On Learning, Fairness, and Complexity

Sílvia Casacuberta Puig

St John's College

University of Oxford

A thesis submitted for the degree of

Master of Science (by Research) in Computer Science

Trinity 2025

On Learning, Fairness, and Complexity

Abstract

In this thesis we study the learning and complexity-theoretic underpinnings of the multigroup fairness framework for prediction algorithms. Multiaccuracy and multicalibration are two primary multigroup fairness notions, which ensure accurate and calibrated predictions, respectively, for every subpopulation that can be identified within a specified class of computations [HKRR18]. They both can be achieved from a single learning primitive: weak agnostic learning. A line of work starting from [GKR$^+$22] has shown that multicalibration implies a very strong indistinguishability-based form of learning called *omniprediction*. The multigroup fairness framework is also deeply connected to complexity theory through the Regularity Lemma and its various implications [CDV24].

We provide a thorough study of the connections between multigroup fairness notions, the central learning primitive of weak agnostic learning, and the fundamental Hardcore Lemma in complexity theory. We find that multiaccuracy in itself is rather weak, but that the addition of global calibration (this notion is called *calibrated multiaccuracy*) boosts its power substantially, enough to recover implications that were previously known only assuming the stronger notion of multicalibration.

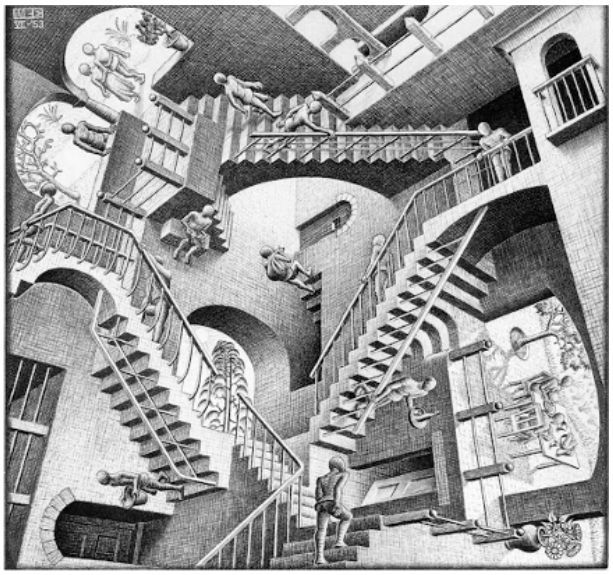
We give evidence that multiaccuracy might not be as powerful as standard weak agnostic learning, by showing that there is no way to post-process a multiaccurate predictor to get a weak learner, even assuming the best hypothesis has correlation $1/2$. However, by also requiring the predictor to be calibrated, we recover not just weak, but strong agnostic learning. A similar picture emerges when we consider the derivation of hardcore measures from predictors satisfying multigroup fairness notions [TTV09; CDV24]. On the one hand, while multiaccuracy only yields hardcore measures of density half the optimal, we show that (a weighted version of) calibrated multiaccuracy achieves optimal density.

Our results yield new insights into the complementary roles played by multiaccuracy and calibration in each setting. They shed light on why multiaccuracy and global calibration, although not particularly powerful by themselves, together yield considerably stronger notions.

We further study the connections between the multigroup fairness framework and the problem of learning selective classifiers, which are predictors that are allowed to abstain on some fraction of the domain. Building on the notion of omniprediction (which is in turn built using tools from the multigroup fairness framework), given a pre-specified class of loss functions, we provide an algorithm for efficiently building a single classifier that learns abstentions and predictions optimally for every loss in the entire class, where the abstentions are decided efficiently for each specific loss function by applying a fixed post-processing function. We call this classifier a *selective omnipredictor*. Our algorithm and theoretical guarantees generalize the previously-known algorithms for learning selective classifiers in formal learning-theoretic models [KKM12].

We then extend the traditional multigroup fairness algorithms to the selective classification setting and show that we can use a calibrated and multiaccurate predictor to efficiently build selective classifiers that abstain optimally not only globally but also locally within each of the groups in any pre-specified collection of possibly intersecting subgroups of the domain, and are also accurate when they do not abstain. This provides yet another use case of the notion of calibrated multiaccuracy. Moreover, we show how our abstention algorithms can be used as conformal prediction methods in the binary classification setting to achieve both marginal and group-conditional coverage guarantees for an intersecting collection of groups. We provide empirical evaluations for all of our theoretical results, demonstrating the practicality of our learning algorithms for the goal of abstaining optimally and fairly.

Contents



M. C. Escher, *Relativity*, 1953.
Inspired by Parikshit Gopalan's talk at the Institute for Advanced Study, April 2022.

The new form of the problem can be described in terms of a game which we call the "imitation game." It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart front the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman.

Alan M. Turing, *Computing Machinery and Intelligence*, 1950.

These new Imitation Games lead to novel, precise, and operative definitions of classical notions, including secret, knowledge, privacy, randomness, proof, fairness, and others. These definitions have in turn led to numerous results, applications, and understanding. (...) Central to each of these settings are computational and information theoretic limitations placed on the referee in the relevant Imitation Game.

Avi Wigderson, 9th Heidelberg Laureate Forum lecture, 2022.

This idea that there is generality in the specific is of far-reaching importance.

Douglas R. Hofstadter, *Gödel, Escher, Bach: An Eternal Golden Braid*, 1979.

Acknowledgments

First and foremost, I want to thank Varun Kanade for supervising my work during my time at Oxford. Varun has been an extraordinary advisor who has allowed me to freely pursue my research interests while giving me endless time to learn from his expertise, rigorous thinking, and brilliant intuition. I feel very lucky to have learned and worked with him.

I want to thank Parikshit Gopalan and Omer Reingold for joining our project from California and for making me grow so much as a researcher. It was my first time working with three professors at once, and it is hard to put into words how much I learned in every single meeting with the four of us, and how incredibly exciting all of the research process was for me throughout the months.

I am thankful to Rahul Santhanam and Igor Carboni Oliveira for providing extensive feedback on this thesis, which has improved my understanding and given me new ideas for future directions.

I am very grateful to Moritz Hardt for hosting me in his Social Foundations of Computation group at the Max Planck Institute for Intelligent Systems in Tübingen, where I spent the summer of 2024 as a research intern. This wonderful group has given me so much hope for what the field of computer science can be, both in terms of research topics and community. I am very grateful to every member of the Social Foundations group for welcoming me so kindly.

I am very thankful to Salil Vadhan and Cynthia Dwork for introducing me to the area of research of this thesis during my undergraduate. With them I explored the complexity-theoretic implications of multicalibration in my undergraduate thesis, and so completing this masters thesis as a continuation of the work that we started, while incorporating the connections to learning theory that I have learned thanks to Varun, has been incredibly intellectually fulfilling for me.

I have also benefited from participating in various research conferences during these two years, including FORC, STOC, ICML, and the World Congress in Probability and Statistics in 2024, FAccT and COLT in 2025, and the Oxbridge Women in Computer Science Conference both years. I thank the participants who have shared useful feedback on the work in this thesis, including Michael P. Kim, Charlotte Peale, Princewill Okoroafor, and Juan Carlos Perdomo.

I want to thank the Rhodes Trust for providing me with the opportunity to come to Oxford in the first place and for sponsoring my studies. The Rhodes community has been the center of my life at Oxford, and for that I am very grateful to the scholarship. I am also thankful to St John's College for having me these two years and to my college advisor Stefan Kiefer for the dinners at the high table. I am also appreciative to the organizers of the Learning Theory and Statistical Optimization reading group, the Oxford Internet Institute, the Institute for Ethics in AI, and the Cosmos Institute. I am grateful that I was able to be a course assistant for Varun's course on computational learning theory twice, where I learned a lot from the students.

Oxford has truly felt like a home to me and for that I am thankful to my friends Amisha Kambath, Brian Wee, Matt Kearney, Astrid Liden, Sophie Huttner, Javi Chico, Isaac Robinson, Hansa Mukherjee, Ann-Hui Ching, Margaret Williams, Rishi Bansal, Kunal Handa, and Bridget Smart. It is a great joy to be able to do computer science research with friends, and for that I thank Emin Berker, Jack Cook, and Tina Behzad. I am also grateful to my friends from back home in Barcelona, whose friendship has remained a constant despite my years living abroad.

Lastly, I want to thank Henry Large and my family for all of their love and for everything that they do for me. To Henry, for making Oxford the happiest place, to my brother, for being my role model ever since I was little, to my father, for instilling in me a passion for mathematics, and to my mother, for always believing in me.

1 Introduction

Undeniably, algorithms are informing decisions that reach ever more deeply into our lives, from news article recommendations to criminal sentencing decisions to healthcare diagnostics. This progress, however, raises (and is impeded by) a host of concerns regarding the societal impact of computation.

TOC for Fairness: a Simons Collaboration Project

The role of computer science in our society has been profoundly reshaped during these past few years. Gone is the time when algorithms only dealt with abstract objects such as graphs or matrices to find the shortest path between two nodes. Today, *people* are at the center of many algorithmic systems: algorithms are now powered by and continuously fed people's (sensitive) data and are widely employed to determine life-altering decisions for individuals in all spheres of society. Automated decision-making has become ubiquitous, and algorithms are being used to decide whether someone should receive a loan, be hired for a job, receive bail, or be treated for a certain illness. These types of decisions usually take the form of *prediction algorithms*, which map individuals to a "risk score" or a "predicted probability". This is a number in the $[0, 1]$ interval understood as a likelihood. For example, we employ algorithms to try to predict the probability that someone repays a loan on time or the probability that someone recommits a crime. These probabilities are then used to make decisions about individuals; for example, to determine how to allocate a scarce resource or to decide who is eligible for certain government benefits.

With great power comes great responsibility, and this responsibility is naturally shared by us computer scientists, given that we are the ones developing the algorithmic toolkit that is then deployed in the world. Unfortunately, we seem to be falling short: along with this rise in automated decision-making, evidence has amounted demonstrating how algorithms can lead to various forms of discrimination and to the enshrinement of present inequalities. Many real-life examples have been documented, filling numerous books, policy reports, and research papers [LMKA16; ONe17; BG18; Bro18; Vin18; Eub18; HTGG22; DK23; Ben23]. While ensuring the responsible use of algorithms in society is a complex and interdisciplinary task, computer science plays a key role in developing the appropriate technical tools and scientific vocabulary for dealing with these issues and understanding

what can go wrong when an algorithm makes decisions about people. For example, in many cases, algorithmic discrimination or bias amounts to an algorithm being inaccurate on a subgroup of the population. In these kind of scenarios, we view algorithmic fairness as requiring more stringent notions of accuracy than the traditional ones, which typically only measure global error. Because a minority subgroup is small, this inaccuracy inside the minority subgroup is barely detectable when looking at the algorithm's global accuracy, but it is nonetheless a flaw of the algorithm that can have major implications for the individuals who are members of these subgroups. Thus, beyond any "fairness" considerations, an algorithm that is inaccurate on a subgroup of the domain is simply a bad algorithm, and us computer scientists should find ways of detecting these flaws and improving the algorithm.

The role of theoretical computer science. A lot of the discussions around algorithmic fairness and responsible machine learning can sometimes feel too vague and complex for us to do rigorous technical work on, especially given the intricacies of the social systems that algorithms are deployed on. But finding good definitions that formalize important concepts and then allow us to build robust methods with provable guarantees is precisely what theoretical computer science has always excelled at. In very broad strokes, our field started with Turing formalizing the notion of *computability*, then established cryptography as a scientific discipline by formalizing the notion of *security*, and then developed the field of differential privacy by formalizing the notion of *privacy*. In trying to keep up with this spirit, over the past few years several theoretical computer scientists have developed the *multigroup fairness framework*, which formalizes the natural idea of requiring a predictor to be accurate or calibrated not just globally, but also locally within any subgroup in a rich and possibly intersecting collection. This thesis is dedicated to the study of the learning and complexity-theoretic underpinnings of this now very rich framework.

While the prevalence of algorithms around us should be reason enough for us to think about local accuracy and calibration, among many other similar problems, good definitions and frameworks do not only allow us to solve the "practical" problems that motivated the creation of the framework in the first place. Good formalizations yield new techniques that are broadly applicable, new and fresh insights into computer science, and fruitful connections with other subfields. We have seen this phenomenon repeatedly; for example, in the connections between cryptographic primitives and complexity theory, or between differential privacy and adaptive data analysis. In this thesis, we will see how the modern multigroup fairness framework elucidates fundamental "classical" concepts in theoretical computer science such as *agnostic learning* and *hardness amplification* in new and illuminating ways.

1.1 The multigroup fairness framework

The *multigroup fairness framework* was proposed in 2018 by Hébert-Johnson, Kim, Reingold, and Rothblum [HKRR18], with a similar approach being studied in [KNRW18; LSH19]. The central idea is the following: we are given a collection $\mathcal{C}$ of subgroups of the population that we wish to protect, which can intersect arbitrarily, and our goal is to build a predictor that satisfies some desired property globally over the domain and also locally when conditioning on any of subgroups in the collection $\mathcal{C}$. When the property that we want to enforce locally is that of accuracy in expectation, then we say that the resulting predictor is $\mathcal{C}$-*multiaccurate*, where "multi" refers to

each of the subgroups $c \in \mathcal{C}$. When the local property corresponds to calibration, we say that the resulting predictor is $\mathcal{C}$-*multicalibrated*, which is a stronger notion. Calibration enforces the predicted probabilities to be "meaningful", in the sense that "they mean what they say": if we look at the set of points where our predictor is predicting 0.7, for example, we want the expectation of the true labels on that set of points to be approximately 0.7.

Multiaccuracy and multicalibration are *definitions* of what we want a "fair" predictor to satisfy. But can we always construct such predictors? Hébert-Johnson et al. answer this in the positive by showing that we can efficiently build multiaccurate and multicalibrated predictors for an arbitrary collection $\mathcal{C}$ using a boosting-based algorithm. Specifically, both algorithms follow the same iterative recipe: at each step, we search for some $c \in \mathcal{C}$ that "witnesses" a violation of the multiaccuracy/multicalibration condition. To do so, we use the learning primitive of a *weak agnostic learner* for the class $\mathcal{C}$, which can identify whether there is some correlation with the residuals. If so, we can then use this same witness to update the predictions by taking a gradient step and show that we have improved the squared loss of our predictor by at least some amount. Then, a potential argument shows that our algorithm terminates within not too many steps. At that point, there are no more violations of the multiaccuracy/multicalibration definition, and thus our predictor satisfies these desired notions for all $c \in \mathcal{C}$.

Since 2018, the multigroup fairness framework has lead to an extremely fruitful line of work, drawing connections to loss minimization [GKR+22; HNRY23; GGKS23; GOR+24], statistical inference [KKG+22; NR23; WLCW24], conformal prediction [JLP+21; JNRR23; GJN+22], confidence scoring in large language models [DBFR24], game theory [LNPR22; NRRX23; HQYZ24; HJZ24], causal inference [KKZ24], and the model multiplicity problem [RTW23; DNW24; BCDT25], among others, as well as enjoying several practical applications [KGZ19; BRA+20; PKD+21; HDNS24]. This framework has been one of the most successful lines of work to come out of the recent efforts to obtain formal definitions and provable guarantees in the field of algorithmic fairness by using the tools and perspectives of theoretical computer science.

1.2 Connections to complexity theory

Recently, the multigroup fairness framework has also been connected to classical problems in complexity theory [DLLT23; Cas23; CDV24; MPV25; HV25]. This relationship is realized by observing that the multiaccuracy theorem (i.e., the claim that we can efficiently build a multiaccurate predictor for a given collection $\mathcal{C}$ of groups [HKRR18]) is exactly equivalent to the so-called *Regularity Lemma* in complexity theory, first showed by Trevisan, Tulsiani, and Vadhan in 2009 [TTV09]. In their setting, the collection $\mathcal{C}$ of minority groups corresponds to a more general notion of *class of distinguishers*, sometimes chosen to be a family of circuits of a certain size. Here, their goal is to construct a "low-complexity" function (i.e., "simple" as measured with respect to the class of distinguishers $\mathcal{C}$) that "simulates" an arbitrarily complex function, in the sense that the two look indistinguishable to all of the distinguishers in the class $\mathcal{C}$. This notion of indistinguishability with respect to the class $\mathcal{C}$ turns out to be exactly equivalent to $\mathcal{C}$-multiaccuracy.[1]

[1] In fact, $\mathcal{C}$-computational indistinguishability and $\mathcal{C}$-multiaccuracy can both be viewed as variations of Turing's remarkable Imitation Game, where the functions $c \in \mathcal{C}$ play the role of the interrogator, the true labels play the role of the human, and the low-complexity simulator/multiaccurate predictor plays the role of the machine.

Crucially, the Regularity Lemma implies central results in complexity theory and related subjects [TTV09; DLLT23; CDV24; HV25]. From it, we can prove all of the following results:

- Impagliazzo's Hardcore Lemma in complexity theory [Imp95; Hol05; TTV09; CDV24].
- The Dense Model Theorem in additive combinatorics [RTTV08; TZ08; GT08; CDV24].
- Characterizations of pseudoentropy in information theory [VZ12; Zhe14; CDV24; MPV25; HV25].
- Yao's XOR Lemma in complexity theory [TTV09; GNW11].
- The Frieze-Kannan Regularity Lemma in graph theory [FK99; TTV09; Skó17; DLLT23].
- Chain rules for computational entropy [GW11; JP14].
- Chang's Inequality in Fourier analysis of Boolean functions [IMR14].
- Equivalences between weak notions of zero knowledge in cryptography [CLP15].

Given that multicalibration (MC) is a *stronger* notion than multiaccuracy (MA), and that the Regularity Lemma is equivalent to the multiaccuracy theorem, a natural question that arises in this context is the following: What stronger and more general versions of these theorems do we obtain if we start from the multicalibration theorem instead? This question was recently studied for the cases of Impagliazzo's Hardcore Lemma, the Dense Model Theorem, characterizations of pseudoentropy, and the Frieze-Kannan Regularity Lemma [DLLT23; Cas23; CDV24]. In this thesis we further study the connections between the multigroup framework and Impagliazzo's Hardcore Lemma (IHCL).

Impagliazzo's Hardcore Lemma. Stated informally, IHCL says that if a function g is somewhat hard to compute on average by a family $\mathcal{C}'$ of Boolean functions (in that the outputs of each of the functions in $\mathcal{C}'$ differ from g on at least a δ fraction of the input points), then there exists a large fixed subset of the inputs (called the "hardcore set") for which the function is very hard to compute, in the sense that g is maximally unpredictable to the family $\mathcal{C}$.[2] That is, no distinguisher in $\mathcal{C}$ can do better than random guessing when trying to guess the output of g. The original paper by Impagliazzo from 1995 provides two proofs of IHCL: one boosting-based proof and one based on the min-max theorem [Imp95]. Both proofs obtain a hardcore set of density (i.e., the size that the set occupies within the domain) δ, where δ is the parameter that quantifies the average hardness of the function in the assumption of the theorem. However, the optimal lower bound on the density of the hardcore set is of 2δ, and it took 10 years for Holenstein to show that this optimal density is indeed achievable [Hol05].

Through multicalibration, Casacuberta, Dwork, and Vadhan showed that we can obtain a stronger and more general version of the Hardcore Lemma, which they call IHCL++ [CDV24].

[2]There is a circuit size difference between the assumption and the conclusion of IHCL: the class $\mathcal{C}'$ is an enlarged class which contains all of the functions that have "low-complexity" relative to $\mathcal{C}$. We require g to be somewhat hard to compute with respect to the enlarged class $\mathcal{C}'$, and then we are able to show that existence of a hardcore set with respect to the smaller class $\mathcal{C}$. It was recently shown that this loss in circuit size is unavoidable regardless of proof strategy [BKST24].

Essentially, we can view the notion multicalibration from a complexity-theoretic point of view as follows: by considering the level sets of a multicalibrated predictor, we have efficiently constructed a low-complexity partition $\mathcal{P}$ such that on every (large enough) piece $P \in \mathcal{P}$, the true labels are $\mathcal{C}$-indistinguishable from a constant function (which is in turn related to the expected value of the labels on the piece P, a quantity that can be viewed as the "balance" of g on P). This allows us to construct a "small" hardcore set within each piece $P \in \mathcal{P}$, whose density is lower-bounded by two times the balance parameter of g on P. Crucially, this statement holds for *any* Boolean function g, unlike the original IHCL statement, where we must require the input function g to be somewhat hard to compute to begin with. If we do bring back this weak hardness assumption, then they show that we recover IHCL with *optimal* density 2δ as a corollary by gluing the per-piece small hardcore sets.

This corollary is very intriguing when viewed side by side with the result of Trevisan, Tulsiani, and Vadhan regarding IHCL, who show that from the Regularity Lemma (i.e., from a multiaccurate predictor) we can construct a hardcore set of density δ, which is suboptimal [TTV09]. Therefore, a glaring question emerges: what is the weakest multigroup fairness definition that allows us to obtain a hardcore set of optimal density 2δ? This question is important to understand the potential limitations of the Regularity Lemma/multiaccuracy, but it is also important to keep in mind that multicalibration comes at a much higher computational price than multiaccuracy: the iterative algorithm for multicalibration requires many more calls to the weak agnostic learner. This implies that, while [CDV24] are able to obtain a hardcore set of optimal density, they incur a much larger circuit size loss in the difference between $\mathcal{C}'$ and $\mathcal{C}$ in IHCL. Therefore, using the multigroup fairness framework, we would like to be able to obtain a Hardcore Lemma of optimal size but with much better circuit size parameters. We remark that we can equivalently move between hardcore *sets* and hardcore *measures* via a probabilistic method argument. In the case of measures, the hardness of the function occurs when sampling according to the hardcore measure (as opposed to when restricting the domain to the hardcore set). We can similarly translate the notion of the density of the set by using an appropriate notion of "density" of the measure.

Boosting and IHCL. A deep insight in theoretical computer science is that hardness and learning are two sides of the same coin. Specifically, the Hardcore Lemma and the notion of *boosting* in learning theory are intimately related. The key idea in boosting is to combine multiple "weak" hypotheses (in the sense that these have only learned the labels slightly better than random) in order to produce a "strong" hypothesis that has very high correlation with the labels. Schapire and Freund were the first to design boosting algorithms, thus proving the equivalence between weak and strong learnability [Sch90; Fre95]. Many practical boosting algorithms followed, such as the celebrated AdaBoost [FS97]. All of the main boosting algorithms tend to use the same recipe: they invoke a weak learner multiple times on distributions that we modify at each iteration, giving more weight to the points that have been hard to learn so far.

The relationship between the Hardcore Lemma and boosting becomes clear when we consider the contrapositive of IHCL, which essentially states that if a function g can be approximated slightly better than random for any dense-enough distribution, then we can find some function that approximates g very well, in the sense that the function is *not* somewhat hard to compute on average. In fact, one of the two original proofs of IHCL by Impagliazzo is a boosting-based proof, where

IHCL is shown by contradiction through this iterative procedure of calling the weak learners, which are guaranteed to exist from the assumption of IHCL. Here, we see yet another re-interpretation of the class $\mathcal{C}$: in the case of IHCL, we understand the functions in $\mathcal{C}$ as distinguishers/circuits of a certain size, whereas from the boosting/learning perspective we view them as *concepts* (or hypotheses) belonging to a concept class $\mathcal{C}$.

This connection between the Hardcore Lemma and boosting was made explicit by Kilvans and Servedio in 2003, who plugged in various boosting algorithms into a general boosting-based proof of IHCL, thus obtaining various IHCL statements with different density guarantees and circuit size parameters [KS03]. In fact, they explicitly relate the smoothness parameter of the boosting algorithm with the density of the hardcore set, but, intriguingly, none of the PAC-based boosting algorithms (i.e., when we work in the realizable setting, where we assume that the underlying distribution is labeled by some concept in $\mathcal{C}$) that they plug in are able to achieve a hardcore sets of optimal density 2δ. Feldman later revisited the connection between boosting and IHCL in the agnostic setting (where we do not have the realizability assumption), showing that hardcore set constructions achieving the optimal density give agnostic boosting and vice-versa [Fel09a]. This connection between hardness and learning is another reason for why we care about the density of the hardcore set: when we view it from this "contrapositive" perspective, more hardness allows us to have better learning. Even if we stay on the "hardness side" of the coin, where we view IHCL as a hardness amplification result, then extracting the maximum possible hardness is useful for the applications of IHCL to cryptography, specifically in the construction of pseudorandom generators [VZ12; VZ13]. Hardness amplification is also very related to the well-studied problem of derandomization [Kab02].

Given these connections between hardcore set constructions and boosting, it is natural to ask what the learning-theoretic/boosting interpretation of IHCL++ is, where IHCL++ corresponds to the stronger and more general version of IHCL shown in [CDV24] that we obtain through multicalibration (rather than from multiaccuracy, as done through the Regularity Lemma [TTV09]). Intuitively, it should correspond to some form of stronger agnostic learning, in light of Feldman's observation.

Class	**Context**	**Interpretation of $\mathcal{C}$**
	Multigroup fairness	Indicator functions of subgroups of interest in the population
$\mathcal{C}$	Complexity theory, IHCL	Distinguishers or bounded-size circuits
	Learning theory, Boosting	Concepts from a concept class

Table 1.1: Interpretations of the class $\mathcal{C}$ across fairness, complexity, and learning.

1.3 Connections to learning theory

From the equivalence between hardcore set constructions and forms of boosting, and in light of the connection between multiaccuracy and IHCL through the Regularity Lemma, it is already clear that there are deep connections between the multigroup fairness notions and learning theory. But

already since the original multicalibration paper, we know that the connection runs deeper.

First, algorithms for building multiaccurate (MA) and multicalibrated (MC) predictors all use a weak agnostic learner to detect whether there is some subgroup $c \in \mathcal{C}$ on which the predictor is not accurate or not calibrated. Therefore, from the multiaccuracy and multicalibration theorems it follows that if the class $\mathcal{C}$ is efficiently weak agnostically learnable, then we can efficiently construct MA and MC predictors. It is important to remark that weak and strong learnability are also equivalent in the agnostic setting, as shown by Kalai and Kanade [KK09] and Feldman [Fel09a]. In the case of multicalibration, Hébert-Johnson et al. show that the reduction goes both ways: we can multicalibrate with respect to $\mathcal{C}$ if and only if we can (weak) agnostically learn $\mathcal{C}$ [HKRR18]. It is natural to ask whether this if and only if relationship also holds in the case of multiaccuracy. A priori, it seems that the answer should be yes, given that testing for multiaccuracy violations corresponds to calling a weak agnostic learner. However, as we will see in this thesis, this relationship between multiaccuracy and weak agnostic learning turns out to be a lot more subtle.

Another powerful connection between the multigroup fairness framework and learning theory came through *omniprediction*, a learning paradigm that was first proposed by Gopalan, Kalai, Reingold, Sharan, and Wieder in 2021 [GKR+22]. In this work, they were concerned with a learning problem that at first does not seem to have any relationship to the multigroup fairness framework. In their case, the notion of omniprediction stems from the following observation: in the agnostic setting, the loss incurred by the predictor that we are building competes with the loss incurred by the best concept $c^* \in \mathcal{C}$. However, the best concept $c^* \in \mathcal{C}$ (i.e., the one that achieves the lowest loss) depends on the chosen loss function. In fact, if we have trained a predictor to optimize for a particular loss, we do not have a way of post-processing it so that it is optimal for a different loss.

It would be much more convenient to instead have a loss minimization paradigm that was somewhat "loss agnostic" in the following sense: we would like to train a *single* predictor that "works" for an entire class of loss functions. For any fixed loss from the class, we would like to efficiently post-process this single predictor such that the error that it incurs is no more than the loss incurred by the optimal predictor c^*_ℓ in $\mathcal{C}$ (with an ϵ slack), where c^*_ℓ depends on the choice of the loss function ℓ. This single predictor corresponds precisely to the definition of an omnipredictor.

Perhaps surprisingly, Gopalan et al. show that we can build an omnipredictor through the technique of multicalibration: if p is multicalibrated for the class $\mathcal{C}$, then p is an omnipredictor for $\mathcal{C}$ and forthe class of all convex, Lipschitz loss functions. This connection is maybe less surprising if we view it in light of the tight relationship between learning, fairness, and complexity that we study in this thesis: namely, a $\mathcal{C}$-multicalibrated predictor appears indistinguishable from the ground truth predictor to the functions $c \in \mathcal{C}$ (where we are implicitly relating the different interpretations of the class $\mathcal{C}$ as summarized in Figure 1.1, as we view $\mathcal{C}$ as a distinguisher class from the MC perspective and as a class of concepts in the setting of omniprediction). Given that the ground truth predictor incurs optimal loss, we would expect the loss incurred by the multicalibrated predictor to also be optimal from the perspective of the concepts in $\mathcal{C}$.

