \mathcal{F}_{\text{lin}})$-calibrated when $1-\gamma \le \beta < 1+\gamma$.

Sigmoid Loss

The sigmoid loss is $\phi(\alpha) = \frac{1}{1+e^\alpha}$. We consider the shifted sigmoid loss as $\phi_\beta(\alpha) = \frac{1}{1+e^{\alpha-\beta}}$ for $\beta \ge 0$. Here, ϕ_β-CCR is

$$\begin{aligned} C_{\phi_\beta}(f,\eta,\mathbf{x}) &= \bar{C}_{\phi_\beta}(f(\mathbf{x}),\eta) \\ &= \frac{\eta}{1+e^{f(\mathbf{x})-\beta}} + \frac{1-\eta}{1+e^{-f(\mathbf{x})-\beta}}. \end{aligned}$$

In addition, $\bar{C}_{\phi_\beta}$ is plotted in Figure 4.15, from which we can see that $\bar{C}_{\phi_\beta}$ is quasiconcave when $\beta \ge 0$.

Minimal inner risk. Based on part 4 of Lemma 4.17, it is easy to check that

$$\begin{aligned} & C^*_{\phi_\beta,\mathcal{F}_{\text{lin}}}(\eta,\mathbf{x}) \\ &= \inf_{f\in\mathcal{F}_{\text{lin}}} C_{\phi_\beta}(f,\eta,\mathbf{x}) \\ &= \inf_{\alpha\in[-\|\mathbf{x}\|_2,\|\mathbf{x}\|_2]} \bar{C}_{\phi_\beta}(\alpha,\eta) \\ &= \min\{\bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta), \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta)\} \\ &= \min\left\{\frac{\eta}{1+e^{\|\mathbf{x}\|_2-\beta}} + \frac{1-\eta}{1+e^{-\|\mathbf{x}\|_2-\beta}}, \frac{\eta}{1+e^{-\|\mathbf{x}\|_2-\beta}} + \frac{1-\eta}{1+e^{\|\mathbf{x}\|_2-\beta}}\right\}. \end{aligned}$$

Calibration function. We focus on the case $\eta > \frac{1}{2}$ owing to the symmetry of C_{ϕ_β}. The minimal inner risk is

$$C^*_{\phi_\beta,\mathcal{F}_{\text{lin}}}(\eta,\mathbf{x}) = \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta) = \frac{\eta}{1+e^{\|\mathbf{x}\|_2-\beta}} + \frac{1-\eta}{1+e^{-\|\mathbf{x}\|_2-\beta}}.$$

We compute $C_{\phi_\beta}(f, \eta, \mathbf{x})$. Because

$$\begin{aligned}\bar{C}_{\phi_\beta}(-\gamma, \eta) - \bar{C}_{\phi_\beta}(\gamma, \eta) &= \left(\frac{\eta}{1+e^{-\gamma-\beta}} + \frac{1-\eta}{1+e^{\gamma-\beta}}\right) - \left(\frac{\eta}{1+e^{\gamma-\beta}} + \frac{1-\eta}{1+e^{-\gamma-\beta}}\right) \\ &= (2\eta - 1)\left(\frac{1}{1+e^{-\gamma-\beta}} - \frac{1}{1+e^{\gamma-\beta}}\right) \\ &> 0, \qquad (\text{since } -\gamma-\beta < \gamma - \beta)\end{aligned}$$

we have $\bar{C}_{\phi_\beta}(\gamma, \eta) < \bar{C}_{\phi_\beta}(-\gamma, \eta)$ for all $\eta > \frac{1}{2}$, implying that $\inf_{|\alpha| \le \gamma} \bar{C}_{\phi_\beta}(\alpha, \eta) = \bar{C}_{\phi_\beta}(\gamma, \eta)$ and $\inf_{f:|f(\mathbf{x})| \le \gamma} C_{\phi_\beta}(f, \eta, \mathbf{x}) = \bar{C}_{\phi_\beta}(\gamma, \eta)$. Note that because we assume $\|x\|_2 > \gamma$, $f \in \mathcal{F}_{\text{lin}}$ exists such that $f(\mathbf{x}) = \gamma$. By contrast, we divide the cases to compute $\inf_{f:f(\mathbf{x}) \le \gamma} C_{\phi_\beta}(f, \eta, \mathbf{x}) = \inf_{\alpha \in [-\|\mathbf{x}\|_2, \gamma]} \bar{C}_{\phi_\beta}(\alpha, \eta)$. Based on part 4 of Lemma 4.17,

$$\inf_{f \in \mathcal{F}_{\text{lin}}: f(\mathbf{x}) \le \gamma} C_{\phi_\beta}(f, \eta, \mathbf{x}) = \begin{cases} \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2, \eta) & \text{if } \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2) < \bar{C}_{\phi_\beta}(\gamma, \eta), \\ \bar{C}_{\phi_\beta}(\gamma, \eta) & \text{if } \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2) \ge \bar{C}_{\phi_\beta}(\gamma, \eta). \end{cases}$$

Thus, based on Lemma 4.9, we can compute δ by evaluating $\bar{\delta}$ and dividing the cases regarding η and x. If $\bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2) < \bar{C}_{\phi_\beta}(\gamma, \eta)$ and $\eta \ge \frac{1+\varepsilon}{2}$,

$$\begin{aligned}\bar{\delta}(\varepsilon, \eta, \mathbf{x}) &= \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2, \eta) - \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2, \eta) \\ &= (2\eta - 1)\left\{\phi_\beta(-\|\mathbf{x}\|_2) - \phi_\beta(\|\mathbf{x}\|_2)\right\},\end{aligned}$$

which is minimized at $\eta = \frac{1+\varepsilon}{2}$ and $\|\mathbf{x}\|_2 = \gamma + \rho$ because $\phi_\beta(-\|\mathbf{x}\|_2) - \phi_\beta(\|\mathbf{x}\|_2) > 0$ is increasing in $\|\mathbf{x}\|_2$, and the constraint

$$\begin{aligned}&\left\{\bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2) < \bar{C}_{\phi_\beta}(\gamma, \eta)\right\} \wedge \left\{\eta \ge \frac{1+\varepsilon}{2}\right\} \\ \iff\ & F(\|\mathbf{x}\|_2) := \frac{\phi_\beta(-\gamma) - \phi_\beta(\|\mathbf{x}\|_2)}{\phi_\beta(-\|\mathbf{x}\|_2) - \phi_\beta(\gamma)} < \frac{1+\varepsilon}{1-\varepsilon}\end{aligned}$$

is always satisfied for any $\varepsilon > 0$ and x such that $\|\mathbf{x}\|_2 > \gamma$. Note that $F(\|\mathbf{x}\|_2)$ is increasing in $\|\mathbf{x}\|_2$ and thereby maximized at $\|\mathbf{x}\|_2 = 1$, where $F(1) < 1 < \frac{1+\varepsilon}{1-\varepsilon}$. Under the choice of the minimizers,

$$\bar{\delta}\left(\varepsilon, \frac{1+\varepsilon}{2}, \mathbf{x}\right) = \varepsilon\left\{\phi_\beta(-\gamma - \rho) - \phi_\beta(\gamma + \rho)\right\} := A_0 \varepsilon,$$

where $A_0 := \phi_\beta(-\gamma - \rho) - \phi_\beta(\gamma + \rho)$.

If $\bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2) \ge \bar{C}_{\phi_\beta}(\gamma, \eta)$ and $\varepsilon \le \eta$, we have $F(\|\mathbf{x}\|_2) \ge \frac{\varepsilon}{2-\varepsilon}$. This constraint with $\|\mathbf{x}\|_2 \ge \gamma + \rho$ is always satisfied because F is increasing and $F(\|\mathbf{x}\|_2) > F(\gamma) = 1 \ge \frac{\varepsilon}{2-\varepsilon}$ for all $\varepsilon \in (0, 1)$. Consequently,

$$\begin{aligned}&\bar{\delta}(\varepsilon, \eta, \mathbf{x}) \\ &= \bar{C}_{\phi_\beta}(\gamma, \eta) - \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2, \eta) \\ &= \left\{\phi_\beta(\gamma) - \phi_\beta(-\gamma) - \phi_\beta(\|\mathbf{x}\|_2) + \phi_\beta(-\|\mathbf{x}\|_2)\right\}\eta + \phi_\beta(-\gamma) - \phi_\beta(-\|\mathbf{x}\|_2),\end{aligned}$$

which is minimized at $\eta = \max\left\{\varepsilon, \frac{1}{2}\right\}$ and $\|\mathbf{x}\|_2 = \gamma + \rho$ because it is nondecreasing in both η and x. Note that $-\bar{C}_{\phi_\beta}(\cdot, \eta)$ is nondecreasing (part 1 of Lemma 4.17). Under the choice of the minimizers,

$$\bar{\delta}\left(\varepsilon, \max\left\{\varepsilon, \tfrac{1}{2}\right\}, \gamma + \rho\right) = A_1\left[\varepsilon - \frac{1}{2}\right]_+ + \delta_0,$$

where

$$A_1 \coloneqq \phi_\beta(\gamma) - \phi_\beta(-\gamma) - \phi_\beta(\gamma+\rho) + \phi_\beta(-\gamma-\rho),$$
$$\delta_0 \coloneqq \frac{\phi_\beta(\gamma) + \phi_\beta(-\gamma) - \phi_\beta(\gamma+\rho) - \phi_\beta(-\gamma-\rho)}{2}.$$

Note that $\delta_0 > 0$ because $\phi_\beta(\alpha) + \phi_\beta(-\alpha)$ is nonincreasing in $\alpha \geq 0$ (part 3 of Lemma 4.17).

If $\bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2) < \bar{C}_{\phi_\beta}(\gamma, \eta)$ and $\varepsilon \leq \eta < \frac{1+\varepsilon}{2}$, the condition is equivalent to $\frac{1+\varepsilon}{1-\varepsilon} < F(\|\mathbf{x}\|_2) \leq \frac{\varepsilon}{2-\varepsilon}$. This condition is never satisfied because $\frac{1+\varepsilon}{1-\varepsilon} \geq \frac{\varepsilon}{2-\varepsilon}$ for all $\varepsilon \in (0,1)$.

By combining these cases, we have

$$\delta_\rho(\varepsilon) = \inf_{\eta\in\left[\frac{1}{2},1\right]} \inf_{\|\mathbf{x}\|_2>\gamma+\rho} \bar{\delta}(\varepsilon,\eta,\mathbf{x}) = \begin{cases} A_0\varepsilon & \text{if } 0 < \varepsilon \leq \varepsilon_0, \\ \delta_0 & \text{if } \varepsilon_0 < \varepsilon \leq \frac{1}{2}, \\ A_1\left(\varepsilon - \frac{1}{2}\right) + \delta_0 & \text{if } \varepsilon > \frac{1}{2}, \end{cases}$$
$$\delta_\rho^{\star\star}(\varepsilon) = \begin{cases} A_0\varepsilon & \text{if } 0 < \varepsilon \leq \frac{1}{2}, \\ A_1\left(\varepsilon - \frac{1}{2}\right) + \delta_0 & \text{if } \frac{1}{2} < \varepsilon, \end{cases}$$

where $\varepsilon_0 \coloneqq \frac{\delta_0}{A_0}$.

Thus, the sigmoid loss is calibrated w.r.t. $(\phi_\gamma, \mathcal{F}_{\text{lin}})$ when $\delta_0 > 0$. This always holds as long as $\beta > 0$.

Modified Squared Loss

We design a bounded and nonincreasing surrogate loss by modifying the squared loss, which we call the modified squared loss herein:

$$\phi(\alpha) = \begin{cases} 1 & \text{if } \alpha \leq 0, \\ (1-\alpha)^2 & \text{if } 0 < \alpha \leq 1, \\ 0 & \text{if } 1 < \alpha, \end{cases}$$

and consider the shifted version $\phi_\beta(\alpha) \coloneqq \phi(\alpha - \beta)$:

$$\phi_\beta(\alpha) = \begin{cases} 1 & \text{if } \alpha \leq \beta, \\ (1-\alpha+\beta)^2 & \text{if } \beta < \alpha \leq 1+\beta, \\ 0 & \text{if } 1+\beta < \alpha. \end{cases}$$

Here, $\bar{C}_{\phi_\beta}$ is plotted in Figure 4.16, from which we can see $\bar{C}_{\phi_\beta}$ is quasiconcave when $\beta \geq 0$.

Calibration function. Now, we consider ϕ_β-CCR $C_{\phi_\beta}(f,\eta,\mathbf{x}) = \bar{C}_{\phi_\beta}(\alpha,\eta) = \eta\phi(\alpha) + (1-\eta)\phi(-\alpha)$, where $\alpha = f(\mathbf{x})$, and focus on the case $\eta > \frac{1}{2}$ owing to the symmetry of $\bar{C}_{\phi_\beta}$ in η (part 1 of Lemma 4.16). Based on part 4 of Lemma 4.17, it is easy to see that

$$C^*_{\phi_\beta,\mathcal{F}_{\text{lin}}}(\eta,\mathbf{x}) = \min\{\bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta), \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta)\} = \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta).$$

We divide into three cases depending on the relationship among $\bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta)$, $\bar{C}_{\phi_\beta}(-\gamma,\eta)$, and $\bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta)$.

(A) When $0 \leq \beta < \gamma$, because

$$\begin{aligned}
&\bar{C}_{\phi_\beta}(-\gamma, \eta) - \bar{C}_{\phi_\beta}(\gamma, \eta) \\
&\qquad\qquad = \left\{\eta \cdot 1 + (1-\eta)(1-\gamma+\beta)^2\right\} - \left\{\eta(1-\gamma+\beta)^2 + (1-\eta)\cdot 1\right\} \\
&\qquad\qquad = (2\eta - 1)(\gamma - \beta)\left\{2 - (\gamma - \beta)\right\} \\
&\qquad\qquad \geq 0,
\end{aligned}$$

we have $\bar{C}_{\phi_\beta}(\gamma, \eta) < \bar{C}_{\phi_\beta}(-\gamma, \eta)$ for all $\eta > \frac{1}{2}$. By contrast, because

$$\begin{aligned}
&\bar{C}_{\phi_\beta}(\gamma, \eta) - \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2, \eta) \\
&\qquad\qquad = -\left\{(1-(1-\gamma+\beta)^2) + (1-(1-\|\mathbf{x}\|_2+\beta)^2)\right\}(\eta - \eta_0(\mathbf{x})) \\
&\qquad\qquad \text{where} \quad \eta_0(\mathbf{x}) := \frac{1-(1-\|\mathbf{x}\|_2+\beta)^2}{(1-(1-\gamma+\beta)^2) + (1-(1-\|\mathbf{x}\|_2+\beta)^2)}
\end{aligned}$$

and $\frac{1}{2} < \eta_0(\mathbf{x}) < 1$, we have $\bar{C}_{\phi_\beta}(\gamma, \eta) \geq \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2, \eta)$ if $\frac{1}{2} < \eta \leq \eta_0(\mathbf{x})$ and $\bar{C}_{\phi_\beta}(\gamma, \eta) < \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2, \eta)$ if $\eta > \eta_0(\mathbf{x})$.

- If $\frac{1}{2} < \eta \leq \eta_0(\mathbf{x})$: By part 4 in Lemma 4.17,

$$\begin{aligned}
\inf_{\substack{f \in \mathcal{F}_{\text{lin}}: \\ |f(\mathbf{x})| \leq \gamma \text{ or } (2\eta-1)f(\mathbf{x}) \leq 0}} C_{\phi_\beta}(f, \eta, \mathbf{x}) &= \inf_{\alpha \in [-\|\mathbf{x}\|_2, \gamma]} \bar{C}_{\phi_\beta}(\alpha, \eta) = \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2, \eta), \\
\inf_{f \in \mathcal{F}_{\text{lin}}: |f(\mathbf{x})| \leq \gamma} C_{\phi_\beta}(f, \eta, \mathbf{x}) &= \inf_{|\alpha| \leq \gamma} \bar{C}_{\phi_\beta}(\alpha, \eta) = \bar{C}_{\phi_\beta}(\gamma, \eta).
\end{aligned}$$

Thus, by Lemma 4.9,

$$\bar{\delta}(\varepsilon, \eta, \mathbf{x}) = \begin{cases} \infty & \text{if } \eta < \varepsilon, \\ \bar{C}_{\phi_\beta}(\gamma, \eta) - \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2, \eta) & \\ \quad = (\|\mathbf{x}\|_2 - \gamma)(2 + 2\beta - \gamma - \|\mathbf{x}\|_2)\eta & \text{if } \varepsilon \leq \eta < \frac{1+\varepsilon}{2}, \\ \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2, \eta) - \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2, \eta) & \\ \quad = (\|\mathbf{x}\|_2 - \beta)(2 - \|\mathbf{x}\|_2 + \beta)(2\eta - 1) & \text{if } \frac{1+\varepsilon}{2} \leq \eta. \end{cases}$$

Hence, we obtain

$$\inf_{\eta \in \left(\frac{1}{2}, \eta_0(\mathbf{x})\right]} \inf_{\|\mathbf{x}\|_2 \geq \gamma + \rho} \bar{\delta}(\varepsilon, \eta, \mathbf{x}) = \begin{cases} A_0 \varepsilon & \text{if } 0 < \varepsilon \leq \varepsilon_0, \\ \delta_0 & \text{if } \varepsilon_0 < \varepsilon \leq \frac{1}{2}, \\ A_1 \varepsilon & \text{if } \frac{1}{2} < \varepsilon, \end{cases}$$

where $A_0 := (\gamma + \rho - \beta)(2 + \beta - \gamma - \rho)$, $A_1 := \rho(2 + 2\beta - 2\gamma - \rho)$, $\delta_0 := \frac{A_1}{2}$, and $\varepsilon_0 := \frac{\delta_0}{A_0}$. Note that the second case will not degenerate ($\delta_0 > 0$).

- If $\eta_0(\mathbf{x}) < \eta \leq 1$, based on part 4 in Lemma 4.17, it follows that

$$\begin{aligned}
\inf_{f \in \mathcal{F}_{\text{lin}}: |f(\mathbf{x})| \leq \gamma \text{ or } (2\eta-1)f(\mathbf{x}) \leq 0} C_{\phi_\beta}(f, \eta, \mathbf{x}) &= \inf_{\alpha \in [-\|\mathbf{x}\|_2, \gamma]} \bar{C}_{\phi_\beta}(\alpha, \eta) = \bar{C}_{\phi_\beta}(\gamma, \eta), \\
\inf_{f \in \mathcal{F}_{\text{lin}}: |f(\mathbf{x})| \leq \gamma} C_{\phi_\beta}(f, \eta, \mathbf{x}) &= \inf_{|\alpha| \leq \gamma} \bar{C}_{\phi_\beta}(\alpha, \eta) = \bar{C}_{\phi_\beta}(\gamma, \eta).
\end{aligned}$$

Thus, by Lemma 4.9,

$$\bar{\delta}(\varepsilon, \eta, \mathbf{x}) = \begin{cases} \infty & \text{if } \eta < \varepsilon, \\ \bar{C}_{\phi_\beta}(\gamma, \eta) - \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2, \eta) & \\ \quad = (\|\mathbf{x}\|_2 - \gamma)(2 + 2\beta - \gamma - \|\mathbf{x}\|_2)\eta & \text{if } \varepsilon \leq \eta. \end{cases}$$

Hence, we obtain

$$\inf_{\eta\in(\eta_0(\mathbf{x}),1]}\inf_{\|\mathbf{x}\|_2\ge\gamma+\rho}\bar{\delta}(\varepsilon,\eta)=\begin{cases}\eta_0(\gamma+\rho)\varepsilon & \text{if } 0<\varepsilon\le\eta_0(\gamma+\rho),\\ A_1\varepsilon & \text{if } \eta_0(\gamma+\rho)<\varepsilon.\end{cases}$$

Note that $\eta_0(\gamma+\rho)>\frac{1}{2}$. Combining the above, we obtain the $(\phi_\gamma,\mathcal{F}_{\text{lin}})$-calibration function from Lemma 4.9:

$$\delta_\rho(\varepsilon)=\begin{cases}A_0\varepsilon & \text{if } 0<\varepsilon\le\varepsilon_0,\\ \delta_0 & \text{if } \varepsilon_0<\varepsilon\le\frac{1}{2},\\ A_1\varepsilon & \text{if } \frac{1}{2}<\varepsilon,\end{cases}$$

where $A_0=(\gamma+\rho-\beta)(2+\beta-\gamma-\rho)$, $A_1=\rho(2+2\beta-2\gamma-\rho)$, $\delta_0=\frac{A_1}{2}$, and $\varepsilon_0=\frac{\delta_0}{A_0}$.

(B) When $\gamma\le\beta<1$, it is easy to see that

$$\inf_{f\in\mathcal{F}_{\text{lin}}:|f(\mathbf{x})|\le\gamma}C_{\phi_\beta}(f,\eta,\mathbf{x})=\inf_{|\alpha|\le\gamma}\bar{C}_{\phi_\beta}(\alpha,\eta)=1,$$

$$\inf_{f\in\mathcal{F}_{\text{lin}}:|f(\mathbf{x})|\le\gamma\text{ or }(2\eta-1)f(\mathbf{x})\le0}C_{\phi_\beta}(f,\eta,\mathbf{x})=\inf_{\alpha\in[-\|\mathbf{x}\|_2,\gamma]}\bar{C}_{\phi_\beta}(\alpha,\eta)=\bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta).$$

Hence, by noting that $\bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta)=\bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta)=1$ for $\|\mathbf{x}\|_2\le\beta$, it follows that

$$\begin{aligned}&\bar{\delta}(\varepsilon,\eta,\mathbf{x})\\ &=\begin{cases}\infty & \text{if } \eta<\varepsilon,\\ 1-\bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta) & \\ \quad=(\|\mathbf{x}\|_2-\beta)(2-\|\mathbf{x}\|_2+\beta)\eta & \text{if } \varepsilon\le\eta<\frac{1+\varepsilon}{2}\text{ and }\|\mathbf{x}\|_2>\beta,\\ \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta)-\bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta) & \\ \quad=(\|\mathbf{x}\|_2-\beta)(2-\|\mathbf{x}\|_2+\beta)(2\eta-1) & \text{if } \frac{1+\varepsilon}{2}\le\eta\text{ and }\|\mathbf{x}\|_2>\beta\\ 0 & \text{if } \varepsilon\le\eta\text{ and }\|\mathbf{x}\|_2\le\beta.\end{cases}\end{aligned}$$

Thus, by Lemma 4.9, we have $\delta_\rho(\varepsilon)=0$ when $\beta\ge\gamma+\rho$, and

$$\delta_\rho(\varepsilon)=\inf_{\eta\in\left(\frac{1}{2},1\right]}\inf_{\|\mathbf{x}\|_2\ge\gamma+\rho}\bar{\delta}(\varepsilon,\eta,\mathbf{x})=A_0\varepsilon$$

when $\beta<\gamma+\rho$, where $A_0=(\gamma+\rho-\beta)(2+\beta-\gamma-\rho)$.

(C) When $1\le\beta$, $\bar{C}_{\phi_\beta}(\alpha,\eta)=1$ for all $\alpha\in[-1,1]$. Hence, $\Delta C_{\phi_\beta,\mathcal{F}_{\text{lin}}}(f,\eta,\mathbf{x})=0$ and $\delta_\rho(\varepsilon)=0$.

To summarize, the $(\phi_\gamma,\mathcal{F}_{\text{lin}})$-calibration function and its Fenchel-Legendre biconjugate of the modified squared loss are as follows:

- If $0\le\beta<\gamma$,

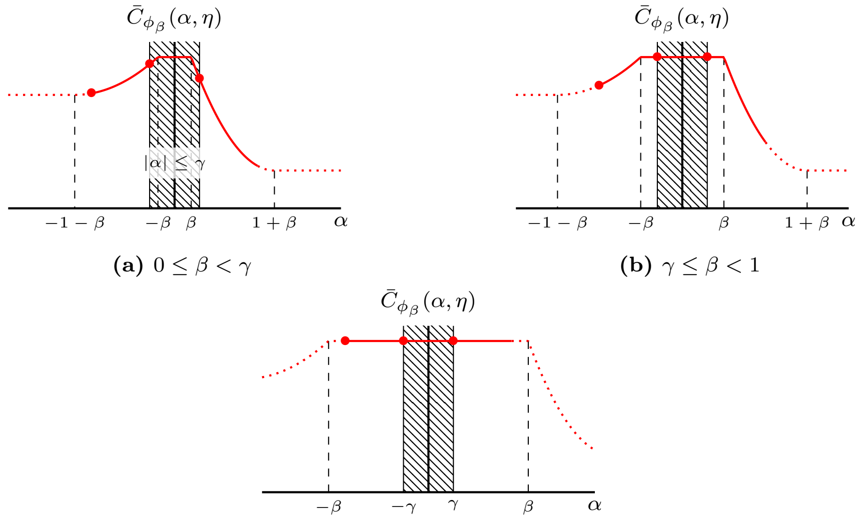
$$\delta_\rho(\varepsilon)=\begin{cases}A_0\varepsilon & \text{if } 0<\varepsilon\le\varepsilon_0,\\ \delta_0 & \text{if } \varepsilon_0<\varepsilon\le\frac{1}{2},\\ A_1\varepsilon & \text{if } \frac{1}{2}<\varepsilon,\end{cases}\quad\text{and}\quad\delta_\rho^{\star\star}(\varepsilon)=A_1\varepsilon,$$

  where $A_0=(\gamma+\rho-\beta)(2+\beta-\gamma-\rho)$, $A_1=\rho(2+2\beta-2\gamma-\rho)$, $\delta_0=\frac{A_1}{2}$, and $\varepsilon_0=\frac{\delta_0}{A_0}$.

- If $\gamma\le\beta<\gamma+\rho$, $\delta_\rho(\varepsilon)=\delta_\rho^{\star\star}(\varepsilon)=A_0\varepsilon$.

- If $\gamma+\rho\le\beta$, $\delta(\varepsilon)=\delta^{\star\star}(\varepsilon)=0$.

We deduce that the modified squared loss is calibrated w.r.t. $(\phi_\gamma,\mathcal{F}_{\text{lin}})$ if $0\le\beta\le\gamma$.

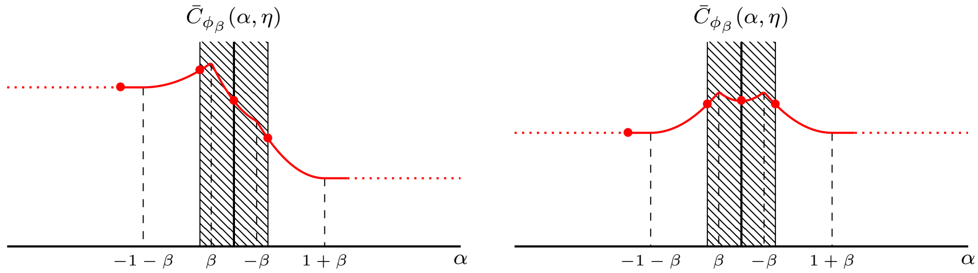


(a) $0 \leq \beta < \gamma$

(b) $\gamma \leq \beta < 1$

(c) $1 \leq \beta$

Figure 4.16: Class-conditional risk of the modified squared loss.



(a) $\eta = 0.7$

(b) $\eta = 0.5$

Figure 4.17: Class-conditional risk of the modified squared loss when $\gamma < \frac{2}{5}$ and $-1-\gamma+\sqrt{1+2\gamma^2} < \beta < 0$.