In a sense, for all of multicalibration, indistinguishability, and omniprediction/agnostic learning, everything we prove is about the relationship between the true labels and the predictor that we construct *as far as the functions in $\mathcal{C}$ can see*. This allows us to see the true labels and our predictions as computationally equivalent with respect to the class $\mathcal{C}$, which yields benefits in all of the

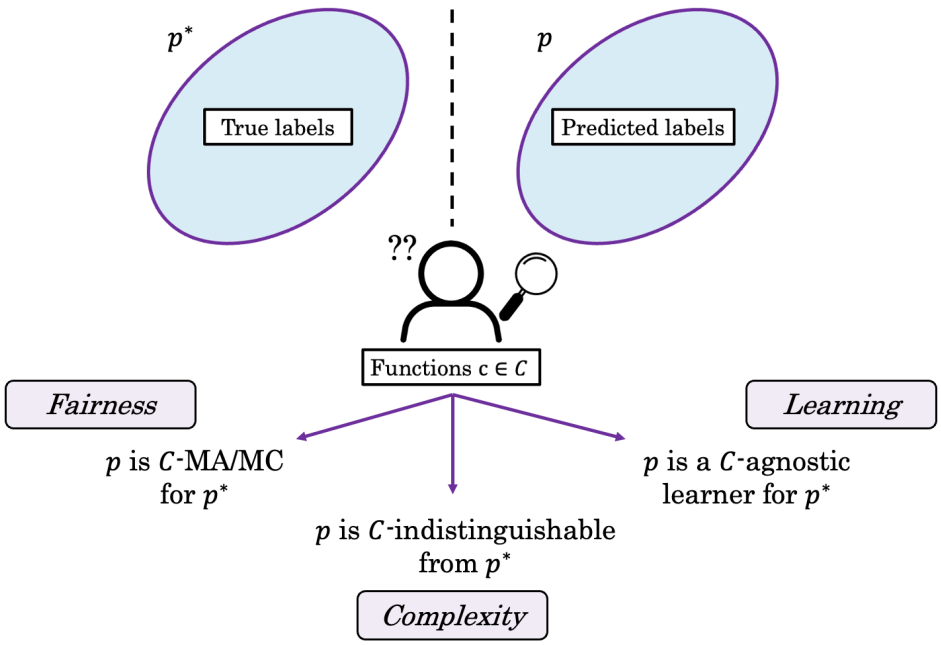


Figure 1.1: Understanding the notions of multigroup fairness, computational indistinguishability, and agnostic learning as an Imitation Game with respect to the class $\mathcal{C}$.

fairness, learning, and complexity-theoretic domains (Figure 1.1). This indistinguishability-based perspective of the multigroup fairness framework has also been formalized through the notion of *Outcome Indistinguishability* [DKR+21], which we can in turn use to understand the loss minimization paradigm in the omnipredictors framework from the indistinguishability lenses [GHK+23].

1.4 Our contributions to this picture

In Part I of this thesis, we complete the picture of relationships between multigroup fairness primitives, complexity-theoretic notions (hardcore sets), and learning-theoretic notions (learning primitives), which we summarize in Figure 1.2.

Hardcore measures of optimal density (Chapter 4). Following the construction of hardcore measures by [TTV09] (for multiaccuracy) and [CDV24] (for multicalibration), we show an explicit form for an optimal hardcore measure of density 2δ, starting from a predictor satisfying calibration and a weighted variant of multiaccuracy that we call *weighted multiaccuracy*. In multiaccuracy, we require that the expected prediction made by our predictor p on every subgroup to be accurate. For weighted multiaccuracy, we take the level sets of p and assign a weight value to each of the level sets, assigning more weight to the individuals with predicted p-value closer to $1/2$. We show how giving more weight to the level sets that have value close to $1/2$ (the specific weight value is provided in the formal statements) strengthens the transformation of a multiaccurate predictor to a hardcore measure first proposed by [TTV09], thus improving their construction. In terms of complexity, obtaining a weighted multiaccurate predictor is roughly equivalent to obtaining a usual multiaccurate predictor. We discuss this hardcore-ness analysis in Section 4.2. The key then comes

from adding *global* calibration to the picture. Global calibration is different from multicalibration in that we only require the predictor to be calibrated on average over the domain, rather than when conditioned on each of the subgroups in $\mathcal{C}$. We perform the density analysis in Section 4.6. Note that both (weighted) multiaccuracy and calibration follow from multicalibration (we always assume that the constant functions are in $\mathcal{C}$), but, importantly, constructing a (weighted) multiaccurate and calibrated predictor is much cheaper than constructing a full-fledged multicalibrated predictor. We compare the specific costs in Section 4.6. Recent work by Gopalan, Hu, Kim, Reingold, and Wieder shows that we can construct a calibrated and multiaccurate predictor at essentially the same cost as multiaccuracy alone [GHK+23]. Therefore, when compared to the result by [CDV24] showing that IHCL++ implies the Hardcore Lemma with density 2δ, we are able to obtain a hardcore set of optimal size with a much better circuit size loss.

Our result is part of a growing body of work demonstrating that calibrated multiaccuracy (and related variants) strikes a very good balance between utility and efficiency: it is much cheaper to construct than multicalibration (in all measures), but it preserves some of the great qualities of multicalibration. In our case, we are able to obtain a hardcore set of optimal density. In the case of Gopalan et al., they show how we can efficiently construct omnipredictors in the case where $\mathcal{C}$ is a family of Boolean functions (rather than bounded) from calibrated multiaccuracy, whereas the original paper on omnipredictors did so from multicalibration [GHK+23]. Recent work by Vadhan and Hu shows how we can provide pseudoentropy characterizations (including for Shannon and min-entropy) from multiaccuracy and a form of global calibration [HV25]. This is in contrast to the result in [CDV24], which did so through the notion of multicalibration. Naturally, multicalibration is still a stronger primitive, and so we cannot hope to recover all of its implications from calibrated multiaccuracy; for example, we cannot obtain IHCL++ from calibrated multiaccuracy.

Lastly, in light of the connections between boosting and the Hardcore Lemma, in Section 4.6 we clarify the stronger form of agnostic boosting that IHCL++ corresponds to (i.e., the ++ version of agnostic learning). It turns out that this corresponds to the notion of *swap agnostic learning* that was recently introduced by Gopalan, Kim, and Reingold [GKR24]. This answer is not so surprising, given that Gopalan et al. showed that swap agnostic learning can be obtained from multicalibration. However, we make this connection through the characterization of multicalibration in terms of hardness of prediction established by [CDV24]. In essence, the complexity-theoretic view of MC in [CDV24] tells us that in a multicalibrated partition $\mathcal{P}$, within every (large enough) piece P, the true labels are $\mathcal{C}$-indistinguishable from a constant function (which is in fact equal to the "balance" of the true labels on the piece). The main theorem in [CDV24] characterizes indistinguishability from a constant function in terms of hardness of prediction (generalizing Yao's equivalence between pseudorandomness and unpredictability). We explain how viewing multicalibration from this perspective of "hardness on each piece" of the partition provides a clear picture of the duality between learning and hardness in the ++ row of our picture (Figure 1.2): we can use it to obtain a "small" hardcore set within every piece (yielding IHCL++), or we can use it to argue that our predictions incur optimal loss in the agnostic sense within every piece (yielding swap agnostic learning). We can view swap agnostic learning as saying that we can learn up to irreducible error, where this irreducibility is from a computational perspective (i.e., it *appears* as true noise to the concepts $c \in \mathcal{C}$). While Gopalan et al. show that we can obtain swap agnostic learning from multicalibration with the ℓ_2 loss, here we show it for the ℓ_1 loss.

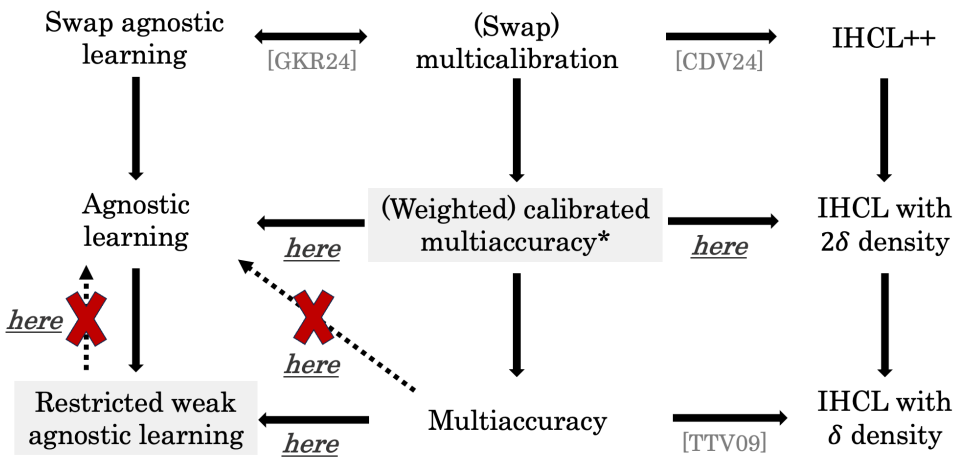


Figure 1.2: In this thesis, we complete the picture connecting learning, fairness, and complexity-theoretic notions.

Multiaccuracy and (weak) agnostic learning (Chapter 3). Given the equivalence between multicalibration and (weak) agnostic learning, we ask whether the same holds for multaccuracy. It would seem that the answer is yes, but somewhat surprisingly, in Section 3.1 we show that we can construct a concept class $\mathcal{C}$, a distribution on the labels, and a $\mathcal{C}$-multiaccurate predictor that has zero correlation with the labels (i.e., it is not even a weak learner). In fact, any post-processing of the multiaccurate predictor continues to achieve zero correlation with the labels. Intuitively, this can occur because multiaccuracy only promises closeness between the true labels and the multiaccurate predictor p in so far as the concepts $c \in \mathcal{C}$ can "see" (more formally, over the span of $\mathcal{C}$). However, the relationship between the labels and p over the *orthogonal* space to the span of $\mathcal{C}$ can mess up the correlation that exists over the span of $\mathcal{C}$, yielding a total of zero correlation.

This observation, however, allows us to prove in Section 3.2 that multiaccuracy does imply weak agnostic learning whenever there is some concept $c \in \mathcal{C}$ that achieves high correlation with the labels – specifically, when this correlation is at least $1/2$. We call this type of learning *restricted weak agnostic learning*, given that we only guarantee that our predictor can learn if the relationship between some concept in $\mathcal{C}$ and the labels is significant enough. Indeed, the traditional notion of weak agnostic learning is actually quite strong, in that an (α, β)-weak agnostic learner, for $\alpha \geq \beta \in [0, 1]$, promises to return a hypothesis h that has correlation at least β with the labels if there is a concept $c \in \mathcal{C}$ that has correlation at least α with the labels. Hence, we promise to output a predictor with some correlation with the labels as long as there is a concept in $\mathcal{C}$ that achieves some positive correlation with the labels, even if this quantity is very small. In our reduction from multiaccuracy to learning, we show that a perfectly multiaccurate predictor yields a weak agnostic learner that has correlation $2\alpha - 1$ with the labels, where α is the correlation achieved with the labels for some $c \in \mathcal{C}$. This guarantee is only non-trivial when $\alpha > 1/2$; this type of guarantee (i.e., when we only promise weak learning if α is high enough) is precisely what we capture through our notion of restricted weak agnostic learning.

Moreover, in Section 3.3 we show that if we also make our multiaccurate predictor *globally calibrated*, then this predictor is not only a weak learner, but in fact a strong agnostic learner. That is, its correlation with the labels is at least as good as that of the *best* hypothesis in $\mathcal{C}$. Our proof clearly delineates the roles played by multiaccuracy and by global calibration, similar to how

in our proof of IHCL with 2δ density we see how multiaccuracy gives the hardcore-ness of the measure whereas global calibration allows us to prove optimal density. Note that the predictions of a calibrated predictor are not necessarily informative: if the labels are balanced, then predicting $1/2$ everywhere satisfies global calibration. However, if a calibrated predictor deviates noticeably from random guessing (i.e., does not predict $1/2$ everywhere), then its predictions are actually quite informative: if we take the level set where we are predicting the value v, then we are guaranteed that the expectation of the true labels on that level set is approximately v. This allows us to argue that if a calibrated predictor deviates noticeably from random guessing, then we can obtain an agnostic learner. Then, the multiaccuracy guarantee allows us to discard the case where the predictor always predicts $1/2$: intuitively, a hypothesis c that is correlated with the labels $\mathbf{y}$ provides a certificate that $\mathbf{y}|\mathbf{x}$ is not uniformly random. A $\mathcal{C}$-multiaccurate predictor has to capture the correlations with c accurately, which forces it to deviate from random guessing. Therefore, when we require both global calibration and multiaccuracy, we obtain a strong agnostic learner.

In conjunction with our result on the Hardcore Lemma, our results demonstrate the power of adding global calibration to the notion of multiaccuracy. Indeed, in the case of IHCL, this allows us to obtain a hardcore measure of optimal density, whereas in the case of learning it allows to go from possibly no learning at all to strong agnostic learning. As we have emphasized, this is yet another piece of evidence in the recent realization of the power of calibrated multiaccuracy as a notion that is much cheaper to attain than full multicalibration, but which retains some of its benefits that multiaccuracy alone cannot provide.

Intuitively, in both the complexity-theoretic (Harcore Lemma) and learning-theoretic (agnostic learning) contexts, global calibration allows us to reason about the outputs of a multiaccurate predictor in a more fine-grained manner. In the case of IHCL, we consider the level sets of the predictor p and show that the assumption that input function g is somewhat hard to compute on average (as parametrized by δ) implies that, on average, the expected value of our predictor is at least δ, given that otherwise we would be able to guess g too well using the appropriate distinguishers (namely, a distinguisher that guesses the majority value on every level set of p). In the case of learning, we similarly consider the level sets of the calibrated predictor p in order to upper-bound its global error.

Beyond multiaccuracy (Chapter 5). Our counter-example shows that a multiaccurate predictor, even when post-processing its outputs, does not always yield learning. This does not entirely conclude the story, given that it does not answer the following more general question: if we can efficiently build a multiaccurate predictor for the class $\mathcal{C}$, does this imply that we can efficiently agnostically learn the class $\mathcal{C}$? For example, we could allow for more sophisticated post-processings of the multiaccurate predictor. What can we do next? We explore multiple different directions, which are essentially independent from each other.

Projecting onto the span of $\mathcal{C}$ (Section 5.1). As we explained, multiaccuracy does ensure closeness between the true labels and our predictor p over the space spanned by $\mathcal{C}$ (it is the orthogonal components that can mess up the learning). Hence, if we knew how to efficiently project a multiaccurate predictor onto the space of $\mathcal{C}$, then this could potentially give us an agnostic learner. In Section 5.1.2 we show that this intuition is correct: if we know how to project onto

the span of $\mathcal{C}$, then multiaccuracy suffices for weak agnostic learning. To prove this, we use a neat characterization of weak agnostic learning as non-trivial squared loss minimization that beats random guessing, which we explain in Section 5.1.1. Due to some technical details in the projection step, we need to require the projection to be ℓ_1 sparse. An example of a case where we can find such ℓ_1 sparse projection is the case of the Goldreich-Levin algorithm, which finds the largest Fourier coefficients of a function given membership access to it.

Relatedly, as we explain in Section 5.1.3, if we have a multiaccurate predictor p, then a crucial observation is that we can query p at any point (as opposed to just being able to sample examples $(\mathbf{x}, \mathbf{y})$ from the distribution). That is, we have query access to p, and because p and the target function p^* are close in the multiaccuracy sense, we show that this is essentially equivalent to having query access to the target function, in terms of learnability of the class. (This is, broadly speaking, how we always use the multiaccuracy/multicalibration guarantee in the proofs: we can appropriately swap the predictor by the target function and vice-versa.) This implies that, if there is a class $\mathcal{C}$ that is hard to learn in the usual agnostic setting but easy to learn if we add membership query access to the target function, then constructing a multiaccurate predictor for $\mathcal{C}$ is as hard as solving the hard-to-learn agnostic learning problem, precisely because having query access to a multiaccurate predictor is equivalent to having membership query access in terms of learnability. A concrete example where this occurs is for the class of parities, where we can apply the Goldreich-Levin algorithm to find the target parity if we have membership query access to it. This implies that finding a multiaccurate predictor for the class of parities is at least as hard as Learning Parities with Noise.

Restricted weak agnostic learning (Section 5.2). Recall that while multiaccuracy doesn't necessarily directly imply learning, it does whenever there is some concept in $\mathcal{C}$ that achieves correlation at least $1/2$ with the labels. This is what motivates our notion of a *restricted weak agnostic learner*, and given that we propose this new notion, we investigate it further. First, while this is a new notion, in Section 5.2.1 we demonstrate how it is closely related to *approximate agnostic learning*. Indeed, we show how an approximation algorithm for a concept class yields a restricted weak agnostic learner for the same class. By using a result by Daniely showing an approximate agnostic learning algorithm for the class of halfspaces, this provides evidence that restricted weak agnostic learning can be efficient for some parameters of α, but no longer efficient as $\alpha \to 0$.

Second, given the equivalence between weak and strong agnostic learning [KK09; Fel09a], it is natural to ask whether we can similarly boost a restricted weak agnostic learner, so that we can decrease the correlation with the labels that we require of some concept in $\mathcal{C}$. That is, if we have a (α, β)-weak agnostic learner, can we build a (α', β')-weak agnostic learner for $\alpha' < \alpha$? This is a different type of boosting than the typical one, where we want to increase the correlation with the labels achieved by the hypothesis that we are producing; that is, the β parameter. We can understand this form of boosting as asking whether we can learn at lower correlation. Unfortunately, in Section 5.2.2 we show that this type of boosting is not possible in general under standard cryptographic assumptions. Specifically, we use a family of pseudorandom functions to define our concept class. Essentially, we divide the domain into two disjoint parts, where one contains an encoding of the key r and the other contains the pseudorandom function f_r. If there is not too much noise, we can learn the class by reconstructing the key from examples, and then identifying

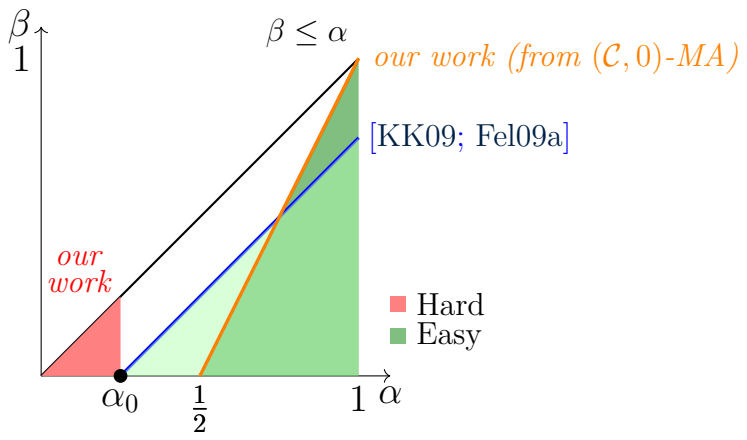


Figure 1.3: The spectrum of weak agnostic learning.

the corresponding function f_r. However, if the noise is high enough so that the labels in the part of the domain that contains the encoding of the key are completely random, then this learning problem becomes equivalent to the problem of distinguishing pseudorandom functions from truly random functions, which is hard under standard cryptographic assumptions. In order to get a tight gap, we encode the key r by using error-correcting codes.

We summarize this impossibility result in Figure 1.3 (in red), where we also plot our reduction from multiaccuracy to a restricted weak agnostic learner (which corresponds the positive result in green below the orange line).

Relatedly, we can ask whether we can construct a multiaccurate predictor from a weak agnostic learner with better error parameters than the ones obtained in the original algorithm by [HKRR18]. That is, the original algorithm uses an (α, β)-weak agnostic learner to obtain a multiaccurate predictor with error α. Could we lessen the error of the multiaccurate predictor while still using the same (α, β)-weak agnostic learner? We also answer this question in the negative in Section 5.2.3, using essentially the same construction with pseudorandom functions. While we can learn the class if there is not too much noise (and thus build a multiaccurate predictor through a weak agnostic learner), we show that constructing a multiaccurate predictor for this class with better error than $\alpha/2$ would again allow us to distinguish pseudorandom functions from truly random functions.

Auditing versus learning for multiaccuracy (Section 5.3). We conclude Part I of the thesis with an interesting future direction, which we preliminarily explore. Recall that in the original algorithm for constructing multiaccurate predictors, we audit for $\mathcal{C}$-multiaccuracy at each step of the algorithm using a $\mathcal{C}$-weak agnostic learner. Auditing for $\mathcal{C}$-multiaccuracy in this search-based way is equivalent to *properly* weak agnostically learning $\mathcal{C}$. Given this equivalence, it is clear that the distinction between the search and decision versions of the problem of auditing for multiaccuracy is closely related to the equivalent question for the problem of weak agnostic learning instead. For the latter, Vadhan [Vad17] and Kothari and Livni [KL18] established an equivalence between weak agnostically learning a class $\mathcal{C}$ and the problem of deciding, given a finite number of samples, whether there is some correlation between the labels and the concepts in $\mathcal{C}$ or whether the labels are truly random (this is called the *refutation* problem for $\mathcal{C}$).

We discuss how we can try to apply their equivalence result to understand whether being able to audit for $\mathcal{C}$-multiaccuracy implies that we can agnostically learn the class $\mathcal{C}$.

1.5 A multigroup perspective on learning with abstentions

As we have seen, the multigroup fairness framework has turned out to be extremely rich: besides its practical applications as a method for avoiding algorithmic discrimination, it has lead to fruitful work in complexity and learning theory (as we explore in Part I of this thesis), statistical inference, game theory, information theory, and causal inference, among many other areas.

All multigroup fairness notions deal with the predictions of a predictor p, which are numbers in the $[0, 1]$ interval that we understand as representing a likelihood. At the end of the day, notions such as multiaccuracy, global calibration, and multicalibration are all about inquiring into these individual probabilities $p(x)$ and ensuring that the algorithm has learnt them in the best way possible. The same is true for most of the work on algorithmic fairness (for example, the works on individual and group fairness notions [DHP+12; HPS16]), which generally deal with detecting and correcting the bias that may be present in a specific subgroup of the population.

But we can take a step back and question a fundamental assumption that we have implicitly been making so far: should the algorithm really be predicting a value $p(x)$ for each individual? If the algorithm is unsure of its predictions, for example because the true labels are hard to learn, then why should we force it to make a prediction on each $x \in \mathcal{X}$? Recent works have shown how typical group fairness metrics can drastically change after we account for the uncertainty present in the predictions and use them to abstain on the fraction of individuals on which the algorithm is most uncertain [LHAC23; CLC+24]. Similar recent papers have advocated for the need to include uncertainty and arbitrariness considerations in the study of algorithmic fairness [ALG21; KKR22; KCGK23; TCL23; LHAC24]. More generally, the use of algorithms in decision-making has been extremely prediction-centered so far. We should start to question this premise and expand our horizons of algorithmic inquiry: our actual goal should be to make the best *decisions* possible down the line, which doesn't necessarily imply that we should center all of our efforts on producing highly accurate predictions $p(x)$ for all individuals x [Per24; SAH24; FKP25].

The main observation here is that it is generally not enough to obtain the probability value that the algorithm outputs for every individual x. We would also like to quantify how *certain* the algorithm is of each prediction. That is, we should take into account the *arbitrariness* or the *uncertainty* of the predictions when thinking about learning problems. This is especially important when using algorithmic decision-making in societal contexts, where forcing the algorithm to predict can cause unnecessary false positives and false negatives that can potentially have devastating consequences for a particular individual. A much better approach would be to instead identify the points in the domain on which the algorithm is uncertain of its predictions, and then deal with those separately. As we argued in the case of the multigroup fairness framework, this is not only a "societal" question: being able to identify the regions of the domain on which the predictor is uncertain is useful and important regardless of whether we view the points x in the domain as people or not. Algorithms that are inaccurate within a subgroup or that produce highly arbitrary predictions are simply bad algorithms.

More broadly, thinking about how algorithms can have adverse societal consequences usually makes us have to think carefully about fundamental questions in computer science that we might not have studied otherwise. For example, in Part I of this thesis, the notion of calibrated multiaccuracy allowed us to obtain a new proof of IHCL with optimal density, and led us to an improvement of the classical construction by [TTV09] in the proof of IHCL through the complexity-theoretic Regularity Lemma, precisely by realizing that global calibration plays a key role in the construction of hardcore measures. We also obtain various new insights into the notion of weak agnostic learning, which has been a central concept in learning theory since the 90s [Hau92; KSS92]. In Part II of this thesis, we observe a similar phenomenon, where our study of uncertainty in predictions allows us to generalize previous results in agnostic learning from the 0-1 loss to a rich class of loss functions [KKM12].

Learning with abstentions (Chapter 6). A natural way in which we can expand our "full prediction" typical approach is by allowing our prediction to abstain on some points of the domain. That is, instead of outputting a prediction value on the $[0, 1]$ range, our predictor p can now output a value in the range $[0, 1] \cup ?$, where we interpret ? as the predictor saying "I don't know." In the literature, this is sometimes knows as *selective classification* [JSK+20; GKKM20], *partial classification* [KT14], or *learning to defer* [MPZ18; CS24].

While selective classification is not a new idea in the literature, most of these methods determine the abstention set as a *posteriori* intervention lacking any theoretical guarantees. For example, several approaches in the literature, particularly those that study the *model multiplicity* problem (which examines classes of classifiers with competing high accuracy but which disagree on individual predictions), quantify arbitrariness by first training many different classifiers and then computing the variance or the number of disagreements between the classifiers on individual points [MCU20; CLC+24]. For example, if we then want to add abstentions, we can do so by abstaining on the individuals that exhibit high variance with respect to the class of models. Various metrics have been proposed for quantifying the variance of the predictions within the class [MCU20], as well as various algorithms for ensembling the competing models in different ways, as to increase reliability [BLF21; RTW23; DNW24; BCDT25]. Several works have studied the relationship between the variance within the class of competing models and group fairness metrics [LHAC23; LHAC24; ALG21; KHS23; JRLT23]. A drawback of these variance-based methods is that they require fitting an entire class of models.

Although somewhat different, a related notion is that of conformal prediction [VGS05; AB21; JNRR23], where the classifier is allowed to output a prediction *set* of possible labels. The goal is for the true label to be in the prediction set with high probability. Some recent works have extended the conformal prediction setting to provide conditional guarantees instead of only marginal guarantees by adapting the multicalibration algorithm [JLP+21; JNRR23]. In the binary classification case, viewing an abstention ? as predicting the set $\{0, 1\}$ relates these two notions tightly [FN24]. We will use this observation in our results to obtain conformal prediction methods as a by-product.

Formal learning-theoretic approaches. Still, we are interested in rigorous approaches to the problem of learning with abstentions that provide provable and mathematical guarantees. Once we start studying abstentions formally, we need to define how exactly we account for the cost of abstention and what the optimal abstention rate looks like. A natural way of going about this is to

have some cost associated with abstaining (for example, the traditional Chow model gives a fixed cost of α to abstentions [Cho57]), and then try to minimize the total loss, which consists of the prediction loss plus the abstention loss [KK21a; KK21b]. Alternatively, rather than having a single loss with a fixed cost for abstention, we can have two different rates: the misclassification rate and the abstention rate [GKKM20].

In the learning theory literature, the formal study of selective classification was initiated in 2009 by Kalai, Kanade, and Mansour, who called it *reliable agnostic learning* [KKM12]. We can view their model as the natural generalization of the usual agnostic learning definition in the setting of selective classification. Their work focuses on binary classification problems using only the 0-1 loss function. Specifically, given a concept class $\mathcal{C}$, the task is to efficiently find a hypothesis h that makes essentially no false positives or false negatives, while abstaining at the minimum number of points. How do we define the optimal abstention rate? Here, in keeping with the agnostic learning approach, the abstention rate of h has to compete with the best abstention rate in $\mathcal{C}$. Because $\mathcal{C}$ is a Boolean concept class with no abstentions, we need to post-process the class as to add abstentions, as we will discuss. Then, the best selective classifier in this post-processed class $\mathcal{C}$ is the one that achieves the lowest abstention rate, subject to also making no false positives or false negatives. Kalai, Kanade, and Mansour showed that if the class $\mathcal{C}$ is agnostically learnable, then it is also reliably agnostically learnable. In the other direction, reliable agnostic learning is believed to be easier than agnostic learning, given that Kanade and Thaler proposed an algorithm for learning majorities in the reliable agnostic setting which is strictly more efficient than the fastest known algorithms in the standard agnostic setting [KT14]. We formally introduce this rigorous model for learning with abstentions in Chapter 6.

Inspired by the reliable agnostic learning framework by [KKM12], in Part II of this thesis we initiate the study of selective classification in the multigroup fairness context. We do this in two ways. First, we introduce abstentions to the powerful learning-theoretic framework of omnipredictors, where the specific loss function can be chosen from a large class of loss functions a *posteriori* after training. To this end, we introduce the notion of of a *selective omnipredictor* and show how to construct them efficiently. Moreover, we show how we can generalize the reliable agnostic results of [KKM12] beyond the 0-1 loss to a rich class of loss functions by using a selective omnipredictor.

Second, we bring the spirit of the multigroup fairness framework into the notion of reliable agnostic learning as introduced in [KKM12]. Specifically, their notion only requires our hypothesis to compete with the abstention rate of the best selective concept. However, if we have a collection $\mathcal{G}$ of subgroups of the domain that we wish to protect, then a reliable agnostic learner could unnecessarily over-abstain on some subgroups. To prevent this, we introduce the notion of a $(\mathcal{C}, \mathcal{G})$-*multigroup selective classifier*, which is required to abstain optimally not only globally but also locally within each of the groups in the pre-specified collection $\mathcal{G}$. Here, calibrated multiaccuracy makes another appearance: we show how to construct a $(\mathcal{C}, \mathcal{G})$-multigroup selective classifier using a calibrated and multiaccurate predictor for a suitable concept class (combining $\mathcal{C}$ and $\mathcal{G}$).

Lastly, we apply our methods to the setting of conformal prediction. We show how our results can be applied in the binary classification setting to obtain conformal prediction algorithms that achieve both marginal and group-conditional coverage guarantees for an intersecting collection of groups. We provide experiments on synthetic data as to visualize our various constructions and demonstrate their feasibility and guarantees in practice.