When $\beta < 0$. In this case, the CCR of the modified squared loss is no longer quasiconcave (see Figure 4.17b). However, ϕ_β is still $(\phi_\gamma, \mathcal{F}_{\mathrm{lin}})$-calibrated under some γ and $\beta < 0$. Here, we show an example.

Assume that $0 < \gamma < \frac{2}{5}$ and $(-\gamma <) - 1 - \gamma + \sqrt{1+2\gamma^2} < \beta < 0$. We focus on $\eta > \frac{1}{2}$ owing to the symmetry of $\bar{C}_{\phi_\beta}(\alpha,\eta)$ in η (part 1 of Lemma 4.16). Because we still have $\eta_0(\mathbf{x}) > \frac{1}{2}$, we can confirm in the same way as in the case (A) that $\bar{C}_{\phi_\beta}(-\gamma,\eta) > \bar{C}_{\phi_\beta}(\gamma,\eta)$, $\bar{C}_{\phi_\beta}(\gamma,\eta) \geq \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta)$ if $\frac{1}{2} < \eta \leq \eta_0(\mathbf{x})$, and $\bar{C}_{\phi_\beta}(\gamma,\eta) < \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta)$ if $\eta_0(\mathbf{x}) < \eta$. In addition, we can see that

$$\begin{aligned}
\bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta) - \bar{C}_{\phi_\beta}(0,\eta) &= \left\{\eta + (1-\eta)(1-\|\mathbf{x}\|_2+\beta)^2\right\} - (1+\beta)^2 \\
&= \eta(1-(1-\|\mathbf{x}\|_2+\beta^2)) - \|\mathbf{x}\|_2(2+2\beta-\|\mathbf{x}\|_2) \\
&> \frac{1}{2}(1-(1-\|\mathbf{x}\|_2+\beta^2)) - \|\mathbf{x}\|_2(2+2\beta-\|\mathbf{x}\|_2) \\
&> \frac{1}{2}(1-(1-\|\mathbf{x}\|_2)^2) - \|\mathbf{x}\|_2(2-\|\mathbf{x}\|_2) \\
&\qquad \text{(nonincreasing in } -1-\gamma+\sqrt{1+2\gamma^2} < \beta < 0) \\
&= \frac{1}{2}\|\mathbf{x}\|_2(\|\mathbf{x}\|_2-2) \\
&< 0,
\end{aligned}$$

$$\begin{aligned}\bar{C}_{\phi_\beta}(0,\eta) - \bar{C}_{\phi_\beta}(\gamma,\eta) &= (1+\beta)^2 - \big\{\eta(1-\gamma+\beta)^2 + (1-\eta)\big\} \\ &= (1+\beta)^2 - 1 + \eta(1-(1-\gamma+\beta)^2) \\ &> (1+\beta)^2 - 1 + \frac{1}{2}(1-(1-\gamma+\beta)^2) \\ &> (1+\beta)^2 - 1 + \frac{1}{2}(1-(1-\gamma)^2) \\ &= (1+\beta)^2 + (1+\gamma)^2 - \frac{1}{2} \\ &> 0,\end{aligned}$$

and

$$\bar{C}_{\phi_\beta}(\gamma,\eta) - \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta) > 0.$$

Thus, we have $\bar{C}_{\phi_\beta}(0,\eta) > \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta)$, $\bar{C}_{\phi_\beta}(0,\eta) > \bar{C}_{\phi_\beta}(\gamma,\eta)$, and $\bar{C}_{\phi_\beta}(\gamma,\eta) > \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta)$ Figure 4.17 and the above comparisons give us

$$\begin{aligned}C^*_{\phi_\beta,\mathcal{F}_{\text{lin}}}(\eta,\mathbf{x}) &= \inf_{\alpha\in[-\|\mathbf{x}\|_2,\|\mathbf{x}\|_2]} \bar{C}_{\phi_\beta}(\alpha,\eta) = C_{\phi_\beta}(\|\mathbf{x}\|_2,\eta), \\ \inf_{f\in\mathcal{F}:|f(\mathbf{x})|\le\gamma} C_{\phi_\beta}(f,\eta,\mathbf{x}) &= \inf_{|\alpha|\le\gamma} \bar{C}_{\phi_\beta}(\alpha,\eta) = \bar{C}_{\phi_\beta}(\gamma,\eta), \\ \inf_{f\in\mathcal{F}:f(\mathbf{x})\le\gamma} C_{\phi_\beta}(f,\eta,\mathbf{x}) &= \inf_{\alpha\in[-\|\mathbf{x}\|_2,\gamma]} \bar{C}_{\phi_\beta}(\alpha,\eta) = \min\{\bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta), \bar{C}_{\phi_\beta}(\gamma,\eta)\}.\end{aligned}$$

By Lemma 4.9, when $\varepsilon \le \eta < \frac{1+\varepsilon}{2}$,

$$\begin{aligned}\bar{\delta}(\varepsilon,\eta,\mathbf{x}) &= \inf_{f\in\mathcal{F}_{\text{lin}}:|f(\mathbf{x})|\le\gamma} \Delta C_{\phi_\beta}(f,\eta,\mathbf{x}) \\ &= \bar{C}_{\phi_\beta}(\gamma,\eta) - \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta) \\ &= \big\{\eta(1-\gamma+\beta)^2 + (1-\eta)\big\} - \big\{\eta(1-\|\mathbf{x}\|_2+\beta)^2 + (1-\eta)\big\} \\ &= (\|\mathbf{x}\|_2-\gamma)(2+2\beta-\gamma-\|\mathbf{x}\|_2)\eta,\end{aligned}$$

and

$$\begin{aligned}&\inf_{\eta\in\left[\varepsilon,\frac{1+\varepsilon}{2}\right]\cap\left(\frac{1}{2},1\right]} \inf_{\|\mathbf{x}\|_2\ge\gamma+\rho} \bar{\delta}(\varepsilon,\eta,\mathbf{x}) \\ &\qquad= \inf_{\|\mathbf{x}\|_2\ge\gamma+\rho} (\|\mathbf{x}\|_2-\gamma)(2+2\beta-\gamma-\|\mathbf{x}\|_2)\max\left\{\varepsilon,\frac{1}{2}\right\} \\ &\qquad= \begin{cases}\delta_0 & \text{if } 0<\varepsilon\le\frac{1}{2}, \\ A_1\varepsilon & \text{if } \frac{1}{2}<\varepsilon.\end{cases}\end{aligned}$$

When $\frac{1+\varepsilon}{2} \le \eta$,

$$\begin{aligned}\bar{\delta}(\varepsilon,\eta) &= \inf_{f\in\mathcal{F}_{\text{lin}}:f(\mathbf{x})\le\gamma} \Delta C_{\phi_\beta}(f,\eta,\mathbf{x}) \\ &= \min\big\{\bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta) - \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta), \bar{C}_{\phi_\beta}(\gamma,\eta) - \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta)\big\} \\ &= \min\big\{(1-(1-\|\mathbf{x}\|_2+\beta))^2(2\eta-1), (\|\mathbf{x}\|_2-\gamma)(2+2\beta-\gamma-\|\mathbf{x}\|_2)\eta\big\},\end{aligned}$$

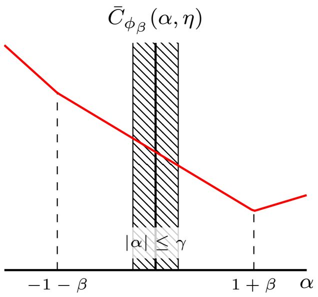


Figure 4.18: Class-conditional risk of the hinge loss.

and

$$
\begin{aligned}
&\inf_{\eta \in \left[\frac{1+\varepsilon}{2}, 1\right] \cap \left(\frac{1}{2}, 1\right]} \inf_{\|\mathbf{x}\|_2 \geq \gamma + \rho} \bar{\delta}(\varepsilon, \eta, \mathbf{x}) \\
&= \inf_{\|\mathbf{x}\|_2 \geq \gamma + \rho} \min \left\{ (1 - (1 - \|\mathbf{x}\|_2 + \beta)^2)\varepsilon, (\|\mathbf{x}\|_2 - \gamma)(2 + 2\beta - \gamma - \|\mathbf{x}\|_2)\frac{1+\varepsilon}{2} \right\} \\
&= \min \left\{ A_0 \varepsilon, A_1 \frac{1+\varepsilon}{2} \right\}.
\end{aligned}
$$

Hence, the $(\phi_\gamma, \mathcal{F}_{\text{lin}})$-calibration function of ϕ_β is

$$
\delta_\rho(\varepsilon) = \begin{cases} A_0 \varepsilon & \text{if } 0 < \varepsilon \leq \varepsilon_0, \\ \delta_0 & \text{if } \varepsilon_0 < \varepsilon \leq \frac{1}{2}, \\ A_1 \varepsilon & \text{if } \frac{1}{2} < \varepsilon, \end{cases}
$$

where $A_0 = (\gamma + \rho - \beta)(2 + \beta - \gamma - \rho)$, $A_1 = \rho(2 + 2\beta - 2\gamma - \rho)$, $\delta_0 = \frac{A_1}{2}$, and $\varepsilon_0 = \frac{\delta_0}{A_0}$. We can see that the second case will not degenerate (i.e., $\delta_0 > 0$) under the range $-1 - \gamma + \sqrt{1 + 2\gamma^2} < \beta < 0$ and $0 < \gamma \leq \frac{2}{5}$.

Hinge Loss

The ϕ_β-CCR is $C_{\phi_\beta}(f, \eta, \mathbf{x}) = \bar{C}_{\phi_\beta}(f(\mathbf{x}), \eta)$, where

$$
\bar{C}_{\phi_\beta}(\alpha, \eta) = \begin{cases} -\eta\alpha + \eta(1 + \beta) & \text{if } \alpha < -(1 + \beta), \\ (1 - 2\eta)\alpha + (1 + \beta) & \text{if } -(1 + \beta) \leq \alpha < 1 + \beta, \\ (1 - \eta)\alpha + (1 - \eta)(1 + \beta) & \text{if } 1 + \beta < \alpha. \end{cases}
$$

Minimal inner risk. When $\eta > \frac{1}{2}$, $\bar{C}_{\phi_\beta}(\alpha, \eta)$ is minimized at $\alpha = \|\mathbf{x}\|_2$, and when $\eta \leq \frac{1}{2}$, $\bar{C}_{\phi_\beta}(\alpha, \eta)$ is minimized at $\alpha = -\|\mathbf{x}\|_2$. Hence,

$$
\begin{aligned}
C^*_{\phi_\beta, \mathcal{F}_{\text{lin}}}(\eta, \mathbf{x}) &= \inf_{\alpha \in [-\|\mathbf{x}\|_2, \|\mathbf{x}\|_2]} \bar{C}_{\phi_\beta}(\alpha, \eta) \\
&= \begin{cases} \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2, \eta) & \text{if } \eta > \frac{1}{2} \\ \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2, \eta) & \text{if } \eta \leq \frac{1}{2} \end{cases} \\
&= -|1 - 2\eta| \cdot \|\mathbf{x}\|_2 + 1 + \beta.
\end{aligned}
$$

Calibration function. We restrict the range of η to $\eta > \frac{1}{2}$ by virtue of part 1 of Lemma 4.16. Then, $C^*_{\phi_\beta, \mathcal{F}_{\text{lin}}}(\eta, \mathbf{x}) = \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2, \eta)$. In addition, $\bar{C}_{\phi_\beta}(\alpha, \eta)$ is

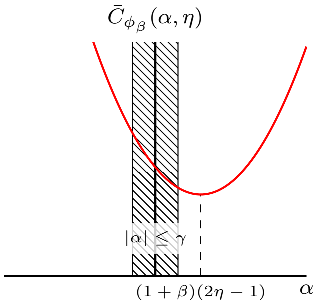


Figure 4.19: Class-conditional risk of the squared loss.

plotted in Figure 4.18 in case of $\eta > \frac{1}{2}$. From the figure, we can see that

$$\begin{aligned}\inf_{f\in\mathcal{F}_{\mathrm{lin}}:f(\mathbf{x})\le\gamma} C_{\phi_\beta}(f,\eta,\mathbf{x}) &= \inf_{\substack{\alpha\in[-\|\mathbf{x}\|_2,\|\mathbf{x}\|_2]:|\alpha|\le\gamma \text{ or}\\(2\eta-1)\alpha\le 0}} \bar{C}_{\phi_\beta}(\alpha,\eta)\\ &= \bar{C}_{\phi_\beta}(\gamma,\eta)\\ &= \inf_{|\alpha|\le\gamma} \bar{C}_{\phi_\beta}(\alpha,\eta)\\ &= \inf_{f\in\mathcal{F}_{\mathrm{lin}}:|f(\mathbf{x})|\le\gamma} C_{\phi_\beta}(f,\eta,\mathbf{x}),\end{aligned}$$

by noting that $\|\mathbf{x}\|_2 > \gamma$ is assumed. Hence, by Lemma 4.9,

$$\bar{\delta}(\varepsilon,\eta,\mathbf{x}) = \begin{cases}\infty & \text{if } \|\mathbf{x}\|_2 \le \gamma \text{ or } \eta < \varepsilon,\\ \bar{C}_{\phi_\beta}(\gamma,\eta) - \bar{C}^*_{\phi_\beta,\mathcal{F}_{\mathrm{lin}}}(\|\mathbf{x}\|_2,\eta) & \\ \quad = (\|\mathbf{x}\|_2-\gamma)(2\eta-1) & \text{if } \|\mathbf{x}\|_2 > \gamma \text{ and } \varepsilon \le \eta,\end{cases}$$

for $\eta > \frac{1}{2}$, and

$$\delta_\rho(\varepsilon) = \inf_{\eta\in\left[\frac{1}{2},1\right]} \inf_{x\in\tilde{\mathcal{X}}_\rho} \bar{\delta}(\varepsilon,\eta,\mathbf{x}) = \begin{cases}0 & \text{if } 0<\varepsilon\le\frac{1}{2},\\ 2\rho\left(\varepsilon-\frac{1}{2}\right) & \text{if } \frac{1}{2}<\varepsilon,\end{cases}$$

and $\delta^{\star\star}_\rho(\varepsilon) = \delta_\rho(\varepsilon)$.

Squared Loss

The ϕ_β-CCR is $C_{\phi_\beta}(f,\eta,\mathbf{x}) = \bar{C}_{\phi_\beta}(f(\mathbf{x}),\eta)$, where

$$\begin{aligned}\bar{C}_{\phi_\beta}(\alpha,\eta) &= \eta(1-\alpha+\beta)^2 + (1-\eta)(1+\alpha+\beta)^2\\ &= \{\alpha-(1+\beta)(2\eta-1)\}^2 + 4(1+\beta)^2\eta(1-\eta).\end{aligned}$$

Let $\alpha_* := (1+\beta)(2\eta-1)$.

Minimal inner risk. When $\eta > \frac{1}{2}$, $\bar{C}_{\phi_\beta}(\alpha,\eta)$ is minimized at $\alpha = \|\mathbf{x}\|_2$ if $\|\mathbf{x}\|_2 \ge \alpha_*$, and at $\alpha = \alpha_*$ if $\|\mathbf{x}\|_2 < \alpha_*$. When $\eta \le \frac{1}{2}$, $\bar{C}_{\phi_\beta}(\alpha,\eta)$ is minimized at $\alpha = -\|\mathbf{x}\|_2$ if $\|\mathbf{x}\|_2 \le -\alpha_*$, and at $\alpha = \alpha_*$ if $\|\mathbf{x}\|_2 > -\alpha_*$. Hence,

$$\begin{aligned}C^*_{\phi_\beta,\mathcal{F}_{\mathrm{lin}}}(\eta,\mathbf{x}) &= \inf_{\alpha\in[-\|\mathbf{x}\|_2,\|\mathbf{x}\|_2]} \bar{C}_{\phi_\beta}(\alpha,\eta)\\ &= \begin{cases}\bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2,\eta) & \text{if } \eta > \frac{1}{2} \text{ and } \|\mathbf{x}\|_2 \ge \alpha_*,\\ \bar{C}_{\phi_\beta}(-\|\mathbf{x}\|_2,\eta) & \text{if } \eta \le \frac{1}{2} \text{ and } \|\mathbf{x}\|_2 \le -\alpha_*,\\ \bar{C}_{\phi_\beta}(\alpha_*,\eta) & \text{otherwise.}\end{cases}\end{aligned}$$

Calibration function. We restrict the range of η to $\eta > \frac{1}{2}$ by virtue of part 1 of Lemma 4.16. Here, $\bar{C}_{\phi_\beta}(\alpha, \eta)$ is plotted in Figure 4.19 in case of $\eta > \frac{1}{2}$. By comparing α_* and $\|\mathbf{x}\|_2$, we have

$$C^*_{\phi_\beta,\mathcal{F}_{\text{lin}}}(\eta, \mathbf{x}) = \begin{cases} \bar{C}_{\phi_\beta}(\alpha_*, \eta) & \text{if } \alpha_* < \|\mathbf{x}\|_2, \\ \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2, \eta) & \text{if } \alpha_* \geq \|\mathbf{x}\|_2. \end{cases}$$

From the figure, we can see that

$$\begin{aligned} \inf_{f \in \mathcal{F}_{\text{lin}} : f(\mathbf{x}) \leq \gamma} C_{\phi_\beta}(f, \eta, \mathbf{x}) &= \inf_{\substack{\alpha \in [-\|\mathbf{x}\|_2, \|\mathbf{x}\|_2] : |\alpha| \leq \gamma \text{ or} \\ (2\eta - 1)\alpha \leq 0}} \bar{C}_{\phi_\beta}(\alpha, \eta) \\ &= \inf_{|\alpha| \leq \gamma} \bar{C}_{\phi_\beta}(\alpha, \eta) \\ &= \inf_{f \in \mathcal{F}_{\text{lin}}} C_{\phi_\beta}(f, \eta, \mathbf{x}) \\ &= \begin{cases} \bar{C}_{\phi_\beta}(\alpha_*, \eta) & \text{if } \gamma > \alpha_*, \\ \bar{C}_{\phi_\beta}(\gamma, \eta) & \text{if } \gamma \leq \alpha_*, \end{cases} \end{aligned}$$

by noting that $\|\mathbf{x}\|_2 > \gamma$ is assumed. Hence, by Lemma 4.9,

$$\begin{aligned} &\bar{\delta}(\varepsilon, \eta, \mathbf{x}) \\ &= \begin{cases} \infty & \text{if } \|\mathbf{x}\|_2 \leq \gamma \text{ or } \eta < \varepsilon, \\ \bar{C}_{\phi_\beta}(\alpha_*, \eta) - \bar{C}_{\phi_\beta}(\alpha_*, \eta) & \text{if } \|\mathbf{x}\|_2 > \gamma \text{ and } \varepsilon \leq \eta \text{ and } \alpha_* < \gamma, \\ \bar{C}_{\phi_\beta}(\gamma, \eta) - \bar{C}_{\phi_\beta}(\alpha_*, \eta) & \text{if } \|\mathbf{x}\|_2 > \gamma \text{ and } \varepsilon \leq \eta \text{ and } \gamma \leq \alpha_* < \|\mathbf{x}\|_2, \\ \bar{C}_{\phi_\beta}(\gamma, \eta) - \bar{C}_{\phi_\beta}(\|\mathbf{x}\|_2, \eta) & \text{if } \|\mathbf{x}\|_2 > \gamma \text{ and } \varepsilon \leq \eta \text{ and } \|\mathbf{x}\|_2 \leq \alpha_*. \end{cases} \end{aligned}$$

By taking the infimum over η and x,

$$\delta_\rho(\varepsilon) = \inf_{\eta \in [\frac{1}{2}, 1]} \inf_{x \in \tilde{\mathcal{X}}_\rho} \bar{\delta}(\varepsilon, \eta, \mathbf{x}) = \begin{cases} 0 & \text{if } 0 < \varepsilon \leq \varepsilon_0, \\ 4(1+\beta)^2(\varepsilon - \varepsilon_0)^2 & \text{if } \varepsilon_0 < \varepsilon, \end{cases}$$

where $\varepsilon_0 \coloneqq \frac{1+\beta+\gamma}{2(1+\beta)}$, and $\delta^{\star\star}_\rho(\varepsilon) = \delta_\rho(\varepsilon)$.

4.12 Conclusion

Calibration analysis was leveraged to analyze the adversarially robust 0-1 loss. Focusing on the class of linear classifiers, we found that no convex surrogate loss is calibrated w.r.t. the adversarially robust 0-1 loss for general distributions. We also established necessary and sufficient conditions for a certain class of nonconvex surrogate losses to be calibrated w.r.t. the adversarially robust 0-1 loss, which includes shifted versions of the ramp and sigmoid losses.

Recently, Tsipras et al. [2019] and Suggala et al. [2019] revealed that there is an intrinsic trade-off between the adversarial robustness and classification accuracy. Tsipras et al. [2019] showed a simple scenario under which the trade-off exists, and Suggala et al. [2019] proposed an alternative evaluation framework of robustness to overcome the trade-off. Since then, many works investigated the trade-off in detail and how to overcome it. Javanmard et al. [2020] showed that linear classifiers do not exhibit the trade-off under the overparametrized regime but the Pareto frontier emerges as the sample size grows. Raghunathan et al. [2020] used self-training as a part of regularization to mitigate the trade-off. Krishnan et al. [2020] showed that there exists a lower bound of the Lipschitz constant of classifiers given a fixed budget of the classification accuracy. Since the Lipschitz constant measures the smoothness of classifiers, this result is also related to the robustness-accuracy trade-off. Dobriban et al. [2021] characterized the Bayes classifier of adversarially robust classification assuming that the data is normally distributed and showed the trade-off from the viewpoint of the Bayes risk. Our proof approach based on calibration analysis could also contribute elucidating and characterizing the robustness-accuracy trade-off and providing better insights, which is left as a future work.

Chapter 5

Connecting Similarity Learning to Classification

> *One needs to look near at hand if one wants to study men; but to study man one must learn to look from afar; one must first observe differences in order to discover attributes.*
>
> — Jean-Jacques Rousseau, *On the Origin of Language*

Similarity learning is a general problem to elicit useful representations by predicting the relationship between a pair of patterns. This problem is related to various important preprocessing tasks such as metric learning, kernel learning, and contrastive learning. Despite the fact that a classifier built upon the representations is expected to perform well in downstream classification, little theory has been provided in literature thus far. Therefore, we tackle a fundamental question: *How is similarity learning relevant to standard classification?* In this chapter, we reveal that a specific formulation of similarity learning is strongly related to an objective of binary classification. This formulation generalizes many types of existing similarity learning, and an excess risk bound shows its explicit connection to classification. Consequently, our results elucidate that similarity learning is capable of solving binary classification by directly eliciting a decision boundary.

5.1 Introduction

Similarity learning is a learning paradigm [Kulis, 2013], that builds a pairwise model to predict whether given paired patterns are similar or dissimilar in the latent classes. We call such a pair of patterns *pairwise supervision*, instead of ordinary *pointwise supervision*, which binds a class label to a single input pattern. Pairwise supervision is commonly available in many domains such as geographical analysis [Wagstaff et al., 2001], chemical experiment [Eisenberg et al., 2000], click-through feedback [Davis et al., 2007], computer vision [Yan et al., 2006, Wang and Gupta, 2015], natural language processing [Mikolov et al., 2013], and crowdsourcing [Gomes et al., 2012]. Similarity learning has therefore been extensively studied, including metric learning [Xing et al., 2003, Bilenko et al., 2004, Davis et al., 2007, Weinberger and Saul, 2009, Bellet et al., 2012, Niu et al., 2014], kernel learning [Cristianini et al., 2002, Bach et al., 2004, Lanckriet et al., 2004, Li and Liu, 2009, Cortes et al., 2010], and $(\varepsilon, \gamma, \tau)$-good similarity [Balcan et al., 2008, Wang et al., 2009, Kar and Jain, 2011, Bellet et al., 2012]. The obtained pairwise model is viewed as a metric function within the pattern space. If a good metric is learned, the model is expected to achieve a good performance in downstream tasks by capturing inherent structures within the data. Correspondingly, it has been widely used for various downstream tasks, such as classification [Cristianini et al., 2002, Balcan et al., 2008, Hsu et al., 2019, Saunshi et al., 2019, Nozawa et al., 2020], clustering [Bromley et al., 1994, Xing et al., 2003, Davis et al., 2007, Weinberger and Saul, 2009], model selection [Lanckriet et al., 2004], and one-shot learning [Koch et al., 2015].

Several studies have theoretically investigated the relationship between similarity learning and downstream classification. Kar and Jain [2011] and Bellet et al. [2012] proved that a feature space built on a learned metric is linearly separable under the framework of $(\varepsilon, \gamma, \tau)$-good similarity. In addition, Saunshi et al. [2019] provided a similar result unique to contrastive learning. These results boil down to two-step learners, which first solve similarity learning and then train the classifiers. However, the latter step often requires as many samples as the former step because the feature space will potentially become high-dimensional to guarantee its separability.

In this chapter, we yield a clue to a fundamental question, i.e., *how is similarity learning relevant to downstream classification*, by revealing that a specific formulation of similarity learning is tightly connected to binary classification. This interrelation provides a new insight in that *a binary decision boundary can essentially be obtained with only pairwise supervision* up to the label flipping. As a result, the post-process resolving the label flipping becomes less label-demanding than the previous formulations [Kar and Jain, 2011, Bellet et al., 2012, Saunshi et al., 2019]. This post-process can further be applied with only pairwise supervision. Our results are notable in that: (i) they show that similarity learning enables us to implicitly elicit a binary decision boundary without any explicit training of classifiers, and (ii) the post-process is less costly in terms of pointwise supervision. Specifically, we will see the following:

- Section 5.4.1: Similarity learning is tied to the binary classification error up to the label flipping.

- Section 5.4.2: The post-process, i.e., how to resolve the label flipping, is discussed. As a by-product, we come across a training method of binary classifiers with only pairwise supervision (Section 5.4.3).

- Section 5.5: A finite-sample excess risk bound is established to connect similarity learning to binary classification. This theoretical finding is numerically demonstrated (Section 5.6).

Remark 5.1 (Multi-class case)**.** *Despite the fact that multi-class classification may be more natural when working with pairwise supervision, our result is still limited to a binary case. In practice, the* one-versus-rest *approach is effective, by generating pairwise supervision from pointwise supervision with treating one class as positive and the rest as negative. Recall that this approach is more beneficial than the similar one-versus-rest approach taken in the previous studies on* $(\varepsilon, \gamma, \tau)$*-good similarity [Kar and Jain, 2011] because the post-process in our formulation is less label-demanding, as we will see in Section 5.4.5.*

5.2 Related Work

A number of studies have tackled pattern recognition with only pairwise supervision. Semi-supervised clustering [Wagstaff et al., 2001, Basu et al., 2002, Bilenko et al., 2004, Zeng and Cheung, 2011] is one of the common approaches, which applies clustering without violating pairwise supervision. Despite its success with domain-specific assumptions [Chapelle et al., 2006], it is not targeted for classification owing to a lack of the generalization guarantee. Recently, a meta-classification approach has emerged [Hsu et al., 2019, Wu et al., 2020], in which a model predicting pairwise labels is decomposed into pointwise classifiers. Although the pairwise generalization performance of the meta-classifier has been studied [Wu et al., 2020], the pointwise counterpart remains unexplored. Zhang and Yan [2007] theoretically justified a similar approach but only for the squared loss and asymptotically. In parallel, several studies have solved classification with pairwise supervision by minimizing unbiased classification risk estimators [Bao et al., 2018, Shimada et al., 2021, Cui et al., 2020, Dan et al., 2021]. Although their approaches are blessed with pointwise generalization error bounds, their performance deteriorates when the class-prior probability is close to uniform.