Selective omniprediction (Chapter 7). As we have discussed, one of the most successful applications of the multigroup fairness framework has been in the creation of the loss minimization paradigm of *omniprediction*. This consists of a single predictor that can be trained for an entire class of loss functions $\mathcal{L}$ (that satisfy some mild conditions), such that we can efficiently post-process our predictor once the loss has been chosen and obtain optimal error in the agnostic sense. As it turns out, we can use the strong indistinguishability properties of the multigroup fairness framework to construct omnipredictors from multicalibration [GKR+22], and it has recently been shown that we can in fact construct them from the weaker primitive of calibrated multiaccuracy [OKK25]. In these connections, we make the collection of groups $\mathcal{C}$ in the multigroup fairness framework equal the concept class $\mathcal{C}$ in omnipredictors (as summarized in Figure 1.1).

We propose to extend the omniprediction paradigm to the setting of selective classification. Specifically, we add loss functions measuring the loss incurred by abstaining to the large class of loss functions $\mathcal{L}$. We compute the total loss of the predictor by adding together the prediction loss and the abstention loss; a function that we call the *generalized Chow loss*. Then, we show how to efficiently construct selective omnipredictors for very general classes of loss functions, provided that $\mathcal{C}$ is agnosticaly learnable. This allows us to build a single classifier that learns abstentions and predictions optimally simultaneously for every loss in the entire class, where the abstentions are decided efficiently for each specific loss function by applying a fixed post-processing function. By "optimally" we mean that, after the efficient post-processing, our selective omnipredictor obtains no worse generalized Chow loss than the best selective classifier in the class. This realizes a very strong form of learning. Moreover, given the connections between omniprediction and the computational indistinguishability [DKR+21; GHK+23], our construction can be viewed as an indistinguishability-based approach to the problem of learning with abstentions.

We show how the p-values of the points of the domain on which our selective omnipredictor abstains in fact form a contiguous interval in $[0, 1]$, which we call I_{abs}. This means that for any generalized Chow loss that is chosen a *posteriori* from the class $\mathcal{L}$, we can directly provide the thresholds indicating which p-values should be sent to ?. We show that I_{abs} can be computed directly from the chosen generalized Chow loss, *independent from the data*. Hence, our post-processing for adding abstentions after training the single selective omnipredictor is extremely efficient, as we can use the corresponding I_{abs} interval off-the-shelf. We provide experiments to demonstrate how we can construct selective omnipredictors in practice.

Moreover, we then show how we can use selective omnipredictors to generalize the reliable agnostic learning results from the original paper by [KKM12]. Specifically, for an agnostically learnable class $\mathcal{C}$, they only show how to efficiently build reliable agnostic learners for the 0-1 loss. We show how to obtain a reliable agnostic learner from a selective omnipredictor. Because the point is precisely that we can construct selective omnipredictors for a very rich class of loss functions, we are thus able to obtain general reliable agnostic learners for all of these loss functions.

Therefore, our connection between the multigroup fairness framework/omniprediction and the problem of learning with abstentions is fruitful in both directions:

- It allows us to extend the multigroup fairness framework and the omniprediction fairness framework from a prediction-only setting to that of learning with abstentions. That is, we extend it from $[0, 1]$ to $[0, 1] \cup \{?\}$. As we have argued, being able to quantify and incorporate the uncertainty of the predictions is necessary and useful, particularly in societal contexts.

Moreover, it provides insightful connections between the notion of (multi)calibration and the problem of uncertainty quantification.

- By using the power of the omniprediction, we are able to learn abstentions optimally in a very general sense. By following the "learn first, optimize later" paradigm, our selective classification algorithm learns abstentions and predictions optimally for every loss in a rich class of loss functions, where we decide the abstentions very efficiently a *posteriori* once the specific loss function has been fixed. Moreover, through selective omniprediction, we are able to generalize the previously-known constructions of reliable (agnostic) learners [KKM12]. The multigroup fairness framework also inspires us to extend the notion of reliable learning to its "multi" version, where we require *local* optimal abstention rates, besides just global. This provides a useful strengthening of the notion of reliable learning, which we show we can attain from calibrated multiaccuracy.

Learning abstentions fairly (Chapter 8). In Chapter 7, we show how we can use the multigroup fairness framework to construct optimal selective classifiers for a very general class of loss functions (i.e., selective omnipredictors), and how we can in turn construct general reliable learners for the same class of loss functions.

However, reliable learners could unnecessarily have very high rates of abstention on some subgroups of the domain. Similar to how multiaccuracy ensures that a predictor is accurate not only globally, but also locally on each group of interest in a collection, and multicalibration ensures that a predictor is calibrated not only globally, but also locally on each group of interest in a collection, we can take a similar multigroup perspective on the problem of selective classification. Specifically, given a rich collection $\mathcal{G}$ of possibly-intersecting subgroups of the domain, we want to build a classifier that is accurate when making a prediction and which moreover abstains optimally not only globally but also locally when conditioning on each of the groups in the collection $\mathcal{G}$. What does "optimally" mean? Again, we take the agnostic learning view, and require our predictor to abstain no more than the best concept in $\mathcal{C}$, where the concept can be chosen tailored to the group $\mathcal{G}$. We can view this as the "multigroup" version of reliable learning, and for this reason we call this primitive $(\mathcal{C}, \mathcal{G})$-*multigroup selective classification*.

Note that, until now, we had always taken $\mathcal{G} = \mathcal{C}$. Indeed, as summarized in Table 1.1 this is how we go from multicalibration to omniprediction: the $\mathcal{C}$-indistinguishability property of the predictor translates into optimal learning with respect to the base class $\mathcal{C}$. However, $\mathcal{C}$ and $\mathcal{G}$ do have different interpretations: one is a concept class and the other is a collection of subgroups of the domain. Indeed, we separate them in our definition of a $(\mathcal{C}, \mathcal{G})$-multigroup selective classifier. On the flip side, in order to efficiently construct $(\mathcal{C}, \mathcal{G})$-multigroup selective classifiers, we then require a weak agnostic learner for the class $\mathcal{C} \cdot \mathcal{G} = \{cg \mid c \in \mathcal{C}, g \in \mathcal{G}\}$. Specifically, we show that if we have access to a $(\mathcal{C} \cdot \mathcal{G})$-calibrated and multiaccurate predictor, which as we know we can construct from a $(\mathcal{C} \cdot \mathcal{G})$-weak agnostic learner, then we can efficiently build a $(\mathcal{C}, \mathcal{G})$-multigroup selective classifier. Perhaps unexpectedly, this construction delineates very clearly the roles played by multiaccuracy and calibration, in a way very similar to our results in Part I regarding strong agnostic learning and the construction of optimal hardcore measures, where we also use the primitive of calibrated multiaccuracy to prove our theorems. Hence, Part II of this thesis provides yet another important use case of the primitive of calibrated multiaccuracy.

Throughout this thesis, we are always interested in understanding the easiest learning primitive that we can use for each task. Hence we ask: Can we construct a $(\mathcal{C}, \mathcal{G})$-multigroup selective classifier from a weaker learning primitive? We show that, in the case where $|\mathcal{G}|$ is small, we can do so from the primitive of a reliable learner. As we have discussed, this is a weaker primitive than that of weak agnostic learning [KT14]. However, we do not know whether weak agnostic learning is necessary for a large collection $\mathcal{G}$, or whether we can also do with reliable learning instead. In Section 8.2, we draw some more learning connections between multigroup fairness primitives, reliable learning, and multigroup selective classifiers.

Lastly, in Section 8.3, we show how the algorithms that we have presented in Chapters 7 and 8 yield useful conformal prediction methods for the binary classification case. Specifically, by establishing a bijective map between ? and the prediction set $\{0, 1\}$, we show how reliable learning is in fact a conformal prediction algorithm that satisfies the required marginal coverage guarantee (i.e., requiring that the true label is almost always included in the prediction set that the predictor outputs). Similarly, with the same bijective map, we show that we can also view a $(\mathcal{C}, \mathcal{G})$-multigroup selective classifier as a conformal prediction algorithm. In this case, it provides not only a global marginal coverage guarantee, but also when conditioned on each group g in $\mathcal{G}$ (this is known as a *group conditional coverage guarantee*).

Moreover, reliable learning and multigroup selective classification provide more useful theoretical guarantees beyond the required coverage guarantees. Typical conformal prediction methods offer no theoretical bounds on the size of the prediction sets, and so, in theory, they do not disallow the case where the prediction trivially outputs the prediction set $\{0, 1\}$ everywhere. However, in our algorithms, our predictor achieves optimal abstention guarantees with respect to a base concept class $\mathcal{C}$ of our choice: it does so globally in the case of reliable learning, and locally within each subgroup $g \in \mathcal{G}$ in the case of multigroup selective classification. Hence, this provides a strong and useful method for conformal prediction that gives provable guarantees on the size of the prediction sets. We provide experiments demonstrating how our methods can be used as conformal prediction algorithms, both for reliable learning and for multigroup selective classification.

We summarize the various constructions that we show in Part II of this thesis in Figure 1.4.

1.6 Thesis structure

Chapter 2. *Notation & Preliminaries.* We provide the notation that we use throughout the thesis. We overview the agnostic learning notions, the multigroup fairness definitions and constructions, the connection to the complexity-theoretic Regularity Lemma, and the connection to the learning-theoretic framework of omniprediction.

Part I of this thesis studies the connections between multigroup fairness notions, learning primitives, and hardcore set constructions.

Chapter 3. *Multiaccuracy & Agnostic learning.* We present our results studying the relationship between multiaccuracy and agnostic learning. Specifically, we show that a multiaccurate predictor itself does not necessarily give weak learning, but it does provide a form of weaker learning that we call *restricted weak agnostic learning*. If we add global calibration, then we do obtain strong agnostic learning.

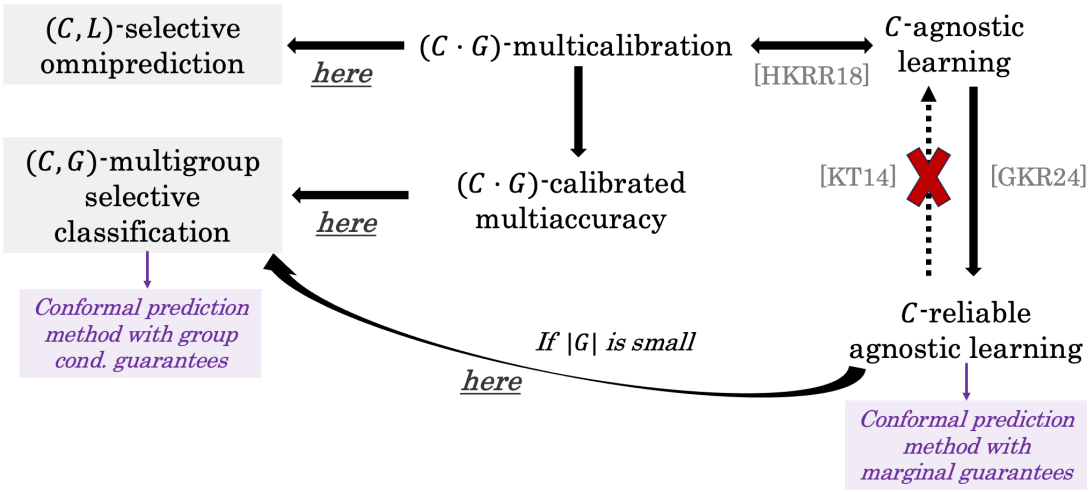


Figure 1.4: Summary of our constructions in Part II of this thesis.

Chapter 4. *Impagliazzo's Hardcore Lemma.* We present our results regarding Impagliazzo's Hardcore Lemma. We show how we can obtain a hardcore measure of optimal density from calibrated multiaccuracy, explore the connection between hardcore measure constructions and boosting, and show how we can connect the construction of hardcore measures and of (swap) agnostic learning in the ++ tier (i.e., the multicalibration tier) through the notion of hardness of prediction.

Chapter 5. *Beyond multiaccuracy.* We further study the learning properties that we can extract from the notion of multiaccuracy. We analyze the problem of projecting a multiaccurate predictor, show how we can view weak learning as squared loss minimization, and explore the consequences of viewing a multiaccurate predictor as a query oracle through the Goldreich-Levin algorithm. We study the notion of restricted weak agnostic learning from the perspective of approximation algorithms and show that we cannot boost them to learn at lower correlations nor construct a multiaccurate predictor with lower error. We also study the related problem of *auditing* for multiaccuracy and its relationship to weak agnostic learning.

In Part II of this thesis, we bring the multigroup fairness framework into the study of selective classification.

Chapter 6. *Learning with abstentions.* We present the formal models of reliable agnostic learning, PQ-learning, the Chow model, and our proposed notion of selective omniprediction, along with the necessary notation for Part II.

Chapter 7. *Selective omniprediction & Reliable learning.* We show how to construct selective omnipredictors efficiently and how we can decide the points of the domain on which to abstain. We illustrate how to use selective omnipredictors in practice with experiments and show how we can construct general reliable agnostic learners from selective omnipredictors.

Chapter 8. *Learning abstentions fairly.* We define the notion of multigroup selective classification, inspired by the multigroup fairness framework. We show how to construct multigroup selective classifiers from calibrated multiaccuracy in the general case and from reliable agnostic learning in the case where the collection of groups is small. We show theoretically and through experiments that we can use both reliable learning and multigroup selective classification as conformal prediction methods with strong guarantees: marginal and conditional coverage guarantees, respectively, and provable guarantees on the optimality of the size of the prediction sets outputted by our predictor, as measured with respect to the base concept class $\mathcal{C}$.

Statement of contribution

- Part I of this thesis is based on the paper "How Global Calibration Strengthens Multiaccuracy", which is joint work with Parikshit Gopalan, Varun Kanade, and Omer Reingold [CGKR25]. This paper has been accepted for publication at the upcoming *66th IEEE Symposium on Foundations of Computer Science* (FOCS 2025).
- Part II of this thesis is based on the paper "Selective Omniprediction and Fair Abstention", which is joint work with Varun Kanade [CK25]. This work has been accepted for publication as a spotlight paper at the upcoming *39th Annual Conference on Neural Information Processing Systems* (NeurIPS 2025). A preliminary version of the paper was presented at the Workshop on Predictions and Uncertainty at the *38th Annual Conference on Learning Theory* (COLT 2025) this past month of June.

In the thesis, we make some brief mentions to the paper "Reconciling Predictive Multiplicity in Practice" [BCDT25], which was also completed by the author during the masters program and presented at the *8th ACM Conference on Fairness, Accountability, and Transparency* (FAccT 2025) this past month of June.

2

Notation & Preliminaries

We develop and study multicalibration —a new measure of algorithmic fairness that aims to mitigate concerns about discrimination that is introduced in the process of learning a predictor from data. Multicalibration guarantees accurate (calibrated) predictions for every subpopulation that can be identified within a specified class of computations.

Hébert-Johnson, Kim, Reingold, and Rothblum [HKRR18]

We begin by providing the notation and main definitions that we will use throughout this thesis. We denote the domain by $\mathcal{X}$ and the labels by $\mathcal{Y} = \{0, 1\}$. We have an underlying distribution $\mathcal{D}$ defined on pairs $(\mathbf{x}, \mathbf{y})$, $\mathcal{X} \times \mathcal{Y}$, where $\mathbf{x} \in \mathcal{X}$ is a feature vector and $\mathbf{y}$ is a label. We denote the marginal distribution over $\mathcal{X}$ by $\mathcal{D}_{\mathcal{X}}$, and we use boldface when referring to random variables. We consider concept classes $\mathcal{C}$ defined on $\mathcal{X}$, where each $c \in \mathcal{C}$ is a function from $\mathcal{X} \to [-1, 1]$. Throughout the thesis, the only places where the range of the concepts is not $[-1, 1]$ are the following: in the case of omnipredictors and selective omnipredictors, their range is all of $\mathbb{R}$. The multigroup fairness definitions and the construction of multigroup fair predictors can be done with concepts that have range $\{0, 1\}$, $[-1, 1]$, or $\mathbb{R}$. In the case of the Hardcore Lemma, reliable agnostic learning, and $(\mathcal{C}, \mathcal{G})$-multigroup selective classification, the concepts are Boolean. Moreover, we always assume that the constant functions $\mathbf{0}$ and $\mathbf{1}$ belong to $\mathcal{C}$. In the setting of the multigroup fairness framework, we usually think of $\mathcal{X}$ as the set of individuals, and of $c \in \mathcal{C}$ as indicator functions of subgroups of the population.[1] In the setting of learning theory, we view the $c \in \mathcal{C}$ as *hypotheses* or *concepts*. In the setting of complexity theory, we view them as general distinguishers or as bounded-size circuits. We summarized these three interpretations in Figure 1.1 in Chapter 1.

Throughout the thesis, we are concerned with building *predictors*. A predictor is a function $p : \mathcal{X} \to [0, 1]$, where we interpret $p(\mathbf{x})$ as the probability that $\mathbf{y}|\mathbf{x}$ is 1. Hence, the ground truth

[1] In this interpretation, it is natural to think of the functions $c \in \mathcal{C}$ as Boolean, but we can generalize them to real-valued functions by interpreting them as encoding fuzzy set membership. More importantly, many follow up works to the original multicalibration paper use that we can use more general real-valued concept classes and achieve the same guarantees.

optimal predictor corresponds to $p^*(\mathbf{x}) = \mathbb{E}[\mathbf{y}|\mathbf{x}]$, which is usually referred to as the *ground truth* predictor. When we speak of the "true labels", then, we refer to the label $\mathbf{y}$ assigned to each point $\mathbf{x}$ by p^*, where $\mathbf{y} \sim \text{Bern}(p^*(\mathbf{x}))$. Note that, in reality, we only observe the Boolean labels $\mathbf{y}$, and we never have access to the "true probabilities" $p^*(\mathbf{x})$. Still, we can approximate the p^* values by taking multiple samples from the distribution; indeed, the multigroup fairness definitions are interchangeably defined using p^* or using the true labels $\mathbf{y}$, and we will swap $\mathbf{y}$ and p^* in our proofs and statements (when considering their expected value). We want the predictor p that we are building to be a "good model" of p^*. What does this mean? Typically, we measure the squared loss of p, which is defined as $\ell_2(y, p) = (y - p)^2$, given that the ground truth predictor achieves 0 loss (e.g., we want the squared loss of our predictor to be less than some threshold ϵ). In the multigroup fairness framework, we take an indistinguishability-based approach, where we are happy with a predictor p that looks "the same" as p^* to a family of distinguishers $\mathcal{C}$. One can view this as a form of statistical-to-computational relaxation, where we go from requiring *statistical* closeness between p^* and p to requiring *computational* closeness between the two predictors (where here "closeness" is measured with respect to a base class $\mathcal{C}$ through multiaccuracy/multicalibration).

Before we formalize the various multigroup fairness notions, we begin by introducing the agnostic learning framework, given that the MA/MC algorithms require using a weak agnostic learner for building the predictors.

2.1 Agnostic learning

A way in which we can measure how well a predictor is learning is by computing its correlation with the true labels:

Definition 2.1 (Correlation)**.** *The* correlation *of a function $c : \mathcal{X} \to [-1, 1]$ with respect to a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$, is defined as* $\text{cor}_{\mathcal{D}}(\mathbf{y}, c) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[(2\mathbf{y} - 1)c(\mathbf{x})]$.

We can think of correlation as the "opposite" of the error. Indeed, for a Boolean predictor $p : \mathcal{X} \to \{0, 1\}$, we have that $\text{cor}_{\mathcal{D}}(\mathbf{y}, 2p - 1) = 1 - 2\text{err}_{\mathcal{D}}(\mathbf{y}, p)$, where $\text{err}_{\mathcal{D}}(\mathbf{y}, p) = \Pr_{\mathcal{D}}[\mathbf{y} \neq p(\mathbf{x})]$. We require the shift from $\mathbf{y}$ to $2\mathbf{y} - 1$ and similarly from p to $2p - 1$ in order to map the $\{0, 1\}$ ranges to $\{-1, +1\}$. Therefore, when we consider the correlation between the $[-1, 1]$-concepts in $\mathcal{C}$ and the labels $\mathbf{y}$, we do so with the shifted version $2\mathbf{y} - 1$. Similarly, because our predictors p take values in $[0, 1]$, when measuring the correlation with the labels we also take the shifted version $2p - 1$, so that the outputs of the predictor also take values in $[-1, 1]$. Here, the strongest possible correlation is 1, whereas a correlation of 0 means that the predicted values and the true labels hold no relationship to each other. Note that this shifting of the ranges from $[0, 1]$ to $[-1, 1]$ translates $p = 1/2$ to 0, translating the fact that guessing randomly at every point gives 0 correlation. It will be repeatedly useful for us to compare our predictor p to the "random guessing predictor" as a baseline (i.e., the predictor that outputs $1/2$ everywhere).

Correlation (or, equivalently, error) allows us to measure the "closeness" between the predictor p that we are building and the true labels. How do we define what a "good" correlation for our predictor p is? I.e., how much correlation are we trying to achieve? In the agnostic setting in learning theory, where we make no assumptions on the data generation process, we define a good correlation with respect to a base concept class $\mathcal{C}$ [Hau92; KSS92]. The strongest version would be to require our predictor to achieve at least as much correlation as the best concept $c \in \mathcal{C}$ (i.e., the

one that achieves the highest correlation with the labels), with some error slack. This corresponds precisely to the notion of strong agnostic learning, which can be formalized as follows:

Definition 2.2 (Strong agnostic learner)**.** *For a hypothesis class $\mathcal{C}$, a* strong agnostic learner *for $\mathcal{C}$ under the marginal distribution $\mathcal{D}_{\mathcal{X}}$ is an algorithm that given any $\varepsilon > 0$ as the target error parameter and random samples from any $\mathcal{D}$ with marginal distribution $\mathcal{D}_{\mathcal{X}}$ returns a hypothesis $h : \mathcal{X} \to [-1, 1]$ such that with probability at least 0.99,*

$$\mathrm{cor}_{\mathcal{D}}(\mathbf{y}, h(\mathbf{x})) \geq \max_{c \in \mathcal{C}} \mathrm{cor}_{\mathcal{D}}(\mathbf{y}, c(\mathbf{x})) - \varepsilon.$$

In the PAC setting, which is also known as the *realizable* setting, we assume that the target function belongs to the concept class $\mathcal{C}$, and therefore there is a $c \in \mathcal{C}$ that achieves correlation 1 with the labels (equivalently, 0 error). Then, in the PAC setting, our predictor would be required to achieve correlation with the labels at least $1 - \epsilon$ [Val84]. The agnostic framework is thus a much stronger learning framework, where we drop any assumptions on the ground-truth target function. Indeed, there are concept classes that are efficiently PAC learnable but not agnostically learnable, such as the class of parities (under standard cryptographic assumptions), which can be learned in the realizable setting through Gaussian Elimination. Because we make no assumptions on the label generation process in the agnostic setting, then our correlation/error only has to compete with how well the concepts in $\mathcal{C}$ are doing at predicting the labels.

One could consider a weakening of strong agnostic learning, where we only promise to output a hypothesis that achieves some correlation with the labels, which does not have to be optimal in the strong agnostic learning sense. This is precisely what motives the introduction of a *weak agnostic learner* [KMV08; KK09; Fel09a]:

Definition 2.3 (Weak agnostic learner)**.** *For $\alpha \geq \beta > 0$, an (α, β)-*weak agnostic learner *for $\mathcal{C}$ under marginal distribution $\mathcal{D}_{\mathcal{X}}$ is an algorithm that when given access to random examples from any $\mathcal{D}$ whose marginal over $\mathcal{X}$ is $\mathcal{D}_{\mathcal{X}}$, if there exists $c \in \mathcal{C}$ such that*

$$\mathrm{cor}_{\mathcal{D}}(\mathbf{y}, c(\mathbf{x})) \geq \alpha,$$

returns a hypothesis $h : \mathcal{X} \to [-1, 1]$ such that with probability at least 0.99,

$$\mathrm{cor}_{\mathcal{D}}(\mathbf{y}, h(\mathbf{x})) \geq \beta.$$

We further say that h is a proper *weak agnostic learner if $h \in \mathcal{C}$, and* improper *otherwise.*

Therefore, weak agnostic learning is easier as α increases, and harder as α decreases. Indeed, if we can construct an (α, β)-weak agnostic learner, then we have an (α', β')-weak agnostic learner for any $\alpha' \geq \alpha$ and $\beta' < \beta$. Note that the case where $\alpha = 1$ corresponds to the realizable setting.

In the case of PAC learning, we similarly have the notion of weak learners, which perform just slightly better than random guessing (i.e., achieve error at least $1/2 + \epsilon$). In a well-known result, Freund and Schapire showed that if we can produce weak learners for a concept class, then we can produce a strong learner for the same class (i.e., achieving error less than ϵ) [Sch90; Fre95; FSA99]. We can show this precisely through the key technique of *boosting*, which consists of finding a way to strongly learn a concept class by means of using several weak learners in a carefully curated way. Usually, these consist of two parts: an iterative part, where we modify the distribution at each

step and call a weak learner to give us a hypothesis for that distribution (note that the definitions require the weak learner to work for any distribution), and an ensembling part, where we combine the multiple weak learners. Typically, we require $O(1/\epsilon^2)$ iterations in the first part, and then aggregate all of the $O(1/\epsilon^2)$ weak hypotheses by taking a weighted majority vote [FSA99]. Recent work demonstrates how we can reduce the number of iterations (i.e., of calls to the weak learner) to $\tilde{O}(1/\epsilon)$, at the expense of then having a complex aggregation rule [AGHM21]. They show that, in fact, a complex aggregation rule is necessary if we want to by-pass the lower bound of $\Theta(1/\epsilon^2)$ number of iterations.

In the case of agnostic learning, a similar result emerges. In 2009, Feldman [Fel09a] and Kalai and Kanade [KK09] concurrently showed that strong agnostic learning reduces to weak agnostic learning. Specifically:

Theorem 2.4 (Agnostic Boosting, [KK09])**.** *An (α, β)-weak agnostic learner for $\mathcal{C}$ can be boosted to a strong agnostic learner under the same marginal distribution on $\mathcal{X}$ to obtain a hypothesis h that satisfies with probability at least $1 - \rho$,*

$$\mathrm{cor}_{\mathcal{D}}(\mathbf{y}, h(\mathbf{x})) \geq \max_{c \in \mathcal{C}} \mathrm{cor}_{\mathcal{D}}(\mathbf{y}, c(\mathbf{x})) - \alpha - \varepsilon$$

for any $\varepsilon > 0$, using poly$(1/\varepsilon, 1/\beta, 1/\rho)$ calls to the weak agnostic learner.

That is, if we have an (α, β)-weak agnostic learner for some α, then for any $\alpha' \geq \alpha$, we can obtain an $(\alpha', \alpha' - \alpha - \epsilon)$-weak agnostic learner for any $\epsilon > 0$. Such a learner can be obtained using at most poly$(1/\epsilon, 1/\beta, 1/\rho)$ calls to the (α, β)-weak agnostic learner. This means that obtaining (improper) strong agnostic learning reduces to the task of constructing an (α, β)-weak agnostic learner for any $\alpha > 0$.

Similar to the realizable setting, the proof of this weak-to-strong equivalence is shown using a boosting algorithm: given a weak agnostic learner for $\mathcal{C}$, we can set up an iterative procedure that calls the weak learner as a subroutine at each iteration and ends by producing a hypothesis that strongly agnostically learns $\mathcal{C}$. A key difference of the agnostic boosting algorithm proposed by Kalai & Kanade algorithm as compared to the usual boosting algorithms (such as Feldman's [Fel09a]) is that theirs is implemented *without* reweighting examples; instead, they are randomly relabeled [KK09].

2.2 Multigroup fairness notions

Multiaccuracy. Now that we have formally introduced the notion of a weak agnostic learner, we can explain how to construct multiaccurate and multicalibrated predictors. For all of the multigroup fairness notions, the concepts c can take values in all of $\mathbb{R}$. Moreover, in the context of multigroup fair predictors, as it is done in the literature we always discretize the predictor p to buckets of width ϵ, so that p has $O(1/\epsilon)$ many level sets. First, we formalize the multigroup fairness notions. We start with the weaker notion of multiaccuracy:

Definition 2.5 (Multiaccuracy [HKRR18])**.** *We say that a predictor p is* $(\mathcal{C}, \epsilon)$-multiaccurate *for a distribution $\mathcal{D}$ if*

$$\max_{c \in \mathcal{C}} \left| \mathop{\mathbb{E}}_{\mathcal{D}} [c(\mathbf{x})(\mathbf{y} - p(\mathbf{x}))] \right| \leq \epsilon. \tag{2.6}$$

Because we take the maximum over all $c \in \mathcal{C}$, this tells us that, if a predictor is $(\mathcal{C}, \epsilon)$-multiaccurate, then for every $c \in \mathcal{C}$ it holds that $\left|\mathbb{E}_{\mathcal{D}}[c(\mathbf{x})(\mathbf{y} - p(\mathbf{x}))]\right| \leq \epsilon$. We interpret the multiplication by c as identifying the region of the domain over which c is "looking at". Returning to the interpretation of Boolean c's as indicator functions for subgroup, the function c cares about the subset of the domain over which $c(x) = 1$. Elsewhere, $c(x)(y - p(x)) = 0$, and so the relationship between y and p can be anything – it will not be factored into the multiaccuracy error (Equation 2.6). Multiaccuracy ensures that, for every $c \in \mathcal{C}$, when we restrict the domain to the points $\mathcal{X}_c = \{x \in \mathcal{X} \mid c(x) = 1\}$, the expected value of $\mathbf{y}$ under $\mathcal{D}$ inside of $\mathcal{X}_c$ is close to the expected value of p under $\mathcal{D}$ inside of $\mathcal{X}_c$. That is,

$$\mathbb{E}[\mathbf{y} \mid \mathbf{x} \in \mathcal{X}_c] \approx \mathbb{E}[p(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}_c]. \tag{2.7}$$

Note that these subgroups represented by the functions $c \in \mathcal{C}$ can intersect, and this is in fact one of the very desirable properties of the multigroup fairness framework. In the general setting where we let the functions c map to $[-1, 1]$, by recalling the definition of correlation (Definition 2.1), one can view the multiaccuracy condition as requiring close to 0 correlation between all of the $c \in \mathcal{C}$ and the residual $\mathbf{y} - p(\mathbf{x})$.