Note that another related learning paradigm is *similarity as features* [Graepel et al., 1999, Balcan et al., 2008, Chen et al., 2009]. This line of research considers a setup in which learners are given labels without features, and similarity relations among data points are utilized as features. By contrast, we treat similarity as a type of supervision, and features are available independently; thus, both feature and similarity information can be incorporated simultaneously into learning.

5.3 Problem Setup

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a d-dimensional pattern space, $\mathcal{Y} = \{\pm 1\}$ be the label space, and $p(\mathbf{x}, y)$ be the density of an underlying distribution over $\mathcal{X} \times \mathcal{Y}$. Denote the positive (negative, resp.) class prior by $\pi_+ := p(y = +1)$ ($\pi_- := p(y = -1)$, resp.). Let $\operatorname{sgn}(\alpha) = 1$ for $\alpha > 0$ and -1 otherwise.

Binary classification. The goal of binary classification is to classify unseen patterns into two classes. This can be formulated as a problem of finding a classifier $h : \mathcal{X} \to \mathcal{Y}$ that minimizes

$$R_{\text{point}}(h) := \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim p(\mathsf{X},\mathsf{Y})} \left[\mathbb{1}_{\{h(\mathsf{X})\neq\mathsf{Y}\}}\right], \tag{5.1}$$

where $\mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim p(\mathsf{X},\mathsf{Y})}[\cdot]$ denotes the expectation with respect to $p(\mathbf{x}, y)$. Typically, we specify a hypothesis class $\mathcal{H}$ beforehand and find a minimizer h^* of R_{point} within it, i.e.,

$$h^* \in \mathop{\arg\min}_{h\in\mathcal{H}} R_{\text{point}}(h).$$

The empirical mean of R_{point} is computed with finite samples.

Similarity learning. Here, we introduce *similarity learning*, which aims to learn the relationship between a pair of patterns. Specifically, we focus on the following formulation to predict whether a pair of patterns belong to the same class. Hereafter, we suppose that $(\mathbf{x}, y)$ and $(\mathbf{x}', y')$ in a pair are independent of each other. Assume that $\mathsf{X} = \mathbf{x}$ and $\mathsf{X}' = \mathbf{x}'$ are observed first, and the pairwise supervision T is drawn from

$$\begin{aligned} &p(\mathsf{T} = \mathsf{Y}\mathsf{Y}' \mid \mathbf{x}, \mathbf{x}') \\ &= \begin{cases} p(\mathsf{Y} = +1 \mid \mathbf{x})p(\mathsf{Y}' = +1 \mid \mathbf{x}') + p(\mathsf{Y} = -1 \mid \mathbf{x})p(\mathsf{Y}' = -1 \mid \mathbf{x}') & \text{if } \mathsf{Y}\mathsf{Y}' = +1, \\ p(\mathsf{Y} = +1 \mid \mathbf{x})p(\mathsf{Y}' = -1 \mid \mathbf{x}') + p(\mathsf{Y} = -1 \mid \mathbf{x})p(\mathsf{Y}' = +1 \mid \mathbf{x}') & \text{if } \mathsf{Y}\mathsf{Y}' = -1. \end{cases} \end{aligned}$$

The product YY' indicates whether Y and Y' are the same or similar, namely, $T = +1$, or not the same or dissimilar, namely, $T = -1$. We are then interested in the minimizer of the following classification error

$$R_{\text{pair}}(h) \coloneqq \mathop{\mathbb{E}}_{\mathsf{X},\mathsf{X}'\sim p(\mathsf{X})} \mathop{\mathbb{E}}_{\mathsf{T}\sim p(\mathsf{T}=\mathsf{YY}'|\mathbf{x},\mathbf{x}')} \left[\mathbb{1}_{\{h(\mathsf{X})\cdot h(\mathsf{X}')\neq\mathsf{T}\}}\right]. \tag{5.2}$$

The similarity model $h(\mathbf{x}) \cdot h(\mathbf{x}')$ is legitimate in the current context because the product form $\mathsf{T} = \mathsf{YY}'$ is used as the target such that we predict the label agreement. We call R_{point} the *pointwise classification error* and R_{pair} the *pairwise classification error*. The empirical mean of R_{pair} is computed with a finite number of triplets $(\mathbf{x}, \mathbf{x}', yy')$.

Remark 5.2 (T is not a hard similarity label)**.** *Even if* $\mathsf{Y} = \mathsf{Y}' = +1$ *(similar) with high probability, we can observe* $\mathsf{T} = -1$ *(dissimilar) with a certain probability. Assume*

$$\eta \coloneqq p(\mathsf{Y} = +1 \mid \mathbf{x}) \in \left(\tfrac{1}{2}, 1\right) \ \textit{and}\ \eta' \coloneqq p(\mathsf{Y}' = +1 \mid \mathbf{x}') \in \left(\tfrac{1}{2}, 1\right).$$

Then, the flipping rate $p(\mathsf{T} = -1 \mid \mathbf{x}, \mathbf{x}')$ *lies in* $(0, \frac{1}{2})$*. In fact,*

$$\begin{aligned} p(\mathsf{T} = -1 \mid \mathbf{x}, \mathbf{x}') &= \eta(1 - \eta') + (1 - \eta)\eta' \\ &= \eta \underbrace{(1 - 2\eta')}_{<0} + \eta' \\ &< \frac{1}{2} \cdot (1 - 2\eta') + \eta' \\ &= \frac{1}{2}. \end{aligned}$$

This means that we do not preclude pairwise supervision with flipping in the probability, which is sufficiently general to cover common annotation scenarios.

5.4 Learning a Binary Classifier with Pairwise Supervision

We draw a connection between the specific formulation of similarity learning (5.2) and binary classification (Theorem 5.1). This linkage enables us to train a pointwise binary classifier using pairwise supervision (Section 5.4.3).

5.4.1 Connection between Similarity Learning and Classification

We first introduce a performance metric for binary classification called a *clustering error* that quantifies the discriminative power of a classifier up to the label flipping:[1]

$$R_{\text{clus}}(h) \coloneqq \min\left\{ R_{\text{point}}(h), R_{\text{point}}(-h) \right\}. \tag{5.3}$$

Here, although R_{clus} is used as an evaluator of binary classifiers, it is usually applied for an evaluation of clustering methods [Fahad et al., 2014]. The clustering error differs from R_{point} in that it dismisses the difference between $+h$ and $-h$; however, a binary decision boundary is still properly evaluated. The clustering error R_{clus} can be tied to the pairwise classification error R_{pair} as follows, which is our primary result.

[1] $1 - R_{\text{clus}}$ is known as the clustering accuracy [Fahad et al., 2014]. The number of clusters is confined to two for our specific purpose.

Theorem 5.1. *Any classifier $h : \mathcal{X} \to \mathcal{Y}$ satisfies*

$$R_{\text{clus}}(h) = \frac{1}{2} - \frac{\sqrt{1 - 2R_{\text{pair}}(h)}}{2}. \tag{5.4}$$

Proof. We derive an equivalent expression of the pairwise classification error R_{pair} as follows:

$$\begin{aligned}
R_{\text{pair}}(h) &= \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim p(\mathsf{X},\mathsf{Y})} \mathop{\mathbb{E}}_{(\mathsf{X}',\mathsf{Y}')\sim p(\mathsf{X},\mathsf{Y})} \left[\mathbb{1}_{\{h(\mathsf{X})\cdot h(\mathsf{X}')\neq \mathsf{Y}\mathsf{Y}'\}}\right] \\
&= \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim p(\mathsf{X},\mathsf{Y})} \mathop{\mathbb{E}}_{(\mathsf{X}',\mathsf{Y}')\sim p(\mathsf{X},\mathsf{Y})} \left[\mathbb{1}_{\{h(\mathsf{X})\neq \mathsf{Y}\}}\mathbb{1}_{\{h(\mathsf{X}')= \mathsf{Y}'\}}\right] \\
&\quad + \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim p(\mathsf{X},\mathsf{Y})} \mathop{\mathbb{E}}_{(\mathsf{X}',\mathsf{Y}')\sim p(\mathsf{X},\mathsf{Y})} \left[\mathbb{1}_{\{h(\mathsf{X})= \mathsf{Y}\}}\mathbb{1}_{\{h(\mathsf{X})\neq \mathsf{Y}'\}}\right] \\
&= \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim p(\mathsf{X},\mathsf{Y})} \left[\mathbb{1}_{\{h(\mathsf{X})\neq \mathsf{Y}\}}\right] \mathop{\mathbb{E}}_{(\mathsf{X}',\mathsf{Y}')\sim p(\mathsf{X},\mathsf{Y})} \left[\mathbb{1}_{\{h(\mathsf{X}')= \mathsf{Y}'\}}\right] \\
&\quad + \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim p(\mathsf{X},\mathsf{Y})} \left[\mathbb{1}_{\{h(\mathsf{X})= \mathsf{Y}\}}\right] \mathop{\mathbb{E}}_{(\mathsf{X}',\mathsf{Y}')\sim p(\mathsf{X},\mathsf{Y})} \left[\mathbb{1}_{\{h(\mathsf{X}')\neq \mathsf{Y}'\}}\right] \\
&= 2 \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim p(\mathsf{X},\mathsf{Y})} \left[\mathbb{1}_{\{h(\mathsf{X})\neq \mathsf{Y}\}}\right] \mathop{\mathbb{E}}_{(\mathsf{X}',\mathsf{Y}')\sim p(\mathsf{X},\mathsf{Y})} \left[\mathbb{1}_{\{h(\mathsf{X}')= \mathsf{Y}'\}}\right] \\
&= 2\, R_{\text{point}}(h)\,(1 - R_{\text{point}}(h))\,.
\end{aligned}$$

We can transform the above equation as

$$R_{\text{point}}(h) = \frac{1}{2} \pm \frac{\sqrt{1 - 2R_{\text{pair}}(h)}}{2}.$$

Then, we also have

$$R_{\text{point}}(-h) = 1 - R_{\text{point}}(h) = \frac{1}{2} \mp \frac{\sqrt{1 - 2R_{\text{pair}}(h)}}{2}.$$

By combining them, we obtain Equation (5.4). □

An immediate corollary is the monotonic relationship

$$R_{\text{clus}}(h_1) < R_{\text{clus}}(h_2) \iff R_{\text{pair}}(h_1) < R_{\text{pair}}(h_2)$$

for any h_1 and h_2. Hence, the minimization of R_{pair} amounts to the minimization of R_{clus}, constituting a decision boundary. That is, *similarity learning can essentially discover a binary decision boundary.* Although similarity learning has previously been connected to downstream classification through intermediate feature spaces [Kar and Jain, 2011, Bellet et al., 2012, Saunshi et al., 2019, Nozawa et al., 2020], our result is the first to explicate that similarity learning is directly related to the construction of a decision boundary.

Surrogate risk minimization. Here, we discuss surrogate losses for similarity learning. We define a hypothesis class by $\mathcal{H} = \{ \operatorname{sgn} \circ f \mid f \in \mathcal{F} \}$, where $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ is a specified class of prediction functions and $\operatorname{sgn} \circ f(\cdot) := \operatorname{sgn}(f(\cdot))$. Theorem 5.1 suggests that we may minimize R_{clus} by minimizing R_{pair} instead. As in the standard binary classification case, the indicator function appearing in R_{pair} is replaced with a surrogate loss $\phi : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ because it is intractable to minimize a discrete objective [Bartlett et al., 2006]. Eventually, the pairwise surrogate risk

$$R^{\phi}_{\text{pair}}(f) := \mathop{\mathbb{E}}_{\mathsf{X},\mathsf{X}'\sim p(\mathsf{X})} \mathop{\mathbb{E}}_{\mathsf{T}\sim p(\mathsf{T}=\mathsf{Y}\mathsf{Y}'|\mathsf{X},\mathsf{X}')} \left[\phi(f(\mathsf{X})f(\mathsf{X}'), \mathsf{T})\right] \tag{5.5}$$

is minimized. If ϕ is classification-calibrated, the minimization of R^{ϕ}_{pair} is expected to lead to minimizing R_{pair} as well. This will be justified by Lemma 5.4 in Section 5.5.

As we will discuss in Section 5.4.5, the formulation (5.5) is general to subsume several existing formulations in similarity learning in terms of the surrogate loss.

5.4.2 Determination of Correct Sign of Classifiers

In Section 5.4.1, we observed that similarity learning can draw a decision boundary up to the label flipping. For a given hypothesis h, we are now interested in its sign, that is, $+h$ or $-h$, leading to a smaller pointwise classification error. We refer to this step as a class assignment. The optimal class assignment is denoted by

$$s^* := \operatorname*{arg\,min}_{s \in \{\pm 1\}} R_{\text{point}}(s \cdot h).$$

We can consider two scenarios. Under both, a class assignment is much cheaper in supervision than training the post-hoc linear separators.

Class assignment with pointwise supervision. If pointwise supervision is available, we can determine the class assignment by minimizing the pointwise classification error R_{point} computed using the additional data. This procedure admits the exponentially small sample complexity [Zhang and Yan, 2007].

Class assignment without pointwise supervision. Herein, we further ask if it is possible to obtain the correct class assignment *without* any class labels. Surprisingly, we found that this is possible if the positive and negative proportions are not equal and we know *which class is the majority*. Based on the equivalent expression of R_{point} [Shimada et al., 2021], this finding is formally stated in the following theorem.

Theorem 5.2. *Assume that the class prior $\pi_+ \neq \frac{1}{2}$. The optimal class assignment s^* can then be represented as $s^* = \operatorname{sgn}(2\pi_+ - 1) \cdot \operatorname{sgn}(1 - 2Q(h))$, where*

$$Q(h) := \mathop{\mathbb{E}}_{\mathsf{X},\mathsf{X}' \sim p(\mathsf{X})} \mathop{\mathbb{E}}_{\mathsf{T} \sim p(\mathsf{T}=\mathsf{YY}' \mid \mathsf{X},\mathsf{X}')} \left[\frac{\mathbb{1}_{\{h(\mathsf{X}) \neq \mathsf{T}\}} + \mathbb{1}_{\{h(\mathsf{X}') \neq \mathsf{T}\}}}{2} \right].$$

The proof is provided in Section 5.7. We approximate Q with a finite number of pairs. As we will see in Lemma 5.6 in Section 5.5, the class assignment error is exponentially small in the number of pairs.

Remark 5.3 (Necessity of $Q(h)$)**.** *If we know which class is the majority, a class assignment may appear to be possible by simply looking at the average of $h(\mathbf{x})$ with unlabeled validation data, instead of Theorem 5.2. Unfortunately, this does not always succeed even asymptotically. This will be discussed in Section 5.4.4.*

5.4.3 Learning a Binary Classifier with Only Pairwise Supervision is Possible

As a result of Theorems 5.1 and 5.2, the following two-stage method can train a pointwise classifier using only pairwise supervision. Assume that the class prior is not $\frac{1}{2}$ and the majority class is known. Let $\mathcal{S}_{\text{train}} := \{(\mathbf{x}_i, \mathbf{x}'_i, \tau_i)\}_{i=1}^{n_{\text{pair}}}$ be a training set, where $\tau_i := y_i y'_i$ and $(\mathbf{x}_i, y_i)$ and $(\mathbf{x}'_i, y'_i)$ are i.i.d. samples following $p(\mathbf{x}, y)$.

We randomly divide n_{pair} pairs in $\mathcal{S}_{\mathrm{train}}$ into two sets $\mathcal{S}_1$ and $\mathcal{S}_2$, where $|\mathcal{S}_1| = m_1$ and $|\mathcal{S}_2| = m_2$ satisfying $m_1 + m_2 = n_{\mathrm{pair}}$.[2]

In Step 1, we obtain a minimizer of the empirical pairwise classification risk with $\mathcal{S}_1$:

$$\widehat{f} \coloneqq \operatorname*{arg\,min}_{f \in \mathcal{F}} \widehat{R}^{\phi}_{\mathrm{pair}}(f), \tag{5.6}$$

where $\widehat{R}^{\phi}_{\mathrm{pair}}$ is the sample mean of R^{ϕ}_{pair} with $\mathcal{S}_1$. In Step 2, we assign classes with $\mathrm{sgn} \circ \widehat{f}$ and $\mathcal{S}_2$:

$$\widehat{s} \coloneqq \mathrm{sgn}(2\pi_+ - 1) \cdot \mathrm{sgn}(1 - 2\widehat{Q}(\mathrm{sgn} \circ \widehat{f})), \tag{5.7}$$

where $\widehat{Q}$ is the sample mean of Q with $\mathcal{S}_2$. After all, $\widehat{s} \cdot \mathrm{sgn} \circ \widehat{f}$ is a desideratum. If a class assignment is not necessary and only separating test patterns into two disjoint groups is the goal, we may simply set $m_1 = n_{\mathrm{pair}}$ and omit Step 2 of finding $\widehat{s}$.

Remark 5.4 (When $\pi = \frac{1}{2}$)**.** *With only pairwise supervision, a class assignment is hopeless because both classes are essentially symmetric; however, it is still possible to draw a decision boundary. A class assignment with pointwise supervision is still possible.*

5.4.4 Class Assignment is Impossible with Only Unlabeled Data

In this subsection, we discuss the impossibility of recovering a class assignment with only unlabeled validation data. Given a real-valued prediction function $f : \mathcal{X} \to \mathbb{R}$ and the class prior π_+, we consider the following class assignment strategy with only unlabeled data, instead of the proposed method:

$$\widetilde{s} \coloneqq \mathrm{sgn}\left(2\pi_+ - 1\right) \cdot \mathrm{sgn}\left(\mathop{\mathbb{E}}_{\mathsf{X} \sim p(\mathsf{X})}\left[\mathrm{sgn}\left(f(\mathsf{X})\right)\right]\right).$$

Our aim is to estimate the optimal class assignment

$$s^* = \operatorname*{arg\,min}_{s \in \{\pm 1\}} R_{\mathrm{point}}\left(s \cdot \mathrm{sgn} \circ f\right).$$

In the following lemma, we claim that the alternative estimator $\widetilde{s}$ cannot recover s^* even with access to an infinite number of data.

Lemma 5.3. *There exists an underlying distribution such that $\widetilde{s} \neq s^*$.*

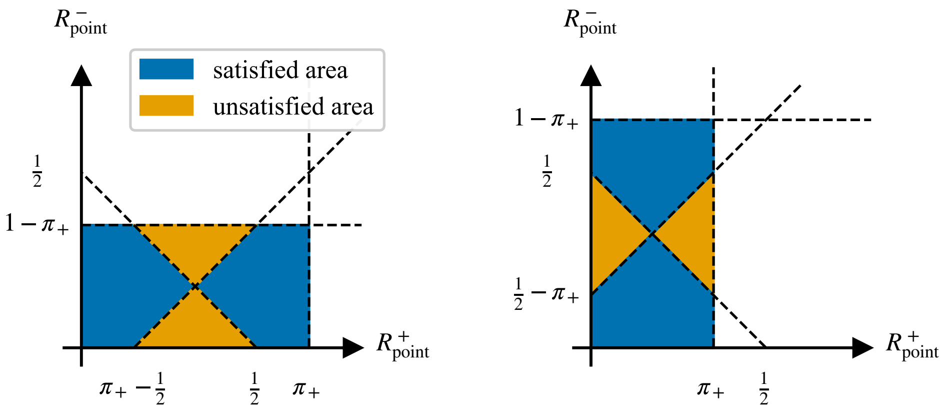
Proof. The optimal class assignment s^* can be expressed by

$$\begin{aligned} s^* &= \mathrm{sgn}\left(R_{\mathrm{point}}\left(-\mathrm{sgn} \circ f\right) - R_{\mathrm{point}}\left(\mathrm{sgn} \circ f\right)\right) \\ &= \mathrm{sgn}\left(\left(1 - R_{\mathrm{point}}\left(\mathrm{sgn} \circ f\right)\right) - R_{\mathrm{point}}\left(\mathrm{sgn} \circ f\right)\right) \\ &= \mathrm{sgn}\left(1 - 2R_{\mathrm{point}}(\mathrm{sgn} \circ f)\right). \end{aligned}$$

Thus, the following condition is necessary and sufficient for $\widetilde{s} = s^*$:

$$\underbrace{\mathrm{sgn}\left(2\pi_+ - 1\right) \cdot \mathrm{sgn}\left(\mathop{\mathbb{E}}_{\mathsf{X} \sim p(\mathsf{X})}\left[\mathrm{sgn}\left(f(\mathsf{X})\right)\right]\right)}_{=\widetilde{s}} \cdot \underbrace{\mathrm{sgn}\left(1 - 2R_{\mathrm{point}}(\mathrm{sgn} \circ f)\right)}_{=s^*} > 0. \tag{5.8}$$

[2] The independent two sets are necessary; otherwise, the errors of Steps 1 and 2 correlate, which leads to an overfitting. Technically, they are required because Theorem 5.7 relies on the union bound.



(a) $\pi_+ > \frac{1}{2}$

(b) $\pi_+ < \frac{1}{2}$

Figure 5.1: Illustration of the areas corresponding to the condition (5.9) (highlighted in blue). $\widetilde{s} = s^*$ in the blue areas; other cases are shown in the orange areas.

We will investigate whether this condition always holds. Denote the following:

$$R^+_{\text{point}} \coloneqq \pi_+ \mathop{\mathbb{E}}_{\mathsf{X}\sim p(\mathsf{X}|\mathsf{Y}=+1)} \left[\mathbb{1}_{\{\text{sgn}(f(\mathsf{X}))\neq +1\}}\right],$$
$$R^-_{\text{point}} \coloneqq (1-\pi_+) \mathop{\mathbb{E}}_{\mathsf{X}\sim p(\mathsf{X}|\mathsf{Y}=-1)} \left[\mathbb{1}_{\{\text{sgn}(f(\mathsf{X}))\neq -1\}}\right].$$

Note that $0 \leq R^+_{\text{point}} \leq \pi_+$ and $0 \leq R^-_{\text{point}} \leq 1-\pi_+$ always hold. Now, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathsf{X}\sim p(\mathsf{X})} \left[\mathbb{1}_{\{\text{sgn}(f(\mathsf{X}))=+1\}}\right] \\
&\quad = \pi_+ \mathop{\mathbb{E}}_{\mathsf{X}|\mathsf{Y}=+1} \left[\mathbb{1}_{\{\text{sgn}(f(\mathsf{X}))=+1\}}\right] + (1-\pi_+) \mathop{\mathbb{E}}_{\mathsf{X}|\mathsf{Y}=-1} \left[\mathbb{1}_{\{\text{sgn}(f(\mathsf{X}))=+1\}}\right] \\
&\quad = \pi_+ \mathop{\mathbb{E}}_{\mathsf{X}|\mathsf{Y}=+1} \left[\left(1 - \mathbb{1}_{\{\text{sgn}(f(\mathsf{X}))=-1\}}\right)\right] + (1-\pi_+) \mathop{\mathbb{E}}_{\mathsf{X}|\mathsf{Y}=-1} \left[\mathbb{1}_{\{\text{sgn}(f(\mathsf{X}))=+1\}}\right] \\
&\quad = -\pi_+ \mathop{\mathbb{E}}_{\mathsf{X}|\mathsf{Y}=+1} \left[\mathbb{1}_{\{\text{sgn}(f(\mathsf{X}))=-1\}}\right] + (1-\pi_+) \mathop{\mathbb{E}}_{\mathsf{X}|\mathsf{Y}=-1} \left[\mathbb{1}_{\{\text{sgn}(f(\mathsf{X}))=+1\}}\right] + \pi_+ \\
&\quad = -R^+_{\text{point}} + R^-_{\text{point}} + \pi_+.
\end{aligned}$$

Similarly, we have

$$\mathop{\mathbb{E}}_{\mathsf{X}\sim p(\mathsf{X})} \left[\mathbb{1}_{\{\text{sgn}(f(\mathsf{X}))=-1\}}\right] = R^+_{\text{point}} - R^-_{\text{point}} + 1 - \pi_+.$$

By combining them, the following expression can be obtained.

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathsf{X}\sim p(\mathsf{X})} \left[\text{sgn}\left(f(\mathsf{X})\right)\right] &= \mathop{\mathbb{E}}_{\mathsf{X}\sim p(\mathsf{X})} \left[\mathbb{1}_{\{\text{sgn}(f(\mathsf{X}))=+1\}}\right] - \mathop{\mathbb{E}}_{\mathsf{X}\sim p(\mathsf{X})} \left[\mathbb{1}_{\{\text{sgn}(f(\mathsf{X}))=-1\}}\right] \\
&= -2R^+_{\text{point}} + 2R^-_{\text{point}} + 2\pi_+ - 1.
\end{aligned}$$

Hence, the necessary and sufficient condition (5.8) is rewritten as

$$\begin{aligned}
&\text{sgn}\left(2\pi_+ - 1\right) \\
&\times \text{sgn}\left(-2R^+_{\text{point}} + 2R^-_{\text{point}} + 2\pi_+ - 1\right) \\
&\times \text{sgn}\left(1 - 2R^+_{\text{point}} - 2R^-_{\text{point}}\right) > 0.
\end{aligned} \tag{5.9}$$

This condition is satisfied when π_+, R^+_{point}, and R^-_{point} satisfy any of the following conditions.

Table 5.1: Comparison of closely related methods used to train classifiers with pairwise supervision. The column "$\pi_+ = \frac{1}{2}$" shows whether the formulation is valid under $\pi_+ = \frac{1}{2}$. In the *sample complexity*, m denotes the number of paired data in Step 1, and either paired or pointwise data in Step 2. The sample complexity analysis of Step 1 is with respect to either pointwise classification or clustering error. In OVPC, the asymptotic convergence analysis is provided for Step 1 but no sample complexity analysis is provided in Zhang and Yan [2007]. To achieve a proper comparison, we assume that the hinge loss is used and eventually the ψ-transform is $\psi(u) = u$. This is detailed in Section 5.5 (Discussion).

		Sample complexity of		
	$\pi_+ = \frac{1}{2}$	Similarity learning (Step 1)	Post-process (Step 2)	Comment
CIPS (Ours)	✓	$O_p(m^{-\frac{1}{4}})$ (Lemma 5.5)	$O_p(e^{-m})$ (Lemma 5.6)	Step 2 is class assignment.
OVPC [Zhang and Yan, 2007]	✓	(N/A)	$O_p(e^{-m})$	Step 2 is a class assignment.
SLLC [Bellet et al., 2012]	✓	$O_p(m^{-\frac{1}{4}})$	$O_p(m^{-\frac{1}{2}})$	Step 2 is SVM training.
MCL [Hsu et al., 2019]	✓	(N/A)	(N/A)	Inner product of classifiers is fitted in Step 1.
SD [Shimada et al., 2021]	✗	$O_p(m^{-\frac{1}{2}})$	(unnecessary)	Step 1 trains the classifiers directly.