One has to be careful when conditioning the multiaccuracy and multicalibration expressions on $\mathcal{X}_c$ or on the level sets of the predictor (which we do for multicalibration). In terms of sample complexity, it is not feasible to try to satisfy the MA/MC condition on subsets of the domain that are too small. To resolve this, we can either: compute the error on average over the distribution (as we do in Definition 2.5), or make the error parameter ϵ depend on the size of the set; specifically, by having the error parameter be $\epsilon / \Pr_{\mathcal{D}}[\mathbf{x} \in \mathcal{X}_c]$. Some works, such as the original paper introducing multiaccuracy and multicalibration [HKRR18], consider the equivalent notion of *approximate* multiaccuracy (and multicalibration), where we only promise the multiaccuracy condition for the sets that have probability mass $\Pr_{\mathcal{D}}[\mathbf{x} \in \mathcal{X}_c]$ at least some chosen constant γ. Usually, it is intuitively helpful to think of the multigroup fairness notions in their conditioned versions, so that we can visually picture a subset of the domain. For the proofs, however, we tend to add an outer expectation, so that in computing the MA/MC error "on average" we already take care of the small sets.

The expression in Equation 2.7 is precisely what we meant in the introduction when explaining the multigroup fairness notions as a Turing test: suppose that we have constructed a predictor p, and we want to know if it is a good enough model for the true labels. To test this, every function $c \in \mathcal{C}$ comes along, looks at the points in the domain that it cares about (i.e., the points where $c(x) = 1$), and sees if it can distinguish between the true labels and the predicted labels. Here, "distinguishing" is mathematically understood as per Equation 2.6 in the multiaccuracy definition (Definition 2.5), where the distinguisher compares the expected value of the true labels with the expected value of the predictions over the points in the domain where $c(x) = 1$. Note that the "accuracy" notion in the definition of multiaccuracy is in fact only accuracy *in expectation*, rather than in the stronger point-wise sense. Continuing with the Imitation Game/Turing Test analogy from Figure 1.1 in the introduction, we visualize the definition of multiaccuracy in Figure 2.1.

Note that we can view multiaccuracy as a computational way (i.e., with respect to a class $\mathcal{C}$) of defining the notion of "closeness" between the true labels $\mathbf{y}$ and our predictor p, as opposed to the usual statistical way (of, for example, measuring the squared loss of the predictor).

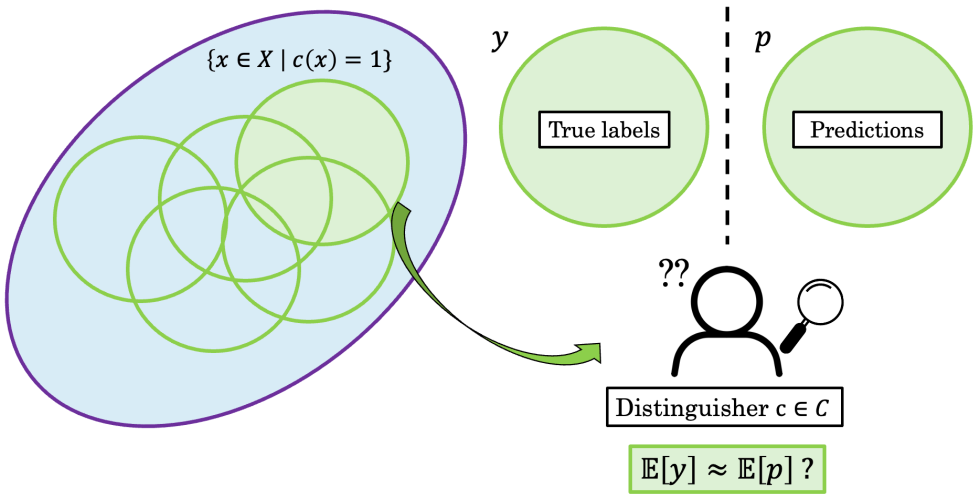


Figure 2.1: Visual depiction of the notion of multiaccuracy as a Turing Test.

Multicalibration. Still, requiring only accuracy in expectation can be quite a weak requirement in some cases. For example, if $p^*(x) = 1/2$ everywhere (recall that $p^*(\mathbf{x}) = \mathbb{E}[\mathbf{y}|\mathbf{x}]$), then a predictor p can predict 0 or 1 (with equal probability) arbitrarily on each point $x \in \mathcal{X}$, and it will satisfy accuracy on average. To disallow this type of scenario, we will consider the notion of *calibration*, which enforces the probabilities predicted by p to "mean what they say" by conditioning on the predicted probabilities. Formally:

Definition 2.8 (Calibration)**.** *We say the predictor is* ϵ-calibrated *for* $\mathcal{D}$ *if*

$$\mathbb{E}\left[\left|\mathbb{E}[\mathbf{y}|p(\mathbf{x})] - p(\mathbf{x})\right|\right] \leq \epsilon. \tag{2.9}$$

As discussed in the case of multiaccuracy, we need to average over the level sets of the predictor p in Equation 2.9 because we need to account for the level sets of p that are too small. But let us take perfect calibration to explain the meaning of calibration: it guarantees that $\mathbb{E}[\mathbf{y}|p(\mathbf{x})] = p(\mathbf{x})$. That is, if we restrict the domain to the points where p is predicting $v \in [0, 1]$ (i.e., we zoom into one of the level sets of p), then the expected value of the true labels $\mathbf{y}$ in that level set is equal to v. In order to measure how close a predictor is to perfect calibration, we use the *expected calibration error* of a predictor:

$$\mathsf{ECE}(p, \mathcal{D}) = \mathbb{E}\left[\left|\mathbb{E}[\mathbf{y}|p(\mathbf{x})] - p(\mathbf{x})\right|\right]. \tag{2.10}$$

We say p is ϵ-calibrated under $\mathcal{D}$ if $\mathsf{ECE}(p, \mathcal{D}) \leq \epsilon$. Equivalently, we can express the expected calibration error using bounded auditing functions [GKSZ22]:

$$\mathsf{ECE}(p, \mathcal{D}) = \max_{v:[0,1]\to[-1,1]} \mathop{\mathbb{E}}_{\mathcal{D}}[v(p(\mathbf{x}))(y - p(\mathbf{x}))]. \tag{2.11}$$

This characterization is very useful, given that it matches the form of the multigroup fairness definitions.

We point out that in recent years, other calibration measures have been proposed beyond ECE,

including *smooth calibration* [KF04], *U-calibration* [KLST23], proper calibration [OKK25], and *subsampled step calibration* [QZ25]. Each of these has interesting (and different) theoretical properties that try to circumvent some of the flaws of ECE. For example, ECE is *discontinuous*, in that small perturbations of p can cause large fluctuations in the expected calibration error [BGHN23].

Before going into the notion of multicalibration, we formally define the notion of calibrated multiaccuracy, which simply puts together the notions of multiaccuracy and global calibration:

Definition 2.12. *We say that p is* $(\mathcal{C}, \epsilon)$-multiaccurate and calibrated *if it is both $(\mathcal{C}, \epsilon)$-multiaccurate and ϵ-calibrated.*

We abbreviate this notion to "Cal-MA" (in conjunction with "MA" for multiaccuracy and "MC" for multicalibration). Note that multiaccuracy is defined with respect to a class $\mathcal{C}$, whereas calibration is not (in contrast to multicalibration, which does require the base class $\mathcal{C}$ as well).

Having understood calibration, we can now formalize the notion of multicalibration. It corresponds exactly to what one would expect having seen the definitions of multiaccuracy and calibration: we require the predictor to be calibrated locally within every $c \in \mathcal{C}$. Formally:

Definition 2.13 (Multicalibration)**.** *We say that p is* $(\mathcal{C}, \epsilon)$-multicalibrated *if*

$$\max_{c \in \mathcal{C}} \mathbb{E}[|c(\mathbf{x})(\mathbb{E}[\mathbf{y}|p(\mathbf{x})] - p(\mathbf{x}))|] \leq \tau.$$

Let us parse this definition slowly. In the case of multiaccuracy, we conditioned on the subgroup represented by c, and compared the expected value of $\mathbf{y}$ and of p. Here, we are conditioning on two subsets: on the subset $\mathcal{X}_c = \{x \in \mathcal{X} \mid c(x) = 1\}$ for each $c \in \mathcal{C}$, and on each level set of p (i.e., on each value in the range of p). In analogy to Equation 2.7 for multiaccuracy, for multicalibration indistinguishability here means that

$$\mathbb{E}[\mathbf{y} \mid \mathbf{x} \in \mathcal{X}_c, p(\mathbf{x}) = v] \approx \mathbb{E}[v \mid \mathbf{x} \in \mathcal{X}_c, p(\mathbf{x}) = v]. \tag{2.14}$$

Note that because we condition on the level sets $p(\mathbf{x}) = v$ for each v in the range of p, we can substitute $p(\mathbf{x})$ for v in the expression.

It can be useful to think about this in two steps: first, we condition on the chosen $c \in \mathcal{C}$ (we do this for every $c \in \mathcal{C}$, so we repeat this process for all of the subgroups). Then, we check whether p is calibrated over $\mathcal{X}_c$. To do so, we further condition on each of the values in the range of p. For each, we check whether $\mathbb{E}[\mathbf{y}] \approx \mathbb{E}[p(\mathbf{x})]$. If this holds for all of the level sets of p, then we conclude that p is calibrated over $\mathcal{X}_c$. We then repeat this process for every $c \in \mathcal{C}$, and if calibration always holds, then we conclude that p is $\mathcal{C}$-multicalibrated. We represent this two-step process visually in Figure 2.2. We remark that in our visual explanations, we have considered the $c \in \mathcal{C}$ to be Boolean functions. In the multigroup fairness literature, most follow-up works use the generalization of real-valued concepts c [KGZ19; GKR+22].

What is a good forecaster? A historical remark here is that multicalibration is in spirit a rediscovery of the notion of *computational calibration* in the literature on forecasting, as first introduced by Philip Dawid [Daw85]. In the forecasting example, we usually think of $p^*(x)$ as the true probability of rain on a given day x. Someone builds a weather forecasting model p and claims that it is a good weather forecaster. How can we check whether p is actually a good forecaster?

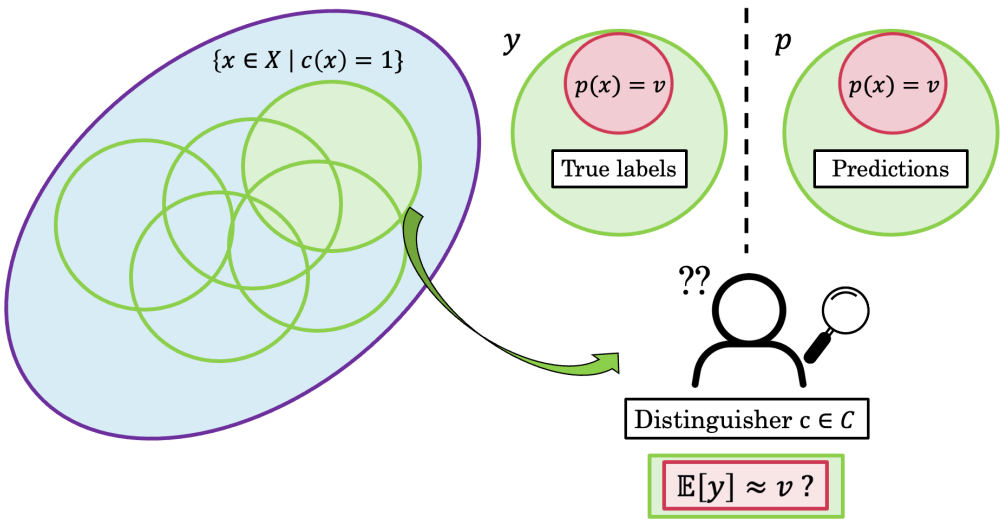


Figure 2.2: Visual depiction of the notion of multicalibration as a Turing Test. We repeat this checking process for every level set of p that intersects with c and for every $c \in \mathcal{C}$.

Dawid and many others have argued that calibration is a very good measure for doing so, which can be seen as a test to check the empirical validity of the forecasts predicted by p [Daw82; FV98; SSV03; San03]. To begin with, note that we never observe the true probabilities p^*; we only observe the Boolean outcomes that are generated from it (i.e., $\mathbf{y} \sim \text{Bern}(p^*(\mathbf{x}))$. The way to "bridge" the true probabilities p^* (which we cannot directly observe but we are interested in mimicking in our predictor p) and the observable Boolean outcomes $\mathbf{y}$ for each $x \in \mathcal{X}$ is by taking groups: if we choose a subset in $\mathcal{X}$, then we can average the Boolean labels in that subset, and then compare this quantity to the average of our predicted probabilities $p(x)$ inside this subset. This is precisely the idea behind recent algorithms explaining how to *reconcile* two predictors p_1, p_2 proposed for the same data: we cannot know the true individual probabilities, but they are contestable via a computationally and data efficient process, which is based on selecting appropriate subgroups of the domain and comparing the true Boolean labels on average on the domain with the average predictions in the same subgroup [RTW23; BCDT25].

In this way, we can test for calibration globally on $\mathcal{X}$ in order to see how well p is doing. But is this enough to determine that p is a good forecaster? We would like it to be calibrated not only globally over $\mathcal{X}$, but also on Mondays, on Fridays, during the month of July, on the group of even days, and so on. This idea should appear very familiar to us by now! Indeed, this is precisely the key insight behind the notion of multicalibration. We can view the group of "Mondays" in a weather forecaster the same way that we view a minority group of the population in algorithm fairness. Then the question becomes: with respect to which collection of groups should we test for calibration? The answer that Dawid and the multicalibration framework gives is the same one: any group that we can efficiently test for (i.e., determine membership). In spirit with Turing's notion of *computable numbers*, Philip Dawid calls this notion *computational calibration* [Daw85], taking $\mathcal{C}$ to be the collection of all "computable selection rules," where a selection rule is a function $c : \mathcal{X} \to \{0, 1\}$. What matters is that these rules are efficiently computable. Similarly, for MA/MC, what we need is for membership testing to each $c \in \mathcal{C}$ to be efficiently computable.

In the case of the algorithmic fairness literature, this provides a computational answer to the usual question of: which are the subgroups in the population that we should "protect" against algorithmic bias? (I.e., for which groups should we ensure and check for high accuracy and calibration?) The answer is again: for any subgroup of the population that is computationally identifiable. Naturally, the hardness of building a $\mathcal{C}$-multicalibrated predictor depends on the complexity of the class $\mathcal{C}$. The more stringent it is, the closer we are requiring p to be to p^* (in the extreme, we reduce computational closeness back to statistical closeness, and only the ground truth predictor p^* would be able to satisfy $\mathcal{C}$-multicalibration). These sorts of discussions on the meaning of individual probabilities are also explicitly explored within the multigroup fairness literature in the framework of *Outcome Indistinguishability* (OI) [DKR+21] and in relation to the *reference class problem* [RTW23]. The OI framework provides a complexity-theoretic definition of what it means for a predictor to yield a generative model for outcomes that cannot be efficiently refuted based on "true" observations produced by "Nature" (i.e., p^*). In essence, we can view computational calibration/multicalibration as a way of getting "close" to p^* in this computational sense even though we never have access to the true individual probabilities, an issue we resolve precisely by evaluating our predictor p on many intersecting subgroups from $\mathcal{C}$ on the domain.

Lastly, we remark that a lot of the modern literature on forecasting studies calibration in relation to decision-making in an online learning context, focusing on the notion of *regret*. Indeed, Foster and Vohra established that agents who best-respond to calibrated forecasts have no swap regret [FV98], and that it is possible to maintain calibrated forecasts in a sequential adversarial setting [FV97; RS24]. Then, we can view calibration as a guarantee that predictions are trustworthy and can be reliably taken at face value for downstream use [FT25], a view that has generated a lot of recent work and a newfound interest in the study of calibration [KLST23; DNW24; FGMS25; DDF+25].

Constructing multiaccurate and multicalibrated predictors. So far, we have presented and parsed the definitions of multiaccuracy, calibrated multiaccuracy, and multicalibration. These are all definitions of notions that we would like our predictor to satisfy. But, given $\mathcal{C}$ and $\mathcal{D}$, can we efficiently construct such predictors? The original paper by [HKRR18] answers the question in the positive for multiaccuracy and multicalibration, demonstrating that we can efficiently build low-complexity predictors (with respect to $\mathcal{C}$) that satisfy these notions. In a follow-up work, Gopalan et al. extend this positive result to the case of calibrated multiaccuracy [GHK+23]. What do we mean by "low-complexity with respect to $\mathcal{C}$"? This is again defined with respect to the underlying class $\mathcal{C}$. To make it precise, we need to introduce the notion of *relative complexity*, which we do through the computational model of circuits. Specifically, we consider circuits with Boolean gates of fan-in at most 2, and the size of a circuit corresponds to the total number of wires in it.

Definition 2.15 (Relative complexity of a function and a function class [JP14, Definition 6])**.** *Let $\mathcal{C}$ be a family of functions $c\colon \mathcal{X} \to [0,1]$. A function h has* complexity (t,q) relative to $\mathcal{C}$ *if it can be computed by an oracle-aided circuit of size t with q oracle gates, where each oracle gate is instantiated with a function from $\mathcal{C}$. We denote by $\mathcal{C}_{t,q}$ the class of functions that have complexity at most (t,q) relative to $\mathcal{C}$.*

Thus, we can always assume that $q \leq t$. We are interested in the parameter q; for the parameter t we just need to be careful about adding the appropriate bit-length parameters and account for the overhead that comes from performing arithmetic computations on a circuit. Indeed, the parameter

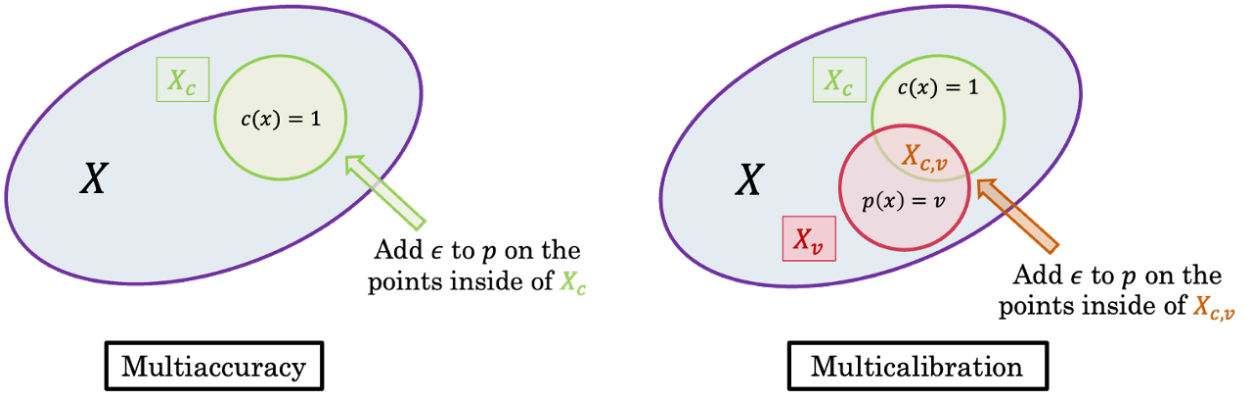


Figure 2.3: Visual depiction of the update step in the iterative algorithm for multiaccuracy (left) and multicalibration (right), using a witness of the violation $c \in \mathcal{C}$. Here, $\mathcal{X}_v = \{x \in \mathcal{X} \mid p(x) = v\}$.

q is what captures the number of calls that we need to perform to the weak agnostic learner during the construction of a multigroup fair predictor. Given a target error ϵ, the algorithms for multiaccuracy and multicalibration follow the same recipe:

1. *Audit for MA/MC violation.* Starting from a trivial predictor p, we check whether there exists any $c \in \mathcal{C}$ that "witnesses" an ϵ-violation of the multiaccuracy/multicalibration condition. That is, some c for which the expression in Definition 2.5/Definition 2.13 is not satisfied. We can view this as checking whether some $c \in \mathcal{C}$ has correlation with the current residuals $\mathbf{y} - p$. This is why it is natural to perform this auditing step with a weak agnostic learner: given a weak agnostic learner for the class $\mathcal{C}$, we call it using the distribution that samples $\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}$ and then assigns label $\mathbf{y} - p(\mathbf{x})$.

2. *Update the predictor using the witness.* If no $c \in \mathcal{C}$ witnesses a MA/MC violation, then our algorithm terminates, and definitionally we have constructed a MA/MC predictor. If we do find a "violation witness" $c \in \mathcal{C}$, then we can use it to update the predictor. Specifically, we perform a gradient step where we add ϵ to our predictions p inside of the region $\mathcal{X}_c$ (capping to $[0, 1]$). In the case of multicalibration, we further condition on the level sets of p, and so we perform the gradient step on the region $\mathcal{X}_c \cap \{p(x) = v\}$. The fact that we need to take this intersection is what makes multicalibration more expensive to achieve, but its algorithm follows this same iterative-based process. We visually represent the update step in Figure 2.3.

The idea at the heart of this algorithm lies in the observation that, whenever we detect a region of the domain on which we are not accurate/calibrated with a witness $c \in \mathcal{C}$, we can use this same witness to update our predictor and make enough progress in squared loss.

Because when this iterative algorithm terminates, we definitionally obtain a MA/MC predictor, what we need to show is that this process terminates within not too many steps (i.e., polynomial in $1/\epsilon$). Hébert-Johnson et al. show that each update step improves the squared loss of p by at least some amount. Therefore, by a potential argument, we are able to upper bound the number of steps in this iterative process. This number of steps is equal to the number of calls to the weak agnostic learner, which is in turn equal to the q parameter in the notion of $\mathcal{C}_{t,q}$. Indeed, note that by the nature of this iterative MA/MC algorithm, the final predictor p is equal to a linear combination of

q many distinguishers $c \in \mathcal{C}$. This is what we mean when we say that p is of "low-complexity with respect to the class $\mathcal{C}$": namely, that if we can efficiently compute the functions $c \in \mathcal{C}$, then p is "not much harder" to compute, where this relative difference is measured precisely by the parameter q (i.e., by how many times we had to call the weak agnostic learner, and so equivalent by how many distinguishers $c \in \mathcal{C}$ appear in the final expression of p as a linear combination of functions $c \in \mathcal{C}$).

We remark that the weak agnostic learner that we use in the MA/MC algorithm need not be proper for the algorithm to work. Indeed, as done in [HKRR18], we simply perform the same gradient step using the hypothesis returned by the weak learner, which need not be in $\mathcal{C}$. Note that all we need is that this gradient update to decrease the potential function (the squared loss in this case) by at least some amount that depends on ϵ; it is not necessary that we actually fix the region $\mathcal{X}_c$ on which the violation occurs at the current iteration.

This fact illustrates why this potential-based iterative algorithm is very powerful and how it enables the stringent notions of multiaccuracy and multicalibration for arbitrary collections $\mathcal{C}$: if we instead tried to fix the predictions of p one $c \in \mathcal{C}$ at a time, then we would only be able to work with families $\mathcal{C}$ of small size. Instead, all we need in the algorithm is to perform a gradient update that will improve the squared loss of p. The power of efficient constructions of multicalibration is realizing that, even though the definition of multicalibration is self-referential, in that we need the predictions of p to define the region on which we condition at each step (as opposed to multiaccuracy, where this always corresponds to the regions defined by the functions $c \in \mathcal{C}$, and $\mathcal{C}$ is a fixed collection all throughout the algorithm), this iterative algorithm still goes through, and we obtain an efficient algorithm that still requires only $\text{poly}(1/\epsilon)$ many calls to the weak agnostic learner. Specifically, from this general iterative algorithm we obtain the following guarantees [HKRR18; GHK+23]:

Theorem 2.16 (MA/Cal-MA/MC theorem [HKRR18; GHK+23])**.** *Let $\mathcal{X}$ be an instance space $\mathcal{X}$, $\mathcal{C}$ a concept class $\mathcal{C}$ of concepts $c : \mathcal{X} \to \mathbb{R}$, $\mathcal{D}$ a distribution on $\mathcal{X} \times \mathcal{Y}$, and $\epsilon > 0$ a parameter. Given access to random samples (x, y) from $\mathcal{D}$ and to a proper weak agnostic learner for $\mathcal{C}$, we can:*

1. *Build a predictor $p_{\text{MA}} \in \mathcal{C}_{t,q}$ that is $(\mathcal{C}, \epsilon)$-multiaccurate for $\mathcal{D}$, with parameters $q = O(1/\epsilon^2)$, $t = ((1/\epsilon^2) \cdot \log(|\mathcal{X}|/\epsilon))$ in time poly(t).*
2. *Build a predictor $p_{\text{Cal}-\text{MA}} \in \mathcal{C}_{t,q}$ that is ϵ-calibrated & $(\mathcal{C}, \epsilon)$-multiaccurate for $\mathcal{D}$, with parameters $q = O(1/\epsilon^2)$, $t = ((1/\epsilon^2 + 1/\epsilon) \cdot \log(|\mathcal{X}|/\epsilon)))$ in time poly(t).*
3. *Build a predictor $h_{\text{MC}} \in \mathcal{C}_{t,q}$ that is $(\mathcal{C}, \epsilon)$-multicalibrated for $\mathcal{D}$, with parameters $q = O(1/\epsilon^6)$, $t = ((1/\epsilon^6) \cdot \log(|\mathcal{X}|/\epsilon))$ in time poly(t).*

For multiaccuracy, the parameter q in Theorem 2.16 follows from the original algorithm in [HKRR18]. For multicalibration, we use the number of calls to the weak agnostic learner used in the multicalibration algorithm presented in [GKR+22], which carries out a more careful analysis of the original multicalibration algorithm from [HKRR18]. The algorithm for constructing a calibrated multiaccurate predictor (which, as we can see Theorem 2.16, has complexity very similar to that of a multiaccurate predictor), was first given by Gopalan, Hu, Kim, Reingold, and Wieder [GHK+23]. The reason why the q parameter for Cal-MA is the same as for MA is because their algorithm

alternates recalibration steps with calls to the weak agnostic learner for multiaccuracy violations. Then, given that the recalibrating step also reduces the squared loss by a similar amount, we can perform the same potential-based argument with a similar amount of total sets.

As for the t parameter, we need to count the number of oracle gates (i.e., the q parameter) plus the number of gates that we need for performing finite-precision arithmetic at a fixed prediction of $\Theta(\epsilon)$. We also need to account for the bit-length of the elements in $\mathcal{X}$, which is where the $\log(1/|\mathcal{X}|)$ term comes from. A detailed explanation for the t parameter for multiaccuracy and multicalibration can be found in [CDV24, §A], which in turn uses the circuit size computation from the work on Outcome Indistinguishability [DKR+21]. As for calibrated multiaccuracy, a formal justification for the t parameter can be found in [CGKR25, §5.3],[2] which analyzes the circuit parameters involved in the calMA algorithm first shown in [GHK+23].

From the iterative algorithm for constructing MA/MC predictors that we have explained in this section, it is clear that if we have access to a weak agnostic learner for the class $\mathcal{C}$, then we can efficiently construct a MA/MC predictor for the same class. In the case of multicalibration, Hébert-Johnson et al. show that this connection goes the other way: if we have access to a multicalibrated predictor for the class $\mathcal{C}$, then we can efficiently obtain a (weak) agnostic learner for $\mathcal{C}$. In Chapter 3, we study whether this is also true for the weaker notion of multiaccuracy.

The richness of the multigroup fairness framework. Ever since [HKRR18] showed that we can efficiently construct multiaccurate and multicalibrated predictors, the literature on multigroup fairness has exploded over the past few years. On the practical side, there is now an R package for multicalibration [PKD+21], as well as a Python package [HDNS24], and multicalibration has been applied to improve the performance of prediction models on minority subgroups in real medical data [BRA+20] and in other types of data [KGZ19], as well as to incorporate human expertise into algorithmic predictions [ARS24]. On the theoretical side, many variants of the initial algorithms have been proposed, including extensions to the multi-class setting and to the online setting [KNRW18; LSH19; ZKS+21; GKSZ22; GRSW22; GJRR24; DRR23; GHR24]. On the learning-theoretic front, a major application of the multigroup fairness framework has been the development of a new loss minimization paradigm called *omniprediction* [GKR+22; GKR23; GGKS23; HNRY23; GKR24; GOR+24; OKK25], which we define in Section 2.4 and explore extensively in Part II of this thesis.

On the complexity-theoretic front, recent works (including this thesis) explore the connections of the multigroup fairness framework to the Regularity Lemma in complexity theory [DLLT23; CDV24; CDV24; MPV25; HV25; CGKR25], which we explain below in Section 2.3. These works explore the connections between the multigroup fairness and Impagliazzo's Hardcore Lemma [Imp95; Hol05], the Dense Model Theorem [RTTV08; TZ08; GT08], characterizations of pseudoentropy [VZ12; VZ13], and the Frieze-Kannan Regularity Lemma in graph theory [FK99; Skó17]. A computational indistinguishability-based perspective on the multigroup fairness framework is also given by *Outcome Indistinguishability*, which sees these predictors as providing a generative model for Nature that cannot be falsified based on empirical data [DKR+21; DKR+22; HPR22].

The multigroup fairness framework has also found recent applications in performative prediction [KP23], statistical inference with covariate shifts [KKG+22; NR23; WLCW24], causal inference [KKZ24], conformal prediction and uncertainty quantification [JLP+21; GJN+22; JNRR23;

[2]The paper the Part I of this thesis is based on.

BGJ+22; AGG+24], confidence scoring in LLMs [DBFR24; LW24], game theory [LNPR22; HJZ24; RS24], the model multiplicity problem [RTW23; DNW24; BCDT25], cryptography [Rot23; DLLT23], ranking algorithms [DKKS24], and language generation [PRR24].

At the heart of many of these applications is the same central idea: we define an appropriate collection of groups, and then we seek to satisfy some condition of interest not only globally, but also locally when conditioned on each of the groups in the collection. We have already discussed the various interpretations of the class $\mathcal{C}$ in the cases of the Regularity Lemma and of learning theory, where the functions $c \in \mathcal{C}$ correspond to distinguishers or bounded-size circuits and to concepts in a concept class. For the connection to the Frieze-Kannan lemma, for example, the functions $c \in \mathcal{C}$ correspond to indicator cut functions in a graph [TTV09; DLLT23]. In applications to LLMs, we take the subgroups c to include concepts that are semantically close to each other [DBFR24]. In the case of the model multiplicity, these correspond to regions of disagreement between two predictors [RTW23; BCDT25]. In applications to the covariate shift problem, the functions $c \in \mathcal{C}$ correspond to score re-weighting functions [KKG+22; KKZ24].

In spirit, all these applications of the multigroup fairness framework are different instantiations of Turing's Imitation Game,[3] also known as the *Turing Test*, where we need to define the appropriate class of distinguishers and the arbitrarily complex function that we wish to simulate. Then, we try to construct a "simple" object that appears indistinguishable from the complex function to the class of distinguishers.

2.3 The Regularity Lemma

To conclude the preliminaries chapter, we further explain the previously-known connections between the multigroup fairness framework and the two areas that we are interested in for this thesis: complexity theory and learning theory. We begin with the former. We remark that we briefly change the distribution $\mathcal{D}$ to be defined over $\mathcal{X}$, in order to match the notion from [TTV09], but then we return to our usual distribution $\mathcal{D}$ defined on $\mathcal{X} \times \mathcal{Y}$.

The starting observation of Casacuberta, Dwork, and Vadhan, inspired by [DLLT23], is that $\mathcal{C}$-multiaccuracy is exactly equivalent to the notion of $\mathcal{C}$-indistinguishability, as studied by Trevisan, Tulsiani, and Vadhan in the complexity theoretic literature [TTV09; CDV24]. Specifically, Trevisan et al. showed the following result, called the *Regularity Lemma*:

Theorem 2.17 (Regularity Lemma [TTV09], informally stated)**.** *For every finite domain $\mathcal{X}$, every function $p^* : \mathcal{X} \to [0,1]$, every collection $\mathcal{C}$ of functions $c : \mathcal{X} \to [0,1]$, every distribution $\mathcal{D}$ on $\mathcal{X}$, and every $\epsilon > 0$, there exists a function $p : \mathcal{X} \to [0,1]$ such that:*

1. *p has "low complexity" relative to $\mathcal{C}$. Specifically, p can be computed by a Boolean circuit that has $O(1/\epsilon^2)$ oracle gates instantiated with functions from $\mathcal{C}$ and has size poly$(\log|\mathcal{X}|, 1/\epsilon)$.*
2. *p is $(\mathcal{C}, \epsilon)$-*indistinguishable *from p^*. That is, for all $c \in \mathcal{C}$, we have:*

$$\left| \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}} [c(\mathbf{x}) \cdot (p^*(\mathbf{x}) - p(\mathbf{x}))] \right| \leq \epsilon. \tag{2.18}$$

[3] As with many central concepts in theoretical computer science, from pseudorandomess to semantic security.

Notice that Condition (2.18) is identical to the notion of multiaccuracy. Then, we can view the Regularity Lemma as equivalent to the "multiaccuracy theorem" (Theorem 2.16). In their case, we view p^* as an arbitrarily complex function, which we want to "simulate" using a low-complexity predictor p.

Trevisan, Tulsiani, and Vadhan prove the Regularity Lemma in two different ways: through a boosting proof using a potential argument and through Von Neumann's min-max theorem. This same two approaches were used by Impagliazzo to prove the Hardcore Lemma [Imp95]. The boosting proof in [TTV09] is equivalent to the approach undertaken in the multigroup fairness literature to construct MA/MC predictors. In recent work, Haghtalab, Jordan, and Zhao provide a unifying game-theoretic approach to multicalibration, where they view multicalibration precisely as a min-max optimization problem [HJZ24].

Bringing this view back into the multigroup fairness framework, we can view multicalibration in the following complexity-theoretic way:

Theorem 2.19 (Complexity-theoretic MC theorem [HKRR18; CDV24])**.** *Given $\mathcal{X}, \mathcal{C}, p^*, \mathcal{D}, \epsilon$, there exists a partition $\mathcal{P}$ of $\mathcal{X}$ such that:*

1. *$\mathcal{P}$ has $k = O(1/\epsilon)$ parts.*
2. *$\mathcal{P}$ has "low complexity" relative to $\mathcal{C}$. Specifically, there is a Boolean circuit $C : \mathcal{X} \to [k]$ of size poly$(1/\epsilon, \log|\mathcal{X}|)$ with $O(1/\epsilon^6)$ oracle gates instantiated with functions from $\mathcal{C}$ such that $\mathcal{P} = \{C^{-1}(1), \ldots, C^{-1}(k)\}$.*
3. *$\mathcal{P}$ is* $(\mathcal{C}, \epsilon)$-multicalibrated *for p^* on $\mathcal{D}$: that is, for all $c \in \mathcal{C}$ and all $P \in \mathcal{P}$, we have*

$$\mathop{\mathbb{E}}_{X_v \sim \mathcal{P}(\mathcal{D})} \left| \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}|_v} [c(\mathbf{x}) \cdot (p^*(\mathbf{x}) - v_P)] \right| \leq \epsilon,$$

where $\mathcal{D}|_v$ denotes the conditional distribution $\mathcal{D}|_{p(x)=v}$ for $v \in [0,1]$ in the support of h, $\mathcal{P}(\mathcal{D})$ denotes the distribution that selects each X_v with probability $\big(\sum_{x \in X_v} \mathcal{D}(x)\big) / \big(\sum_{x \in \mathcal{X}} \mathcal{D}(x)\big)$, and where $v_P := \mathbb{E}_{x \sim \mathcal{D}|_v}[g(x)]$.

Where is the partition coming from? We simply need to take the level sets of a multicalibrated predictor p. Then, Theorem 2.19 can be understood as follows: multicalibration allows us to efficiently build a low-complexity partition of the domain such that on every large enough $P \in \mathcal{P}$, the arbitrarily complex function p^* is $\mathcal{C}$-indistinguishable from the constant function v_P. Casacuberta, Dwork, and Vadhan call these functions *constant-Bernoulli* functions, given that we can equivalently see this within-piece condition as saying that p^* is $\mathcal{C}$-indistinguishable from a Bernoulli random variable of parameter v_P. Hence, an MC partition provides a decomposition of the domain into not too many pieces such that the arbitrarily complex function p^* is a constant-Bernoulli function on every piece.

Of interest to us in our discussion to the Hardcore Lemma is what [CDV24] calls the *balance* of p^* on a piece $P \in \mathcal{P}$:

Definition 2.20 (Balance of p^* [CDV24])**.** *Given an arbitrary function $p^* : \mathcal{X} \to [0,1]$ and a partition $\mathcal{P} = \{P\}$ of $\mathcal{X}$, we let $v_P = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}|_P}[p^*(\mathbf{x})]$ for each $P \in \mathcal{P}$ and $b_P = \min\{v_P, 1 - v_P\} \leq 1/2$. We call b_P the* balance *of p^* on P.*

As we will see with the Hardcore Lemma, by generalizing Yao's equivalence between pseudorandomness and unpredictability [Yao82], [CDV24] show how we can characterize indistinguishability from a constant-Bernoulli function (which we obtain on every $P \in \mathcal{P}$ from the multicalibration theorem) as hardness of prediction, through the balance parameter b_P. In turn, by adapting the proof of Trevisan, Tulsiani, and Vadhan showing that we can prove the Hardcore Lemma from the Regularity Lemma, we can use this hardness present on every piece to construct a "small" hardcore set on each $P \in \mathcal{P}$. This construction is what allows us to prove a stronger and more general version of IHCL through multicalibration, which [CDV24] call "IHCL++". We make all of these intuitions more concrete in Chapter 4, where we show how to prove the original IHCL statement with optimal set density through calibrated (and weighted) multiaccuracy.

The broader key insight in this complexity-theoretic understanding of multigroup fairness is that multicalibration gives us a collection of polynomially many parameters b_P which characterize the average hardness of the input function p^*, which is *arbitrary*. It is important to remark that the field of cryptography also has their version of the Regularity Lemma, in their case called the *leakage simulation lemma* [JP14; DLLT23], originating in the field of leakage-resilient cryptography [DP08].

2.4 Omnipredictors

Lastly, we explain how the multigroup fairness framework has led to the development to a very powerful new learning-theoretic approach to the problem of loss minimization. As motivated in Chapter 1, in the agnostic setting we are tasked with producing a predictor that achieves loss no worse than the loss incurred by the best concept $c \in \mathcal{C}$ for some base concept class $\mathcal{C}$ (plus an ϵ slack). However, the best concept in $\mathcal{C}$ depends on the choice of the loss function, and current approaches to loss minimization heavily depend on the a *priori* chosen loss. Indeed, we do not have known ways of transforming a predictor optimized for the ℓ_1 loss, for example, into a predictor that is optimal for the ℓ_2 loss. So what should we do if we do not know the specific loss at the time of training, or if we want to change it at a later time without having the re-train from scratch?

This observation motivated Gopalan, Kalai, Reingold, Sharan, and Wieder to introduce the notion of an *omnipredictor*, which formalizes this idea of a "loss agnostic" approach to loss minimization:

Definition 2.21 (Omniprediction [GKR+22])**.** *Given a class of loss functions $\mathcal{L}$ and a concept class $\mathcal{C}$ of concepts $c : \mathcal{X} \to \mathbb{R}$, a predictor $h : \mathcal{X} \to [0,1]$ is an* $(\mathcal{L}, \mathcal{C}, \epsilon)$-omnipredictor *if for every $\ell \in \mathcal{L}$ there exists a function $k_\ell : [0,1] \to \mathbb{R}$ so that*

$$\ell_{\mathcal{D}}(k_\ell \circ h) \leq \min_{c \in \mathcal{C}} \ell_{\mathcal{D}}(c) + \epsilon.$$

That is, for every loss $\ell \in \mathcal{L}$, there exists a simple (univariate) transformation k_ℓ of the predictions h (chosen tailored to ℓ) such that $k_\ell \circ f$ has loss comparable to the best hypothesis $c \in \mathcal{C}$, which is chosen dependent on ℓ. That is, we can train a *single* predictor h that is able to do as well as the best hypothesis in $\mathcal{C}$ separately for every loss function in $\mathcal{L}$. This realizes a very strong learning guarantee, which we summarize in Figure 2.4. Note that for every $\mathcal{C}, \mathcal{L}$, the ground truth predictor p^* is an $(\mathcal{L}, \mathcal{C}, 0)$-omnipredictor. As shown in [GKR+22], the right post-processing function k_ℓ turns out to be the minimizer of the expected loss under the Bernoulli distribution. The key idea behind

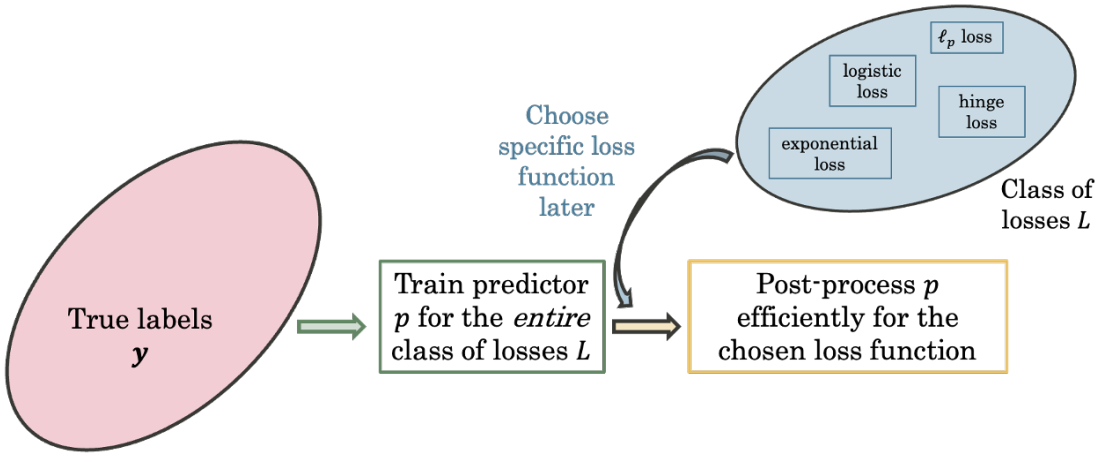


Figure 2.4: Diagram representing the notion of omniprediction.

the omnipredictors framework is to follow a new mantra for loss minimization: namely, "learn first, and optimize for a specific loss function later." For example, through the omnipredictors framework, besides choosing the loss of interest a *posteriori* (and being able to change it at any time), we can also add constraints after training [HNRY23].

In their main result, Gopalan et al. show that we can construct omnipredictors efficiently using the technique of multicalibration:

Theorem 2.22 (Building omnipredictors from multicalibration [GKR+22])**.** *Let $\mathcal{D}$ be a distribution on $\mathcal{X} \times \{0,1\}$, $\mathcal{C}$ a family of real-valued functions on $\mathcal{X}$, and $\mathcal{L}$ the family of all B-Lipschitz, convex loss functions. Then, a $(\mathcal{C}, \epsilon)$-multicalibrated predictor p is an $(\mathcal{L}, \mathcal{C}, 2\epsilon B)$-omnipredictor.*

A priori, this connection between multicalibration and loss minimization can seem surprising, given that the multigroup fairness framework does not explicitly deal with any notion related to this problem. However, after having thoroughly investigated the ideas that lie at the heart of the notion of multicalibration in Chapters 1 and 2, Theorem 2.22 should be intuitively true. This is because, as we have repeatedly pointed out, a $\mathcal{C}$-multicalibrated predictor appears indistinguishable from the ground truth predictor p^* to the functions $c \in \mathcal{C}$, and, crucially, the ground truth predictor is an $(\mathcal{L}, \mathcal{C}, 0)$-omnipredictor. Therefore, for any reasonable loss function ℓ, the loss incurred by a $\mathcal{C}$-multicalibrated predictor will be comparable to the loss incurred by the ground truth optimal predictor, which is always optimal by definition, as far as the concepts $c \in \mathcal{C}$ are concerned. But because we are operating within the agnostic learning framework, learnability is precisely defined relative to the base concept class $\mathcal{C}$ – i.e., precisely "as far as the concepts $c \in \mathcal{C}$ are concerned." Understanding loss minimization in this computational indistinguishability-based perspective has been formalized through the Outcome Indistinguishability framework [DKR+21; GHK+23].

In recent work, it has been shown that we can efficiently construct omnipredictors from the weaker primitive of calibrated multiaccuracy, rather than through full multicalibration [GHK+23; OKK25]. The omnipredictors theorem (Theorem 2.22) has also been extended to include loss functions beyond convex and Lipschitz losses: we can now efficiently build omnipredictors for loss functions such as the exponential loss, GLM losses, 1-Lipschitz losses, proper losses, and bounded variation losses [GKR+22; OKK25].

I

Agnostic Learning, Multigroup Fairness, and Hardcore Measures

3

Multiaccuracy & Agnostic Learning

The ultimate goal in this direction is informally termed agnostic learning, in which we make virtually no assumptions on the target function. The name derives from the fact that as designers of learning algorithms, we give up the belief that Nature (as represented by the target function) has a simple or succinct explanation.

Kearns, Schapire, and Sellie [KSS92]

In this chapter, we focus on the connections between multigroup fairness notions and agnostic learning. Our goal in this thesis is to complete the picture relating (1) multigroup fairness notions, (2) learning primitives, and (3) hardcore set constructions. Here, we study the relationship between (1) and (2). In the next chapter, we study the relationship between (1) and (3), as well as the relationship between (2) and (3) through the technique of boosting.

3.1 Multiaccuracy does not always yield learning

We begin by inquiring into the relationship between multiaccuracy and weak agnostic learning. As explained in Chapter 2, known algorithms for both multiaccuracy and multicalibration consist of an iterative process that calls a weak agnostic learner at each step. Therefore, we know that if a class $\mathcal{C}$ is weak agnostically learnable, then we can efficiently construct MA/MC predictors. In the original work on multicalibration, the authors show that the implication goes the other way: from a multicalibrated predictor we can obtain a (strong) agnostic learner [HKRR18]. Crucially, we obtain strong learnability from the *same* multicalibrated predictor.

Is the same true for multiaccuracy? Perhaps surprisingly, given that a multicalibrated predictor is itself a strong agnostic learner, the answer is no:

Theorem 3.1. *There exists a class $\mathcal{C}$, a distribution $\mathcal{D}$ on $\mathcal{X} \times \{0,1\}$, and a predictor $p : \mathcal{X} \to [0,1]$ such that*

- *p is $(\mathcal{C}, 0)$-multiaccurate.*
- *For any $k : [0,1] \to [-1,1]$, $k(p)$ is not a $(1/2, \beta)$-weak agnostic learner for $\mathcal{C}$ for any $\beta > 0$.*

Note that this rules out the possibility of being able in general to obtain a weak agnostic learner directly from a multiaccurate predictor p, or by post-processing its ouputs. However, it does not rule out the possibility that we could still agnostically learn the class $\mathcal{C}$ by post-processing the multiaccurate predictor in more complex ways.

Proof. We provide a counter-example by constructing a specific class $\mathcal{C}$, distribution $\mathcal{D}$, and predictor p. We take the uniform distribution on the domain $\{\pm 1\}^n$. Intuitively, we will want to play with the values of the domain that the concepts in $\mathcal{C}$ are not "looking at", so that we can construct a $\mathcal{C}$-multiaccurate predictor but then mess up the learning on the parts of the domain that $\mathcal{C}$ is not concerned with. For this reason, we choose the class $\mathcal{C} = \{c : \{\pm 1\}^{n-1} \to [-1,1]\}$ to be the set of all functions of the first $n-1$ bits. We define our predictor p using p^* as follows:

$$p(x_1, \ldots, x_n) = p^*(x_1, \ldots, x_{n-1}, -x_n).$$

That is, p outputs the value that p^* would have output, had it been given the same string but with the last bit flipped. First, we claim that p is $(\mathcal{C}, 0)$-multiaccurate for $\mathcal{D}$.

p is $(\mathcal{C}, 0)$-multiaccurate. To see that, let us consider p^* and p in their multilinear expansions. Namely, we can write p^* by separating the last bit x_n as

$$p^*(x_1, \ldots, x_n) = p_0^*(x_1, \ldots, x_{n-1}) + x_n p_1^*(x_1, \ldots, x_{n-1}).$$

This is because, by the fundamental Fourier expansion theorem in Fourier analysis, we know that every function $f : \{\pm 1\}^n \to \mathbb{R}$ can be uniquely expressed as a multilinear polynomial

$$f(x) = \sum_{S \subseteq [n]} \hat{f}(S) x^S,$$

where $x^S = \prod_{i \in S} x_i$ (with $x^\emptyset = 1$ by convention) [ODo14]. Equivalently, if we use the parity function $\chi_S(x) = \prod_{i \in S} x_i$ we can write the expansion as $f(x) \sum_{S \subseteq [n]} \hat{f}(S) \chi_S(x)$.

Hence, we can separate the multilinear polynomial corresponding to p^* into two parts: the polynomial that contains the monomials x^S such that $n \notin S$ (which corresponds to p_0^*) and the polynomial that does depend on the x_n; i.e., the one grouping all of the terms x^S in the expansion of p^* such that $n \in S$ (which corresponds to p_1^*). If we wanted to compute p_0^* and p_1^* explicitly, one can see that

$$p_0^*(x_1, \ldots, x_{n-1}) = \frac{1}{2}\big(p^*(x_1, \ldots, x_{n-1}, +1) + p^*(x_1, \ldots, x_{n-1}, -1)\big),$$

$$p_0^*(x_1,\ldots,x_{n-1}) = \frac{1}{2}\big(p^*(x_1,\ldots,x_{n-1},+1) - p^*(x_1,\ldots,x_{n-1},-1)\big).$$

If we consider the multilinear expansion of p, and separate the terms x^S depending on whether n is in S or not in the same way, we similarly have that

$$p(x_1,\ldots,x_n) = p_0(x_1,\ldots,x_{n-1}) + x_n p_1(x_1,\ldots,x_{n-1}).$$

Given that, by definition, $p(x_1,\ldots,x_n) = p^*(x_1,x_{n-1},-x_n)$, it follows that we can re-write the multilinear expansion of p by using p_0^* and p_1^* instead of p_0 and p_1 as follows:

$$p(x_1,\ldots,x_n) = p_0^*(x_1,\ldots,x_{n-1}) - x_n p_1^*(x_1,\ldots,x_{n-1}).$$

Indeed, if $n \notin S$ then $\chi_S(x_1,\ldots,-x_n) = \chi_S(x_1,\ldots,x_n)$, and if $n \in S$ then $\chi_S(x_1,\ldots,-x_n) = \chi_{S\setminus\{n\}}(x_1,\ldots,x_{n-1})\cdot(-x_n)$. Therefore,

$$\begin{aligned} p^*(x) - p(x) &= p_0^*(x_1,\ldots,x_{n-1}) + x_n p_1^*(x_1,\ldots,x_{n-1}) - (p_0^*(x_1,\ldots,x_{n-1}) - x_n p_1^*(x_1,\ldots,x_{n-1})) \\ &= 2x_n p_1^*(x_1,\ldots,x_{n-1}). \end{aligned}$$

It follows that for any function $c(x_1,\ldots,x_{n-1})$ of the first $n-1$ bits,

$$\mathbb{E}[c(\mathbf{x})(\mathbf{y}-p(\mathbf{x}))] = \mathbb{E}[c(\mathbf{x})(p^*(\mathbf{x})-p(\mathbf{x}))] = \mathbb{E}[c(\mathbf{x})2\mathbf{x}_n p_1^*(\mathbf{x})] = 0$$

since both c and p_1^* only depend on the first $n-1$ bits. So p is $(\mathcal{C},0)$-multiaccurate for $\mathcal{D}$.

$k(p)$ is not a weak agnostic learner. Next, we need to show that for any post-processing function k applied to the outputs of our multiaccurate predictor p, we cannot obtain not even a weak agnostic learner.

To do this, we will instantiate p^* to be a concrete function. Specifically, we use the *majority function* $\mathrm{MAJ}:\{\pm1\}^3\to\{0,1\}$, which is defined as

$$\mathrm{MAJ}(x_1,x_2,x_3) = \begin{cases} 1 & x_1+x_2+x_3 > 0 \\ 0 & \text{otherwise} \end{cases}$$

We define p^* using the MAJ function as:

$$p^*(x) = \mathrm{MAJ}(x_1,x_2,x_n).$$

We define the distribution $\mathcal{D}$ that one would expect: the marginal is uniform over $\mathcal{X}$, and once $\mathbf{x}$ is picked, its label $\mathbf{y}$ is given by $p^*(\mathbf{x})$. Then, recall that we defined our function p to output whatever p^* would output on the same string but with the last bit flipped, and so

$$p(x) = \mathrm{MAJ}(x_1,x_2,-x_n).$$

While we have shown that p is $(\mathcal{C},0)$-multiaccurate for any p^*, and so in particular when p^* corre-

sponds to the majority function on three bits, the values of p and p^* are very uncorrelated. Indeed:

$$\Pr[p^*(\mathbf{x}) = p(\mathbf{x})] = \Pr[\mathbf{x}_1 = \mathbf{x}_2] = \frac{1}{2}.$$

$$\Pr[p^*(\mathbf{x}) \neq p(\mathbf{x})] = \Pr[\mathbf{x}_1 \neq \mathbf{x}_2] = \frac{1}{2}.$$

It can be helpful to visualize the values of p^* and p with the following table:

x_1	x_2	$x_3, \ldots, x_{n-1}$	x_n	$-x_n$	$p^*(x)$	$p(x)$
-1	-1	$\ldots$	-1	$+1$	$\mathrm{MAJ}(-1,-1,-1) = 0$	$\mathrm{MAJ}(-1,-1,+1) = 0$
-1	-1	$\ldots$	$+1$	-1	$\mathrm{MAJ}(-1,-1,+1) = 0$	$\mathrm{MAJ}(-1,-1,-1) = 0$
-1	$+1$	$\ldots$	-1	$+1$	$\mathrm{MAJ}(-1,+1,-1) = 0$	$\mathrm{MAJ}(-1,+1,+1) = 1$
-1	$+1$	$\ldots$	$+1$	-1	$\mathrm{MAJ}(-1,+1,+1) = 1$	$\mathrm{MAJ}(-1,+1,-1) = 0$
$+1$	-1	$\ldots$	-1	$+1$	$\mathrm{MAJ}(+1,-1,-1) = 0$	$\mathrm{MAJ}(+1,-1,+1) = 1$
$+1$	-1	$\ldots$	$+1$	-1	$\mathrm{MAJ}(+1,-1,+1) = 1$	$\mathrm{MAJ}(+1,-1,-1) = 0$
$+1$	$+1$	$\ldots$	-1	$+1$	$\mathrm{MAJ}(+1,+1,-1) = 1$	$\mathrm{MAJ}(+1,+1,+1) = 1$
$+1$	$+1$	$\ldots$	$+1$	-1	$\mathrm{MAJ}(+1,+1,+1) = 1$	$\mathrm{MAJ}(+1,+1,-1) = 1$

We can see from the table that p^* and p agree on exactly half the inputs (marked in green), and disagree on the other half (marked in red), and thus their correlation is 0. Moreover, p^* and p have 0 correlation even when conditioned on the predicted values of p (this is why we chose the MAJ function as opposed to, for example, the AND function). Indeed, we can see from the table that, when conditioned on the value $p(x) = 0$, p^* and p still agree on half the inputs and disagree in the other half. The same holds when we condition on the value $p(x) = 1$.

That is, we have that

$$\mathbb{E}[(2\mathbf{y} - 1) \mid p(\mathbf{x}) = 0] = \mathbb{E}[(2\mathbf{y} - 1) \mid p(\mathbf{x}) = 1] = 0. \tag{3.2}$$

Recall from Chapter 2 that we measure the correlation with the $\{\pm 1\}$ and $[-1, 1]$ versions of the predictors, which is why we use $2\mathbf{y} - 1$. Therefore, we conclude that for any $k : [0, 1] \to [-1, 1]$,

$$\mathrm{cor}_{\mathcal{D}}(\mathbf{y}, k \circ p(\mathbf{x})) = \frac{k(0)}{2}\,\mathbb{E}[(2\mathbf{y} - 1) \mid p(\mathbf{x}) = 0] + \frac{k(1)}{2}\,\mathbb{E}[(2\mathbf{y} - 1) \mid p(\mathbf{x}) = 1] = 0.$$

We are not done yet, because if all the concepts in $\mathcal{C}$ also achieved correlation 0 with the labels, then we would not have a contradiction. Per the definition of weak agnostic learning, it remains to show that some concept in $\mathcal{C}$ achieves positive correlation with the labels.

And yet, there is high correlation between $c \in \mathcal{C}$ and the labels. In fact, not only does there exist a concept in $\mathcal{C}$ that achieves positive correlation; it achieves correlation $1/2$. Indeed, consider the dictator function $c_1(x) = x_1$, which belongs to $\mathcal{C}$, given that it is only a function of the first $n - 1$ bits (in fact, only of the 1st bit). Then, we can see that

$$\mathbb{E}[(2\mathbf{y} - 1)x_1 \mid x_1 = 1] = \mathbb{E}[(2\mathbf{y} - 1)x_1 \mid x_1 = -1] = \frac{3}{4} - \frac{1}{4} = \frac{1}{2}.$$

$$\mathrm{cor}_{\mathcal{D}}(\mathbf{y}, c_1(\mathbf{x})) = \frac{1}{2}\,\mathbb{E}[(2\mathbf{y} - 1)x_1 \mid x_1 = 1] + \frac{1}{2}\,\mathbb{E}[(2\mathbf{y} - 1)x_1 \mid x_1 = -1] = \frac{1}{2}.$$

That is, the dictator function $c_1 \in \mathcal{C}$ achieves correlation $1/2$ with the labels. We can again visualize this in our table: if we look at the last 4 rows, which correspond to the cases where $x_1 = 1$, then we see that these agree with the p^* value on 3 out of the 4 cases (which corresponds to a correlation of $1/2$ given that we measure correlation in the $[-1, 1]$ range). □

Note that in this construction, the predictor p is maximally *anti-calibrated*, in the sense that $\mathbb{E}[\mathbf{y}|p(\mathbf{x}) = v] = 1/2$ for any v in the range of p. That is, the predictor p is doing no better than random guessing in terms of squared loss. This construction fits nicely with our result below in Section 3.3 showing that if, besides multiaccuracy, we also require non-trivial global calibration, then this enforces some correlation between the predictor p and the target labels. This allows us to obtain the flip positive result: if p is multiaccurate *and calibrated*, then $\text{sign}(2p - 1)$ is a strong agnostic learner.