- $\pi_+ \geq \frac{1}{2}$, $R^-_{\text{point}} \geq R^+_{\text{point}} + \frac{1}{2} - \pi_+$, and $R^-_{\text{point}} \leq -R^+_{\text{point}} + \frac{1}{2}$,
- $\pi_+ \geq \frac{1}{2}$, $R^-_{\text{point}} < R^+_{\text{point}} + \frac{1}{2} - \pi_+$, and $R^-_{\text{point}} > -R^+_{\text{point}} + \frac{1}{2}$,
- $\pi_+ < \frac{1}{2}$, $R^-_{\text{point}} \geq R^+_{\text{point}} + \frac{1}{2} - \pi_+$, and $R^-_{\text{point}} > -R^+_{\text{point}} + \frac{1}{2}$,
- $\pi_+ < \frac{1}{2}$, $R^-_{\text{point}} < R^+_{\text{point}} + \frac{1}{2} - \pi_+$, and $R^-_{\text{point}} \leq -R^+_{\text{point}} + \frac{1}{2}$.

The conditions (5.9) are depicted in Figure 5.1. As can be seen from this figure, for any binary classification problem (i.e., for any π_+), there exists a case in which the class assignment with unlabeled data fails ($\widetilde{s} \neq s^*$). □

5.4.5 Benefits of Our Formulation over Existing Similarity Learning

We reiterate that similarity learning in our formulation directly elicits a boundary and the post-process (class assignment) is cheaper than training a classifier. Indeed, classifier training requires the usual $O_p(m^{-\frac{1}{2}})$ sample complexity [Bellet et al., 2012, Theorem 1]. Table 5.1 provides an overview of the comparison with previous related studies. We remark that the sample complexity of SLLC is transformed into the complexity in terms of paired data (Step 1) from the original complexity in the pointwise data [Bellet et al., 2012, Theorem 3].[3] In addition, although our Step 1 is worse than SD, our formulation is valid even when $\pi_+ = \frac{1}{2}$ with pointwise supervision. Subsequently, we discuss the other perspectives of our formulation.

Generalization in terms of surrogate losses. Several existing formulations can be regarded as special cases of our formulation (5.5). Kernel alignment [Cristianini et al., 2002] measures the similarity between a kernel and a target function, as defined by the cosine similarity. If the product $f(\mathbf{x}) \cdot f(\mathbf{x}')$ is used as a kernel and the similarity between data points is regarded as a target function, a kernel alignment is equivalent (up to the normalization factor) to minimizing

[3] Given m pointwise data, $O(m^2)$ pairs can be generated and thereby the sample complexity is transformed.

Equation (5.5) with the linear loss $\phi_{\text{lin}}(z,t) := -zt$. By contrast, ($\varepsilon$,$\gamma$,$\tau$)-good similarity [Balcan et al., 2008] regards a similarity function inducing a good linear separator as a good similarity. Here, the linear separability is defined through the hinge loss $\phi_{\text{hinge}}(z,t) := [1-zt]_+$. Bellet et al. [2012] formulated the learning of a bilinear similarity $\mathbf{x}^\top A\mathbf{x}'$ by minimizing the hinge loss, which is equivalent to the minimization of Equation (5.5) with ϕ_{hinge} and the choice of $A = \mathbf{w}\mathbf{w}^\top$ such that $f(\mathbf{x}) = \mathbf{w}^\top\mathbf{x}$. In addition to these examples, recent contrastive learning [Logeswaran and Lee, 2018, Saunshi et al., 2019] can be regarded as learning the good similarity $f(\mathbf{x}) \cdot f(\mathbf{x}')$ by minimizing the logistic loss $\phi_{\log}(z,t) := \log(1+e^{-zt})$, although they are subtly different in that triplets (or N-tuples, in general) are used as supervision.

Explicit relation to classification. Hsu et al. [2019] formulated similarity learning in a slightly different way, i.e., as a maximum likelihood estimation of the pairwise label $S_\tau := \frac{\tau+1}{2}$:[4]

$$\min_{f\in\mathcal{F}} \frac{1}{m_1} \sum_{(\mathbf{x},\mathbf{x}',\tau)\in\mathcal{S}_1} -S_\tau \log(\widetilde{q}(f(\mathbf{x}), f(\mathbf{x}')) - (1-S_\tau)\log(1-\widetilde{q}(f(\mathbf{x}), f(\mathbf{x}'))), \quad (5.10)$$

where

$$\widetilde{q}(z,z') := \begin{bmatrix} q(z) \\ 1-q(z) \end{bmatrix}^\top \begin{bmatrix} q(z') \\ 1-q(z') \end{bmatrix}$$

is the inner product of two binary probability vectors, and $q(z) := (1+\exp(-z))^{-1}$ denotes the logistic model. By contrast, our formulation (5.6) with the logistic loss $\phi_{\log}(z,t) = -S_t \log(q(z)) - (1-S_t)\log(1-q(z))$ is

$$\min_{f\in\mathcal{F}} \frac{1}{m_1} \sum_{(\mathbf{x},\mathbf{x}',\tau)\in\mathcal{S}_1} -S_\tau \log(q(f(\mathbf{x}) \cdot f(\mathbf{x}'))) - (1-S_\tau)\log(1-q(f(\mathbf{x}) \cdot f(\mathbf{x}'))). \quad (5.11)$$

In the formulation (5.10), the similarity is defined by the inner product of the class probabilities, whereas it is defined by the inner product of f in the formulation (5.11). The latter definition is often called the *inner product similarity* (IPS) model [Okuno and Shimodaira, 2020].[5] Although both are valid similarity learning methods, the IPS model (5.11) would be a more natural extension of the classification risk minimization, i.e., one can choose arbitrary loss functions; in addition, the pairwise classification risk minimization (5.6) admits an excess risk bound (Lemma 5.4 in Section 5.5). For this reason, we call our formulation a *Classifier with Inner Product Similarity (CIPS)* from this point on.

5.5 Excess Risk and Sample Complexity Analysis

In this section, we provide the missing sample complexity analyses of CIPS in Table 5.1. The proofs for the lemmas and theorems in this section are deferred to

[4] The multi-class formulation in Hsu et al. [2019] was simplified in binary classification herein for comparison.

[5] The IPS model originally defined similarity between two vector data representations, and hence is called an *inner* product similarity. However, the IPS model is applied to a one-dimensional prediction $f(\mathbf{x})$ in our context. The IPS model has been used in several domains [Tang et al., 2015, Logeswaran and Lee, 2018, Saunshi et al., 2019, Okuno and Shimodaira, 2020].

Section 5.7. In addition, the excess risk is obtained to claim that CIPS does solve binary classification. Let $\widehat{f}$ and $\widehat{s}$ be the solutions to Equations (5.6) and (5.7), respectively. The excess risk for similarity learning is denoted by

$$\varepsilon(\widehat{s}, \widehat{f}) \coloneqq R_{\text{point}}(\widehat{s} \cdot \operatorname{sgn} \circ \widehat{f}) - R^*_{\text{point}}, \tag{5.12}$$

where $R^*_{\text{point}} \coloneqq \inf_f R_{\text{point}}(\operatorname{sgn} \circ f)$, and $\inf_f$ indicates the infimum over all measurable functions. To derive the excess risk bound on Equation (5.12), we need to handle errors in clustering error minimization and class assignment independently, which are described in Lemmas 5.5 and 5.6, respectively. As an important insight when combining two errors, if the class assignment is successful, the excess risk $\varepsilon(\widehat{s}, \widehat{f})$ is equivalent to the one with respect to the clustering error minimization. That is,

$$\widehat{s} = \mathop{\arg\min}_{s \in \{\pm 1\}} R_{\text{point}}(s \cdot \operatorname{sgn} \circ \widehat{f}) \implies \varepsilon(\widehat{s}, \widehat{f}) = R_{\text{clus}}(\operatorname{sgn} \circ \widehat{f}) - R^*_{\text{clus}}, \tag{5.13}$$

where $R^*_{\text{clus}} \coloneqq \inf_f R_{\text{clus}}(\operatorname{sgn} \circ f)$. To derive a probabilistic guarantee for $R_{\text{clus}}(\operatorname{sgn} \circ \widehat{f}) - R^*_{\text{clus}}$, we use the Rademacher complexity [Bartlett and Mendelson, 2002] defined on the class $\{ (\mathbf{x}, \mathbf{x}') \mapsto f(\mathbf{x}) \cdot f(\mathbf{x}') \mid f \in \mathcal{F} \}$

$$\mathfrak{R}_m(\mathcal{F}) \coloneqq \mathop{\mathbb{E}}_{\{(\mathbf{x}_i, \mathbf{x}'_i)\}} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(\mathbf{x}_i) \cdot f(\mathbf{x}'_i) \right],$$

where $\{\sigma_i\}_{i=1}^m$ are Rademacher variables. Before obtaining an excess risk bound of R_{clus}, we need to bridge R_{pair} and the surrogate risk R^ϕ_{pair}.

Lemma 5.4. *If a surrogate loss ϕ is classification-calibrated [Bartlett et al., 2006], then there exists a convex, non-decreasing, and invertible $\psi : [0, 1] \to [0, +\infty)$ such that for any sequence (u_i) in $[0, 1]$,*

$$\psi(u_i) \to 0 \quad \textit{if and only if} \quad u_i \to 0,$$

and for any measurable function f and probability distribution on $\mathcal{X} \times \mathcal{Y}$,

$$\psi\left(R_{\text{pair}}(\operatorname{sgn} \circ f) - R^*_{\text{pair}}\right) \le R^\phi_{\text{pair}}(f) - R^{\phi,*}_{\text{pair}},$$

where $R^*_{\text{pair}} \coloneqq \inf_f R_{\text{pair}}(\operatorname{sgn} \circ f)$ and $R^{\phi,*}_{\text{pair}} \coloneqq \inf_f R^\phi_{\text{pair}}(f)$.

Although this lemma is similar to Bartlett et al. [2006, Theorem 1], the proof for R_{pair}, instead of for R_{point}, requires special care to properly treat the product of the prediction functions.

Then, the excess risk bound for R_{clus} is derived based on Lemma 5.4 and the uniform bound.

Lemma 5.5. *Let $f^* \in \mathcal{F}$ be a minimizer of R^ϕ_{pair}, and $\widehat{f} \in \mathcal{F}$ be a minimizer of $\widehat{R}^\phi_{\text{pair}}$ defined in Equation (5.6). Assume that $\phi(\cdot, \pm 1)$ is ρ-Lipschitz ($0 < \rho < \infty$), and that $\|f\|_\infty \le C_b$ for any $f \in \mathcal{F}$ for some C_b. Let $C_\phi \coloneqq \sup_{t \in \{\pm 1\}} \phi(C_b^2, t)$. For any $\delta > 0$, with a probability of at least $1 - \delta$,*

$$\begin{aligned} &R_{\text{clus}}(\operatorname{sgn} \circ \widehat{f}) - R^*_{\text{clus}} \\ &\qquad \le \sqrt{\frac{1}{2} \psi^{-1} \left(R^\phi_{\text{pair}}(f^*) - R^{\phi,*}_{\text{pair}} + 4\rho \mathfrak{R}_{m_1}(\mathcal{F}) + \sqrt{\frac{2C_\phi^2 \log(2/\delta)}{m_1}} \right)}. \end{aligned}$$

Next, the class assignment error probability using pairwise supervision is analyzed.

Lemma 5.6. *Assume that $\pi_+ \neq \frac{1}{2}$. Let $\widehat{s}$ be the solution defined in Equation (5.7). We then have*

$$\Pr\left(\widehat{s} \neq \operatorname*{arg\,min}_{s\in\{\pm 1\}} R_{\text{point}}(s \cdot \operatorname{sgn} \circ \widehat{f})\right)$$
$$\leq \exp\left(-\frac{m_2}{2}(2\pi_+ - 1)^2\left(2R_{\text{point}}(\operatorname{sgn} \circ \widehat{f}) - 1\right)^2\right).$$

Several observations from Lemma 5.6 follow. As $\pi_+ \to \frac{1}{2}$, the upper bound becomes looser. This is derived from the fact that the estimation of the pointwise classification error with pairwise supervision becomes more difficult as $\pi_+ \to \frac{1}{2}$ [Shimada et al., 2021]. Moreover, the discriminability of function $\widehat{f}$, i.e., $R_{\text{point}}(\operatorname{sgn} \circ \widehat{f})$, appears in the inequality and thus is directly related to the error rate. Intuitively, if a given function classifies a large portion of data correctly, the optimal sign can be identified easily.

Finally, an overall excess risk bound is derived by combining Lemmas 5.5 and 5.6 and the fact (5.13).

Theorem 5.7. *Suppose that we have $\pi_+ \neq \frac{1}{2}$. Let*

$$r := \exp\left(-\frac{m_2}{2}(2\pi_+ - 1)^2(2R_{\text{point}}(\operatorname{sgn} \circ \widehat{f}) - 1)^2\right).$$

Under the same assumptions as Lemma 5.5, for any $\delta > r$, with a probability of at least $1 - \delta$,

$$R_{\text{point}}(\widehat{s} \cdot \operatorname{sgn} \circ \widehat{f}) - R^*_{\text{point}}$$
$$\leq \sqrt{\frac{1}{2}\psi^{-1}\left(R^{\phi}_{\text{pair}}(f^*) - R^{\phi,*}_{\text{pair}} + 4\rho\mathfrak{R}_{m_1}(\mathcal{F}) + \sqrt{\frac{2C_\phi^2 \log\frac{2}{\delta - r}}{m_1}}\right)}.$$

If $\mathfrak{R}_{m_1}(\mathcal{F}) = o(1)$, the upper bound asymptotically approaches the approximation error (i.e., $R^{\phi}_{\text{pair}}(f^*) - R^{\phi,*}_{\text{pair}}$) in probability. For example, linear-in-parameter model $\mathcal{F} = \left\{ f(\mathbf{x}) = \mathbf{w}^\top \boldsymbol{\varphi}(\mathbf{x}) + b \right\}$ satisfies $\mathfrak{R}_{m_1}(\mathcal{F}) = O({m_1}^{-\frac{1}{2}})$, as shown in Kuroki et al. [2019, Lemma 5], where $\mathbf{w} \in \mathbb{R}^k$ and $b \in \mathbb{R}$ are weights and bias parameters, and $\boldsymbol{\varphi} : \mathbb{R}^d \to \mathbb{R}^k$ are mapping functions. Note that our result is stronger than Zhang and Yan [2007] because they only provided the asymptotic convergence, whereas Theorem 5.7 provides a finite sample guarantee.

Discussion. Because a class assignment admits the exponential decay of the error probability (Lemma 5.6) under the moderate condition ($\pi_+ \neq \frac{1}{2}$), we may set $m_2 \ll m_1$ in practice. By contrast, our excess risk bound of clustering error minimization (Lemma 5.5) is governed in part by a ψ-transform. The explicit rate requires specific choices of loss functions: e.g., the hinge loss gives $\psi(u) = u$. Hence, under the assumption $\mathfrak{R}_{m_1}(\mathcal{F}) = O({m_1}^{-\frac{1}{2}})$, the explicit rate is $O_p({m_1}^{-\frac{1}{4}})$ for the hinge loss.[6] This rate is no slower than the pointwise supervised case

[6] As another example, the logistic loss gives $\psi(u) = \Omega(u^2)$, entailing the explicit rate $O_p({m_1}^{-\frac{1}{8}})$ for the excess risk bound (Lemma 5.5). For more examples of ψ, please refer to Steinwart [2007, Table 1].

$O_p(m^{-\frac{1}{2}})$ because $O(m^2)$ pairwise supervision can be generated if m pointwise labels are available.

Note again that CIPS assumes $\pi_+ \neq \frac{1}{2}$ only in a class assignment (Step 2 & Lemma 5.6), not in the clustering error minimization (Step 1 & Lemma 5.5). This is a subtle but notable difference from earlier similarity learning methods based on unbiased classification risk estimators, which requires $\pi_+ \neq \frac{1}{2}$ even under risk minimization (see, e.g., Shimada et al. [2021, Theorem 3]).

Our excess risk bound (Theorem 5.7) resembles transfer bounds among binary classification, class probability estimation (CPE), and bipartite ranking. Narasimhan and Agarwal [2013] reduced classification and CPE to ranking and showed that the excess risks of both classification and CPE can be bounded from above through that of ranking. As can be seen in Narasimhan and Agarwal [2013, Theorems 4 and 14], the excess risk of classification/CPE slows down to be $O(\lambda(m)^{-\frac{1}{2}})$ when supposing that the excess risk of ranking is $\lambda(m)$. The same decay is observed in Theorem 5.7 as well, reducing classification to similarity learning. This decay $O((\cdot)^{-\frac{1}{2}})$ can be regarded as a *cost arising from a problem reduction.*

5.6 Experiments

This section describes the simulation results confirming our findings.

⟨♣⟩ Sample complexity of the clustering error minimization through similarity learning (Lemma 5.5).

⟨♡⟩ The class-prior effect in similarity learning (Discussion in Section 5.5).

⟨♠⟩ Class assignment without pointwise supervision (Lemma 5.6).

In addition, we conducted a comparison with the baselines using benchmark and real-world datasets (PubMed-Diabetes). All experiments except PubMed-Diabetes were carried out using a 3.60-GHz Intel® Core™ i7-7700 CPU and a GeForce GTX 1070. Experiments using the PubMed-Diabetes dataset were carried out using 1.40-GHz Intel® Xeon Phi™ 7250 CPU.

5.6.1 Clustering Error Minimization on Benchmark Datasets

Tabular datasets from LIBSVM [Chang and Lin, 2011] and UCI [Dua and Graff, 2017] repositories and three image classification datasets, i.e., MNIST [LeCun, 2013], Fashion-MNIST [Xiao et al., 2017], and Kuzushiji-MNIST [Clanuwat et al., 2018] were used for the benchmarks. For the image classification datasets, the original ten class categories were converted into positive/negative labels by grouping even/odd class labels. Pairwise supervision was generated through a random coupling of the pointwise data in the original datasets. We describe the details of the baselines and the implementation details below.

We briefly introduce the baselines as follows.

- CIPS (Ours): The empirical pairwise classification risk R_{pair} (5.6) was computed with the logistic loss. The linear model $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ was applied. The risk was optimized using the stochastic gradient descent with the following hyperparameters.
  - minibatch size: 64
  - learning rate: 10^{-2}

  - ℓ_2-regularization parameter: 10^{-4}
  - training epochs: 500

- MCL (Meta-Classification Likelihood) [Hsu et al., 2019]: This is an approach based on the maximum likelihood estimation over pairwise labels. The loss function is based on the maximum likelihood, that is, the logistic loss as in the original study. The model and optimization setup were the same as those of CIPS.

- Similar-Dissimilar (SD) classification [Shimada et al., 2021]: This is a classification method using pairwise supervision, based on an unbiased risk estimator of the classification risk. Their proposed classification risk was computed using the logistic loss. The model and optimization setup were the same as those of CIPS.

- OVPC [Zhang and Yan, 2007]: This is another classification method using pairwise supervision. We followed the authors and used the squared loss and evaluated the closed-form minimizer.

- Semi-supervised Spectral (SSP) clustering [von Luxburg, 2007]: This is a semi-supervised clustering method based on spectral clustering. Pairwise data were used as hard constraints. To construct the neighborhood sets for the Laplacian matrix, 5-nearest neighbors were used. The features are obtained through a propagation of the constraints. To apply the final k-means clustering on the obtained features, scikit-learn implementation [Pedregosa et al., 2011] was used with the default parameters.

- Constrained k-Means (CKM) clustering [Wagstaff et al., 2001]: This is another semi-supervised clustering method, in which k-means clustering is conducted iteratively until hard constraints are satisfied. Pairwise data were used as the hard constraints. Clustering was carried out using 10 different random initializations, and the best one was reported. For each initialization, the number of maximum iterations was set to 300, and the tolerance parameter was set to 10^{-4}.

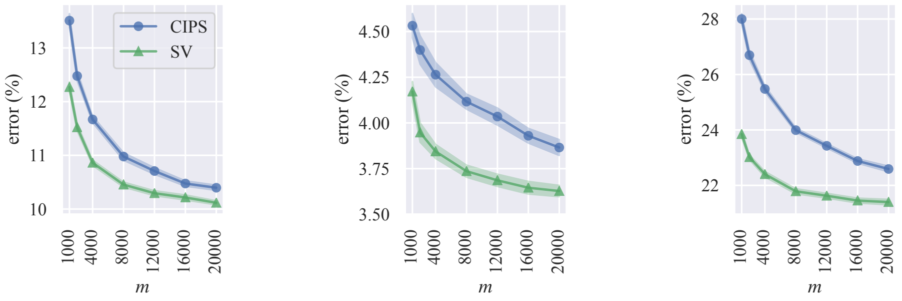
- k-Means (KM) clustering [MacQueen, 1967]: Pairwise data were used for training without all link information. We used the k-means clustering implementation provided by scikit-learn [Pedregosa et al., 2011] with the default parameters.

- Supervised (SV): The true class labels were revealed during training. The model and optimization setup were the same as those of CIPS.

⟨♣⟩ First, noting the sample complexity behavior in Lemma 5.5, the classifiers were trained using MNIST, Fashion-MNIST, and Kuzushiji-MNIST. The numbers of pairwise data m were set to each of

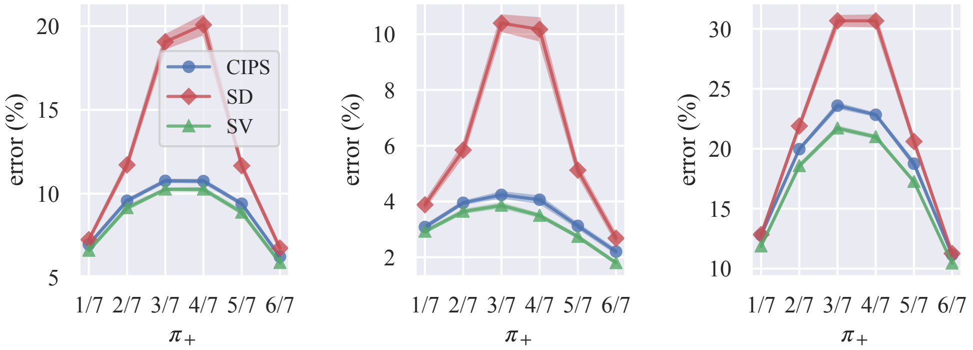
$$m \in \{ 1{,}000, 2{,}000, 4{,}000, 8{,}000, 12{,}000, 16{,}000, 20{,}000 \} .$$

Figure 5.2 presents the performances of CIPS and SV. This demonstrates that the clustering error of CIPS constantly decreases as m increases, which is consistent with Lemma 5.5. Moreover, CIPS performed more efficiently than expected in terms of the sample complexity, i.e., as we discussed in Section 5.5, we expected that CIPS with $O(m^2)$ pairs would perform comparably to SV with m data points.



(a) MNIST **(b)** Fashion-MNIST **(c)** Kuzushiji-MNIST

Figure 5.2: Mean clustering error and standard error (shaded areas) over 20 trials on image classification datasets.



(a) MNIST **(b)** Fashion-MNIST **(c)** Kuzushiji-MNIST

Figure 5.3: Mean clustering error and standard error (shaded areas) over 10 trials on image classification datasets under controlled class priors.

⟨♡⟩ Next, to see the effect of the class prior, we compared CIPS, SD, and SV with various class priors. During this experiment, training and test data were generated from MNIST under the controlled class prior π_+, where π_+ was set to each of $\{\frac{1}{7}, \ldots, \frac{6}{7}\}$ individually. For each trial, 10,000 pairs were randomly subsampled from MNIST for training and the performance was evaluated with another 10,000 labeled examples. The average clustering errors and standard errors over 10 trials are plotted in Figure 5.3. This result indicates that CIPS is less affected in comparison with the SD.

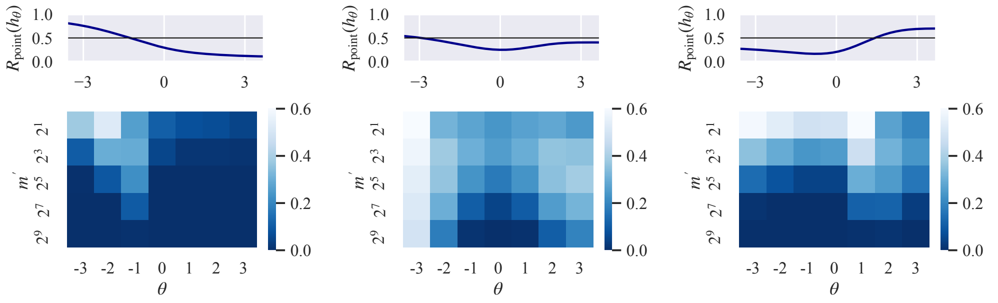
Finally, we show the benchmark performances of each method on the tabular datasets in Table 5.2, where each cell contains the average clustering error and the standard error over 20 trials. For each trial, we randomly subsampled $m \in \{100, 1000\}$ pairs for the training data and 1,000 pointwise examples for the evaluation. This result demonstrates that CIPS performs better with a sufficient number of data than most of the baselines and comparably to MCL. In particular, the performance difference between CIPS and the clustering methods implies that larger samples do improve the downstream classification performance of CIPS thanks to its generalization guarantee (Theorem 5.7).