3.2 Multiaccuracy gives restricted weak agnostic learning

In our majorities example, we are able to produce a $(\mathcal{C}, 0)$-multiaccurate predictor that has 0 correlation with the labels, even though there is a concept $c \in \mathcal{C}$ that achieves correlation $\alpha = 1/2$ with the labels (we use the parameter α given the definition of a weak agnostic learner in Definition 2.3, where we used parameters $\alpha \geq \beta$). Conceptually, we can understand what is going on through Figure 3.1: it is possible for c to have the same correlation with both p^* and p (as required by $\mathcal{C}$-multiaccuracy), yet for p^* and p to be orthogonal.

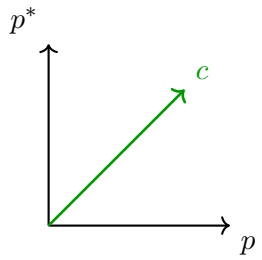


Figure 3.1: $\mathcal{C}$-multiaccuracy does not necessarily imply that the predictions of p and p^* are correlated.

However, is $1/2$ the maximum correlation that we could hope to obtain between c and the labels, while keeping p^* and p orthogonal and ensuring that p is multiaccurate for p^*? Because we compute correlations in the $[-1, 1]$ range, let us consider the $[-1, 1]$ versions of p^* and p, and call them q^* and q instead. That is, we let $q^*(x) = 2p^*(x) - 1$ and $q(x) = 2p(x) - 1$. Suppose that p is $(\mathcal{C}, 0)$-multiaccurate for p^*, and suppose that there exists a $c \in \mathcal{C}$ that has correlation at least α be the labels; that is, such that $\text{cor}(\mathbf{y}, c) = \langle q^*, c\rangle \geq \alpha$, where $\langle q^*, c\rangle$ denotes the inner product between q^* and c. Then, we are asking: What is the largest that α can be, such that we continue to ensure that: (1) q^* and q are orthogonal (that is, $\langle q^*, q\rangle = 0$, which means that the two predictors are completely uncorrelated), and (2) q is a $(\mathcal{C}, 0)$-multiaccurate predictor for q^*, but (3) q is not a weak learner for $\mathcal{C}$?

Perfect multiaccuracy ensures that

$$\langle q^*, c\rangle = \langle q, c\rangle.$$

Then, the assumption that $\langle q^*, c\rangle \geq \alpha$ implies that $\langle q, c\rangle \geq \alpha$ as well. For simplicity, assume that all of q^*, q, c are vectors in $\{\pm 1\}^n$. Because $\langle q^*, q\rangle = 0$ by assumption, from the definition of the inner product and given that q^* and q only take values in $\{\pm 1\}^n$, it must be that q^* and q agree on exactly half the bits and disagree on the other half. Then, we can re-arrange the n bits of q^* and q so that we first put the $n/2$ bits on which q^* and q agree, and then the $n/2$ bits on which q^* and q disagree (we can assume n is a multiple of four for simplicity). We provide an example of this construction in Figure 3.2.

Next, we provide an example of a vector c that achieves correlation $1/2$ with q^*, while q and q^* continue to be orthogonal. We already gave one such example by using the majority function (Section 3.1). Let d_H denote the normalized Hamming distance in this space. Note that for any two vectors x, y in $\{\pm 1\}^n$, we can relate their inner product and their Hamming distance as follows:

$$\langle x, y\rangle = 1 - 2d_H(x, y), \qquad d_H(x, y) = \frac{1 - \langle x, y\rangle}{2}.$$

This is because, by definition.

$$\langle x, y\rangle = \frac{1}{n}\sum_{i=1}^{n} x_i y_i, \qquad d_H(x, y) = \frac{1}{n}\big|\{i : x_i \neq y_i\}\big|.$$

Then, given that

$$x_i y_i = \begin{cases} 1 & \text{if } x_i = y_i, \\ -1 & \text{if } x_i \neq y_i, \end{cases}$$

it follows that

$$\langle x, y\rangle = \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \frac{1}{n}\cdot\sum_{i=1}^{n}\big(1 - 2\cdot \mathbb{1}[x_i \neq y_i]\big) = 1 - 2d_H(xy, y).$$

Note that this relation is analogous to the usual relation between correlation and error (which we stated in the beginning of Chapter 2). Indeed, the inner product corresponds to the correlation, whereas the Hamming distance counts the number of disagreements and thus corresponds to the error.

We let $c \in \{\pm 1\}^n$ be a mid-point of q, q^* in the Hamming metric. We can do so by making c be equal to $q = q^*$ on the first $n/2$ bits, and then on the second half of the bits, we make c agree with q^* on half of those bits, and with q on the other half. We provide one such example in Figure 3.2.

q^*	+1	+1	−1	−1	+1	+1	−1	−1
q	+1	+1	−1	−1	−1	−1	+1	+1
c	+1	+1	−1	−1	+1	+1	+1	+1

Figure 3.2: Bitwise comparison of q^*, q, and the midpoint vector c.

Then, per our construction, it follows that

$$\langle q^*, c\rangle = \langle q, c\rangle = 1/2,$$

given that c disagrees with both q^* and q on a $(1/4)$-th fraction of the bits. However, we still have that $\langle q^*, q\rangle = 0$, as we wanted to show.

Moreover, $\alpha = 1/2$ is a tight bound: if some $c \in \mathcal{C}$ achieves correlation $\alpha > 1/2$ with q^*, then q being $\mathcal{C}$-multiaccurate will imply that q and q^* have positive correlation, and thus that q is a weak learner in this case. Before we give the proof in full generality, we use Figure 3.2 for the simplified case where q and q^* are Boolean to understand why this is the case.

So we are assuming that $\langle q^*, q\rangle = 0$ and that q is $(\mathcal{C}, 0)$-multiaccurate. The former means again means that q^* and q agree on exactly $n/2$ bits, so we can place them as in Figure 3.2. The latter again implies that $\langle q^*, c\rangle = \langle q, c\rangle$. Suppose that $\alpha = \langle q^*, c\rangle > 1/2$. We denote the correlation over the first half of the n bits (from the construction as per Figure 3.2) with a subscript L (for "left"), and over the second half with R (for "right"). Then, we can write the total correlation as the sum of the correlation on the two parts:

$$\langle q^*, c\rangle = \frac{1}{2}\langle q^*, c\rangle_L + \frac{1}{2}\langle q^*, c\rangle_R,$$

$$\langle q, c\rangle = \frac{1}{2}\langle q, c\rangle_L + \frac{1}{2}\langle q, c\rangle_R.$$

From the $(\mathcal{C}, 0)$-multiaccuracy condition, it then follows that

$$\frac{1}{2}\langle q^*, c\rangle_L + \frac{1}{2}\langle q^*, c\rangle_R = \frac{1}{2}\langle q, c\rangle_L + \frac{1}{2}\langle q, c\rangle_R.$$

Because c agrees with q and q^* on the first half of the bits, we have that $\langle q^*, c\rangle_L = \langle q, c\rangle_L$. In order to preserve the orthogonality between q and q^*, we moreover need that

$$\langle q^*, c\rangle_R + \langle q, c\rangle_R = 0.$$

But if are to satisfy this, as well as the multiaccuracy condition that $\langle q^*, c\rangle = \langle q, c\rangle$, then the only way we can guarantee this is by having $\langle q^*, c\rangle_R = \langle q, c\rangle_R = 0$. Looking at Figure 3.2 again, we see that there is no way in which we can flip one of the bits of c in the R side to match with q^* on more than 2 bits (so that we obtain positive correlation between c and q^* on the R side), while still satisfying that $\langle q^*, c\rangle_R = \langle q, c\rangle_R$, which is required by multiacuracy. Then, the correlation of c with q^* remains bounded by $1/2$; as soon as $\alpha > 1/2$, it must be that $\langle q^*, q\rangle > 0$.

We can alternatively phrase this argument again using the normalized Hamming metric. As we showed above, the assumption that $\langle q^*, c\rangle \geq \alpha$ is equivalent to saying that $d_H(q^*, c) \leq (1-\alpha)/2$. By the perfect multiaccuracy guarantee, we also then have that $\langle q, c\rangle \geq \alpha$, and equivalently that $d_H(q, c) \leq (1-\alpha)/2$. Then, by the triangle inequality, we must have

$$d_H(q^*, q) \leq d_H(q^*, c) + d_H(q, c) \leq 1 - \alpha.$$

But translating back from Hamming distance to inner product, this means that the correlation

between q and q^* is lower bounded by

$$\langle q^*, q \rangle \geq 1 - 2(1-\alpha) = 2\alpha - 1. \tag{3.3}$$

This is a non-trivial guarantee whenever $\alpha > 1/2$ (that is, we only get learning if the correlation between q^* and q is positive).

Having given intuition for why $\alpha = 1/2$ is the right threshold, we now provide the formal statment. Essentially, we need to generalize this argument to the non-Boolean case.

Theorem 3.4. *For any $\mathcal{C}, \mathcal{D}$, we can post-process a $(\mathcal{C}, \tau)$-multiaccurate predictor to obtain a $(\alpha, 2\alpha - 1 - 2\tau)$-weak agnostic learning for $\mathcal{C}$ under the distribution $\mathcal{D}$. In particular, any $(\mathcal{C}, \tau)$-multiaccurate predictor p for $\mathcal{D}$ satisfies*

$$\mathrm{cor}_{\mathcal{D}}(\mathbf{y}, 2p(\mathbf{x}) - 1) \geq 2 \max_{c \in \mathcal{C}} \mathrm{cor}_{\mathcal{D}}(\mathbf{y}, c(\mathbf{x})) - 1 - 2\tau.$$

Proof. Let p be a $(\mathcal{C}, \tau)$-multiaccurate predictor with respect to $\mathcal{D}$. For any $c \in \mathcal{C}$, we have that:

$$\begin{aligned}
\mathrm{cor}_{\mathcal{D}}(\mathbf{y}, c) &= \mathrm{cor}_{\mathcal{D}}(p^*, c) \\
&= \mathop{\mathbb{E}}_{\mathcal{D}}[(2p^*(\mathbf{x}) - 1)c(\mathbf{x})] \\
&= \mathop{\mathbb{E}}_{\mathcal{D}}[c(\mathbf{x})(p^*(\mathbf{x}) - (1 - p^*(\mathbf{x})))(p(\mathbf{x}) + (1 - p(\mathbf{x})))] \\
&= \mathop{\mathbb{E}}_{\mathcal{D}}[c(\mathbf{x})(p^*(\mathbf{x})p(\mathbf{x}) - (1 - p^*(\mathbf{x}))(1 - p(\mathbf{x}))] + \mathop{\mathbb{E}}_{\mathcal{D}}[c(\mathbf{x})(p^*(\mathbf{x}) - p(\mathbf{x}))] \\
&\leq \mathop{\mathbb{E}}_{\mathcal{D}}[|c(\mathbf{x})| \cdot |p^*(\mathbf{x})p(\mathbf{x}) - (1 - p^*(\mathbf{x}))(1 - p(\mathbf{x}))|] + \tau \\
&\leq \mathop{\mathbb{E}}_{\mathcal{D}}[p^*(\mathbf{x})p(\mathbf{x}) + (1 - p^*(\mathbf{x}))(1 - p(\mathbf{x}))] + \tau.
\end{aligned}$$

The second equality follows because we measure the correlation with the $[-1, 1]$-versions of the predictors. In the next two equation, we split the terms so that we can separate the term $\mathbb{E}_{\mathcal{D}}[c(\mathbf{x})(p^*(\mathbf{x}) - p(\mathbf{x}))]$, and then apply to it the $(\mathcal{C}, \tau)$-multiaccuracy guarantee of p on $\mathcal{D}$, which tells us that this expression is upper bounded by τ. In the last equality, we use the fact that $|c(\mathbf{x})| \leq 1$, and that

$$|p^*(x)p(x) - (1 - p^*(x))(1 - p(x))| \leq p^*(x)p(x) + (1 - p^*(x))(1 - p(x)),$$

since both $p(x)p^*(x) \in [0, 1]$.

We can generalize the correspondence between the correlation/inner product and the normalized Hamming metric that we used in our Boolean example above with the following identity for $a, b \in \mathbb{R}$:

$$(2a - 1)(2b - 1) = 2(ab + (1 - a)(1 - b)) - 1.$$

We use this identity to express the correlation between $\mathbf{y}$ and p in this way:

$$\mathrm{cor}_{\mathcal{D}}(\mathbf{y}, 2p - 1) = \mathrm{cor}_{\mathcal{D}}(p^*, 2p - 1)$$

$$= \mathop{\mathbb{E}}_{\mathcal{D}}[(2p^*(\mathbf{x}) - 1)(2p(\mathbf{x}) - 1)]$$
$$= 2\mathop{\mathbb{E}}_{\mathcal{D}}[p^*(\mathbf{x})p(\mathbf{x}) + (1 - p^*(\mathbf{x}))(1 - p(\mathbf{x}))] - 1.$$

We can re-arrange the first inequality to match the form of the second one:

$$2\mathrm{cor}_{\mathcal{D}}(\mathbf{y}, c) - 1 \leq 2\mathop{\mathbb{E}}_{\mathcal{D}}[p^*(\mathbf{x})p(\mathbf{x}) + (1 - p^*(\mathbf{x}))(1 - p(\mathbf{x}))] - 1 + 2\tau.$$

Then, combining the two inequalities, we get that,

$$\mathrm{cor}_{\mathcal{D}}(\mathbf{y}, 2p - 1) = 2\mathop{\mathbb{E}}_{\mathcal{D}}[p^*(\mathbf{x})p(\mathbf{x}) + (1 - p^*(\mathbf{x}))(1 - p(\mathbf{x}))] - 1$$
$$\geq 2\max_{c \in \mathcal{C}} \mathrm{cor}_{\mathcal{D}}(\mathbf{y}, c) - 1 - 2\tau,$$

which proves the claim. Note that this last inequality is precisely the continuous-version of the lower bound that we showed in the Boolean case above using the normalized Hamming distance (Equation 3.3). □

Theorem 3.4 illustrates what we mean by a *restricted* weak agnostic learner: it is a weaker notion of the usual definition of a weak agnostic learner (Definition 2.3). For a typical (α, β)-weak agnostic learner, the learning guarantee is required to hold for any values of $\alpha \geq \beta > 0$. For a restricted weak agnostic learner, it is only required to hold if the value of α is at least some value γ_0; i.e., if some concept $c \in \mathcal{C}$ has correlation at least γ_0 with the labels. In the case of Theorem 3.4, $\gamma_0 = 1/2$.

Decomposition of p, p^* into two orthogonal components. As we have seen in our examples, intuitively, the reason why $\mathcal{C}$-multiaccuracy does not directly imply weak agnostic learning for the class $\mathcal{C}$ is that multiaccuracy only implies a certain type of closeness between the predictor p and the labels $\mathbf{y}$ over the span of the concept class $\mathcal{C}$. We conclude this section by making this intuition more precise.

As before, let q, q^* denote the $[-1, 1]$-versions of p, p^*. Let $\mathsf{span}(\mathcal{C})$ denote the vector space spanned by $\mathcal{C}$. Then, we can decompose each of q and q^* into two components: the part that lies in $\mathsf{span}(\mathcal{C})$, and the part that is orthogonal to it. Namely:

$$q = q_{\mathcal{C}} + q_{\perp} \quad \text{and} \quad q^* = q^*_{\mathcal{C}} + q^*_{\perp},$$

where $q_{\mathcal{C}}, q^*_{\mathcal{C}}$ are the components of q and q^* in the $\mathsf{span}(\mathcal{C})$ respectively, and $q_{\perp}, q*_{\perp}$ are the components of q, q^* orthogonal to $\mathsf{span}(\mathcal{C})$. For example, in our majorities example (Section 3.1), the projection of p onto $\mathsf{span}(\mathcal{C})$ is precisely p_0, and similarly for p^*, whereas $x_n p_1$ corresponds to $p_{\perp}$. Recall that the inner product $\langle f, g \rangle$ for functions $f, g : \mathcal{X} \to \mathbb{R}$ is defined by $\mathbb{E}_{\mathcal{D}_{\mathcal{X}}}[f(\mathbf{x})g(\mathbf{x})]$. Then, we can view the multiaccuracy condition in terms of an inner product (as we did with our Hamming example); specifically, if p is $(\mathcal{C}, 0)$-multiaccurate for $\mathcal{D}$, then it follows that

$$\langle c, q^* - q \rangle = 2\,\mathbb{E}[c(\mathbf{x})(p^*(\mathbf{x}) - p(\mathbf{x}))] = 0$$

for every $c \in C$. That is, $q^* - q$ is orthogonal to $\mathsf{span}(\mathcal{C})$. This is precisely what the notion of

multiaccuracy is meant to capture: if a predictor p is perfectly multiaccurate, then no $c \in \mathcal{C}$ should have correlation with the residual $q^* - q$, and so the two must be orthogonal.

Moreover, we can decompose the correlation between q^* and q in terms of their components:

$$\begin{aligned}\mathrm{cor}(\mathbf{y}, p) = \langle q^*, q\rangle = \mathbb{E}[q^*(\mathbf{x})q(\mathbf{x})] &= \mathbb{E}[(q^*_{\mathcal{C}} + q^*_{\perp})(q_{\mathcal{C}} + q_{\perp})] \\ &= \langle q^*_{\mathcal{C}}, q_{\mathcal{C}}\rangle + \langle q^*_{\mathcal{C}}, q_{\perp}\rangle + \langle q^*_{\perp}, q_{\mathcal{C}}\rangle + \langle q^*_{\perp}, q_{\perp}\rangle \\ &= \langle q^*_{\mathcal{C}}, q_{\mathcal{C}}\rangle + \langle q^*_{\perp}, q_{\perp}\rangle,\end{aligned}$$

where in the last equality we us the fact that the orthogonal components of the functions are orthogonal to $\mathsf{span}(\mathcal{C})$, and thus $\langle q^*_{\mathcal{C}}, q_{\perp}\rangle$ and $\langle q^*_{\perp}, q_{\mathcal{C}}\rangle$ are both equal to 0. From this expression we can understand how cases like our majorities example in Section 3.1 can arise: while $\mathcal{C}$-multiaccuracy ensures that the first term $\langle q^*_{\mathcal{C}}, q_{\mathcal{C}}\rangle$ is positive (i.e., $\mathcal{C}$-multiaccuracy ensures closeness between q and q^* within $\mathsf{span}(\mathcal{C})$, and so $q^*_{\mathcal{C}}$ and $q_{\mathcal{C}}$ are highly correlated), it says nothing about what is going on in the orthogonal space. Then, as in our majorities example, the correlation $\langle q^*_{\perp}, q_{\perp}\rangle$ can undermine the correlation present in $\mathsf{span}(\mathcal{C})$, so that $\mathrm{cor}(\mathbf{y}, p) = \langle q^*_{\mathcal{C}}, q_{\mathcal{C}}\rangle + \langle q^*_{\perp}, q_{\perp}\rangle = 0$, and therefore q and q^* end up being completely uncorrelated.

In the case of $(\mathcal{C}, 0)$-multiaccuracy, $q^*_{\mathcal{C}}$ and $q_{\mathcal{C}}$ are in fact equal, and so they have maximal correlation. This is because we can write

$$q^* - q = q^*_{\mathcal{C}} - q_{\mathcal{C}} + q^*_{\perp} - q_{\perp}.$$

As $q^*_{\perp}, q_{\perp}$ are both orthogonal to $\mathsf{span}(\mathcal{C})$, and $q^*_{\mathcal{C}} - q_{\mathcal{C}} \in \mathsf{span}(\mathcal{C})$, we must have $q^*_{\mathcal{C}} = q_{\mathcal{C}}$, given that $q^* - q$ is orthogonal to $\mathsf{span}(\mathcal{C})$. In turn, by the definition of the inner product and the norm, it follows that

$$\langle q^*_{\mathcal{C}}, q_C\rangle = \langle q^*_{\mathcal{C}}, q^*_{\mathcal{C}}\rangle = \|q^*_{\mathcal{C}}\|_2^2.$$

Summarizing, it can be very helpful to understand multiaccuracy by separating the correlation that is ensured over $\mathsf{span}(\mathcal{C})$, which multiaccuracy controls, and the orthogonal component, about which we cannot promise anything in general, and can behave in any way. Formally:

Lemma 3.5. *Let q, q^* denote the $[-1, 1]$-versions of predictors p, p^*. If the predictor p is $(\mathcal{C}, 0)$-multiaccurate for $\mathcal{D}$, then:*

1. $\mathrm{cor}(c, p^* - p^*) = \langle c, q^* - q\rangle = 0$.
2. $\mathrm{cor}(\mathbf{y}, p) = \langle q^*, q\rangle = \langle q^*_{\mathcal{C}}, q_{\mathcal{C}}\rangle + \langle q^*_{\perp}, q_{\perp}\rangle = \langle q^*_{\mathcal{C}}, q^*_{\mathcal{C}}\rangle + \langle q^*_{\perp}, q_{\perp}\rangle = \|q^*_{\mathcal{C}}\|_2^2 + \langle q^*_{\perp}, q_{\perp}\rangle$, *given that $(\mathcal{C}, 0)$-multiaccuracy implies that $q^*_{\mathcal{C}} = q_{\mathcal{C}}$.*

With our majorities example, we have already discarded the scenario where q has some correlation with the labels (and thus gives us weak learning). But given that perfect multiaccuracy does ensure that $q^*_{\mathcal{C}} = q_{\mathcal{C}}$, a natural hope here would be to see whether $q_{\mathcal{C}}$ (that is, the projection of q onto $\mathsf{span}(\mathcal{C})$ rather than q itself) can provide us with some learning. I.e., is $q_{\mathcal{C}}$ correlated with the labels? Through a similar decomposition calculation, we can see that:

$$\begin{aligned}\mathrm{cor}(\mathbf{y}, p_{\mathcal{C}}) = \langle q^*, q_{\mathcal{C}} \rangle &= \langle q^*_{\mathcal{C}} + q^*_{\perp}, q_{\mathcal{C}} \rangle \\ &= \langle q^*_{\mathcal{C}}, q_{\mathcal{C}} \rangle + \langle q^*_{\perp}, q_{\mathcal{C}} \rangle \\ &= \|q^*_{\mathcal{C}}\|_2^2,\end{aligned}$$

where we again use the fact that $q^*_{\perp}$ is orthogonal to $\mathsf{span}(\mathcal{C})$ and that $q^*_{\mathcal{C}} = q_{\mathcal{C}}$. Hence, if $\|q^*_{\mathcal{C}}\|_2^2 > 0$, then we would get learning from $p_{\mathcal{C}}$. We will pursue this direction in Chapter 5, showing that an ℓ_1 sparse projection of p onto $\mathsf{span}(\mathcal{C})$ does give us weak agnostic learning.

3.3 Global calibration to the rescue

So far, we have seen the following:

- A $\mathcal{C}$-multiaccurate predictor p is not necessarily a weak agnostic learner for $\mathcal{C}$ (majorities example, Section 3.1).
- If some concept $c \in \mathcal{C}$ has correlation $\alpha > 1/2$ with the labels, then multiaccuracy does give learning: if p is $(\mathcal{C}, 0)$-multiaccurate, then $\mathrm{sign}(p - 1/2)$ achieves correlation $\beta = 2\alpha - 1$ with the labels (Section 3.2). We call this reduced version of learning *restricted weak agnostic learning*, where we only promise to output a hypothesis that has correlation at least β with the labels if α is at least some threshold value γ_0 (in this particular case, $\gamma_0 = 1/2$). This stands in contrast to the usual notion of weak agnostic learning, where we promise learning even if α is very small.
- We have used the decomposition of p, p^* onto $\mathsf{span}(\mathcal{C})$ and its orthogonal subspace to explain how p, p^* can be uncorrelated even though p is $(\mathcal{C}, 0)$-multiaccurate. This has lead us to the hope that if we know how to efficiently project p onto $\mathsf{span}(\mathcal{C})$, then $p_{\mathcal{C}}$ could give us learning (we will return to this idea in Section 5.1).

However, in general we do not have efficient algorithms for projecting a predictor onto a class $\mathcal{C}$. Instead, we can ask: is there some minimal property that we can add to the predictor p (on top of multiaccuracy) so that it gives us learning?

In this section, we show that the answer is yes: If p is both $\mathcal{C}$-multiaccurate and calibrated, then $\mathrm{sign}(p - 1/2)$ is in fact not only a weak agnostic learner, but a *strong* agnostic learner. Note that we are only talking about *global* calibration here (i.e., defined without any class $\mathcal{C}$), not the much stronger notion of $\mathcal{C}$-multicalibration. This is one of the key points of our work, which we will see reproduced in Chapter 4 when we discuss the Hardcore Lemma: global calibration adds a lot of power to a multiaccurate predictor. Here, it is allowing us to go from 0 correlation (as in the majorities example) to strong agnostic learning. Moreover, our proof very clearly delineates the roles played by multiaccuracy and calibration respectively. The ideas that we discuss in this section are intimately related to the proof showing that from a (weighted) multiaccurate predictor we can construct a hardcore set of optimal density, as we will see in Chapter 4.

As we pointed out in the introduction, a calibrated predictor is not always very informative: for example, if the labels $\mathbf{y}$ are balanced (i.e., half of them are 0 and half are 1), then the random-guessing predictor that predicts $1/2$ everywhere is calibrated. But whenever a calibrated predictor

predicts a value $v \in [0,1]$ bounded away from $1/2$, it is informative: we know that within the region $\mathcal{X}_v = \{x \in \mathcal{X} \mid p(x) = v\}$, the expected value of the labels on $\mathcal{X}_v$ is approximately v. That is, when it predicts $v \in [0,1]$, the expected value of the label is indeed v. In this case, a calibrated predictor is doing much better than random guessing. By how much? We will now see that we can quantify this through the notion of the *balance* parameter of the true labels on each of the level sets of p.

For the purposes of this explanation, we assume that our predictor p is perfectly calibrated on each of its level sets $p(x) = v$. As we discussed in the introduction, in the actual proof we need to relax this in two ways so that we can actually attain this notion in practice: (1) we relax the calibration requirement to hold on average across the level sets of p (i.e., where each level set $\mathcal{X}_v$ is weighted according to its size as defined by $\mathcal{D}$; that is, according to $\Pr_{\mathcal{D}}[x \in \mathcal{X}_v]$), and (2) we ask for τ-calibration, rather than for perfect calibration.

In this "perfect calibration" setting, we can split the domain $\mathcal{X}$ into disjoint pieces $\mathcal{X}_v$, which correspond to the level sets of the calibrated predictor p. This yields a partition $\mathcal{P}$ of the domain $\mathcal{X}$. By calibration, within every piece $P = \mathcal{X}_v \in \mathcal{P}$, the expected value of the labels is exactly v. Following the notation in [CDV24] and in Definition 2.20, we sometimes denote this value by v_P, to make it clear that we mean the expected value of the labels over the piece $P \in \mathcal{P}$ (but because we defined the partition precisely by taking the level sets of p, these two views are exactly equivalent, given that $P = \mathcal{X}_v$). Moreover, we let $b_P = \min\{v_P, 1 - v_P\}$, so that $b_P \leq 1/2$.

That is, as we represent with an example in Figure 3.3, on the level set $P = \mathcal{X}_v$, we know that a v-fraction of the points in $\mathcal{X}_v$ are equal to 1, and a $(1-v)$-fraction of the points in $\mathcal{X}_v$ are equal to 0, whereas we are always predicting v. But this means that calibration allows to say a lot about how well p is doing; i.e., we can exactly compute its error. Indeed, when conditioning on the level set $P = \mathcal{X}_v$, in the case where $v \leq 1/2$ we have that

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{D}}\left[|\mathbf{y} - p(\mathbf{x})| | \mathbf{x} \in \mathcal{X}_v\right] &= |1 - v| \cdot \Pr_{\mathcal{D}}[\mathbf{y} = 1 | \mathbf{x} \in \mathcal{X}_v] + |0 - v| \cdot \Pr_{\mathcal{D}}[\mathbf{y} = 0 | \mathbf{x} \in \mathcal{X}_v] \\
&= (1 - v)v + v(1 - v) \\
&= 2v(1 - v).
\end{aligned}$$

Symmetrically, the same expression follows for the case where $v > 1/2$. Moreover, note that

$$\min\{v, 1 - v\} \leq 2v(1 - v) \leq 2\min\{v, 1 - v\}, \tag{3.6}$$

and so, equivalently,

$$b \leq 2v(1 - v) \leq 2b, \tag{3.7}$$

implying that

$$\mathop{\mathbb{E}}_{\mathcal{D}}\left[|\mathbf{y} - p(\mathbf{x})| \mid \mathbf{x} \in \mathcal{X}_v\right] \leq 2b.$$

In other words, the fact that p is calibrated ensures that, on every level set $P = \mathcal{X}_v$, the error of the predictor is upper-bounded by $2b_P$, where b_P corresponds to the "balance" of the labels on that piece P (indeed, $b_P \approx 1/2$ if the true function behaves randomly on P, whereas $b_P \approx 0$ indicates that the the true function is essentially the constant $\mathbf{0}$ or the constant $\mathbf{1}$ function on that piece). This is why earlier we said that the balance b_P measures how close we are to the random guessing predictor; indeed, given that $b_P = \min\{v_P, 1 - v_P\}$, we can re-write $|v_P - 1/2|$ as $1/2 - b_P$.