5.6.2 Class Assignment on Synthetic Dataset

The performance of the proposed class assignment method was empirically investigated on synthetic dataset. The class-conditional distributions with the standard Gaussian distributions were used as the underlying distribution: $p(x|y = +1) = \mathcal{N}(x \mid \mu_+, \sigma_+)$ and $p(x|y = -1) = \mathcal{N}(x \mid \mu_-, \sigma_-)$. Throughout this experiment,

Table 5.2: Mean clustering error and standard error on different benchmark datasets over 20 trials. Bold font indicates outperforming methods, which were chosen by a one-sided t-test with a significance level of 5%.

dataset (dim., π_+)	m	CIPS (Ours)	MCL	SD	OVPC	SSP	CKM	KM	(SV)
adult	100	39.8 (1.6)	38.4 (2.1)	30.8 (0.9)	45.0 (0.9)	**24.7 (0.3)**	28.9 (0.8)	**24.9 (0.5)**	21.9 (0.4)
(123, 0.24)	500	21.5 (1.0)	**19.3 (0.4)**	23.2 (0.4)	44.7 (0.9)	24.3 (0.3)	28.2 (0.4)	27.5 (0.5)	16.9 (0.3)
	1000	**17.6 (0.3)**	**17.2 (0.3)**	20.5 (0.3)	45.5 (0.7)	24.2 (0.3)	27.9 (0.4)	27.9 (0.5)	15.9 (0.3)
banana	100	**43.6 (0.6)**	**44.5 (0.6)**	45.3 (0.6)	46.0 (0.7)	**43.0 (1.0)**	46.4 (0.7)	45.8 (0.7)	44.6 (0.6)
(2, 0.45)	500	43.1 (0.8)	43.3 (0.6)	45.1 (0.7)	46.0 (0.7)	**14.3 (0.7)**	45.5 (0.6)	44.4 (0.4)	45.1 (0.6)
	1000	44.4 (0.6)	44.3 (0.7)	44.4 (0.5)	46.2 (0.5)	**11.0 (0.2)**	45.0 (0.7)	44.0 (0.3)	45.1 (0.7)
codrna	100	**24.7 (1.8)**	32.3 (1.4)	**28.0 (1.3)**	32.0 (2.0)	45.5 (1.5)	46.7 (0.6)	42.5 (1.0)	11.0 (0.6)
(8, 0.33)	500	**6.4 (0.2)**	10.6 (0.3)	12.0 (0.6)	28.0 (2.1)	48.6 (0.3)	46.2 (0.3)	44.0 (0.7)	6.6 (0.2)
	1000	**6.3 (0.2)**	**6.5 (0.2)**	8.8 (0.4)	28.3 (2.0)	44.8 (1.6)	46.1 (0.4)	45.4 (0.6)	6.3 (0.2)
ijcnn1	100	16.6 (2.3)	24.9 (2.9)	**10.7 (0.3)**	41.1 (1.1)	31.6 (2.0)	40.0 (1.3)	31.9 (2.4)	9.1 (0.2)
(22, 0.10)	500	**7.7 (0.2)**	8.2 (0.2)	8.3 (0.2)	41.6 (1.3)	33.0 (2.5)	45.4 (0.8)	41.7 (0.7)	7.9 (0.2)
	1000	**7.7 (0.2)**	**7.9 (0.2)**	**8.1 (0.2)**	42.0 (1.4)	34.9 (1.7)	45.9 (0.8)	43.4 (0.7)	7.6 (0.2)
magic	100	**24.9 (1.3)**	**28.7 (1.8)**	30.7 (1.3)	41.9 (1.0)	47.1 (0.5)	45.5 (1.2)	44.0 (1.2)	21.8 (0.4)
(10, 0.35)	500	**21.5 (0.3)**	**21.3 (0.3)**	25.5 (0.8)	39.6 (1.5)	46.8 (0.5)	46.8 (0.4)	44.4 (0.4)	20.8 (0.3)
	1000	**21.3 (0.3)**	**20.9 (0.3)**	23.8 (0.4)	39.5 (1.7)	43.6 (0.9)	46.8 (0.3)	44.6 (0.4)	20.7 (0.3)
phishing	100	**12.7 (2.3)**	**12.8 (2.3)**	34.6 (1.8)	41.7 (1.0)	46.6 (0.5)	24.4 (3.4)	47.0 (0.5)	7.6 (0.2)
(44, 0.68)	500	7.2 (0.2)	**6.6 (0.1)**	26.9 (1.4)	42.9 (0.8)	46.0 (0.5)	16.9 (2.6)	46.4 (0.5)	6.5 (0.2)
	1000	**6.5 (0.2)**	**6.3 (0.2)**	22.0 (1.0)	43.8 (1.1)	45.5 (0.5)	15.2 (2.7)	46.4 (0.5)	6.3 (0.2)
phoneme	100	**28.2 (1.2)**	33.1 (1.9)	**29.1 (1.2)**	38.4 (1.3)	31.0 (1.3)	**28.0 (1.0)**	32.9 (1.2)	25.7 (0.4)
(5, 0.71)	500	**25.0 (0.4)**	**24.2 (0.5)**	26.1 (0.6)	38.6 (1.9)	25.5 (0.5)	28.0 (0.8)	32.7 (0.3)	25.0 (0.3)
	1000	**25.2 (0.4)**	**25.0 (0.4)**	26.0 (0.4)	39.8 (1.5)	**24.5 (0.5)**	30.2 (0.6)	32.7 (0.3)	25.3 (0.2)
spambase	100	**13.8 (1.0)**	**13.3 (1.3)**	31.6 (1.5)	39.7 (1.3)	40.5 (0.4)	**15.9 (2.0)**	39.7 (1.3)	10.5 (0.3)
(57, 0.39)	500	9.4 (0.2)	**8.6 (0.2)**	22.6 (0.9)	38.0 (1.6)	40.8 (0.3)	11.5 (0.2)	37.4 (2.3)	8.5 (0.2)
	1000	8.3 (0.2)	**7.6 (0.1)**	19.7 (0.8)	39.3 (1.2)	40.2 (0.4)	11.5 (0.2)	39.7 (1.3)	7.8 (0.2)
w8a	100	31.5 (1.9)	31.4 (2.1)	11.8 (0.3)	39.7 (1.4)	**5.3 (1.2)**	**6.8 (1.9)**	**5.5 (1.3)**	10.3 (0.4)
(300, 0.03)	500	5.6 (0.7)	4.2 (0.5)	**3.2 (0.1)**	38.3 (1.3)	**3.5 (0.1)**	14.0 (3.1)	5.5 (1.1)	2.6 (0.1)
	1000	2.6 (0.2)	**2.2 (0.1)**	2.6 (0.2)	43.1 (0.8)	3.0 (0.1)	8.9 (2.6)	3.7 (0.5)	2.0 (0.1)
waveform	100	**18.2 (0.3)**	**17.7 (0.3)**	26.4 (0.9)	41.9 (1.6)	44.1 (0.6)	41.0 (1.3)	45.1 (0.6)	16.2 (0.2)
(21, 0.33)	500	15.8 (0.2)	**15.1 (0.2)**	20.2 (0.5)	38.9 (1.3)	44.9 (0.7)	45.1 (0.6)	47.1 (0.4)	14.8 (0.2)
	1000	**14.9 (0.2)**	**14.7 (0.2)**	18.4 (0.3)	37.0 (1.7)	45.5 (0.5)	44.9 (0.5)	47.8 (0.4)	14.4 (0.2)



(a) $\pi_+ = 0.1$ **(b)** $\pi_+ = 0.4$ **(c)** $\pi_+ = 0.7$

Figure 5.4: Classification error for each threshold classifier (upper) and the error probability of the proposed class assignment method over 10,000 trials (bottom) on the synthetic Gaussian dataset with $\pi_+ \in \{ 0.1, 0.4, 0.7 \}$.

Table 5.3: Mean clustering error and standard error on the Pubmed-Diabetes dataset over 20 trials. Bold font indicates the outperforming method (excluding SV), which were chosen by the one-sided t-test in the same way as in Table 5.2.

CIPS (Ours)	MCL	DML	(SV)
86.9 (0.4)	**86.6 (0.4)**	85.1 (0.2)	94.7 (0.1)

we fixed $(\mu_+, \sigma_+, \mu_-, \sigma_-)$ to $(1, 1, -1, 2)$. Here, we consider a 1-D thresholded classifier denoted by $h_\theta(x) = 1$ if $x \geq \theta$, and -1, otherwise. Given the class prior $\pi_+ \in (0, 1)$, we generated m' pairwise examples from the above distributions and applied the proposed class assignment method for a fixed classifier h_θ. We then evaluated whether the estimated class assignment is optimal or not. Each parameter was set as follows: $m' \in \{2^1, 2^3, 2^5, 2^7, 2^9\}$, $\pi_+ \in \{ 0.1, 0.4, 0.7 \}$, and $\theta \in \{-3, -2, \ldots, 3\}$. For each (θ, π_+, m'), we repeated these data generation processes, class assignment, and evaluation procedure for 10,000 times.

⟨♠⟩ The error probabilities are depicted in Figure 5.4. We found that the performance of the proposed class assignment method improves as (i) the number of pairwise examples m' grows and (ii) the classification error for a given classifier $R_{\text{point}}(h_\theta)$ moves away from $\frac{1}{2}$. These results are aligned with our analysis in Section 5.5. Moreover, we observed that class assignment improves as the class prior π_+ reaches farther from $\frac{1}{2}$.

5.6.3 Clustering Error Minimization on a Real-world Dataset

Finally, we show experimental results on a citation network dataset, PubMed-Diabetes.[7] The aim of this experiment is to verify whether CIPS is sufficiently robust against real-world noises in pairwise supervision.

The Pubmed-Diabetes dataset is a citation network dataset consisting of 19,717 nodes representing scientific publications related to diabetes and 44,338 (directed) edges representing citing relationships. Each node is described by 500-dimensional TF/IDF features, and categorized into three classes, among which we chose classes 1 (“Diabetes Mellitus, Experimental”) and 3 (“Diabetes Mellitus Type 2”) to convert it into a binary-labeled dataset.

We compare CIPS (proposed) with three baselines, MCL (described above), deep metric learning (DML), and SV (supervised). The implementation details of CIPS and the baselines were as follows.

[7] Available at https://linqs.soe.ucsc.edu/data.

- CIPS (Ours): The 4-layer perceptron (500-8-8-8-1) with the softplus activation [Dugas et al., 2000] was used. The softmax cross entropy was optimized using Adam [Kingma and Ba, 2015], with the following hyperparameters.
    - minibatch size: 4,096
    - learning rate: 10^{-3}
    - training epochs: 100

  The ℓ_2-regularization parameter is chosen from $\{ 10^{-2}, 10^{-4}, 10^{-6} \}$ through a five-fold cross-validation. The early stopping is applied with the patience of 10 epochs. We randomly extracted 20% of the nodes as the test data. The pairwise supervision was generated as follows: First, the edges having both ends in the training data were extracted as similar, and randomly coupled non-connected nodes were then extracted as dissimilar, with the same numbers of similar and dissimilar pairs. Approximately 19,000 pairs were obtained.

- MCL [Hsu et al., 2019]: The setup of model, optimization, and data generation were the same as with CIPS.

- DML [Chopra et al., 2005]: DML combines metric learning and k-means clustering: We first train the embeddings such that their ℓ_2 distances are close for similar pairs and vice versa, and k-means clustering is applied on the embeddings. The metric loss function proposed by Chopra et al. [2005] was used. The model was the same as CIPS except for the last layer, and 8-dimensional outputs of the penultimate layer were used as the embeddings, upon which k-means clustering was performed. The scikit-learn implementation [Pedregosa et al., 2011] of k-means clustering was used with the default parameters. The setup of optimization and data generation setup was the same as in CIPS.

- SV (Supervised): Labeled 7,889 nodes ($\pi_+ \approx 0.65$) were used during training. The setup of model and optimization was the same as CIPS.

The results are reported in Table 5.3, from which we can see that CIPS can train a meaningful classifier even under the presence of real-world noises, and operates comparably to MCL and better than DML.

5.7 Proofs

5.7.1 Proof of Theorem 5.2

The optimal sign s^* can be written as

$$s^* = \operatorname*{arg\,min}_{s \in \{\pm 1\}} R_{\text{point}}(s \cdot h) = \operatorname{sgn}\left(R_{\text{point}}(-h) - R_{\text{point}}(h)\right). \tag{5.14}$$

According to Shimada et al. [2021], R_{point} is equivalently expressed as follows.

Lemma 5.8 (Shimada et al. [2021]). *Assume that $\pi_+ \neq \frac{1}{2}$. The pointwise classification error for a given classifier $h : \mathsf{X} \to \mathsf{Y}$ can then be equivalently represented as*

$$\begin{aligned} &R_{\text{point}}(h) \\ &= \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim p(\mathsf{X},\mathsf{Y})} \mathop{\mathbb{E}}_{(\mathsf{X}',\mathsf{Y}') \sim p(\mathsf{X},\mathsf{Y})} \left[\frac{\mathbb{1}_{\{h(\mathsf{X}) \neq \mathsf{Y}\mathsf{Y}'\}} + \mathbb{1}_{\{h(\mathsf{X}') \neq \mathsf{Y}\mathsf{Y}'\}}}{2\,(2\pi_+ - 1)} \right] - \frac{1 - \pi_+}{2\pi_+ - 1}. \end{aligned} \tag{5.15}$$

By plugging Equation (5.15) into Equation (5.14), we obtain

$$\begin{aligned}
&R_{\text{point}}(-h) - R_{\text{point}}(h) \\
&= \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim p(\mathsf{X},\mathsf{Y})} \mathop{\mathbb{E}}_{(\mathsf{X}',\mathsf{Y}')\sim p(\mathsf{X},\mathsf{Y})} \left[\frac{\mathbb{1}_{\{-h(\mathsf{X})\neq \mathsf{Y}\mathsf{Y}'\}} + \mathbb{1}_{\{-h(\mathsf{X}')\neq \mathsf{Y}\mathsf{Y}'\}}}{2\ (2\pi_+ - 1)}\right] \\
&\quad - \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim p(\mathsf{X},\mathsf{Y})} \mathop{\mathbb{E}}_{(\mathsf{X}',\mathsf{Y}')\sim p(\mathsf{X},\mathsf{Y})} \left[\frac{\mathbb{1}_{\{h(\mathsf{X})\neq \mathsf{Y}\mathsf{Y}'\}} + \mathbb{1}_{\{h(\mathsf{X}')\neq \mathsf{Y}\mathsf{Y}'\}}}{2\ (2\pi_+ - 1)}\right] \\
&= \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim p(\mathsf{X},\mathsf{Y})} \mathop{\mathbb{E}}_{(\mathsf{X}',\mathsf{Y}')\sim p(\mathsf{X},\mathsf{Y})} \left[\frac{1 - 2\cdot\mathbb{1}_{\{h(\mathsf{X})\neq \mathsf{Y}\mathsf{Y}'\}} + 1 - 2\cdot\mathbb{1}_{\{h(\mathsf{X}')\neq \mathsf{Y}\mathsf{Y}'\}}}{2\ (2\pi_+ - 1)}\right] \\
&= \frac{1}{2\pi_+ - 1} \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim p(\mathsf{X},\mathsf{Y})} \mathop{\mathbb{E}}_{(\mathsf{X}',\mathsf{Y}')\sim p(\mathsf{X},\mathsf{Y})} \left[1 - \mathbb{1}_{\{h(\mathsf{X})\neq \mathsf{Y}\mathsf{Y}'\}} - \mathbb{1}_{\{h(\mathsf{X}')\neq \mathsf{Y}\mathsf{Y}'\}}\right] \\
&= \frac{1}{2\pi_+ - 1}(1 - 2Q(h)).
\end{aligned}$$

Thus, we derive the following result:

$$\begin{aligned}
s_h^* &= \operatorname{sgn}\left(R_{\text{point}}(-h) - R_{\text{point}}(h)\right) \\
&= \operatorname{sgn}\left(\frac{1}{2\pi_+ - 1}\right) \cdot \operatorname{sgn}(1 - 2Q(h)) \\
&= \operatorname{sgn}(2\pi_+ - 1) \cdot \operatorname{sgn}(1 - 2Q(h)),
\end{aligned}$$

which completes the proof of Theorem 5.2. Note that s^* can be either ±1 when $Q(h) = \frac{1}{2}$, which is equivalent to $R_{\text{point}}(h) = R_{\text{point}}(-h) = \frac{1}{2}$. Herein, we arbitrarily set $s^* = -\operatorname{sgn}(2\pi_+ - 1)$ in this case. □

5.7.2 Proof of Lemma 5.4

We introduce the following notation:

$$\begin{aligned}
S^{\phi}_{\text{point}}(\alpha, \eta) &:= \eta\phi(\alpha, +1) + (1 - \eta)\phi(\alpha, -1), \\
H^{\phi}_{\text{point}}(\eta) &:= \inf_{\alpha\in\mathbb{R}} S^{\phi}_{\text{point}}(\alpha, \eta), \\
H^{\phi,-}_{\text{point}}(\eta) &:= \inf_{\alpha:\alpha(2\eta-1)\leq 0} S^{\phi}_{\text{point}}(\alpha, \eta).
\end{aligned}$$

Here, S^{ϕ}_{point} represents the conditional ϕ-risk in the following sense:

$$\mathop{\mathbb{E}}_{\mathsf{X}}[S^{\phi}_{\text{point}}(f(\mathsf{X}), p(\mathsf{Y} = +1 \mid \mathsf{X}))] = R^{\phi}_{\text{point}}(f),$$

where

$$R^{\phi}_{\text{point}}(f) := \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim p(\mathsf{X},\mathsf{Y})}\left[\phi(f(\mathsf{X}), \mathsf{Y})\right].$$

Define the function $\psi_{\text{point}} : [0, 1] \to [0, +\infty)$ by $\psi_{\text{point}} = \widetilde{\psi}^{\star\star}_{\text{point}}$, where $\widetilde{\psi}^{\star\star}_{\text{point}}$ is the Fenchel-Legendre biconjugate of $\widetilde{\psi}_{\text{point}}$, and

$$\widetilde{\psi}_{\text{point}}(\varepsilon) := H^{\phi,-}_{\text{point}}\left(\frac{1+\varepsilon}{2}\right) - H^{\phi}_{\text{point}}\left(\frac{1+\varepsilon}{2}\right).$$

In addition, ψ_{point} corresponds exactly to a ψ-transform introduced by Bartlett et al. [2006] exactly.

We will show that the statement of the lemma is satisfied by $\psi = \psi_{\text{point}}$ based on the calibration analysis. We further introduce the following notation:

$$
\begin{aligned}
S_{\text{pair}}(\alpha, \alpha', \eta, \eta') \coloneqq{} & \eta\eta' \mathbb{1}_{\{\operatorname{sgn}(\alpha)\operatorname{sgn}(\alpha') \neq +1\}} \\
& + \eta(1-\eta')\mathbb{1}_{\{\operatorname{sgn}(\alpha)\operatorname{sgn}(\alpha') \neq -1\}} \\
& + (1-\eta)\eta'\mathbb{1}_{\{\operatorname{sgn}(\alpha)\operatorname{sgn}(\alpha') \neq -1\}} \\
& + (1-\eta)(1-\eta')\{\operatorname{sgn}(\alpha)\operatorname{sgn}(\alpha') \neq +1\}, \\
S^{\phi}_{\text{pair}}(\alpha, \alpha', \eta, \eta') \coloneqq{} & \eta\eta'\phi(\alpha\alpha', +1) + \eta(1-\eta')\phi(\alpha\alpha, -1) \\
& + (1-\eta)\eta'\phi(\alpha\alpha, -1) + (1-\eta)(1-\eta')\phi(\alpha\alpha, +1), \\
H_{\text{pair}}(\eta, \eta') \coloneqq{} & \inf_{\alpha, \alpha' \in \mathbb{R}} S_{\text{pair}}(\alpha, \alpha', \eta, \eta'), \\
H^{\phi}_{\text{pair}}(\eta, \eta') \coloneqq{} & \inf_{\alpha, \alpha' \in \mathbb{R}} S^{\phi}_{\text{pair}}(\alpha, \alpha', \eta, \eta').
\end{aligned}
$$

Here, S^{ϕ}_{pair} represents the conditional ϕ-risk in the following sense:

$$
\mathop{\mathbb{E}}_{\mathsf{X},\mathsf{X}'} \left[S^{\phi}_{\text{pair}}(f(\mathsf{X}), f(\mathsf{X}'), p(\mathsf{Y} = +1|\mathsf{X}), p(\mathsf{Y}' = +1|\mathsf{X}')) \right] = R^{\phi}_{\text{pair}}(f),
$$

and

$$
\mathop{\mathbb{E}}_{\mathsf{X},\mathsf{X}'} \left[S_{\text{pair}}(f(\mathsf{X}), f(\mathsf{X}'), p(\mathsf{Y} = +1|\mathsf{X}), p(\mathsf{Y}' = +1|\mathsf{X}')) \right] = R_{\text{pair}}(\operatorname{sgn} \circ f).
$$

Let $\widetilde{\psi}_{\text{pair}} : [0, 1] \to [0, +\infty)$ be the calibration function defined by

$$
\begin{aligned}
\widetilde{\psi}_{\text{pair}}(\varepsilon) \coloneqq \inf_{\eta, \eta \in [0,1]} \inf_{\alpha, \alpha' \in \mathbb{R}} & S^{\phi}_{\text{pair}}(\alpha, \alpha', \eta, \eta') - H^{\phi}_{\text{pair}}(\eta, \eta') \\
& \text{s.t. } S_{\text{pair}}(\alpha, \alpha', \eta, \eta') - H_{\text{pair}}(\eta, \eta') \geq \varepsilon.
\end{aligned}
$$

Based on the consequence of Lemma 2.9 of Steinwart [2007], $\widetilde{\psi}_{\text{pair}}(\varepsilon) > 0$ for all $\varepsilon > 0$ implies that $R^{\phi}_{\text{pair}}(f) \to R^{*}_{\text{pair}} \implies R_{\text{pair}}(\operatorname{sgn} \circ f) \to R^{*}_{\text{pair}}$. Further, under this condition, Theorem 2.13 of Steinwart [2007] implies that $\widetilde{\psi}_{\text{pair}}$ is non-decreasing, invertible, and satisfies

$$
\widetilde{\psi}^{\star\star}_{\text{pair}}(R_{\text{pair}}(\operatorname{sgn} \circ f) - R^{*}_{\text{pair}}) \leq R^{\phi}_{\text{pair}}(f) - R^{\phi,*}_{\text{pair}}
$$

for any measurable function f. Hence, it is sufficient to show that $\widetilde{\psi}_{\text{pair}}(\varepsilon) > 0$ for all $\varepsilon > 0$. Indeed, $\widetilde{\psi}_{\text{pair}} = \widetilde{\psi}_{\text{point}}$, and $\widetilde{\psi}_{\text{point}}(\varepsilon) > 0$ for all $\varepsilon > 0$ because ϕ is classification-calibrated [Bartlett et al., 2006, Lemma 2]. From now on, we will see $\widetilde{\psi}_{\text{pair}} = \widetilde{\psi}_{\text{point}}$.

First, we simplify the constraint part of $\widetilde{\psi}_{\text{pair}}$. Because

$$
\begin{aligned}
S_{\text{pair}}(\alpha, \alpha', \eta, \eta') &= (1 - \eta - \eta' + 2\eta\eta')\mathbb{1}_{\{\operatorname{sgn}(\alpha)\operatorname{sgn}(\alpha') = -1\}} \\
&\quad + (\eta + \eta' - 2\eta\eta')\mathbb{1}_{\{\operatorname{sgn}(\alpha)\operatorname{sgn}(\alpha') = +1\}} \\
&= \widetilde{\eta}\mathbb{1}_{\{\operatorname{sgn}(\alpha)\operatorname{sgn}(\alpha') = +1\}} + (1 - \widetilde{\eta})\mathbb{1}_{\{\operatorname{sgn}(\alpha)\operatorname{sgn}(\alpha') = -1\}},
\end{aligned}
$$

where $\widetilde{\eta} \coloneqq 1 - \eta - \eta' + 2\eta\eta'$, we have $H_{\text{pair}}(\eta, \eta') = \min\{\widetilde{\eta}, 1 - \widetilde{\eta}\}$. Similarly,

$$
S^{\phi}_{\text{pair}}(\alpha, \alpha', \eta, \eta') = \widetilde{\eta}\phi(\alpha\alpha', +1) + (1 - \widetilde{\eta})\phi(\alpha\alpha', -1).
$$

With a slight abuse of notation, we may write $S_{\text{pair}}(\alpha, \alpha', \widetilde{\eta}) = S_{\text{pair}}(\alpha, \alpha', \eta, \eta')$ (the same for S^{ϕ}_{pair}, H_{pair}, and H^{ϕ}_{pair}). Through simple algebra, we obtain

$$
S_{\text{pair}}(\alpha, \alpha', \widetilde{\eta}) - H_{\text{pair}}(\widetilde{\eta}) = |2\widetilde{\eta} - 1| \cdot \mathbb{1}_{\{(2\widetilde{\eta} - 1)\operatorname{sgn}(\alpha)\operatorname{sgn}(\alpha') \leq 0\}}.
$$

Noting that $\widetilde{\eta}$ ranges over $[0,1]$ with $\eta, \eta' \in [0,1]$, we have

$$\begin{aligned} \widetilde{\psi}_{\text{pair}}(\varepsilon) = \inf_{\widetilde{\eta}\in[0,1]} \inf_{\alpha,\alpha'\in\mathbb{R}} & S^{\phi}_{\text{pair}}(\alpha,\alpha',\widetilde{\eta}) - H^{\phi}_{\text{pair}}(\widetilde{\eta}) \\ \text{s.t.} \quad & |2\widetilde{\eta}-1| \cdot \mathbb{1}_{\{(2\widetilde{\eta}-1)\text{sgn}(\alpha)\text{sgn}(\alpha')\le 0\}} \ge \varepsilon. \end{aligned}$$

If $\varepsilon = 0$, $\widetilde{\psi}_{\text{pair}}(0) = 0$, and the infimum is attained by $\widetilde{\eta} = \frac{1}{2}$ and an arbitrary α and α'. If $\varepsilon > 0$, $\widetilde{\eta} = \frac{1}{2}$ cannot satisfy the constraint. Hence, we assume from here that $\widetilde{\eta} \neq \frac{1}{2}$. When $\widetilde{\eta} > \frac{1}{2}$, the constraint reduces to

$$\left\{\alpha\alpha' \le 0 \wedge (\alpha,\alpha') \neq (0,0)\right\} \vee \widetilde{\eta} \ge \frac{1+\varepsilon}{2}.$$

Because S^{ϕ}_{pair} contains α and α' only in the form of $\alpha\alpha'$, the infimum over

$$\left\{\, \alpha,\alpha' \in \mathbb{R} \;\middle|\; \alpha\alpha' \le 0 \wedge (\alpha,\alpha') \neq (0,0) \,\right\}$$

is equal to that over

$$\left\{\, \alpha,\alpha' \in \mathbb{R} \;\middle|\; \alpha\alpha' \le 0 \,\right\}.$$

If we write $\alpha\alpha' \coloneqq \widetilde{\alpha}$, then

$$\begin{aligned} \widetilde{\psi}_{\text{pair}}(\varepsilon) &= \inf_{\widetilde{\eta}\in\left[\frac{1+\varepsilon}{2},1\right]} \inf_{\widetilde{\alpha}\in\mathbb{R}:\widetilde{\alpha}\le 0} S^{\phi}_{\text{pair}}(\alpha,\alpha',\widetilde{\eta}) - H^{\phi}_{\text{pair}}(\widetilde{\eta}) \\ &= \inf_{\widetilde{\eta}\in\left[\frac{1+\varepsilon}{2},1\right]} \inf_{\widetilde{\alpha}\in\mathbb{R}:\widetilde{\alpha}\le 0} S^{\phi}_{\text{point}}(\widetilde{\alpha},\widetilde{\eta}) - H^{\phi}_{\text{point}}(\widetilde{\eta}) \\ &= \inf_{\widetilde{\alpha}\in\mathbb{R}:\widetilde{\alpha}\le 0} S^{\phi}_{\text{point}}\left(\widetilde{\alpha}, \frac{1+\varepsilon}{2}\right) - H^{\phi}_{\text{point}}\left(\frac{1+\varepsilon}{2}\right) \\ &= H^{\phi,-}_{\text{point}}\left(\frac{1+\varepsilon}{2}\right) - H^{\phi}_{\text{point}}\left(\frac{1+\varepsilon}{2}\right) \\ &= \widetilde{\psi}_{\text{point}}(\varepsilon). \end{aligned}$$

When $\widetilde{\eta} < \frac{1}{2}$, $\widetilde{\psi}_{\text{pair}} = \widetilde{\psi}_{\text{point}}$ can be shown in the same way. Hence, the statement is proven. □

5.7.3 Proof of Lemma 5.5

We start by introducing the following statement.

Lemma 5.9. *For real values α and β satisfying $0 \le \alpha \le \beta \le 1$, we have*

$$\sqrt{\beta} - \sqrt{\alpha} \le \sqrt{\beta - \alpha}.$$

Proof.

$$(\beta - \alpha) - (\sqrt{\beta} - \sqrt{\alpha})^2 = 2\sqrt{\alpha\beta} - 2\alpha = 2\sqrt{\alpha}\left(\sqrt{\beta} - \sqrt{\alpha}\right) \ge 0.$$

Thus, we have $(\beta - \alpha) \ge (\sqrt{\beta} - \sqrt{\alpha})^2$, which completes the proof of Lemma 5.9. □

With this lemma, an excess risk of a clustering error can be connected with that of a pairwise classification error as follows. From the equation in Equation (5.4), we have

$$R^{*}_{\text{clus}} = \frac{1}{2} - \frac{\sqrt{1 - 2R^{*}_{\text{pair}}}}{2}.$$

Thus, we can bound excess risk on the clustering error as follows.

$$
\begin{aligned}
R_{\text{clus}}(\operatorname{sgn}\circ\widehat{f}) - R^*_{\text{clus}} &= \left(\frac{1}{2} - \frac{\sqrt{1-2R_{\text{pair}}(\operatorname{sgn}\circ\widehat{f})}}{2}\right) - \left(\frac{1}{2} - \frac{\sqrt{1-2R^*_{\text{pair}}}}{2}\right) \\
&= \frac{1}{2}\left\{\sqrt{1-2R^*_{\text{pair}}} - \sqrt{1-2R_{\text{pair}}(\operatorname{sgn}\circ\widehat{f})}\right\} \\
&\leq \sqrt{\frac{R_{\text{pair}}(\operatorname{sgn}\circ\widehat{f}) - R^*_{\text{pair}}}{2}} \\
&\leq \sqrt{\frac{1}{2}\psi^{-1}\left(R^\phi_{\text{pair}}(\widehat{f}) - R^{\phi,*}_{\text{pair}}\right)},
\end{aligned}
\tag{5.16}
$$

where Lemma 5.9 and Lemma 5.4 were applied to obtain the penultimate and the last inequalities, respectively. The excess risk with respect to pairwise surrogate risk, i.e., $R^\phi_{\text{pair}}(\widehat{f}) - R^{\phi,*}_{\text{pair}}$, can be decomposed into *approximation error* and *estimation error* as

$$
R^\phi_{\text{pair}}(\widehat{f}) - R^{\phi,*}_{\text{pair}} = \underbrace{R^\phi_{\text{pair}}(f^*) - R^{\phi,*}_{\text{pair}}}_{\text{approximation error}} + \underbrace{R^\phi_{\text{pair}}(\widehat{f}) - R^\phi_{\text{pair}}(f^*)}_{\text{estimation error}}, \tag{5.17}
$$

where f^* is the minimizer of $R^\phi_{\text{pair}}(f)$ in a specified function space $\mathcal{F}$. Now, we provide the following upper bound for the estimation error with the Rademacher complexity.

Lemma 5.10. *Let $f^* \in \mathcal{F}$ be a minimizer of R^ϕ_{pair}, and $\widehat{f} \in \mathcal{F}$ be a minimizer of the empirical risk $\widehat{R}^\phi_{\text{pair}}$. Assume that the loss function ϕ is a ρ-Lipschitz function with respect to the first argument ($0 < \rho < \infty$), and all functions in the model class $\mathcal{F}$ are bounded, i.e., there exists a constant C_b such that $\|f\|_\infty \leq C_b$ for any $f \in \mathcal{F}$. Let $C_\phi := \sup_{t\in\{\pm1\}} \phi(C_b^2, t)$. For any $\delta > 0$, with a probability of at least $1-\delta$,*

$$
R^\phi_{\text{pair}}(\widehat{f}) - R^\phi_{\text{pair}}(f^*) \leq 4\rho\mathfrak{R}_{m_1}(\mathcal{F}) + \sqrt{\frac{2C_\phi^2 \log\frac{2}{\delta}}{m_1}}. \tag{5.18}
$$

Proof. The estimation error can be bounded as

$$
\begin{aligned}
R^\phi_{\text{pair}}(\widehat{f}) - R^\phi_{\text{pair}}(f^*) &\leq \left(R^\phi_{\text{pair}}(\widehat{f}) - \widehat{R}^\phi_{\text{pair}}(\widehat{f})\right) + \left(\widehat{R}^\phi_{\text{pair}}(f^*) - R^\phi_{\text{pair}}(f^*)\right) \\
&\leq 2\sup_{f\in\mathcal{F}}\left|R^\phi_{\text{pair}}(\widehat{f}) - \widehat{R}^\phi_{\text{pair}}(\widehat{f})\right|.
\end{aligned}
\tag{5.19}
$$

With the Rademacher complexity, the following inequalities hold with probability at least $1-\delta$.

$$
\left|R^\phi_{\text{pair}}(\widehat{f}) - \widehat{R}^\phi_{\text{pair}}(\widehat{f})\right| \leq 2\mathfrak{R}_{m_1}(\phi\circ\mathcal{F}) + \sqrt{\frac{C_\phi^2\log\frac{2}{\delta}}{2m_1}}, \tag{5.20}
$$

where $\phi\circ\mathcal{F}$ indicates a class of composite functions defined by $\{\,\phi\circ f \mid f\in\mathcal{F}\,\}$. By applying Talagrand's lemma, the Rademacher complexity of $\phi\circ\mathcal{F}$ can be bounded as

$$
\mathfrak{R}_{m_1}(\phi\circ\mathcal{F}) \leq \rho\mathfrak{R}_{m_1}(\mathcal{F}). \tag{5.21}
$$

The proofs of Equations (5.20) and (5.21) can be found in Mohri et al. [2018, Theorem 3.1 and Lemma 4.2], respectively. By plugging Equations (5.20) and (5.21) into Equation (5.19), we obtain the result in Equation (5.18). □

By combining Equations (5.16) and (5.17) and Lemma 5.10, we obtain the following inequality with probability at least $1-\delta$,

$$R_{\text{clus}}(\text{sgn}\circ\widehat{f}) - R^*_{\text{clus}} \leq \sqrt{\frac{1}{2}\psi^{-1}\left(R^{\phi}_{\text{pair}}(f^*) - R^{\phi,*}_{\text{pair}} + 4\rho\mathfrak{R}_{m_1}(\mathcal{F}) + \sqrt{\frac{2C_\phi^2\log\frac{2}{\delta}}{m_1}}\right)}. \tag{5.22}$$

□

5.7.4 Proof of Lemma 5.6

We first derive a sufficient condition for the proposed class assignment failures. Let $\widehat{s}$ be a estimated class assignment for a given hypothesis $h : \mathcal{X} \to \mathcal{Y}$.

$$\begin{aligned}
&\Pr\left(\widehat{s} \neq \operatorname*{arg\,min}_{s\in\{\pm1\}} R_{\text{point}}(s\cdot h)\right)\\
&\quad= \Pr\left(\text{sgn}\left(1-2\widehat{Q}(h)\right) \neq \text{sgn}\left(1-2Q(h)\right)\right)\\
&\quad= \begin{cases}\Pr\left(2\widehat{Q}(h)-1>0\right) & (1-2Q(h)>0),\\ \Pr\left(2\widehat{Q}(h)-1\leq 0\right) & (\text{otherwise})\end{cases} \qquad (5.23)\\
&\quad= \begin{cases}\Pr\left(\widehat{Q}(h)-Q(h)>\frac{1}{2}-Q(h)\right) & (1-2Q(h)>0),\\ \Pr\left(Q(h)-\widehat{Q}(h)\geq Q(h)-\frac{1}{2}\right) & (\text{otherwise})\end{cases}
\end{aligned}$$

By applying Hoeffding's inequality [Hoeffding, 1963], we obtain the following bounds.

$$\Pr\left(\widehat{Q}(h)-Q(h)>\frac{1}{2}-Q(h)\right) \leq \exp\left(-2m_2\left(Q(h)-\frac{1}{2}\right)^2\right), \tag{5.24}$$

$$\Pr\left(Q(h)-\widehat{Q}(h)\geq Q(h)-\frac{1}{2}\right) \leq \exp\left(-2m_2\left(Q(h)-\frac{1}{2}\right)^2\right), \tag{5.25}$$

where m_2 is the number of pairwise examples to compute $\widehat{Q}(h)$. Therefore, we can bound the error probability of the proposed class assignment method regardless of the value of $Q(h)$ as

$$\Pr\left(\widehat{s} \neq \operatorname*{arg\,min}_{s\in\{\pm1\}} R_{\text{point}}(s\cdot h)\right) \leq \exp\left(-2m_2\left(Q(h)-\frac{1}{2}\right)^2\right). \tag{5.26}$$

Now, we further explore how the term $Q(h)-\frac{1}{2}$ can be expressed. From the definition of Q and the equivalent risk expression in Equation (5.15), we have

$$Q(h) = (2\pi_+ - 1)R_{\text{point}}(h) + 1 - \pi_+. \tag{5.27}$$

Therefore,

$$Q(h)-\frac{1}{2} = (2\pi_+-1)\left(R_{\text{point}}(h)-\frac{1}{2}\right). \tag{5.28}$$

By plugging Equation (5.28) into Equation (5.26), we finally obtain

$$\Pr\left(\widehat{s} \neq \operatorname*{arg\,min}_{s\in\{\pm1\}} R_{\text{point}}(s\cdot h)\right) \leq \exp\left(-\frac{m_2}{2}(2\pi_+-1)^2\left(2R_{\text{point}}(h)-1\right)^2\right), \tag{5.29}$$

which completes the proof of Lemma 5.6. □

5.8 Conclusion

In this chapter, we presented the underlying relationship between similarity learning and binary classification (Theorem 5.1). Eventually, the two-step similarity learning procedure for binary classification with only pairwise supervision was obtained. Our similarity learning can elicit the underlying decision boundary and is less affected by the class prior. The post-processing class assignment is less costly than training a new classifier. Our framework subsumes many existing similarity learning methods with specific losses. A parallel connection for multi-class classification remains open for discussion.

Chapter 6

Conclusions and Future Prospects

> *All models are wrong but some are useful.*
>
> — George E. P. Box, *Science and Statistics*

In this chapter, we summarize the conclusions of this dissertation and discuss areas of future research.

6.1 Summary

This dissertation was devoted to investigating a new perspective of learning theory, i.e., excess risk transfer, to thoroughly understand the gap between learning and evaluation criteria thoroughly. This insight will help us verify whether a learning criterion has been appropriately designed in light of the evaluation criteria, leading to reliable machine learning algorithms. In addition, an investigation into the relationship between learning and evaluation criteria will promote our understanding of the underlying hierarchy among machine learning problems, revealing what we can learn from a specific learning problem. Following Vapnik's principle, these perspectives will lead us to the design of a "minimally sufficient" learning criterion given a set of our final evaluation criteria, providing a guideline for how the learned knowledge can help us solve the other problem. We may express this belief through the following declaration, inspired by Wittgenstein's famous aphorism:

> *The understanding of knowledge is its use in the learning.*

Subsequently, we summarize the contributions of this dissertation in light of both reliable machine learning and knowledge transfer.

Contributions to reliable machine learning. In Chapter 3, we focused on the family of complex classification performance metrics commonly used in class imbalance classification, called the linear-fractional metrics, and proposed learning criteria calibrated to the linear-fractional metrics. The proposed learning criteria are computationally feasible, and we empirically observed that the sample efficiency was better than that of the plug-in classifiers. Because the linear-fractional metrics have significant practical meaning in real-world scenarios such as information retrieval and computer vision, the proposed learning criteria promote the reliability of a learned predictor in light of the target linear-fractional metrics.

In Chapter 4, we used calibration analysis to investigate whether existing learning criteria are calibrated to the adversarially robust classification risk and revealed that commonly used convex surrogate losses are not calibrated in adversarially robust classification. We also proposed several nonconvex calibrated losses under this setup. This chapter provided an important insight indicating that the learning criteria will not lead to truly robust solutions without a careful design. The analysis in this chapter is meaningful by itself because calibration analysis was used to analyze the robustness of the classifiers, which in existing studies has thus far been used to analyze the consistency in terms of the classification accuracy.

Contribution to knowledge transfer. In Chapter 5, we elucidated that a specific formulation of similarity learning has a substantial connection to binary classification, meaning that solving the similarity learning directly elicits a binary decision boundary. Although it has yet to be clarified what knowledge a learner elicits through similarity learning in most of the previous studies, this chapter contributed to this area by showing that a learner can acquire a decision boundary through similarity learning. This contribution takes the understanding of the problem relationship one step further to reveal that binary classification and similarity learning in the specific formulation are equivalent.

6.2 Future Prospects

A better understanding of the relationship between two learning problems will bring us the hierarchy of learning problems, leading to a design of learning criteria oriented to one's multiple ultimate goals (see Section 1.4.4 for more details). Under the current paradigm of machine learning, we still do not have satisfactory tools to validate whether our model achieves the optimal solution, or the Pareto optimal solution at least, to which our perspective provides an approach.

What will this novel perspective of machine learning eventually play a key role for? We suppose that even we humans do not know what we want to solve in the end. For this reason, the human-in-the-loop approach can be a promising way to elucidate the *underlying* evaluation criteria, where we repeatedly iterate the following steps: the machine provides multiple candidate models to the user based on the *guessed* evaluation criteria, and the user chooses the best model among them. The machine will be able to accurately guess how the user expects to evaluate models after the repeated queries. In this regard, we argue that the machine ought to submit the optimal models to the user by dealing with the underlying diverse objectives and constraints.

Although our contributions in this dissertation made indispensable steps towards this goal, we suppose that further research is awaited. Subsequently, we show several important future research directions to conclude this dissertation.

6.2.1 Calibration Analysis with Restricted Function Spaces

The calibration analysis we introduced in Section 2.3 considers the convergence towards Bayes risk. Let us recapitulate this here. Given a target loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ and a surrogate loss function $\phi : \mathcal{T} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$, the Bayes ℓ- and ϕ-risks are defined as

$$R_\ell^* := \inf_{f \in \mathcal{F}_{\mathrm{all}}} R_\ell(f), \quad R_\phi^* := \inf_{f \in \mathcal{F}_{\mathrm{all}}} R_\phi(f),$$

respectively, where $\mathcal{F}_{\text{all}} \subseteq \mathbb{R}^{\mathcal{X}}$ is the set of all measurable functions. We say that ϕ is ℓ-consistent if for any distributions and sequence of functions $\{f_i\}_i \subseteq \mathcal{F}_{\text{all}}$,

$$R_\phi(f_i) \to R_\phi^* \implies R_\ell(f_i) \to R_\ell^* \quad \text{as } i \to \infty.$$

Although the ℓ-consistency is one of the key properties throughout this dissertation, it is limited to the convergence to Bayes risks. Because we usually seek risk minimizers from a restricted score function space, such as linear-in-input models, this characterization of the target consistency may be inappropriate. Indeed, [Long and Servedio, 2013] provided theoretical evidence indicating that a finer notion of consistency can explain the empirical success of the Crammer-Singer loss [Crammer and Singer, 2001] in multi-class SVMs, whereas the classical calibration analysis concludes that the Crammer-Singer loss is not calibrated to the multi-class 0-1 loss [Zhang, 2004b].

This relaxation, which can be seen in Section 4.5 as well, makes the analysis of the adversarially robust 0-1 loss easier by restricting the score function space to the halfspaces. Correspondingly, the calibration function and the notion of calibrated losses can be relaxed to a restricted score function space. However, if the score function space is not $\mathcal{F}_{\text{all}}$, it is not straightforward to guarantee the minimizability of the loss functions (Definition 2.4), which are sufficient conditions for the calibration to imply the target consistency. Whereas the existing research on the relaxed calibration analysis [Long and Servedio, 2013, Zhang and Agarwal, 2020] relies heavily on the assumption of realizability, i.e., the existence of the Bayes minimizer within the restricted score function space, the general analysis remains open.

6.2.2 Calibration Analysis with Flexible Objectives and Constraints

Although a supervised learning problem was formulated as a minimization problem of the target risk functional,

$$\min_{g \in \mathcal{G}} R_\ell(g),$$

for some hypothesis space $\mathcal{G} \subseteq \mathcal{Y}^{\mathcal{X}}$, some real-world problems may need more complicated formulations. Narasimhan [2018] provided many examples as follows:

- Coverage constraint: The recall must be larger than a given threshold.
- Fairness constraint: Demographic parity requires that the recall must be equalized across all groups, for example [Kleinberg et al., 2017].
- Quantification constraint: The predicted class distribution must be close to the true class-prior probability in terms of a certain probability divergence.

To handle these flexible constraints, Narasimhan [2018] extended the above problem formulation to incorporate into the following constraints:

$$\min_{g \in \mathcal{G}} R_\ell(g) \quad \text{s.t.} \quad Q_k(g) \le \varepsilon_k, \forall k \in [K],$$

where $Q_k : \mathcal{G} \to \mathbb{R}$ characterizes a constraint with the admissible level ε_k.

An extension of calibration analysis would be both theoretically interesting and practically significant. We expect to have

$$R_\ell(g_i) \to R_\ell^* \quad \text{and} \quad \mathbb{P}(Q_k(g_i) > \varepsilon_k) \to 0 \quad \forall k \in [K] \quad \text{as } i \to \infty$$

for a sequence $\{g_i\}_i$ by designing a learning criterion. To date, Narasimhan [2018] has transformed the constrained (variational) problem into a constrained optimization problem over the confusion matrix. That being said, their results are only applicable to algorithms based on a plug-in classifier. Because we believe that a learning criterion independent of a class-posterior probability estimator is better in terms of the sample complexity, as was observed in Chapter 3, an extension of calibration analysis to deal with general learning criteria and constraints deserves further study.

6.2.3 Calibration Analysis with Generalized Risk Measures

We introduced calibration analysis by focusing on the surrogate and target risks, which are the expectation of the loss functions, following the classical formulation of statistical machine learning [Vapnik, 1998]. By contrast, some recent studies have considered alternative risk-averse measures such as

- conditional value-at-risk (CVaR) [Kashima, 2007, Sinha et al., 2018, Soma and Yoshida, 2020, Holland and Haress, 2021] and
- coherent risk measures [Tamar et al., 2015, Lee et al., 2020].

For example, given a loss function $\ell(f(\mathsf{X}), \mathsf{Y})$, its CVaR is defined as follows:

$$\mathrm{CVaR}_\alpha(f) \coloneqq \underset{(\mathsf{X},\mathsf{Y})}{\mathbb{E}}[\ell(f(\mathsf{X}), \mathsf{Y}) \mid \ell(f(\mathsf{X}), \mathsf{Y}) \geq \mathrm{VaR}_\alpha(f)],$$

where VaR_α is the value-at-risk, i.e., the $(1-\alpha)$-quantile of the random variable:

$$\mathrm{VaR}_\alpha(f) \coloneqq \inf \left\{ \tau \in \mathbb{R} \mid \mathbb{P}_{(\mathsf{X},\mathsf{Y})}(\ell(f(\mathsf{X}), \mathsf{Y}) \leq \tau) \geq 1 - \alpha \right\}.$$

When the CVaR is employed as a target risk functional, the classical calibration analysis is no longer applicable. Because the CVaR and coherent risk measure have attracted increasing attention in finance, insurance, and firm management [Artzner et al., 1999], an extension for such risk measures must be valuable.

6.2.4 Hybrid Learning Theory of Calibration Analysis and Optimization Perspective

The calibration analysis introduced in this dissertation only focused on a statistical perspective. More specifically, we assume that the exact minimizer of a given surrogate risk can be attained. From an optimization perspective, however, this is difficult in general with a finite number of computational resources. For this reason, it is meaningful to allow a small admissible error for the surrogate risk.

Osokin et al. [2017] introduced the following variant of the target consistency.

Definition 6.1 (Level-η consistency [Osokin et al., 2017])**.** *A surrogate loss ϕ is* consistent up to level $\eta \geq 0$ *with respect to a target loss ℓ and a score function space $\mathcal{F}$ if and only if the (uniform) calibration function $\delta(\varepsilon)$ satisfies $\delta(\varepsilon) > 0$ for all $\varepsilon > \eta$ and there exists $\widehat{\varepsilon} > \eta$ such that $\delta(\widehat{\varepsilon})$ is finite.*

Recall that the necessary and sufficient condition of a (uniform) calibration is $\delta(\varepsilon) > 0$ for all $\varepsilon > 0$. The level-η consistency is its relaxation. Figure 6.1 illustrates a usual calibration function and a calibration function under the level-η consistency: The calibration function represented by the solid line is not calibrated in the strict sense but satisfies level-η consistency, whereas the calibration function represented by the dashed line is calibrated. Nevertheless, the solid line converges

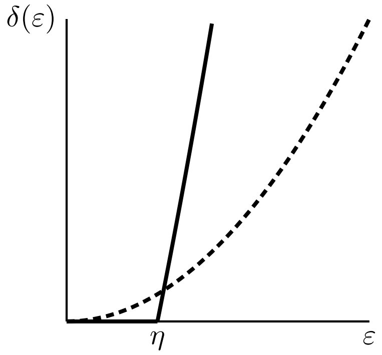


Figure 6.1: Calibration function under level-η consistency.

faster towards η, meaning that the same number of optimization steps with a solid line can contribute more up to the admissible target risk η.

This relaxation may be useful when we incorporate the optimization perspective into the analysis because each function class entails a different convergence rate. For example, the optimal convergence rates for strongly convex and smooth functions, smooth convex functions, non-smooth convex functions, and quasiconvex functions are $O(e^{-k})$, $O(1/k^2)$, $O(1/\sqrt{k})$, and $O(1/\sqrt{k})$, respectively, where k is the number of iterations [Nesterov, 2004]. It is generally known that the convergence rate of the stochastic gradient descent (SGD) is $O(1/k)$ for strongly convex functions [Rakhlin et al., 2012], whereas the rate of a variant of the SGD is $O(1/\sqrt{k})$ for strictly quasiconvex functions [Hazan et al., 2015]. Given these convergence rates, a design of learning criteria should pay more attention to which function classes we focus on. Because Osokin et al. [2017] only studied a specific setting in a structured prediction, a more general study is important.

6.2.5 Quantitative Comparison of Surrogate Risk Minimization and Plug-in Classifiers

When we design a target consistent algorithm, there are roughly two types of approaches: a surrogate-loss-based learning algorithm and plug-in classifiers. The plug-in classifiers estimate the class-posterior probability $\mathbb{P}(\mathsf{Y} \mid \mathsf{X} = \mathbf{x})$ first and then plug the estimate into the closed-form of the Bayes classifier. As we criticized and empirically observed in Chapter 3, the plug-in classifiers may not be a good approach particularly when the threshold for $\mathbb{P}(\mathsf{Y} \mid \mathsf{X} = \mathbf{x})$ might vary, as in the case of the linear-fractional metrics.[1] This is because the class-posterior probability needs to be estimated as accurately as possible within the entire range of $[0, 1]$, which requires a high sample complexity. Hence, it is important to quantitatively compare the two approaches from the perspective of the sample complexity.

6.2.6 Characterization of Irreducibility between Learning Problems

As we introduced in Chapter 1, excess risk transfer can be a powerful tool to characterize the relationship among machine learning problems. For example, Narasimhan and Agarwal [2013] showed that problems of bipartite ranking and binary CPE can be reduced to each other. That said, if we have a good bipar-

[1] If the optimal threshold of the Bayes classifier does not vary as in the usual binary classification, even the plug-in classifiers can achieve faster learning rates under a strong assumption regarding the underlying probability distribution (called the strong density assumption) [Audibert and Tsybakov, 2007].

tite ranking model, there is a procedure for transferring the model into a good class-posterior probability estimator, and vice versa. As in this example, the reducibility from one learning problem to another can be established through excess risk transfer. However, we still do not have a satisfactory theoretical tool to establish the irreducibility, which means that for any procedure, it is impossible to construct a good model for one learning problem based on another learning problem.

Irreducibility should be a prominent concept for characterizing the hierarchical relationship among machine learning problems. If one knows that problem A cannot be reduced to problem B, then we must be careful in the design of the learning criteria such that a learner is exposed to an equivalently difficult learning problem to problem B. For example, if we need a class-posterior probability estimate, we should not attempt to obtain a classifier first because it is impossible to recover the probability. Charoenphakdee et al. [2019] discussed this perspective with a broader class of loss functions called symmetric losses and showed the inability to recover the class-posterior probability from symmetric loss minimizers. Their approach is based on the form of the Bayes risk and demonstrates the inability through the fact that the symmetric loss minimizers are proportional to

$$f^{\phi,*}(\mathbf{x}) \propto \operatorname{sgn}\left(\mathbb{P}(\mathsf{Y} = 1 \mid \mathsf{X} = \mathbf{x}) - \frac{1}{2}\right)$$

under binary classification, and it is impossible to recover $\mathbb{P}(\mathsf{Y} = 1 \mid \mathsf{X} = \mathbf{x})$ from $f^{\phi,*}(\mathbf{x})$. We can also extend this discussion to wider situations. Here, the non-invertibility of the Bayes score functions may play a key role in characterizing the irreducibility.

The (ir)reducibility may appear similar to a Turing reduction in computational complexity theory [Arora and Barak, 2009].[2] The notion of the reducibility informally introduced herein is rather statistical in the sense that we ask if knowledge from one learning problem is richer than and/or equivalent to knowledge that can be obtained from another learning problem. Because there is no consideration of the optimization and computational perspective, this notion characterizes the statistical learnability.

One related paradigm is a *property elicitation*, a framework used to seek a loss function whose minimization eventually leads to recovering a specific functional of a probability distribution, called a property [Osband, 1985, Fissler, 2017]. Property elicitation has gradually gained attention in mathematical statistics [Gneiting, 2011, Fissler and Ziegel, 2016] and economics [Lambert et al., 2008]. Conceptually, the minimizer of every loss function is associated with a property [Agarwal and Agarwal, 2015, Finocchiaro et al., 2019]. Hence, we may organize learning problems hierarchically in terms of the properties by regarding two properties as being hierarchically ordered if one property is obtained through a non-invertible functional transform of the other [Frongillo et al., 2016]. However, this still requires more investigation because we only know the elicitability of simple properties [Steinwart et al., 2014] and the general characterization remains open [Frongillo et al., 2016].

[2] In comparison with proofs of the reducibility, proving the irreducibility is not straightforward at all even in the field of computational complexity theory. A few proof techniques such as diagonalization and relativization have been provided [Arora and Barak, 2009], leading to nice irreducibility results such as Ladner's Theorem [Ladner, 1975], which ensures the existence of NP-intermediate problems in between P and NP unless P = NP.

6.2.7 Connection between Similarity Learning and Multi-class Classification

In Chapter 5, we elucidated the equivalence between similarity learning and binary classification under a specific formulation. Specifically, a pair of labeled examples are regarded as similar when $y = y'$, and dissimilar when $y \neq y'$. With this similarity, similarity learning used to predict whether a given pair of examples belong to the same class or not can be related to binary classification through a simple relationship:

$$R_{\text{pair}}(h) = 2R_{\text{point}}(h)(1 - R_{\text{point}}(h)),$$

where R_{pair} is the similarity classification risk and R_{point} is the *binary* classification risk.

However, we do not have any such relationships for multi-class classification. Because the similarity has been popularly used in the recent self-supervised representation learning [Jaiswal et al., 2021], it has a significant impact on knowing whether multi-class classification is fundamentally possible through similarity learning or not. Thus far, we conjecture that a parallel connection to multi-class classification will never exist and that a similarity is fundamentally insufficient to elicit sufficiently useful knowledge in solving multi-class classification. Let us first consider binary classification. It is easy to observe that combinations of label pairs are mutually exclusive between the similar and dissimilar cases:

- Similar: $(y, y') \in \{ (1,1), (2,2) \}$
- Dissimilar: $(y, y') \in \{ (1,2), (2,1) \}$

Here, the class labels are denoted by 1 and 2. By contrast, they are no longer mutually exclusive for one-versus-rest multi-class classification. For example, if the number of classes is three, then the label combinations are as follows:

- Similar: $(y, y') \in \{ (1,1), (2,2), (3,3) \}$
- Dissimilar: $(y, y') \in \{ (1,2), (2,3), (3,1), (2,1), (3,2), (1,3) \}$

If we are to apply one-versus-rest multi-class classifiers, we expect that similar and dissimilar label combinations are mutually exclusive for binary labels c and $\bar{c}$ ($c \in \mathcal{Y}$), where $\bar{c}$ is the complementary class label reducing all examples without class c. For example, if $c = 1$, we have the following:

- Similar: $(y, y') \in \{ (1,1), (\bar{1},\bar{1}) \}$
- Dissimilar: $(y, y') \in \{ (1,\bar{1}), (\bar{1},1) \}$

This does not correspond to the aforementioned label combinations for multi-class classification because $(2,3)$ should be dissimilar but falsely classified as similar $(\bar{1},\bar{1})$. Of course, this intuition only suggests the impossibility of one-versus-rest multi-class classifiers and does not say much about other multi-class surrogate models such as the Crammer-Singer SVMs and the softmax cross-entropy loss. A solid study on these factors could reveal the fundamental (im)possibility of similarity learning.