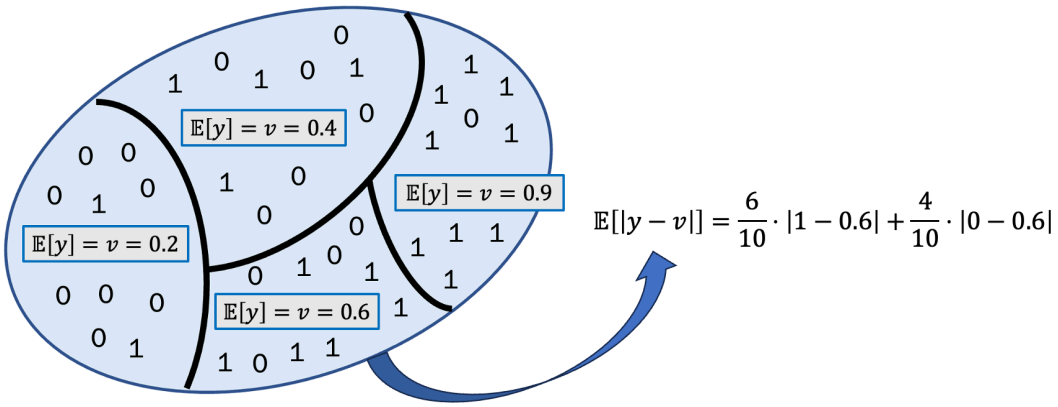


Figure 3.3: Visual representation of how calibration allows us to argue about the error of our predictor on its level sets.

By flipping this argument from error to calibration and swapping the ranges of the predictors as usual, we can equivalently obtain a lower bound on the correlation between p and the labels $\mathbf{y}$ on each level set $\mathcal{X}_v$. Then, we can compute the total correlation between p and the labels (i.e., over the entire domain $\mathcal{X}$) by expressing the total correlation as a weighted average of the correlation between p and the labels achieved at every piece (i.e., level set of p), and use the calibration condition to lower bound the correlation present on each piece.

We do so in the proof of the following lemma, which illustrates the typical way of using calibration: first express the quantity of interest as a weighted average over the level sets of p (this is simply using conditional expectation), and then apply the calibration guarantee on every level set, effectively swapping $\mathbf{y}$ by $p(\mathbf{x}) = v$ on each level set. In the following lemma we return to the usual relaxed (and achievable) notion of calibration, rather than assuming perfect calibration.

Lemma 3.8. *Let p be τ-calibrated for some distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$. Then*

$$\begin{aligned}\mathrm{cor}_{\mathcal{D}}(\mathbf{y}, \mathrm{sign}(2p(\mathbf{x}) - 1)) &\geq 2\,\mathbb{E}\left[\left|p(\mathbf{x}) - \frac{1}{2}\right|\right] - 2\tau \\ &= 2\,\mathbb{E}\left[1/2 - b_P\right] - 2\tau.\end{aligned}$$

Proof. We can write

$$\begin{aligned}\mathrm{cor}_{\mathcal{D}}(\mathbf{y}, \mathrm{sign}(2p(\mathbf{x}) - 1)) &= \mathop{\mathbb{E}}_{\mathcal{D}}[\mathrm{sign}(2p(\mathbf{x}) - 1)(2\mathbf{y} - 1)] \\ &= \mathop{\mathbb{E}}_{\mathcal{D}}[\mathrm{sign}(2p(\mathbf{x}) - 1)(2\,\mathbb{E}[\mathbf{y}|p(\mathbf{x})] - 1)] \\ &= \mathop{\mathbb{E}}_{\mathcal{D}}[\mathrm{sign}(2p(\mathbf{x}) - 1)(2p(\mathbf{x}) - 1)] + 2\mathop{\mathbb{E}}_{\mathcal{D}}[\mathrm{sign}(2p(x) - 1)(\mathbb{E}[\mathbf{y}|p(\mathbf{x})] - p(\mathbf{x}))] \\ &\geq \mathop{\mathbb{E}}_{\mathcal{D}}[|2p(\mathbf{x}) - 1|] - 2\,\mathbb{E}[|\,\mathbb{E}[\mathbf{y}|p(\mathbf{x})] - p(\mathbf{x})|] \\ &\geq 2\mathop{\mathbb{E}}_{\mathcal{D}}\left[\left|p(\mathbf{x}) - \frac{1}{2}\right|\right] - 2\tau \\ &= 2\mathop{\mathbb{E}}_{P \sim \mathcal{P}(\mathcal{D}_{\mathcal{X}})}\left[1/2 - b_P\right] - 2\tau.\end{aligned}$$

The last inequality follows from the definition of τ-calibration. □

This alone does not ensure that p gives us learning; to get learning, we need to achieve positive correlation with the labels. So far, we have only lower-bounded the correlation through the quantity $|p - 1/2|$; i.e., through how far p is to a random guessing predictor. Here is where multiaccuracy comes into the picture. The key observation is that if $\mathcal{C}$ contains a hypothesis that correlates well with the labels, and if p is $(\mathcal{C}, \tau)$-multiaccurate, then p cannot always predict values close to $1/2$. That is, it forces the global balance parameter $\mathbb{E}[b_P]$ to be bounded away from $1/2$. In other words, for a general set of labels, a calibrated p can in fact be very close to a random predictor. E.g., in the case where the labels $\mathbf{y}$ themselves are truly random. But if some hypothesis $c \in \mathcal{C}$ is correlated with the labels, then this precisely provides a certificate that $\mathbf{y}|\mathbf{x}$ is in fact *not* uniformly random. This is because a $\mathcal{C}$-multiaccurate predictor has to capture the correlations with c accurately, which forces it to deviate from random guessing. Formally:

Lemma 3.9. *If p is $(\mathcal{C}, \tau)$-multiaccurate for the distribution $\mathcal{D}$, then for every $c \in \mathcal{C}$,*

$$\mathop{\mathbb{E}}_{\mathcal{D}}\left[\left|p(\mathbf{x}) - \frac{1}{2}\right|\right] \geq \frac{\mathrm{cor}_{\mathcal{D}}(\mathbf{y}, c(\mathbf{x}))}{2} - \tau.$$

Proof. Fix any $c \in C$. Since p is $(\mathcal{C}, \tau)$-multiaccurate, by definition we have that

$$\mathop{\mathbb{E}}_{\mathcal{D}}[c(\mathbf{x})(\mathbf{y} - p(\mathbf{x}))] \leq \tau. \tag{3.10}$$

We can write the correlation between $\mathbf{y}$ and c equivalently as:

$$\mathop{\mathbb{E}}_{\mathcal{D}}\left[c(\mathbf{x})\left(\mathbf{y} - \frac{1}{2}\right)\right] = \frac{\mathrm{cor}_{\mathcal{D}}(\mathbf{y}, c(\mathbf{x}))}{2}. \tag{3.11}$$

Subtracting Equation (3.10) from (3.11) gives

$$\mathop{\mathbb{E}}_{\mathcal{D}}\left[c(\mathbf{x})\left(p(\mathbf{x}) - \frac{1}{2}\right)\right] \geq \frac{\mathrm{cor}_{\mathcal{D}}(\mathbf{y}, c(\mathbf{x}))}{2} - \tau.$$

The claim now follows by Hölder's inequality and the fact that $|c(\mathbf{x})| \leq 1$ for all $c \in \mathcal{C}$:

$$\mathop{\mathbb{E}}_{\mathcal{D}}\left[c(\mathbf{x})\left(p(\mathbf{x}) - \frac{1}{2}\right)\right] \leq \max_{\mathbf{x}} |c(\mathbf{x})| \cdot \mathbb{E}\left[\left|p(\mathbf{x}) - \frac{1}{2}\right|\right] \leq \mathbb{E}\left[\left|p(\mathbf{x}) - \frac{1}{2}\right|\right],$$

□

We can now put together Lemma 3.8 and Lemma 3.9 to obtain our Theorem 3.12 below, showing that if p is $(\mathcal{C}, \tau)$-multiaccurate and calibrated, then $\mathrm{sign}(2p - 1)$ is a strong agnostic learner.

Theorem 3.12. *Let $\mathcal{C}$ be a hypothesis class and $\tau > 0$. Suppose p is a predictor that is $(\mathcal{C}, \tau)$-multiaccurate and τ-calibrated with respect to $\mathcal{D}$. Then,* $\text{sign}(2p-1)$ *satisfies,*

$$\text{cor}(\mathbf{y}, \text{sign}(2p(\mathbf{x}) - 1))] \geq \max_{c \in \mathcal{C}} \text{cor}(\mathbf{y}, c(\mathbf{x})) - 4\tau.$$

Proof. Putting together our two lemmas, we get that:

$$\begin{aligned}
\text{cor}(\mathbf{y}, \text{sign}(2p(\mathbf{x}) - 1)) &\geq 2\,\mathbb{E}\left[\left|p(\mathbf{x}) - \frac{1}{2}\right|\right] - 2\tau && p \text{ is } \tau\text{-calibrated (Lemma 3.8)} \\
&\geq 2 \cdot \left(\frac{\text{cor}_{\mathcal{D}}(\mathbf{y}, c(\mathbf{x}))}{2} - \tau\right) - 2\tau && p \text{ is } (\mathcal{C}, \tau)\text{-multiaccurate (Lemma 3.9)} \\
&\geq \text{cor}_{\mathcal{D}}(\mathbf{y}, c(\mathbf{x})) - 4\tau.
\end{aligned}$$

□

4

Impagliazzo's Hardcore Lemma

Consider a decision problem that cannot be $1-\delta$ approximated by circuits of a given size in the sense that any such circuit fails to give the correct answer on at least a δ fraction of instances. We show that for any such problem there is a specific "hard-core" set of inputs which is at least a δ fraction of all inputs and on which no circuit of a slightly smaller size can get even a small advantage over a random guess.

Russell Impagliazzo [Imp95]

Having explored the relationships between the multigroup fairness framework and agnostic learning in Chapter 3, here we continue to explore the fascinating picture relating multigroup fairness notions, learning primitives, and hardcore set constructions. In this chapter, we explain Impagliazzo's Hardcore Lemma and explore how it relates to fairness and learning.

4.1 The original IHCL statement

Impagliazzo's Hardcore Lemma (IHCL) is a fundamental result in complexity theory which studies the average-case hardness of functions, as first shown by Russell Impagliazzo in 1995 [Imp95]. It studies how the hardness of prediction of any Boolean function g that is somewhat hard to predict on average is "spread" among the domain $\mathcal{X}$. Specifically, it turns out that this mild hardness on average implies that there exists a *fixed* hardcore set H in the domain where the input function is maximally unpredictable, in the sense that no distinguisher can do better than random guessing (this is why it is called a "hardcore" set – inside it, g appears maximally hard to the distinguishers). That is, g is very hard to compute inside of the hardcore set H. Indeed, this can be viewed as a form of hardness amplification, given that we go from mild average hardness to strong hardness. The hardness of prediction of the input function is always measured with respect to a base class $\mathcal{C}$ of distinguishers. As we have seen, this is one of the key recurring themes of this thesis, where we measure hardness of prediction, learning of the labels, and unbiased-ness all with respect to a base class $\mathcal{C}$. This is the only chapter in Part I of this thesis where the functions $c \in \mathcal{C}$ take values in $\{0,1\}$ rather than in $[-1,1]$, as it is done in the Hardcore Lemma literature.

Moreover, in the IHCL statement we also show a lower bound on the size of the hardcore set (otherwise, having a very small hardcore set would not be interesting), which is referred to as the *density* of the hardcore set. We would thus like the hardcore set to be as dense as possible, so that we extract all of the possible strong hardness from the input function g. We will carefully analyze the density of the set in our results. While it is visually more intuitive to think of IHCL in terms of sets, all of the statements and proofs deal with hardcore *measures* instead. A measure on $\mathcal{X}$ is simply a function $\mu : \mathcal{X} \to [0,1]$ that is not identically 0. We can turn into a probability distribution $\bar{\mu}$ by re-weighting μ by $\mathcal{D}_{\mathcal{X}}$:

$$\bar{\mu}(\mathbf{x}) = \frac{\Pr_{\mathcal{D}_{\mathcal{X}}}[\mathbf{x} = x]\mu(\mathbf{x})}{\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}}[\mu(\mathbf{x})]}.$$

What does hardcore-ness mean in the context of measures? It means that g is maximally unpredictable (in the same sense as we have discussed) when we *sample* according to μ, rather than according to $\mathcal{D}$. We can likewise translate the notion of "density" from sets to measures. Any IHCL statement using hardcore measures implies an IHCL statement using hardcore sets instead, by using a probabilistic method argument. Essentially, we perform the following randomized procedure: for each $x \in \mathcal{X}$, we include in the hardcore set with probability $\bar{\mu}$. Then, if the initial measure is dense enough, it follows that the hardcore set is dense as well (i.e., we have added to it enough points $x \in \mathcal{X}$). For details on this transformation, see [KS03, §4.4] and [CDV24, §B].

We now make all of these notions precise and formally state IHCL. In the case of IHCL, the concepts in $\mathcal{C}$ are all Boolean and should be thought of as bounded-size circuits. We start by formalizing what we mean by the *density* of a measure:

Definition 4.1 (Density of a measure)**.** *Given a measure $\mu : \mathcal{X} \to [0,1]$ and a base distribution $\mathcal{D}_{\mathcal{X}}$ on $\mathcal{X}$, the* density *of μ in $\mathcal{D}_{\mathcal{X}}$ is given by*

$$\mathsf{dns}(\mu) = \min_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}}[\mu(\mathbf{x})].$$

We drop $\mathcal{D}_{\mathcal{X}}$ from the $\mathsf{dns}(\mu)$ notation for convenience, but the density of a measure is always defined with respect to a base distribution. We say that μ is δ-*dense* if $\mathsf{dns}(\mu) = \delta$. Sometimes, as in [TTV09, §5], one will instead find a point-wise definition of density of a distribution. However, the two definitions are equivalent, given that

$$\max_{x\in\mathcal{X}} \frac{\bar{\mu}(x)}{\Pr_{\mathcal{D}_{\mathcal{X}}}[\mathbf{x} = x]} = \max_{x\in\mathcal{X}} \frac{\mu(x)}{\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}}[\mu(\mathbf{x})]} = \frac{\max_x \mu(x)}{\mathsf{dns}(\mu)} \leq \frac{1}{\mathsf{dns}(\mu)}. \tag{4.2}$$

This is why some definitions say that $\bar{\mu}$ is δ-dense in $\mathcal{D}_{\mathcal{X}}$ if $\bar{\mu}(x) \leq \frac{1}{\delta} \cdot \Pr_{\mathcal{D}_{\mathcal{X}}}[\mathbf{x} = x]$ for every $x \in \mathcal{X}$. We obtain this definition from re-arranging Definition 4.2: indeed, from the inequality

$$\max_{x\in\mathcal{X}} \frac{\bar{\mu}(x)}{\Pr_{\mathcal{D}_{\mathcal{X}}}[\mathbf{x} = x]} \leq \frac{1}{\mathsf{dns}(\mu)}.$$

Hence, we can see the notion of the density of a distribution as measuring its min-entropy, and that this latter definition of density is equivalent to Definition 4.1. Now we know what it means for a measure to be δ-dense. What does it mean for it to be "hardcore"? For that, we need to define the *hardness of prediction* with respect to a distinguisher class $\mathcal{C}$:

Definition 4.3 (Hardness of a function)**.** *Given a class $\mathcal{C} = \{c : \mathcal{X} \to \{0,1\}\}$, a function $g : \mathcal{X} \to \{0,1\}$, a distribution $\mathcal{D}_{\mathcal{X}}$ on $\mathcal{X}$, and $\delta > 0$, we say that g is* $(\mathcal{C}, \delta)$-hard *on $\mathcal{D}_{\mathcal{X}}$ if for all $c \in \mathcal{C}$,*

$$\Pr_{\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}}[c(\mathbf{x}) = g(\mathbf{x})] \leq 1 - \delta.$$

That is, each distinguisher $c \in \mathcal{C}$ errs when trying to compute g on at least a δ-fraction of the domain. Naturally, this δ-fraction subset of the domain can be different for each distinguisher. We are only asking that no distinguisher is able to fully compute g on $\mathcal{X}$. Moreover, note that the maximal possible hardness occurs when $\delta = 1/2 - \epsilon$, given that being $(\mathcal{C}, 1/2 - \epsilon)$-hard corresponds to stating that

$$\Pr_{\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}}[c(\mathbf{x}) = g(\mathbf{x})] \leq 1/2 + \epsilon.$$

That is, no distinguisher in $\mathcal{C}$ can guess g noticeably better than a random bit. This is why we sometimes refer to being $(\mathcal{C}, 1/2 - \epsilon)$-hard as being ϵ-*strongly hard*, whereas being $(\mathcal{C}, \delta)$-hard can be referred to as being δ-*weakly hard*. We say that a measure μ is *hardcore* for g if it induces a hardcore distribution $\bar{\mu}$ on which g is ϵ-strongly hard to predict. Recall also the notion of *relative complexity* that we introduced in Definition 2.15 in Chapter 2, where the class $\mathcal{C}_{t,q}$ denotes the class of functions that have complexity at most (t, q) relative to $\mathcal{C}$.

We can now formally state IHCL:

Theorem 4.4 (IHCL, [Imp95; Hol05])**.** *Let $\mathcal{C} = \{c : \mathcal{X} \to \{0,1\}\}$ be a family of functions, let $\mathcal{D}_{\mathcal{X}}$ be a probability distribution over $\mathcal{X}$, and let $\epsilon, \delta > 0$. There exist $t = \mathrm{poly}(\log|\mathcal{X}|, 1/\epsilon, 1/\delta)$ and $q = \mathrm{poly}(1/\epsilon, 1/\delta)$ such that the following holds: If $g : \mathcal{X} \to \{0,1\}$ is $(\mathcal{C}_{t,q}, \delta)$-hard on $\mathcal{D}_{\mathcal{X}}$, then there is a measure μ satisfying:*

- *Hardness: g is $(\mathcal{C}, 1/2 - \epsilon)$-hard on $\bar{\mu}$.*
- *Optimal density:* $\mathsf{dns}(\mu) = 2\delta$.

Note that there is a circuit size difference between the assumption and the conclusion of the theorem: the average mild hardness of g is with respect to $C_{t,q}$, whereas the hardcore strong hardness is with respect to $\mathcal{C}$. This circuit size difference makes a lot of sense if we think of IHCL as the "counterpart" of boosting: as we explained in Chapter 1, the contrapositive of IHCL is effectively a boosting algorithm, where we go from computing g slightly better than random ($> 1/2 + \epsilon$) to computing g almost perfectly ($> 1 - \delta$). Indeed, one of the two original proofs of IHCL by Impagliazzo is a boosting-based proof, which we will return to in Section 4.8.[1] It consists of an iterative procedure where at every step we use the assumption in the contrapositive of IHCL that g is weakly learnable on a new distribution, ensembling the $O(1/(\epsilon^2\delta^2))$ weak learners through a majority vote at the end [Imp95, Thm. 1]. The ensembling of these various learners is precisely what causes the circuit-size parameter loss that we see in the statement of IHCL. While this degradation in the circuit-size parameters seems natural in any boosting-based proof IHCL, Blanc, Koch, Strassle and Tan recently showed that this difference in circuit size is unavoidable, regardless of whether the IHCL proof is based on boosting or not [BKST24].

[1] The other proof (due to Nisan) uses the min-max theorem, similar to how Trevisan, Tulsiani, and Vadhan also prove the Regularity Lemma with a proof through boosting and a proof through the min-max theorem [TTV09].

Another important remark is that the two original proofs by Impagliazzo obtain a hardcore measure of density δ, which is suboptimal. Indeed, note that the optimal density is 2δ: if there exists a hardcore distribution for g of density ρ, then g is $\big(\rho(1/2-\epsilon)\big)$-weakly hard on average on $\mathcal{D}$ with respect to $\mathcal{C}$. More intriguingly for us, as we explain in the next section (Section 4.2) Trevisan, Tulsiani, and Vadhan provided a proof of IHCL through the Regularity Lemma (which we view as the multiaccuracy theorem, as we explained in Section 2.3), but they also only obtained the suboptimal δ density parameter. However, if we go higher in the tiers of multigroup fairness notions, [CDV24] show that we can obtain a stronger and more general version of the hardcore lemma (which they call IHCL++) from multicalibration. In turn, they show that IHCL++ implies IHCL with optimal density 2δ. In other words, while multiaccuracy (through the lenses of the Regularity Lemma by [TTV09]) gets suboptimal δ-density, multicalibration is able to recover the optimal 2δ density parameter. However, because they go through the more expensive notion of multicalibration, this incurs a much bigger loss in circuit size; namely, by the multicalibration theorem (Theorem 2.16), they get $q = O(1/(\epsilon^{12}\delta^6))$. Through multiaccuracy, [TTV09] instead incur a circuit loss size of $q = O(1/(\epsilon^2\delta^2))$, as in Impagliazzo's boosting proof.

This is precisely what motivates us to study whether obtaining suboptimal density is inherent to multiaccuracy, or whether we can obtain an optimal hardcore measure from it. Perhaps surprisingly, as it happened in Chapter 3, global calibration will come to the rescue. Indeed, in the learning setting, adding calibration to multiaccuracy allowed us to go from no learning (Section 3.1) to strong agnostic learning (Section 3.3). Here, we will show that adding calibration to (weighted) multiaccuracy allows us to go from suboptimal density to the optimal 2δ.

Beyond the original proofs by Impagliazzo [Imp95] and the proof through the Regularity Lemma [TTV09], there are many other versions of the Hardcore Lemma. Ten years after Impagliazzo's original paper, Holenstein gave a boosting proof of the hardcore lemma that achieved the optimal 2δ density parameter for the first time. While the proof is also boosting-based, the circuits that weak learn g are combined differently than in the case of Impagliazzo [Hol05]. A few years later, Barak, Hardt, and Kale gave an alternative construction of an optimal hardcore set using a generalized multiplicative update rule combined with a natural notion of approximate Bregman projection [BHK09].

4.2 From multigroup fair predictors to hardcore measures

Given that we are studying the relationships between multigroup fairness notions, learning primitives, and hardcore set constructions, the proof of IHCL that we are most interested in is the one shown by Trevisan, Tulsiani, and Vadhan through the Regularity Lemma. We begin by explaining how we can show IHCL with density δ from a multiaccurate predictor, following [TTV09].

Given a distribution $\mathcal{D}_{\mathcal{X}}$ on $\mathcal{X}$ and $g : \mathcal{X} \to \{0,1\}$, define the distribution $\mathcal{D}^g$ on $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \{0,1\}$ where $\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}$ and $\mathbf{y} = g(\mathbf{x})$. We start from a $(\mathcal{C}, \epsilon\delta)$-multiaccurate predictor p for $\mathcal{D}^g$. With this multiaccurate predictor at hand, [TTV09] construct the following measure:

$$\mu_{\mathsf{TTV}}(x) = |g(x) - p(x)|.$$

Note that, unlike boosting-based approaches to IHCL, here we are defining the hardcore measure directly and explicitly, rather than through an iterative process (in a sense, the boosting has already

occurred when calling the Regularity Lemma/multiaccuracy theorem).

Trevisan, Tulsiani, and Vadhan then show that this measure possesses the two properties that we require for IHCL: (1) hardness, and (2) density. Crucial in the future insight by [CDV24] is the fact that we can study these properties separately:

1. *Hardness:* In all of [TTV09] (multiaccuracy), [CDV24] (multicalibration) and in our result that we present in this section (weighted multiaccuracy), we show the hardcore-ness of the measure from the indistinguishability endowed by the multigroup fairness definition.

2. *Density:* here, [TTV09] (multiaccuracy) get δ directly from the average-case of the input function g, whereas we use global calibration to get 2δ, and [CDV24] uses multicalibration.

This is the intuition for why we will be able to get a hardcore measure from (weighted) multiaccuracy and global calibration: in essence, we can use the [TTV09] proof for the indistinguishability property (for which we do not need the stronger multicalibration indistinguishability properties, given that we are still trying to prove IHCL, not IHCL++), and the [CDV24] proof for the density property. For the latter, our observation is that their density proof when showing IHCL from IHCL++ didn't need the full power of multicalibration; only global calibration is needed. Indeed, we only need to be able to argue about the error of the predictor on each of its level sets, similar to the analysis we performed in Section 3.3 when showing that adding calibration to multiaccuracy gives us strong agnostic learning.

IHCL from multiaccuracy. In the case of μ_{TTV}, [TTV09] show its hardcoreness and density properties as follows.

(1) Hardness. Because p is a $(\mathcal{C}, \epsilon\delta)$-multiaccurate predictor for $\mathcal{D}^g$, and multiaccuracy offers strong indistinguishability guarantees (i.e., p is $(\mathcal{C}, \epsilon\delta)$-indistinguishable from g), we want to argue that the $\mathcal{C}$-indistinguishability of p from g "transfers" to μ_{TTV}, endowing the measure with hardness. Then, we want to relate the multiaccuracy guarantee on p with the hardness of prediction of g. This is precisely what [TTV09] do, proceeding in two steps. First, they use the identity:

$$|g(x) - p(x)| \cdot \mathbb{1}_{[c(x)=g(x)]} = \left[\left(c(x) - \frac{1}{2}\right) \cdot (g(x) - p(x)) + \frac{1}{2} \cdot |g(x) - p(x)|\right]. \tag{4.5}$$

Then, we can take the expectation on both sides, so that we can use the $(\mathcal{C}, \epsilon\delta)$-multiaccuracy guarantee of p on the RHS. Indeed:

$$\mathop{\mathbb{E}}_{\mathcal{D}_{\mathcal{X}}}[|g(\mathbf{x}) - p(\mathbf{x})| \cdot \mathbb{1}_{[c(\mathbf{x})=g(\mathbf{x})]}] \leq \epsilon\delta + \frac{1}{2} \mathop{\mathbb{E}}_{\mathcal{D}_{\mathcal{X}}}[|g(\mathbf{x}) - p(\mathbf{x})|]. \tag{4.6}$$

Note that the term $\mathbb{E}_{\mathcal{D}_{\mathcal{X}}}[|g(x) - p(x)|]$ corresponds precisely to the density of the measure μ_{TTV} (Definition 4.1); i.e., is equal to $\mathsf{dns}(\mu_{\mathsf{TTV}})$.

We can now relate Equation 4.6 to the hardness of prediction of g when sampling according to the distribution $\bar{\mu}_{\mathsf{TTV}} = |g(x) - p(x)| / \sum_{y \in \mathcal{X}} |g(y) - p(y)|$, which is the normalized version of the measure μ_{TTV}, obtaining that, for every $c \in \mathcal{C}$,

$$\Pr_{\mathbf{x} \sim \bar{\mu}_{\mathsf{TTV}}}[c(\mathbf{x}) = g(\mathbf{x})] = \frac{\mathbb{E}_{\mathcal{D}_{\mathcal{X}}}[|g(\mathbf{x}) - p(\mathbf{x})| \cdot \mathbb{1}_{[c(\mathbf{x})=g(\mathbf{x})]}]}{\mathbb{E}_{\mathcal{D}_{\mathcal{X}}}[|g(\mathbf{x}) - p(\mathbf{x})|]}. \tag{4.7}$$

Then, by plugging Equation 4.6 into Equation 4.7, we get that

$$\Pr_{\mathbf{x}\sim\bar{\mu}_{\mathsf{TTV}}}[c(\mathbf{x}) = g(\mathbf{x})] \leq \frac{\epsilon\delta + \frac{1}{2}\,\mathbb{E}_{\mathcal{D}_{\mathcal{X}}}[|g(\mathbf{x}) - p(\mathbf{x})|]}{\mathbb{E}_{\mathcal{D}_{\mathcal{X}}}[|g(\mathbf{x}) - p(\mathbf{x})|]} = \frac{\epsilon\delta + \mathsf{dns}(\mu_{\mathsf{TTV}})/2}{\mathsf{dns}(\mu_{\mathsf{TTV}})} = \frac{1}{2} + \frac{\epsilon\delta}{\mathsf{dns}(\mu_{\mathsf{TTV}})}. \tag{4.8}$$

Hence, we have related the multiaccuracy error of p (namely, $\epsilon\delta$), with the hardness of prediction of g with respect to $\mathcal{C}$ and μ_{TTV}. Moreover, the density of the measure appears in this relationship, as we can see in Equation 4.8. To conclude, we need to know what this density is.

(2) Density. The argument for why μ_{TTV} is δ-dense follows from the fact that g is assumed to be $(\mathcal{C}_{t,q}, \delta)$-hard on $\mathcal{D}_{\mathcal{X}}$. The key observation is that $\mathcal{C}_{t,q}$ is expressive enough to compute $\mathbf{1}[p(x) \geq \psi]$ for $\psi \in [0,1]$. This is because the (multiaccurate) predictor p is discretized to precision $\Theta(\epsilon)$, and by the Regularity Lemma/Theorem 2.16, p can be written as a linear combination of $O(1/\epsilon^2)$ many hypotheses from $\mathcal{C}$ (i.e., $p \in \mathcal{C}_{t,q}$). Therefore, it is enough to compute the threshold ψ with precision $\Theta(\epsilon)$ as well, which ensures that the function $\mathbf{1}[p(x) \geq \psi]$ is also in $\mathcal{C}_{t,q}$. Then, we can view the quantity $|g(\mathbf{x}) - p(\mathbf{x})|$ is the error probability of the randomized function which outputs 1 with probability $p(\mathbf{x})$. By averaging, there exists a deterministic threshold $\psi \in [0,1]$ so that the deterministic function $\tilde{p}(x) = \mathbf{1}[p(x) \geq \psi]$ does as well. But since $\tilde{p} \in \mathcal{C}_{t,q}$, this implies that $\mathsf{dns}_{\mathcal{D}}(\mu_{\mathsf{TTV}}) \geq \delta$. Otherwise, we would be contradicting the δ-hardness of g.