6.2.8 Comparison between Similarity Learning and Other Binary Machine Learning Problems

The similarity learning discussed in Chapter 5 deals with relationships between two (binary) labeled data points. This setup is somewhat similar to bipartite ranking because a ranking model compares two given examples and ranks them in bipartite ranking. It is then natural to seek for the underlying relationship between similarity learning and bipartite ranking. If we know the mutual reducibility relationship (see Section 6.2.6) between them, we can not only relate similarity learning to bipartite ranking but also to binary CPE because bipartite ranking and binary CPE are equivalent in a certain sense [Gao and Zhou, 2015]. This is valuable because CPE is one of the most fundamental learning problems in supervised learning.

In computer vision, Parikh and Grauman [2011] introduced a learning problem called learning from relative attributes, which is a ranking problem with untraditional supervision. They considered ordered pairs and unordered pairs: The former tells us which data point should be ranked higher, and the latter tells us that two data points have similar strengths. Parikh and Grauman [2011] empirically showed that relative attributes can lead to a good few-shot learner. This problem can be seen as a combination of (bipartite) ranking and similarity learning.

6.2.9 Closer Look at Strongly and Strictly Proper Losses

This open problem is somewhat related to the irreducibility and problem hierarchy (see Section 6.2.6). As introduced in Section 2.4.2, proper losses are a class of loss functions that play an important role in (binary) CPE. We recall the notation again. Let us focus on the binary case $y \in \{1, -1\}$ here. A loss $\ell : [0,1] \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is introduced to measure the quality of a probability estimate $\widehat{\eta} \in [0,1]$ for $\mathbb{P}(\mathsf{Y} = y \mid \mathsf{X} = \mathbf{x})$. The class-conditional ℓ-risk is

$$C_\ell(\widehat{\eta}, \eta) = \eta\ell(\widehat{\eta}, 1) + (1-\eta)\ell(\widehat{\eta}, -1)$$

with the conditional Bayes ℓ-risk $C^*_\ell(\eta) = \inf_{\widehat{\eta}\in[0,1]} C_\ell(\widehat{\eta}, \eta)$. We state that ℓ is strictly proper if and only if $C_\ell(\widehat{\eta}, \eta) = C^*_\ell(\eta)$ only when $\widehat{\eta} = \eta$ for all $\eta \in [0,1]$.

Agarwal [2014] introduced a stronger class of proper losses for the convenience of proofs.

Definition 6.2 (Strongly proper losses [Agarwal, 2014]). *Let $\lambda > 0$. We state that a binary CPE loss ℓ is λ-strongly proper if for all $\eta, \widehat{\eta} \in [0,1]$,*

$$C_\ell(\widehat{\eta}, \eta) - C^*_\ell(\eta) \geq \frac{\lambda}{2}(\eta - \widehat{\eta})^2.$$

This definition is of course related to strongly convex functions. Indeed, $-C^*_\ell$ is strongly convex for a strongly proper ℓ [Agarwal, 2014].[3] Many common proper losses such as the logistic loss and squared loss are strongly proper.

Strong properness immediately implies strict properness, but to date we do not know anything about the converse implication. Intuitively, there should be a loss function that is strictly proper but not strongly proper. If such a loss function is found, we may have better characterizations of proper losses; for example, we may

[3] By contrast, $-C^*_\ell$ is strictly convex when ℓ is strictly proper [Savage, 1971, Buja et al., 2005, Gneiting and Raftery, 2007, Reid and Williamson, 2009]. Historically, Savage [1971] imposed this strict convexity to uniquely elicit personal preferences.

overcome the square-root excess risk transfer rate of strongly proper losses [Agarwal, 2014, Frongillo and Waggoner, 2021] using the new loss.[4] Otherwise, we can show the mutual reducibility between strongly proper and strictly proper losses, and the hierarchy collapses. This hierarchy collapse will free us from considering strictly convex functions.

6.2.10 Systematic Design of Learning Criteria from User Feedback

This is the most important question. Throughout this dissertation, we discussed calibration analysis and excess risk transfer, supposing that a target loss of an evaluation criterion has already been given. However, this is often unrealistic because even humans may not be sure of the best way to evaluate computer systems in an objective manner. How should we characterize algorithmic fairness? Although fairness is one of the hottest trends in the machine learning community, numerous definitions of fairness have been proposed during the last couple of years and we still do not have any consensus [Verma and Rubin, 2018]. What type of attackers should we suppose when we protect our models from adversarial attacks? Indeed, every time a defense methods is proposed, new attack methods are proposed immediately after for a while [Athalye et al., 2018]. Because defense methods cannot avoid supposing a certain attack model, some researchers have been skeptical regarding the existence of a true robustness measure [Shafahi et al., 2018]. In addition, common definitions of adversarial robustness have thus far had a trade-off with accuracy, which would be contradictory to human perception because we seem to perceive targets both accurately and robustly [Suggala et al., 2019]. In any case, designing the evaluation criteria is not a straightforward task for us at all.

One promising direction is metric elicitation, which is a framework proposed by Hiranandani et al. [2019] for determining the best performance metric of classifiers from a pairwise comparison oracle. We iteratively show two classifiers to a human expert and ask which classifier is preferable. A target performance metric is then obtained through a derivative-free optimization. Currently, metric elicitation has been studied for classification [Hiranandani et al., 2019] and fairness metrics [Hiranandani et al., 2020]. This direction can also be studied for wider applications.

[4] In particular, Frongillo and Waggoner [2021] showed that a surrogate risk has a lower bound of the square-root excess risk with respect to the classification risk under certain conditions on the surrogate loss such as the strong convexity and smoothness.

References

A. Agarwal and S. Agarwal. On consistent surrogate risk minimization and property elicitation. In *Proceedings of the 28th Conference on Learning Theory*, pages 4–22, 2015. (cite on page 145)

S. Agarwal. Surrogate regret bounds for bipartite ranking via strongly proper losses. *Journal of Machine Learning Research*, 15(1):1653–1674, 2014. (cite on page 41, 147, 148)

F. Ahmed, D. Tarlow, and D. Batra. Optimizing expected Intersection-over-Union with candidate-constrained CRFs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1850–1858, 2015. (cite on page 54)

D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2(4): 343–370, 1988. (cite on page 78)

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. (cite on page 7)

S. Arora and B. Barak. *Computational Complexity: A Modern Approach*. Cambridge University Press, 2009. (cite on page 34, 145)

P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999. (cite on page 143)

K. Asif, W. Xing, S. Behpour, and B. D. Ziebart. Adversarial cost-sensitive classification. In *Proceedings of the Conference on Uncertainty in Artifitial Intelligence*, pages 92–101, 2015. (cite on page 7, 56)

A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 274–283, 2018. (cite on page 148)

J.-Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007. (cite on page 17, 57, 144)

D. Aussel, J. Corvellec, and M. Lassonde. Subdifferential characterization of quasiconvexity and convexity. *Journal of Convex Analysis*, 1(2):195–201, 1994. (cite on page 73, 74)

B. Ávila Pires and C. Szepesvári. Multiclass classification calibration functions. *arXiv preprint arXiv:1609.06385*, 2016. (cite on page 10, 78)

B. Ávila Pires, C. Szepesvári, and M. Ghavamzadeh. Cost-sensitive multiclass classification risk bounds. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1391–1399, 2013. (cite on page 78)

P. Awasthi, N. Frank, A. Mao, M. Mohri, and Y. Zhong. Calibration and consistency of adversarial surrogate losses. In *Advances in Neural Information Processing Systems 34 (to appear)*, 2021a. (cite on page 78, 89, 91)

P. Awasthi, A. Mao, M. Mohri, and Y. Zhong. A finer calibration analysis for adversarial robustness. *arXiv preprint arXiv:2105.01550*, 2021b. (cite on page 78)

R. Azen and D. V. Budescu. The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods*, 8(2):129, 2003. (cite on page 13)

F. R. Bach, G. R. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the 21st International Conference on Machine Learning*, page 6, 2004. (cite on page 116)

C. Baier and J.-P. Katoen. *Principles of Model Checking*. MIT press, 2008. (cite on page 11)

M.-F. Balcan, A. Blum, and N. Srebro. A theory of learning with similarity functions. *Machine Learning*, 72(1-2):89–112, 2008. (cite on page 116, 117, 118, 125)

H. Bao and M. Sugiyama. Calibrated surrogate maximization of linear-fractional utility in binary classification. In *Proceedings of the 23th International Conference on Artificial Intelligence and Statistics*, 2020. (cite on page 17, 78)

H. Bao and M. Sugiyama. Fenchel-Young losses with skewed entropies for class-posterior probability estimation. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pages 1648–1656, 2021. (cite on page 57)

H. Bao, G. Niu, and M. Sugiyama. Classification from pairwise similarity and unlabeled data. In *Proceedings of the 35th International Conference on Machine Learning*, pages 461–470, 2018. (cite on page 118)

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002. (cite on page 7, 28, 126)

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006. (cite on page 9, 10, 18, 26, 27, 34, 35, 46, 50, 51, 76, 78, 80, 82, 120, 126, 134, 135)

K. Bascol, R. Emonet, E. Fromont, A. Habrard, G. Metzler, and M. Sebban. From cost-sensitive to tight F-measure bounds. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 1245–1253, 2019. (cite on page 57)

S. Basu, A. Banerjee, and R. Mooney. Semi-supervised clustering by seeding. In *Proceedings of the 19th International Conference on Machine Learning*, 2002. (cite on page 117)

A. Bellet, A. Habrard, and M. Sebban. Similarity learning for provably accurate sparse linear classification. In *Proceedings of the 29th International Coference on Machine Learning*, pages 1491–1498, 2012. (cite on page 19, 116, 117, 120, 124, 125)

S. Ben-David, N. Eiron, and P. M. Long. On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, 66(3):496–514, 2003. (cite on page 46)

S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, volume 19, page 137, 2007. (cite on page 8)

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010. (cite on page 78)

A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*, volume 28. Princeton University Press, 2009. (cite on page 77)

M. Berman, A. R. Triki, and M. B. Blaschko. The Lovász-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. (cite on page 16, 17, 43, 54, 56)

D. Bertsimas, D. B. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011. (cite on page 77)

M. Bilenko, S. Basu, and R. J. Mooney. Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of the 21st International Conference on Machine Learning*, pages 839–846, 2004. (cite on page 116, 117)

P. Billingsley. *Probability and Measure*. John Wiley & Sons, 2008. (cite on page 7)

M. Blondel. Structured prediction with projection oracles. In *Advances in Neural Information Processing Systems 32*, pages 12145–12156, 2019. (cite on page 78)

A. Blum, A. Frieze, R. Kannan, and S. Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1):35–52, 1998. (cite on page 2, 6)

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. (cite on page 26, 48, 50, 73, 100)

G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. (cite on page 26, 36)

K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann. The balanced accuracy and its posterior distribution. In *ICPR*, pages 3121–3124, 2010. (cite on page 43)

J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature verification using a "siamese" time delay neural network. In *Advances in Neural Information Processing Systems 7*, pages 737–744, 1994. (cite on page 117)

S. Bubeck, Y. T. Lee, E. Price, and I. Razenshteyn. Adversarial examples from computational constraints. In *Proceedings of the 36th International Conference on Machine Learning*, pages 831–840, 2019. (cite on page 75)

A. Buja, W. Stuetzle, and Y. Shen. Loss functions for binary class probability estimation and classification: Structure and applications. *Technical Report*, 2005. (cite on page 26, 36, 37, 39, 147)

R. Busa-Fekete, B. Szörényi, K. Dembczyński, and E. Hüllermeier. Online F-measure optimization. In *Advances in Neural Information Processing Systems 28*, pages 595–603, 2015. (cite on page 57)

V. Cabannnes, A. Rudi, and F. Bach. Structured prediction with partial labelling through the infimum loss. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1230–1239, 2020. (cite on page 10)

F. P. Cantelli. Sulla determinazione empirica delle leggi di probabilit'a. *Giorn. Ist. Ital. Attuari*, 4:421–424, 1933. (cite on page 7)

R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997. (cite on page 8)

D. V. Carvalho, E. M. Pereira, and J. S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019. (cite on page 13)

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. (cite on page 6)

V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):1–58, 2009. (cite on page 5)

C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2011. URL http://www.csie.ntu.edu.tw/~cjlin/libsvm. ACM Transactions on Intelligent Systems and Technology. (cite on page 58, 128)

O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. MIT Press, 2006. (cite on page 5, 118)

N. Charoenphakdee, J. Lee, and M. Sugiyama. On symmetric losses for learning from corrupted labels. In *Proceedings of the 36th International Conference on Machine Learning*, 2019. (cite on page 43, 78, 84, 85, 145)

Y. Chen, E. K. Garcia, M. R. Gupta, A. Rahimi, and L. Cazzanti. Similarity-based classification: Concepts and algorithms. *Journal of Machine Learning Research*, 10(3), 2009. (cite on page 19, 118)

S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 539–546. IEEE, 2005. (cite on page 133)

J. Cid-Sueiro, D. García-García, and R. Santos-Rodríguez. Consistency of losses for learning from weak labels. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 197–210. Springer, 2014. (cite on page 10)

M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning*, pages 854–863, 2017. (cite on page 71, 77)

T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha. Deep learning for classical Japanese literature. In *NeurIPS Workshop on Machine Learning for Creativity and Design*, 2018. (cite on page 128)

F. H. Clarke. *Optimization and Nonsmooth Analysis*, volume 5. SIAM, 1990. (cite on page 73, 74)

J. Cohen, E. Rosenfeld, and Z. Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1310–1320, 2019. (cite on page 18, 77)

S. M. Cohen and C. D. C. Reeve. Aristotle's Metaphysics. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2020 edition, 2020. URL `https://plato.stanford.edu/archives/win2020/entries/aristotle-metaphysics/`. (cite on page 1)

M. Collins and N. Duffy. Convolution kernels for natural language. In *Advances in Neural Information Processing Systems*, pages 625–632, 2001. (cite on page 3)

R. Collobert, F. Sinz, J. Weston, and L. Bottou. Trading convexity for scalability. In *Proceedings of the 23rd international conference on Machine learning*, pages 201–208, 2006. (cite on page 83)

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273–297, 1995. (cite on page 3, 9, 27)

C. Cortes, M. Mohri, and A. Rostamizadeh. Two-stage learning kernel algorithms. In *Proceedings of the 27th International Conference on Machine Learning*, pages 239–246, 2010. (cite on page 116)

K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001. (cite on page 142)

Z. Cranko, A. Menon, R. Nock, C. S. Ong, Z. Shi, and C. Walder. Monge blunts Bayes: Hardness results for adversarial training. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1406–1415, 2019. (cite on page 77)

N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. S. Kandola. On kernel-target alignment. In *Advances in Neural Information Processing Systems 15*, pages 367–373, 2002. (cite on page 116, 117, 124)

G. Csurka, D. Larlus, F. Perronnin, and F. Meylan. What is a good evaluation measure for semantic segmentation? In *BMVC*, pages 1–11, 2013. (cite on page 54)

Z. Cui, N. Charoenphakdee, I. Sato, and M. Sugiyama. Classification from triplet comparison data. *Neural Computation*, 32(3):659–681, 2020. (cite on page 118)

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989. (cite on page 2)

C. Dan, Y. Wei, and P. Ravikumar. Sharp statistical guaratees for adversarially robust gaussian classification. In *Proceedings of the 37th International Conference on Machine Learning*, pages 2345–2355, 2020. (cite on page 77)

S. Dan, H. Bao, and M. Sugiyama. Learning from noisy similar and dissimilar data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 233–249. Springer, 2021. (cite on page 118)

J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th International Conference on Machine Learning*, pages 209–216, 2007. (cite on page 116, 117)

R. Dechter. Learning while searching in constraint-satisfaction-problems. In *Proceedings of the 5th AAAI National Conference on Artificial Intelligence*, pages 178–183, 1986. (cite on page 3)

M. Delange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. (cite on page 8)

K. Dembczynski, W. Kotłowski, and E. Hüllermeier. Consistent multilabel ranking through univariate loss minimization. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1347–1354, 2012. (cite on page 10, 78)

K. Dembczyński, A. Jachnik, W. Kotłowski, W. Waegeman, and E. Hüllermeier. Optimizing the F-measure in multi-label classification: Plug-in rule approach versus structured loss minimization. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1130–1138, 2013. (cite on page 56, 57)

K. Dembczyński, W. Kotłowski, O. Koyejo, and N. Natarajan. Consistency analysis for binary classification revisited. In *Proceedings of the 34th International Conference on Machine Learning*, pages 961–969, 2017. (cite on page 45, 56)

L. Derczynski. Complementarity, F-score, and NLP evaluation. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 261–266, 2016. (cite on page 16)

R. Descartes. The world. *The Philosophical Writings of Descartes*, 1:81–98, 1985. (cite on page 1)

E. Dobriban, H. Hassani, D. Hong, and A. Robey. Provable tradeoffs in adversarially robust classification. In *Proceedings of the 9th International Conference on Learning Representations*, 2021. (cite on page 115)

T. Dreossi, S. Ghosh, A. L. Sangiovanni-Vincentelli, and S. A. Seshia. A formalization of robustness for deep neural networks. In *Proceedings of the AAAI Spring Symposium Workshop on Verification of Neural Networks (VNN)*, 2019. (cite on page 77)

D. Dua and C. Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`. (cite on page 128)

J. C. Duchi, L. W. Mackey, and M. I. Jordan. On the consistency of ranking algorithms. In *Proceedings of the 27th International Conference on Machine Learning*, pages 327–334, 2010. (cite on page 78)

R. M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967. (cite on page 7, 28, 29)

C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia. Incorporating second-order functional knowledge for better option pricing. *Advances in Neural Information Processing Systems*, 13:472–478, 2000. (cite on page 133)

A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956. (cite on page 7)

E. Eban, M. Schain, A. Mackey, A. Gordon, R. A. Saurous, and G. Elidan. Scalable learning of non-decomposable objectives. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 832–840, 2017. (cite on page 56)

D. Eisenberg, E. M. Marcotte, I. Xenarios, and T. O. Yeates. Protein function in the post-genomic era. *Nature*, 405(6788):823–826, 2000. (cite on page 116)

M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. (cite on page 54)

A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, and A. Bouras. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing*, 2 (3):267–279, 2014. (cite on page 119)

F. Farnia and D. Tse. A minimax approach to supervised learning. In *Advances in Neural Information Processing Systems 29*, pages 4240–4248, 2016. (cite on page 78)

R. Fathony and Z. Kolter. AP-Perf: Incorporating generic performance metrics in differentiable learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pages 4130–4140, 2020. (cite on page 56)

R. Fathony, A. Liu, K. Asif, and B. Ziebart. Adversarial multiclass classification: A risk minimization perspective. In *Advances in Neural Information Processing Systems 29*, pages 559–567, 2016. (cite on page 78)

R. Fathony, K. Asif, A. Liu, M. A. Bashiri, W. Xing, S. Behpour, X. Zhang, and B. D. Ziebart. Consistent robust adversarial prediction for general multiclass classification. *arXiv preprint arXiv:1812.07526*, 2018. (cite on page 56)

V. Feldman, V. Guruswami, P. Raghavendra, and Y. Wu. Agnostic learning of monomials by halfspaces is hard. *SIAM Journal on Computing*, 41(6):1558–1590, 2012. (cite on page 9, 46, 76)

C. Finn. *Learning to Learn with Gradients.* PhD thesis, UC Berkeley, 2018. (cite on page 14)

J. Finocchiaro, R. Frongillo, and B. Waggoner. An embedding framework for consistent polyhedral surrogates. In *Advances in Neural Information Processing Systems 33*, volume 32, pages 10781–10791, 2019. (cite on page 26, 56, 145)

T. Fissler. *On higher order elicitability and some limit theorems on the Poisson and Wiener space.* PhD thesis, University of Bern, 2017. (cite on page 145)

T. Fissler and J. F. Ziegel. Higher order elicitability and Osband's principle. *The Annals of Statistics*, 44(4):1680–1707, 2016. (cite on page 145)

R. Fletcher. *Practical Methods of Optimization.* John Wiley & Sons, 2013. (cite on page 50)

J. A. Fodor and Z. W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988. (cite on page 2, 13)

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997. (cite on page 27)

J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2):337–407, 2000. (cite on page 9)

R. Frongillo and I. A. Kash. Vector-valued property elicitation. In *Proceedings of the 28th Conference on Learning Theory*, pages 710–727, 2015. (cite on page 26)

R. Frongillo and B. Waggoner. Surrogate regret bounds for polyhedral losses. In *Advances in Neural Information Processing Systems 34 (to appear)*, 2021. (cite on page 148)

R. Frongillo, I. Kash, and S. Becker. Open problem: Property elicitation and elicitation complexity. In *Proceedings of the 29th Conference on Learning Theory*, pages 1655–1658, 2016. (cite on page 145)

K. Fukushima and S. Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and Cooperation in Neural Nets*, pages 267–285. Springer, 1982. (cite on page 3)

K.-I. Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2(3):183–192, 1989. (cite on page 2)

W. Gao and Z.-h. Zhou. On the consistency of multi-label learning. In *Proceedings of 24th Annual Conference on Learning*, 2011. (cite on page 10, 78)

W. Gao and Z.-H. Zhou. On the consistency of AUC pairwise optimization. In *IJCAI*, pages 939–945, 2015. (cite on page 10, 36, 51, 52, 78, 147)

T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola. Multi-instance kernels. In *Proceedings of the 19th International Conference on Machine Learning*, pages 179–186, 2002. (cite on page 3)

A. Geiger, I. Cases, L. Karttunen, and C. Potts. Posing fair generalization tasks for natural language inference. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4485–4495, 2019. (cite on page 2)

P. Germain, A. Habrard, F. Laviolette, and E. Morvant. A PAC-Bayesian approach for domain adaptation with specialization to linear classifiers. In *Proceedings of the 30th International Conference on Machine Learning*, pages 738–746, 2013. (cite on page 78)

A. Ghosh, N. Manwani, and P. Sastry. Making risk minimization tolerant to label noise. *Neurocomputing*, 160:93–107, 2015. (cite on page 72, 78, 84)

V. I. Glivenko. Sulla determinazione empirica di probabilit'a. *Giorn. Ist. Ital. Attuari*, 4:92–99, 1933. (cite on page 7)

T. Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011. (cite on page 145)

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007. (cite on page 36, 37, 147)

N. Golowich, A. Rakhlin, and O. Shamir. Size-independent sample complexity of neural networks. In *Proceedings of the 31st Conference on Learning Theory*, pages 297–299, 2018. (cite on page 29)

R. Gomes, P. Welinder, A. Krause, and P. Perona. Crowdclustering. In *Advances in Neural Information Processing Systems 25*, 2012. (cite on page 116)

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 24*, 2014. (cite on page 4)

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015. (cite on page 12, 18, 71, 77)

J. C. Gower and P. Legendre. Metric and euclidean properties of dissimilarity coefficients. *Journal of Classification*, 3(1):5–48, 1986. (cite on page 43, 46)

T. Graepel, R. Herbrich, B. Schölkopf, A. Smola, P. Bartlett, K.-R. Müller, K. Obermayer, and R. Williamson. Classification on proximity data with LP-machines. 1999. (cite on page 118)

S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386, 2020. (cite on page 4)

S. Gu and L. Rigazio. Towards deep neural network architectures robust to adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations Workshop*, 2015. (cite on page 77)

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330, 2017. (cite on page 11, 12, 21)

P. R. Halmos. The theory of unbiased estimation. *The Annals of Mathematical Statistics*, pages 34–43, 1946. (cite on page 23)

E. Hazan, K. Levy, and S. Shalev-Shwartz. Beyond convexity: Stochastic quasi-convex optimization. In *Advances in Neural Information Processing Systems 28*, pages 1594–1602, 2015. (cite on page 48, 144)

T. Hazan, J. Keshet, and D. A. McAllester. Direct loss minimization for structured prediction. In *Advances in Neural Information Processing Systems 23*, pages 1594–1602, 2010. (cite on page 78)

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. (cite on page 4)

M. Hein and M. Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems 30*, pages 2266–2276, 2017. (cite on page 77)

L. Henderson. The Problem of Induction. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2020 edition, 2020. (cite on page 6)

G. E. Hinton and R. S. Zemel. Autoencoders, minimum description length, and Helmholtz free energy. In *Advances in Neural Information Processing Systems 6*, pages 3–10, 1994. (cite on page 3)

G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006. (cite on page 3)

G. Hiranandani, S. Boodaghians, R. Mehta, and O. Koyejo. Performance metric elicitation from pairwise classifier comparisons. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 371–379, 2019. (cite on page 148)

G. Hiranandani, H. Narasimhan, and S. Koyejo. Fair performance metric elicitation. In *Advances in Neural Information Processing Systems 33*, 2020. (cite on page 148)

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. (cite on page 3)

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. (cite on page 138)

M. Holland. Classification using margin pursuit. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 712–720, 2019. (cite on page 78, 84)

M. Holland and E. M. Haress. Learning with risk-averse feedback under potentially heavy tails. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pages 892–900, 2021. (cite on page 143)

J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79 (8):2554–2558, 1982. (cite on page 3)

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. (cite on page 2)

Y.-C. Hsu, Z. Lv, J. Schlosser, P. Odom, and Z. Kira. Multi-class classification without multi-class labels. In *Proceedings of the 7th International Conference on Learning Representations*, 2019. (cite on page 117, 118, 124, 125, 129, 133)

W. Hu, G. Niu, I. Sato, and M. Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *Proceedings of the 35th International Conference on Machine Learning*, pages 2034–2042, 2018. (cite on page 78)

P. J. Huber. *Robust Statistics*. Springer, 2011. (cite on page 78)

A. Hyvärinen. Independent component analysis: recent advances. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110534, 2013. (cite on page 5)

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 448–456, 2015. (cite on page 4)

T. S. Jaakkola, D. Haussler, et al. Exploiting generative models in discriminative classifiers. pages 487–493, 1999. (cite on page 3)

P. Jaccard. Étude de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901. (cite on page 16, 46, 54)

M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, pages 427–435, 2013. (cite on page 57)

A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2021. (cite on page 5, 19, 21, 146)

N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–449, 2002. (cite on page 16)

A. Javanmard, M. Soltanolkotabi, and H. Hassani. Precise tradeoffs in adversarial training for linear regression. In *Proceedings of the 33rd Conference on Learning Theory*, pages 2034–2078, 2020. (cite on page 114)

Q. Jiang, O. Adigun, H. Narasimhan, M. M. Fard, and M. Gupta. Optimizing black-box metrics with adaptive surrogates. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4784–4793, 2020. (cite on page 57)

C. Jin, L. T. Liu, R. Ge, and M. I. Jordan. On the local minima of the empirical risk. In *Advances in Neural Information Processing Systems 32*, 2018. (cite on page 25)

T. Joachims. Transductive learning via spectral graph partitioning. In *Proceedings of the 20th International Conference on Machine Learning*, pages 290–297, 2003. (cite on page 14)

T. Joachims. A support vector method for multivariate performance measures. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 377–384, 2005. (cite on page 43, 56)

J. W. Johnson. A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research*, 35(1):1–19, 2000. (cite on page 13)

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. (cite on page 4)

A. T. Kalai and R. Sastry. The Isotron algorithm: High-dimensional isotonic regression. In *Proceedings of the 21st Conference on Learning Theory*, 2009. (cite on page 2, 40)

I. Kant. *Critique of Pure Reason.* 1781. (cite on page 21)

P. Kar and P. Jain. Similarity-based learning via data driven embeddings. In *Advances in Neural Information Processing Systems 24*, pages 1998–2006, 2011. (cite on page 116, 117, 120)

P. Kar, H. Narasimhan, and P. Jain. Online and stochastic gradient methods for non-decomposable loss functions. In *Advances in Neural Information Processing Systems 27*, pages 694–702, 2014. (cite on page 56)

H. Kashima. Risk-sensitive learning via minimization of empirical conditional value-at-risk. *IEICE Transactions on Information and Systems*, 90(12):2043–2052, 2007. (cite on page 143)

M. Kearns. Thoughts on hypothesis boosting. *Unpublished Manuscript*, 45:105, 1988. (cite on page 6)

M. J. Kearns, R. E. Schapire, and L. M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994. (cite on page 2, 6, 9, 26, 34)

N. S. Keskar, J. Nocedal, P. T. P. Tang, D. Mudigere, and M. Smelyanskiy. On large-batch training for deep learning: Generalization gap and sharp minima. In *Proceedings of the 5th International Conference on Learning Representations*, 2017. (cite on page 28)

J. Khim and P.-L. Loh. Adversarial risk bounds via function transformation. In *Advances in Neural Information Processing Systems 32*, 2019. (cite on page 77)

G. King and L. Zeng. Logistic regression in rare events data. *Political Analysis*, 9(2):137–163, 2001. (cite on page 57)

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015. (cite on page 133)

D. P. Kingma and M. Welling. Auto-encoding variational bayes. 2014. (cite on page 4)

J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017. (cite on page 142)

G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2. Lille, 2015. (cite on page 117)

W. Kotlowski and K. Dembczyński. Surrogate regret bounds for generalized classification performance metrics. In *Proceedings of the 8th Asian Conference on Machine Learning*, pages 301–316, 2016. (cite on page 41)

O. Koyejo, N. Natarajan, P. Ravikumar, and I. S. Dhillon. Consistent multilabel classification. In *Advances in Neural Information Processing Systems 28*, volume 29, pages 3321–3329, 2015. (cite on page 17)

O. O. Koyejo, N. Natarajan, P. K. Ravikumar, and I. S. Dhillon. Consistent binary classification with generalized performance metrics. In *Advances in Neural Information Processing Systems 27*, pages 2744–2752, 2014. (cite on page 17, 43, 44, 52, 53, 54, 57, 58)

V. Krishnan, A. Makdah, A. AlRahman, and F. Pasqualetti. Lipschitz bounds and provably robust training by laplacian smoothing. In *Advances in Neural Information Processing Systems 33*, volume 33, pages 10924–10935, 2020. (cite on page 114)

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105, 2012. (cite on page 1, 4)

B. Kulis. Metric learning: A survey. *Foundations and Trends® in Machine Learning*, 5(4):287–364, 2013. (cite on page 19, 116)

A. Kumar, H. Narasimhan, and A. Cotter. Implicit rate-constrained optimization of non-decomposable objectives. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5861–5871, 2021. (cite on page 56)

A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016. (cite on page 11, 18)

S. Kuroki, N. Charoenphakdee, H. Bao, J. Honda, I. Sato, and M. Sugiyama. Unsupervised domain adaptation based on source-guided discrepancy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4122–4129, 2019. (cite on page 78, 127)

R. E. Ladner. On the structure of polynomial time reducibility. *Journal of the ACM*, 22(1):155–171, 1975. (cite on page 145)

B. Lake and M. Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International conference on machine learning*, pages 2873–2882, 2018. (cite on page 2)

N. S. Lambert, D. M. Pennock, and Y. Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138, 2008. (cite on page 145)

G. R. Lanckriet, L. E. Ghaoui, C. Bhattacharyya, and M. I. Jordan. A robust minimax approach to classification. *Journal of Machine Learning Research*, 3: 555–582, 2002. (cite on page 77)

G. R. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004. (cite on page 116, 117)

J. Langford and J. Shawe-Taylor. PAC-Bayes & margins. In *Advances in Neural Information Processing Systems*, pages 439–446, 2003. (cite on page 8)

Y. LeCun. The MNIST database of handwritten digits, 2013. URL `http://yann.lecun.com/exdb/mnist`. (cite on page 128)

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten ZIP code recognition. *Neural Computation*, 1(4):541–551, 1989. (cite on page 2)

M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana. Certified robustness to adversarial examples with differential privacy. In *IEEE Symposium on Security and Privacy (SP)*, pages 656–672. IEEE, 2019. (cite on page 77)

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 1991. (cite on page 28, 66, 67, 68, 70)

J. Lee, S. Park, and J. Shin. Learning bounds for risk-sensitive learning. In *Advances in Neural Information Processing Systems 33*, 2020. (cite on page 143)

Z. Li and J. Liu. Constrained clustering by spectral kernel learning. In *Proceedings of the 12th IEEE International Conference on Computer Vision*, pages 421–427, 2009. (cite on page 116)

M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`. (cite on page 58)

Y. Lin. A note on margin-based loss functions in classification. *Statistics & Probability Letters*, 68(1):73–82, 2004. (cite on page 9, 10, 27, 34, 78)

M. Liu, X. Zhang, X. Zhou, and T. Yang. Faster online learning of optimal threshold for consistent F-measure optimization. In *Advances in Neural Information Processing Systems 31*, pages 3893–3903, 2018. (cite on page 57)

H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2: 419–444, 2002. (cite on page 3)

L. Logeswaran and H. Lee. An efficient framework for learning sentence representations. In *Proceedings of the 6th International Conference on Learning Representations*, 2018. (cite on page 125)

P. Long and R. Servedio. Consistency versus realizable H-consistency for multiclass classification. In *Proceedings of the 30th International Conference on Machine Learning*, pages 801–809, 2013. (cite on page 10, 32, 77, 78, 142)

P. M. Long and R. A. Servedio. Random classification noise defeats all convex potential boosters. *Machine Learning*, 78(3):287–304, 2010. (cite on page 78)

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967. (cite on page 19, 129)

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the 6th International Conference on Learning Representations*, 2018. (cite on page 77)

R. P. C. Manning and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. (cite on page 16, 53)

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Proceedings of 22th Annual Conference on Learning*, 2009. (cite on page 78)

H. Masnadi-Shirazi and N. Vasconcelos. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. In *Advances in Neural Information Processing Systems 22*, pages 1049–1056, 2009. (cite on page 78, 84)

P. Massart and É. Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006. (cite on page 6, 72, 83, 88)

D. A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the 12th Annual Conference on Computational Learning Theory*, pages 164–170, 1999. (cite on page 8)

J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon. A proposal for the Dartmouth summer research project on artificial intelligence. `http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html`, 1955. (cite on page 1)

P. McCullagh and J. A. Nelder. *Generalized Linear Models.* Chapman & Hall/CRC, 1989. (cite on page 26)

W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133, 1943. (cite on page 1)

C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, 141(1):148–188, 1989. (cite on page 67, 68)

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021. (cite on page 11)

A. Menon, H. Narasimhan, S. Agarwal, and S. Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In *Proceedings of the 30th International Conference on Machine Learning*, pages 603–611, 2013. (cite on page 43)

A. Menon, B. van Rooyen, C. S. Ong, and R. Williamson. Learning from corrupted binary labels via class-probability estimation. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 125–134, 2015. (cite on page 43, 45)

A. K. Menon and R. C. Williamson. Bipartite ranking: a risk-theoretic perspective. *Journal of Machine Learning Research*, 17(1):6766–6867, 2016. (cite on page 39)

A. K. Menon, X. J. Jiang, S. Vembu, C. Elkan, and L. Ohno-Machado. Predicting accurate probabilities with a ranking loss. In *Proceedings of the 27th International Conference on Machine Learning*, volume 2012, page 703, 2012. (cite on page 57)

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119, 2013. (cite on page 116)

F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 4th International Conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. (cite on page 16, 17)

M. Minsky and S. A. Papert. *Perceptrons: An Introduction to Computational Ceometry.* MIT Press, 1972. (cite on page 2)

L. Mohammadi and S. van de Geer. Asymptotics in empirical risk minimization. *Journal of Machine Learning Research*, 6(12), 2005. (cite on page 25)

M. Mohri, A. Rostamizadeh, F. Bach, and A. Talwalkar. *Foundations of Machine Learning*. MIT Press, 2018. (cite on page 26, 30, 46, 70, 137)

M. Nagao. Informatics is the forefront of philosophy. *LRG: Linerary Resource Guide*, 27:10–76, 2019. URL `http://hdl.handle.net/2433/244172`. (cite on page 21)

V. Nagarajan and J. Z. Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems*, 2019. (cite on page 8)

H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems 29*, pages 2208–2216, 2016. (cite on page 78)

H. Namkoong and J. C. Duchi. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems 30*, pages 2971–2980, 2017. (cite on page 78)

Y. Nan, K. M. Chai, W. S. Lee, and H. L. Chieu. Optimizing F-measure: A tale of two approaches. In *Proceedings of the 29th International Conference on Machine Learning*, pages 289–296, 2012. (cite on page 43, 53, 57)

H. Narasimhan. Learning with complex loss functions and constraints. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, pages 1646–1654, 2018. (cite on page 12, 21, 57, 142, 143)

H. Narasimhan and S. Agarwal. On the relationship between binary classification, bipartite ranking, and binary class probability estimation. In *Advances in Neural Information Processing Systems 26*, pages 2913–2921, 2013. (cite on page 14, 19, 21, 38, 39, 40, 128, 144)

H. Narasimhan, R. Vaish, and S. Agarwal. On the statistical consistency of plug-in classifiers for non-decomposable performance measures. In *Advances in Neural Information Processing Systems 27*, pages 1493–1501, 2014. (cite on page 17, 44, 57, 58)

H. Narasimhan, P. Kar, and P. Jain. Optimizing non-decomposable performance measures: a tale of two classes. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 199–208, 2015. (cite on page 44, 54, 57)

H. Narasimhan, W. Pan, P. Kar, P. Protopapas, and H. G. Ramaswamy. Optimizing the multiclass F-measure via biconcave programming. In *ICDM*, pages 1101–1106, 2016. (cite on page 57)

H. Narasimhan, A. Cotter, and M. Gupta. Optimizing generalized rate metrics with three players. volume 32, pages 10747–10758, 2019. (cite on page 57)

N. Natarajan, O. Koyejo, P. Ravikumar, and I. Dhillon. Optimal classification with multivariate losses. In *Advances in Neural Information Processing Systems 29*, pages 1530–1538, 2016. (cite on page 44)

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer, 2004. (cite on page 144)

B. Neyshabur, S. Bhojanapalli, D. Mcallester, and N. Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems 30*, pages 5947–5956, 2017. (cite on page 8)

A. Y. Ng, S. J. Russell, et al. Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning*, page 2, 2000. (cite on page 6)

G. Niu, B. Dai, M. Yamada, and M. Sugiyama. Information-theoretic semi-supervised metric learning via entropy regularization. *Neural Computation*, 26(8):1717–1762, 2014. (cite on page 116)

M. Nordström, H. Bao, F. Löfman, H. Hult, A. Maki, and M. Sugiyama. Calibrated surrogate maximization of Dice. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 269–278. Springer, 2020. (cite on page 17)

A. B. Novikoff. On convergence proofs for perceptrons. Technical report, Stanford Research Institute, 1963. (cite on page 2)

K. Nozawa, P. Germain, and B. Guedj. PAC-Bayesian contrastive unsupervised representation learning. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 21–30, 2020. (cite on page 117, 120)

A. Okuno and H. Shimodaira. Hyperlink regression via Bregman divergence. *Neural Networks*, 126:362–383, 2020. (cite on page 125)

S. Ontañón. An overview of distance and similarity functions for structured data. *Artificial Intelligence Review*, 53(7):5309–5351, 2020. (cite on page 19)

K. Osband. *Providing Incentives for Better Cost Forecasting*. PhD thesis, University of California, Berkeley, 1985. (cite on page 145)

A. Osokin, F. Bach, and S. Lacoste-Julien. On structured prediction theory with calibrated convex surrogate losses. In *Advances in Neural Information Processing Systems 30*, pages 302–313, 2017. (cite on page 50, 78, 143, 144)

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2009. (cite on page 8, 13)

S. P. Parambath, N. Usunier, and Y. Grandvalet. Optimizing F-measures by cost-sensitive classification. In *Advances in Neural Information Processing Systems 27*, pages 2123–2131, 2014. (cite on page 44, 53, 57, 58)

D. Parikh and K. Grauman. Relative attributes. In *Proceedings of the 14th IEEE International Conference on Computer Vision*, pages 503–510. IEEE, 2011. (cite on page 147)

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. (cite on page 129, 133)

F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, 2019. (cite on page 13)

R. Pinot, L. Meunier, A. Araujo, H. Kashima, F. Yger, C. Gouy-Pailler, and J. Atif. Theoretical evidence for adversarial robustness through randomization. In *Advances in Neural Information Processing Systems 32*, pages 11838–11848, 2019. (cite on page 77)

D. A. Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, 3(1):88–97, 1991. (cite on page 6)

A. Raghunathan, J. Steinhardt, and P. Liang. Certified defenses against adversarial examples. In *Proceedings of the 6th International Conference on Learning Representations*, 2018a. (cite on page 71, 77)

A. Raghunathan, J. Steinhardt, and P. S. Liang. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems 31*, pages 10877–10887, 2018b. (cite on page 77)

A. Raghunathan, S. M. Xie, F. Yang, J. Duchi, and P. Liang. Understanding and mitigating the tradeoff between robustness and accuracy. In *Proceeedings of the 37th Proceedings of Machine Learning Research*, 2020. (cite on page 114)

H. Rahimian and S. Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019. (cite on page 11)

A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1571–1578, 2012. (cite on page 144)

H. G. Ramaswamy and S. Agarwal. Classification calibration dimension for general multiclass losses. In *Advances in Neural Information Proceedings Systems 25*, pages 2078–2086, 2012. (cite on page 78)

H. G. Ramaswamy and S. Agarwal. Convex calibration dimension for multiclass loss matrices. *The Journal of Machine Learning Research*, 17(1):397–441, 2016. (cite on page 78)

H. G. Ramaswamy, S. Agarwal, and A. Tewari. Convex calibrated surrogates for low-rank loss matrices with applications to subset ranking losses. In *Advances in Neural Information Processing Systems 26*, pages 1475–1483, 2013. (cite on page 78)

P. Ravikumar, A. Tewari, and E. Yang. On NDCG consistency of listwise ranking methods. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 618–626, 2011. (cite on page 10, 78)

I. Redko, E. Morvant, A. Habrard, M. Sebban, and Y. Bennani. A survey on domain adaptation theory. *arXiv preprint arXiv:2004.11829*, 2020. (cite on page 8)

M. D. Reid and R. C. Williamson. Surrogate regret bounds for proper losses. In *Proceedings of the 26th International Conference on Machine Learning*, pages 897–904, 2009. (cite on page 37, 38, 57, 59, 147)

M. D. Reid and R. C. Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010. (cite on page 31, 37, 77)

D. Rezende and S. Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1530–1538, 2015. (cite on page 4)

R. T. Rockafellar. *Convex Analysis*, volume 28. Princeton University Press, 1970. (cite on page 23)

F. Rosenblatt. *The Perceptron, a Perceiving and Recognizing Automaton.* Cornell Aeronautical Laboratory, 1957. (cite on page 2, 36)

F. Rosenblatt. *Principles of neurodynamics. perceptrons and the theory of brain mechanisms.* Spartan Books, 1961. (cite on page 2)

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. (cite on page 2, 4)

O. Russakovsky, J. Deng, Z. Huang, A. C. Berg, and L. Fei-Fei. Detecting avocados to zucchinis: what have we done, and where are we going? In *Proceedings of the 16th IEEE International Conference on Computer Vision*, pages 2064–2071, 2013. (cite on page 4)

B. Russell. *The Problems of Philosophy.* Oxford University Press, 1912. (cite on page 15)

S. S. M. Salehi, D. Erdogmus, and A. Gholipour. Tversky loss function for image segmentation using 3D fully convolutional deep networks. In *International Workshop on Machine Learning in Medical Imaging*, pages 379–387. Springer, 2017. (cite on page 17)

H. Salman, J. Li, I. Razenshteyn, P. Zhang, H. Zhang, S. Bubeck, and G. Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems 32*, pages 11289–11300, 2019. (cite on page 77)

A. Sanyal, P. Kumar, P. Kar, S. Chawla, and F. Sebastiani. Optimizing non-decomposable measures with deep networks. *Machine Learning*, 107(8-10): 1597–1620, 2018. (cite on page 57)

N. Saunshi, O. Plevrakis, S. Arora, M. Khodak, and H. Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5628–5637, 2019. (cite on page 117, 120, 125)

L. J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971. (cite on page 36, 37, 147)

R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990. (cite on page 6)

C. Scott. Surrogate losses and regret bounds for cost-sensitive classification with example-dependent costs. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 153–160, 2011. (cite on page 10, 78)

C. Scott. Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics*, 6:958–992, 2012. (cite on page 10, 44, 58, 78)

S. A. Seshia, D. Sadigh, and S. S. Sastry. Towards verified artificial intelligence. *arXiv preprint arXiv:1606.08514*, 2016. (cite on page 11)

S. A. Seshia, A. Desai, T. Dreossi, D. J. Fremont, S. Ghosh, E. Kim, S. Shivakumar, M. Vazquez-Chanlatte, and X. Yue. Formal specification for deep neural networks. In *International Symposium on Automated Technology for Verification and Analysis*, pages 20–34. Springer, 2018. (cite on page 77)

A. Shafahi, W. R. Huang, C. Studer, S. Feizi, and T. Goldstein. Are adversarial examples inevitable? *arXiv preprint arXiv:1809.02104*, 2018. (cite on page 148)

U. Shaham, Y. Yamada, and S. Negahban. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. *Neurocomputing*, 307:195–204, 2018. (cite on page 12, 18, 77)

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. (cite on page 9, 28)

J. Shawe-Taylor, N. Cristianini, et al. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004. (cite on page 3)

Z. Shen, J. Liu, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021. (cite on page 11)

T. Shimada, H. Bao, I. Sato, and M. Sugiyama. Classification from pairwise similarities/dissimilarities and unlabeled data via empirical risk minimization. *Neural Computation*, 33(5):1234–1268, 2021. (cite on page 118, 121, 124, 127, 128, 129, 133)

H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90 (2):227–244, 2000. (cite on page 8)

P. K. Shivaswamy, C. Bhattacharyya, and A. J. Smola. Second order cone programming approaches for handling missing and uncertain data. *Journal of Machine Learning Research*, 7:1283–1314, 2006. (cite on page 77)

E. H. Shuford, A. Albert, and H. E. Massengill. Admissible probability measurement procedures. *Psychometrika*, 31(2):125–145, 1966. (cite on page 37)

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529 (7587):484–489, 2016. (cite on page 1, 4)

A. Sinha, H. Namkoong, and J. Duchi. Certifying some distributional robustness with principled adversarial training. In *Proceedings of the 6th International Conference on Learning Representations*, 2018. (cite on page 78, 143)

N. K. Smith. *A Commentary to Kant's "Critique of Pure Reason"*. 1918. (cite on page 21)

P. Smolensky. *Information Processing in Dynamical Systems: Foundations of Harmony Theory*. 1986. (cite on page 3, 4)

M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009. (cite on page 16, 45)

T. Soma and Y. Yoshida. Statistical learning with conditional value at risk. *arXiv preprint arXiv:2002.05826*, 2020. (cite on page 143)

I. Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1):128–142, 2005. (cite on page 9)

I. Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287, 2007. (cite on page 18, 31, 32, 33, 36, 72, 76, 78, 79, 80, 99, 127, 135)

I. Steinwart, C. Pasin, R. Williamson, and S. Zhang. Elicitation and identification of properties. In *Proceedings of the 27th Conference on Learning Theory*, pages 482–526, 2014. (cite on page 145)

A. S. Suggala, A. Prasad, V. Nagarajan, and P. Ravikumar. Revisiting adversarial risk. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 2331–2339, 2019. (cite on page 114, 148)

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018. (cite on page 5)

A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor. Policy gradient for coherent risk measures. In *Advances in Neural Information Processing Systems 28*, pages 1468–1476, 2015. (cite on page 143)

J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. LINE: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1067–1077, 2015. (cite on page 125)

T. T. Tanimoto. Elementary mathematical theory of classification and prediction. *IBM Internal Report*, 1958. (cite on page 16)

S. K. Tavker, H. G. Ramaswamy, and H. Narasimhan. Consistent plug-in classifiers for complex objectives and constraints. In *Advances in Neural Information Processing Systems 33*, volume 33, 2020. (cite on page 57)

A. Tewari and P. L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025, 2007. (cite on page 10, 36, 78)

S. Thrun and L. Pratt. *Learning to Learn*. Springer Science & Business Media, 1998. (cite on page 14)

D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. In *Proceedings of the 7th International Conference on Learning Representations*, 2019. (cite on page 114)

Y. Tsuzuku, I. Sato, and M. Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *Advances in Neural Information Processing Systems 31*, pages 6541–6550, 2018. (cite on page 71, 75, 77)

A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950. (cite on page 1)

A. Tversky. Features of similarity. *Psychological Review*, 84(4):327, 1977. (cite on page 17)

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11): 1134–1142, 1984. ISSN 0001-0782. (cite on page 2, 6)

S. van de Geer. *Empirical Processes in M-estimation*. Cambridge University Press, 2000. (cite on page 8, 43, 46)

L. van der Maaten, E. Postma, and J. van den Herik. Dimensionality reduction: a comparative. *Journal of Machine Learning Research*, 10(66-71):13, 2009. (cite on page 5)

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000. (cite on page 44, 55)

C. J. van Rijsbergen. *Foundation of Evaluation*. Number 4. 1974. (cite on page 16, 43, 46)

B. van Rooyen, A. Menon, and R. C. Williamson. Learning with symmetric label noise: The importance of being unhinged. In *Advances in Neural Information Processing Systems 28*, pages 10–18, 2015. (cite on page 78, 84)

V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998. (cite on page 5, 6, 7, 9, 46, 143)

V. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer Science & Business Media, 2006. (cite on page 14)

V. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971. (cite on page 7, 28, 29)

V. Vapnik and R. Izmailov. Rethinking statistical learning theory: learning using statistical invariants. *Machine Learning*, 108(3):381–423, 2019. (cite on page 7)

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008, 2017. (cite on page 4)

S. Verma and J. Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (Fairware)*, pages 1–7. IEEE, 2018. (cite on page 148)

S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgwardt. Graph kernels. *Journal of Machine Learning Research*, 11:1201–1242, 2010. (cite on page 3)

U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17 (4):395–416, 2007. (cite on page 129)

K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl. Constrained k-means clustering with background knowledge. In *Proceedings of the 18th International Conference on Machine Learning*, volume 1, pages 577–584, 2001. (cite on page 116, 117, 129)

L. Wang, M. Sugiyama, C. Yang, K. Hatano, and J. Feng. Theory and algorithm for learning with dissimilarity functions. *Neural Computation*, 21(5):1459–1484, 2009. (cite on page 19, 116)

X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the 18th IEEE International Conference on Computer Vision*, pages 2794–2802, 2015. (cite on page 116)

K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009. (cite on page 19, 116, 117)

R. C. Williamson, E. Vernet, and M. D. Reid. Composite multiclass losses. *Journal of Machine Learning Research*, 17:1–52, 2016. (cite on page 38)

R. L. Winkler and A. H. Murphy. “Good” probability assessors. *Journal of Applied Meteorology and Climatology*, 7(5):751–758, 1968. (cite on page 36)

E. Wong and Z. Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5283–5292, 2018. (cite on page 18, 71, 77)

S. Wu, X. Xia, T. Liu, B. Han, M. Gong, N. Wang, H. Liu, and G. Niu. Multi-class classification from noisy-similarity-labeled data. *arXiv preprint arXiv:2002.06508*, 2020. (cite on page 118)

H. Xiao, B. Biggio, B. Nelson, H. Xiao, C. Eckert, and F. Roli. Support vector machines under adversarial label contamination. *Neurocomputing*, 160:53–62, 2015. (cite on page 18)

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. (cite on page 128)

E. P. Xing, M. I. Jordan, S. J. Russell, and A. Y. Ng. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems 16*, pages 521–528, 2003. (cite on page 19, 116, 117)

H. Xu and S. Mannor. Robustness and generalization. *Machine Learning*, 86(3): 391–423, 2012. (cite on page 71)

H. Xu, C. Caramanis, and S. Mannor. Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10:1485–1510, 2009. (cite on page 71, 75, 77)

R. Xu and D. Wunsch. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005. (cite on page 5)

B. Yan, O. Koyejo, K. Zhong, and P. Ravikumar. Binary classification with Karmic, threshold-quasi-concave metrics. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5531–5540, 2018. (cite on page 17, 44, 52, 53, 57, 58, 62)

R. Yan, J. Zhang, J. Yang, and A. G. Hauptmann. A discriminative learning framework with pairwise constraints for video object classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):578–593, 2006. (cite on page 116)

H. Yanaka, K. Mineshima, D. Bekki, and K. Inui. Do neural models learn systematicity of monotonicity inference in natural language? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6105–6117, 2020. (cite on page 2)

J. Yu and M. Blaschko. Learning submodular losses with the Lovász hinge. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1623–1631, 2015. (cite on page 56)

H. Zeng and Y.-M. Cheung. Semi-supervised maximum margin clustering with pairwise constraints. *IEEE Transactions on Knowledge and Data Engineering*, 24(5):926–939, 2011. (cite on page 117)

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021. (cite on page 8)

H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7472–7482, 2019a. (cite on page 78)

J. Zhang and R. Yan. On the value of pairwise constraints in classification and consistency. In *Proceedings of the 24th International Conference on Machine Learning*, pages 1111–1118. ACM, 2007. (cite on page 118, 121, 124, 127, 129)

M. Zhang and S. Agarwal. Bayes consistency vs. H-consistency: The interplay between surrogate loss functions and the scoring function class. In *Advances in Neural Information Processing Systems 33*, 2020. (cite on page 32, 77, 142)

M. Zhang, H. G. Ramaswamy, and S. Agarwal. Convex calibrated surrogates for the multi-label F-measure. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11246–11255, 2020. (cite on page 10, 56)

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004a. (cite on page 9, 10, 26, 27, 50, 78)

T. Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004b. (cite on page 10, 36, 50, 78, 142)

Y. Zhang, T. Liu, M. Long, and M. Jordan. Bridging theory and algorithm for domain adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7404–7413, 2019b. (cite on page 78)

K. Zhao, S. Gao, W. Wang, and M.-M. Cheng. Optimizing the F-measure for threshold-free salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8849–8857, 2019. (cite on page 56)

Z.-H. Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2018. (cite on page 5)