Coming back to Equation 4.8, we can now use the fact that $\mathsf{dns}_{\mathcal{D}}(\mu_{\mathsf{TTV}}) \geq \delta$ to conclude that

$$\Pr_{\mathbf{x}\sim\bar{\mu}_{\mathsf{TTV}}}[c(\mathbf{x}) = g(\mathbf{x})] \leq \frac{1}{2} + \frac{\epsilon\delta}{\delta} = \frac{1}{2} + \epsilon,$$

thus proving that g is $(\mathcal{C}, 1/2 - \epsilon)$-hard on $\bar{\mu}$, as desired. From this expression it is very clear why we had to call the multiaccuracy error $\epsilon\delta$ (so that we obtain ϵ-strong hardness), and thus why this approach obtains a Hardcore Lemma with q parameter $q = O(1/(\epsilon^2\delta^2))$.

4.3 Improving the TTV construction

The key in the [TTV09] construction is the mapping from a multiaccurate predictor p to a (hardcore) measure $\mu_{\mathsf{TTV}}(x) = |g(x) - p(x)|$. Especially given that this measure has density δ rather than the optimal 2δ, it is natural to ask: Is this the optimal way in which we could have mapped p to a (hardcore) measure μ? By optimal we mean the one that would produce the densest measure.

To study this, let's understand more carefully what the measure μ_{TTV} is doing. Essentially, it is an error function, computing the difference between g and p. Consider the partition of the domain $\mathcal{X}$ into the level sets of p; that is, we let $\mathcal{X}_v = \{x \in \mathcal{X} \mid p(x) = v\}$ for each v in the range of p. Then, we can write the global error function $\mathbb{E}[|g(x) - p(x)|]$ as a weighted sum of the error function on each level set $\mathcal{X}_v$ (simply by a conditional expectation). Once we have localized on a level set $\mathcal{X}_v$, we can understand μ_{TTV} as follows. Let $\mathcal{X}_v^1$ be the set of points in $\mathcal{X}_v$ that have g-value equal to 1, and symmetrically let $\mathcal{X}_v^0$ be the set of points in $\mathcal{X}_v$ that have g-value equal to 0. Then, what weight is the measure μ_{TTV} assigning to each of $\mathcal{X}_v^0$ and $\mathcal{X}_v^1$?

- The points in $\mathcal{X}_v^0$ get $\mu_{\mathsf{TTV}}(x) = |0 - v| = v$ weight.
- The points in $\mathcal{X}_v^1$ get $\mu_{\mathsf{TTV}}(x) = |1 - v| = 1 - v$ weight.

In order to continue the error analysis, let's suppose that p is calibrated. Then, we are able to compute $\mathbb{E}[\mu_{\mathsf{TTV}}]$ conditioned on $x \in \mathcal{X}$, given that calibration tells us that in $\mathcal{X}_v$, a v-fraction of the g-values are 1. Indeed:

$$\mathop{\mathbb{E}}_{\mathcal{D}_{\mathcal{X}}}[\mu_{\mathsf{TTV}}(\mathbf{x})|\mathbf{x} \in \mathcal{X}_v] = v \Pr_{\mathcal{D}_{\mathcal{X}}}[g(\mathbf{x}) = 0|\mathbf{x} \in \mathcal{X}_v] + (1-v) \Pr_{\mathcal{D}_{\mathcal{X}}}[g(\mathbf{x}) = 1|\mathbf{x} \in \mathcal{X}_v] = 2v(1-v). \tag{4.9}$$

This calculation should look familiar! It is exactly what we did in Section 3.3 when showing that multiaccuracy plus global calibration gives us agnostic learning. Indeed, in that case, we used global calibration to compute the error of a calibrated predictor by conditioning on its level sets and then arguing that, on each level set, the error of the predictor is upper-bounded by $2b$, where $b = \min\{v, 1-v\}$. We can see that μ_{TTV} corresponds precisely to the same error function. Together, these learning and theoretic perspectives help illuminate the power of calibration:

If a predictor p is calibrated, we can analyze its error by conditioning on its level sets and applying the calibration condition on each level set.

Indeed, without a calibration guarantee or any more information about the multiaccurate predictor p, we are not able to say anything else about its values, other than knowing that the MA condition is satisfied on each $c \in \mathcal{C}$. But with added calibration, we can now perform quite a fine-grained analysis of the error of p.

Having understood why global calibration gives us quite a lot of leverage, both in the learning context and in the context of hardcore measures, let's return to μ_{TTV}. From Equation 4.9 we can see that μ_{TTV} assigns equal weight $v(1-v)$ to both $\mathcal{X}_v^0$ and $\mathcal{X}_v^1$. Recall that a hardcore measure is meant to be such that g appears maximally hard to $\mathcal{C}$ when sampling according to the measure. If $v \leq 1/2$, then the points with g-value equal to 1 are in the minority group within $\mathcal{X}_v$, whereas the points with g-value equal to 0 are in the majority group within $\mathcal{X}_v$. The measure μ_{TTV} gives weight v to the minority elements in $\mathcal{X}_v^1$, and weight $1-v$ to the majority elements in $\mathcal{X}_v^0$ (this is in the case where $v \leq 1/2$; if $v > 1/2$, then the majority group is $\mathcal{X}_v^1$ and the minority group is $\mathcal{X}_v^0$). But, if we want our measure to make g appear as hard as possible to predict, then we should be giving much more weight to the minority elements in $\mathcal{X}_v^1$, so that it becomes harder to predict g on each $\mathcal{X}_v$.

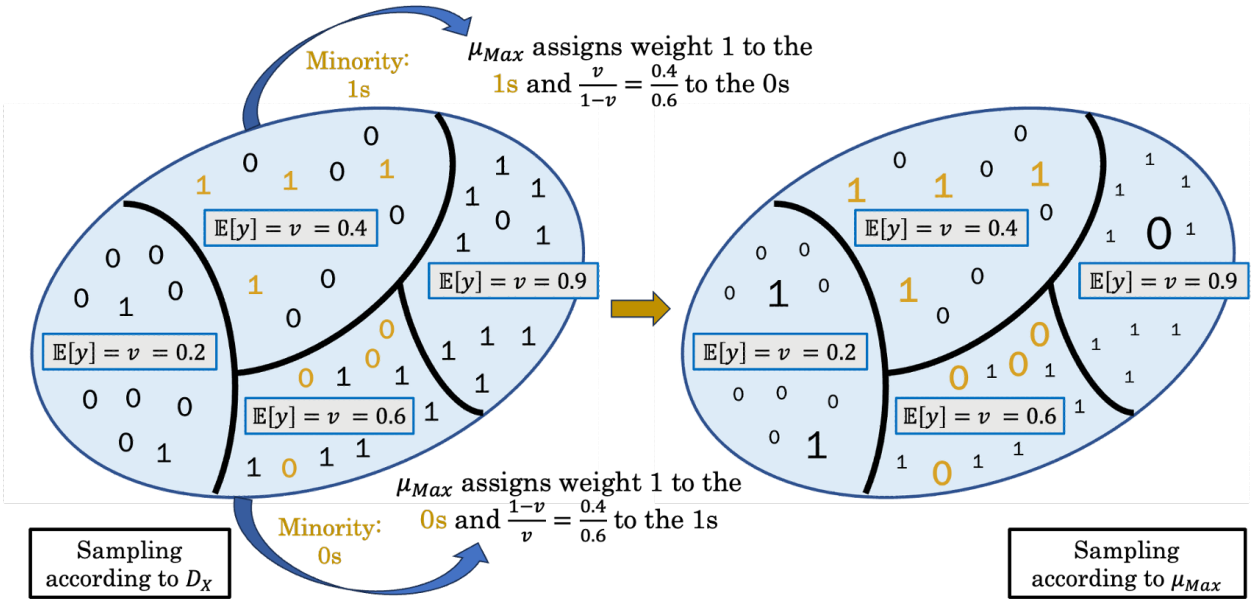


Figure 4.1: Example of how the measure μ_{Max} assigns weights on each level set $\mathcal{X}_v$.

The optimal way in which we can do this, given that a measure μ is constrained in taking values in $[0, 1]$ is the following:

$$\mu_{\mathsf{Max}}(x) = \begin{cases} 1 & \text{if } x \in \mathcal{X}_v^1 \text{ (minority group)} \\ \dfrac{v}{1-v} & \text{if } x \in \mathcal{X}_v^0 \text{ (majority group)} \end{cases}$$

In the case where $v > 1/2$, we do the opposite:

$$\mu_{\mathsf{Max}}(x) = \begin{cases} \dfrac{1-v}{v} & \text{if } x \in \mathcal{X}_v^1 \text{ (majority group)} \\ 1 & \text{if } x \in \mathcal{X}_v^0 \text{ (minority group)} \end{cases}$$

We represent μ_{Max} pictorially in Figure 4.1 using the same level set partition that we used in Figure 3.3 in Section 3.3 (when showing that multiaccuracy and calibration gives strong agnostic learning). Now, the total weight given to each of $\mathcal{X}_v^0$ and $\mathcal{X}_v^1$ is $\min\{v, 1-v\} = b$. Note that for both μ_{TTV} and μ_{Max}, the measures are balanced under calibration: on each level set $\mathcal{X}_v$, it assigns the same total weight to each of $\mathcal{X}_v^0$ and $\mathcal{X}_v^1$.

A natural way to guess g is to output the majority value on each level set $\mathcal{X}_v$; that is, we consider the majority predictor $\tilde{p} = \mathbf{1}[p \geq 1/2]$. Then, μ_{Max} assigns weight 1 to the points where $\tilde{p}$ is incorrect. For example, in Figure 4.1, the majority predictor would predict 0 on the level set $\mathcal{X}_{0.4}$. This would cause us to be wrong on all of the 1s, which is precisely the points where μ_{Max} is assigning weight 1 (i.e., the maximum possible). The rest of the weights are assigned so that S_v is balanced, assuming that p is in fact calibrated. Thus, this ensures that $\tilde{p}$ is incorrect with probability $1/2$, which is what we want in order to maximize hardness.

Note that if we put together the cases where $v \leq 1/2$ and $v > 1/2$, the way we described our

measure μ_{Max} by assigning weights 1 and $v/(1-v)$ or $(1-v)/v$ is equivalent to writing it as

$$\mu_{\mathsf{Max}}(x) = \frac{|g(x) - p(x)|}{\max\{p, 1-p\}}.$$

Compared to the μ_{TTV} measure, here we still use the error function $|g(x) - p(x)|$, weighted by the term $1/\max\{p, 1-p\}$. Indeed, μ_{Max} should be assigning more weight to the level sets with p-value closer to $1/2$, in order to maximize hardness (i.e., the level sets with balance parameter b_P close to $1/2$ are the ones on which g is harder to predict, so we want to give more mass to them).

Why does μ_{Max}, and not μ_{TTV}, achieve optimal density? As we formally analyze in Section 4.6, our measure μ_{Max} achieves optimal density, while μ_{TTV} does not. We explain why this is the case, assuming perfect calibration. First let's assume that $v \leq 1/2$. Our calibrated predictor p always predicts v inside $\mathcal{X}_v$ (by definition of a level set). As we already discussed, in $\mathcal{X}_v$, the measure μ_{TTV} assigns weight $|1-v| = 1-v$ to the points where g is 1, which occurs in a v-fraction of points in $\mathcal{X}_v$, and weight $|0-v| = v$ to the points where g is 0, which occurs in a $(1-v)$-fraction of points in $\mathcal{X}_v$. Therefore, as we computed in Equation 4.9,

$$\mathop{\mathbb{E}}_{\mathcal{D}_{\mathcal{X}}}[\mu_{\mathsf{TTV}}(\mathbf{x})|\mathbf{x} \in \mathcal{X}_v] = v|1-v| + (1-v)|0-v| = 2v(1-v).$$

In the symmetric case where $v > 1/2$ we likewise get $2v(1-v)$. Hence, by averaging over the level sets, we get that

$$\mathop{\mathbb{E}}_{\mathcal{D}_{\mathcal{X}}}[\mu_{\mathsf{TTV}}(\mathbf{x})] = 2 \mathop{\mathbb{E}}_{\mathcal{D}_{\mathcal{X}}}[p(\mathbf{x})(1-p(\mathbf{x}))] = 2 \mathop{\mathbb{E}}_{\mathcal{P}(\mathcal{D}_{\mathcal{X}})}[v_P(1-v_P)].$$

Meanwhile, since $\max\{v, 1-v\} = 1-v$ in the case where $v \leq 1/2$, the measure μ_{Max} assigns weight $|1-v|/(1-v) = 1$ to the points in $\mathcal{X}_v$ where g is 1, which occurs in a v-fraction of them, and weight $|0-v|/(1-v) = v/(1-v)$ to the points in $\mathcal{X}_v$ where g is 0, which occurs in a $(1-v)$-fraction of them. Therefore,

$$\mathop{\mathbb{E}}_{\mathcal{D}_{\mathcal{X}}}[\mu_{\mathsf{Max}}(\mathbf{x})|\mathbf{x} \in \mathcal{X}_v] = \frac{|1-v|}{1-v} \cdot v + \frac{|0-v|}{1-v} \cdot (1-v) = v + v = 2v.$$

In the case where $v > 1/2$, we have

$$\mathop{\mathbb{E}}_{\mathcal{D}_{\mathcal{X}}}[\mu_{\mathsf{Max}}(\mathbf{x})|\mathbf{x} \in \mathcal{X}_v] = \frac{|1-v|}{v} \cdot v + \frac{|0-v|}{v} \cdot (1-v) = 1 - v + 1 - v = 2(1-v).$$

Therefore, putting the two cases together, we get that the density of the measure conditioned on $\mathcal{X}_v$ is equal to $2\min\{v, 1-v\} = 2b$. Hence, by averaging over the level sets, we get that

$$\mathop{\mathbb{E}}_{\mathcal{D}_{\mathcal{X}}}[\mu_{\mathsf{Max}}(\mathbf{x})] = 2 \mathop{\mathbb{E}}_{\mathcal{D}_{\mathcal{X}}}[\min\{p(\mathbf{x}), 1-p(\mathbf{x})\}] = 2 \mathop{\mathbb{E}}_{\mathcal{P}(\mathcal{D}_{\mathcal{X}})}[\min\{v_P, 1-v_P\}] = 2 \mathop{\mathbb{E}}_{\mathcal{P}(\mathcal{D}_{\mathcal{X}})}[b_P].$$

As we have discussed, both μ_{TTV} and μ_{Max} represent error functions. In the case of μ_{TTV}, we can think of the measure as measuring the error of a randomized predictor that, on every level set $\mathcal{X}_v$, outputs 1 with probability v, and 0 with probability $1-v$. Then, the error of this predictor is

precisely $2v(1-v)$, as we have computed. On the other hand, we can think of μ_{Max} as measuring the error of the majority predictor $\tilde{p} = \mathbf{1}[p \geq 1/2]$ which, on every level sets $\mathcal{X}_v$, outputs 1 if $v \geq 1/2$, and 0 if $v < 1/2$. In this case, the error of the predictor is precisely $\min\{v, 1-v\}$, as we have computed.

Note that in all of the arguments in this section, we have only used the fact that p is calibrated; we haven't yet dealt with any indistinguishability or multiaccuracy properties. As in the case of [TTV09] and [CDV24], a key observation is that we can treat the indistinguishability and the density analysis separately. Summarizing, we have obtained that:

$$\mathop{\mathbb{E}}_{\mathcal{D}_\mathcal{X}}[\mu_{\mathsf{TTV}}] = 2 \mathop{\mathbb{E}}_{\mathcal{P}(\mathcal{D}_\mathcal{X})}[v_P(1-v_P)], \qquad \mathop{\mathbb{E}}_{\mathcal{D}_\mathcal{X}}[\mu_{\mathsf{Max}}] = 2 \mathop{\mathbb{E}}_{\mathcal{P}(\mathcal{D}_\mathcal{X})}[b_P]. \tag{4.10}$$

How do these relate? Recall the inequality that we used in Section 3.3 (Equation 3.6):

$$\min\{v_P, 1-v_P\} \leq 2v_P(1-v_P) \leq 2\min\{v_P, 1-v_P\}, \tag{4.11}$$

or, equivalently,

$$b_P \leq 2v_P(1-v_P) \leq 2b_P. \tag{4.12}$$

We can plot these three functions to see their relationship visually in Figure 4.2.

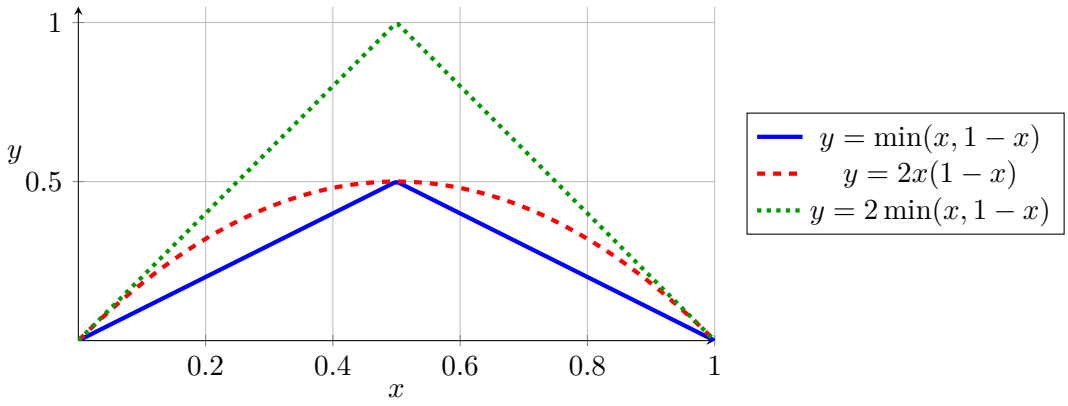


Figure 4.2: Visual representation of the chain of inequalities $b_P \leq 2v_P(1-v_P) \leq 2b_P$.

Bringing in the δ parameter. Now that we have computed the densities of μ_{TTV} and μ_{Max}, how does the δ parameter come in? We promised that μ_{Max} would get 2δ density. Recall that in IHCL we assume the input function g to be $(\mathcal{C}_{t,q}, \delta)$-hard to compute, and recall the majority predictor $\tilde{p} = 1(p \geq 1/2)$. The key observation here is that $\tilde{p} \in C_{t,q}$, and so $\tilde{p}$ must have error at least δ. Otherwise, it would contradict the δ-weakly hardness of g. And, as we have explained, the error of $\tilde{p}$ is precisely $\min\{v, 1-v\}$ on each level set $\mathcal{X}_v$. Therefore:

$$\tilde{p} \in C_{t,q} \implies \mathop{\mathbb{E}}_{\mathcal{D}_\mathcal{X}}[\min\{p(\mathbf{x}), 1-p(\mathbf{x})\}] \geq \delta \implies \mathop{\mathbb{E}}_{\mathcal{P}(\mathcal{D}_\mathcal{X})}[b_P] \geq \delta.$$

Plugging this into Equation 4.10 and using the inequality shown in Equation 4.12, it follows that

$$\mathop{\mathbb{E}}_{\mathcal{D}_{\mathcal{X}}}[\mu_{\mathsf{TTV}}] = 2 \mathop{\mathbb{E}}_{\mathcal{P}(\mathcal{D}_{\mathcal{X}})}[v_P(1-v_P)] \geq \mathop{\mathbb{E}}_{\mathcal{P}(\mathcal{D}_{\mathcal{X}})}[b_P] \geq \delta, \qquad \mathop{\mathbb{E}}_{\mathcal{D}_{\mathcal{X}}}[\mu_{\mathsf{Max}}] = 2 \mathop{\mathbb{E}}_{\mathcal{P}(\mathcal{D}_{\mathcal{X}})}[b_P] \geq 2\delta.$$

Note that the measure μ_{TTV} can get density better than δ in some cases (e.g., when p is far from $1/2$), given that we are using b_P as a lower bound to $2v_P(1-v_P)$.

This also establishes an interesting counter-point to our proof in Section 3.3 showing that a calibrated and $\mathcal{C}$-multiaccurate predictor yields strong agnostic learning. In that case, calibration also allowed us to analyze the error/correlation of the predictor through the balance parameter b_P, by analyzing the predictor on each of its level sets. Specifically:

The usefulness of calibration:

- *Section 3.3.* Calibration allowed us to lower-bound the total correlation of the predictor with the labels by $2\,\mathbb{E}[1/2 - b_P]$.
- *Here.* Calibration allows us to compute the density of μ_{Max}: it is equal to $2\,\mathbb{E}[b_P]$.

The average balance parameter $\mathbb{E}[b_P]$:

- *Section 3.3.* Moreover, if the predictor is $\mathcal{C}$-multiaccurate and if there is some concept in $\mathcal{C}$ that has correlation with the labels, then this pushes $\mathbb{E}[b_P]$ away from $1/2$ and closer to 0. Then, $2\,\mathbb{E}[1/2 - b_P]$ is bounded away from 0, and so we ensure that the predictor has positive correlation with the labels and is thus a learner.
- *Here.* Moreover, if the input function g is δ-weakly hard, then this pushes $\mathbb{E}[b_P]$ away from 0 and closer to $1/2$. Then, $2\,\mathbb{E}[b_P]$ is δ-bounded away from 0, and so we ensure that the measure μ_{Max} is 2δ-dense.

We conclude this section by giving a concrete example comparing the densities of μ_{TTV} and μ_{Max}, where it is clear to see how μ_{Max} is a denser measure.

We define a predictor $p : \mathcal{X} \to [0,1]$ that predicts only two values on $\mathcal{X}$: $p(x) = 1/2$ and $p(x) = \eta$ for some small value $0 < \eta < 1/2$. Let $\mathcal{X}_{1/2} = \{x : p(x) = 1/2\}$ and $\mathcal{X}_{\eta} = \{x : p(x) = \eta\}$ be the level sets of p; their sizes are such that $\Pr_{\mathcal{D}_{\mathcal{X}}}[\mathbf{x} \in \mathcal{X}_{1/2}] = 2\delta$ and $\Pr_{\mathcal{D}_{\mathcal{X}}}[\mathbf{x} \in \mathcal{X}_{1/2}] = 1 - 2\delta$, where $\delta > 0$ and $\eta \ll \delta$. Moreover, p is perfectly calibrated for $\mathcal{D}^g$, where g is the Boolean input function. Given these p, g, and $\mathcal{D}$, we compare the densities of the measures μ_{TTV} and μ_{Max}. Figure 4.3 summarizes the construction (drawn for example values $\eta = 0.05, \delta = 0.2$).

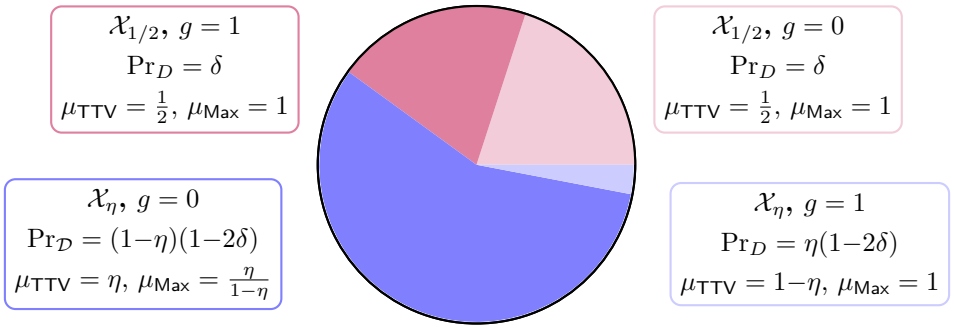


Figure 4.3: Weights assigned by μ_{TTV} and μ_{Max} on each region of $\mathcal{X}$. The two measures nearly agree on the *easy* blue region, but μ_{Max} is twice μ_{TTV} on the *hard* red region.

- μ_{TTV} : In $\mathcal{X}_{1/2}$, $p = 1 - p = 1/2$, and so μ_{TTV} assigns weight $1/2$ to all $\mathbf{x} \in \mathcal{X}_{1/2}$. By the calibration guarantee on p, within $\mathcal{X}_{1/2}$ half the points have g-value equal to 0 and half the points have g-value equal to 1. In $\mathcal{X}_\eta$, μ_{TTV} gives weight $p = \eta$ to the points where $g(x) = 0$ and weight $1 - p = 1 - \eta$ to the points where $g(x) = 1$. By the calibration guarantee on p, within $\mathcal{X}_\eta$ a $(1-\eta)$-fraction of points have g-value equal to 0, and a η-fraction of points have g-value equal to 1. Therefore,

$$\begin{aligned}
\mathsf{dns}(\mu_{\mathsf{TTV}}) &= \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} [\mu_{\mathsf{TTV}}(x)] \\
&= \frac{1}{2} \cdot 2\delta + \eta \cdot (1-\eta)(1-2\delta) + (1-\eta) \cdot \eta(1-2\delta) \\
&= \delta + 2\eta(1-\eta)(1-2\delta) \\
&= \delta + O(\eta).
\end{aligned}$$

- μ_{Max} : In $\mathcal{X}_{1/2}$, $p = 1 - p$, so μ_{Max} assigns weight 1 to all $\mathbf{x} \in \mathcal{X}_{1/2}$. In $\mathcal{X}_\eta$, μ_{Max} gives weight $\eta(1-\eta)$ to the points where $g(x) = 0$ weight 1 to the points where $g(x) = 1$. Again using the calibration guarantee on p within each of $\mathcal{X}_{1/2}$ and $\mathcal{X}_\eta$ it follows that

$$\begin{aligned}
\mathsf{dns}(\mu_{\mathsf{Max}}) &= \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} [\mu_{\mathsf{Max}}(x)] \\
&= 1 \cdot 2\delta + \frac{\eta}{1-\eta} \cdot (1-\eta)(1-2\delta) + 1 \cdot \eta(1-2\delta) \\
&= 2\delta + 2\eta(1-2\delta) \\
&= 2\delta + O(\eta).
\end{aligned}$$

Given that $\eta \ll \delta$, it follows that μ_{Max} is nearly twice as dense as μ_{TTV}.

In this section, we analyze the hardcore-ness of μ_{Max}. Before we proceed, we will make our arguments more general and encompass a whole family of measures. Recall that both μ_{TTV} and μ_{Max} involve the error function $|g(x) - p(x)|$ and are balanced under calibration (i.e., on each level set $\mathcal{X}_v$, they assign equal weight to $\mathcal{X}_v^0$ and to $\mathcal{X}_v^1$). First, we can formally express our μ_{Max} measure as a weighted version of the μ_{TTV} measure:

Definition 4.13. *We define the maximal* balanced under calibration *measure* μ_{Max} *as*

$$\mu_{\mathsf{Max}}(x) = w_{\mathsf{Max}}(p(x)) \cdot |g(x) - p(x)|,$$

where $w_{\mathsf{Max}}(p) = 1/\max(p, 1-p)$ *is a weight function mapping* $[0,1]$ *to* $[1,2]$.

We can now get a family of balanced under calibration measures by allowing any (reasonable) weight function $w(p(x))$. Specifically:

Definition 4.14 (Balanced under calibration measures)**.** *Consider the weight function family*

$$W = \{w : [0,1] \to [1,2] \text{ s.t. } w(p) \in [1, w_{\mathsf{Max}}(p)]\}.$$

Given $g : \mathcal{X} \to \{0,1\}$ *and* $p : \mathcal{X} \to [0,1]$*, we define the set* $M(p,g)$ *of* balanced under calibration *measures to be* $\{\mu_w\}_{w \in W}$ *where*

$$\mu_w(x) = w(p(x)) \cdot |g(x) - p(x)| \text{ for } w \in W.$$

Note that the minimal measure in $M(p,g)$ corresponds to μ_{TTV}, where w corresponds to the constant 1 function. Because a measure $\mu(x)$ always takes values in $[0,1]$, this is why we require the constraint $w(p(x)) \leq w_{\mathsf{Max}}(p(x)) \leq 2$. We add the condition $w(p) \geq 1$ because we want to find dense measures, and so there is no point in choosing weights less than 1. As we saw in our example in Figure 4.3, when $p \approx 1/2$, the measure μ_{Max} nearly twice as much mass as μ_{TTV}. On the flip side, when p is close to $\{0,1\}$, we have that $\mu_{\mathsf{Max}} \approx \mu_{\mathsf{TTV}}$, given that $w_{\mathsf{Max}} \approx 1$.

As w_μ increases from 1 to $w_{\mathsf{Max}}(p)$, $\mu(x)$ and hence $\mathsf{dns}(\mu) = \mathbb{E}_{\mathcal{D}_\mathcal{X}}[\mu(\mathbf{x})]$ increases. We will now analyze the hardcore-ness properties of the measures in the family $M(p,g)$. One way to show the hardcore-ness of μ_{Max} is to carry out the [TTV09] proof with the added weight function $w(p(x))$ in all steps. Here, we carry out the analysis in more generality so that we can clearly observe the roles played by each of the weighted multiaccuracy error, the density of the measure, and the correlation between g and $\mathcal{C}$. Instead of using the fact that p is multiaccurate, here we need to adapt the analysis to the weighted versions of the μ_{TTV} measure, which is why we are going to require a weighted version of multiaccuracy.

Specifically, we incorporate the weight function $w(\cdot)$ to multiaccuracy as one would expect:

Definition 4.15 (Weighted multiaccuracy)**.** *Let* $\mathcal{D}$ *be a distribution on* $\mathcal{X} \times \{0,1\}$*, let* $\mathcal{C} = \{c : \mathcal{X} \to \{0,1\}\}$ *be a class of functions, let* $p : \mathcal{X} \to [0,1]$ *be predictor, and* $w : [0,1] \to \mathbb{R}^+$ *a weight function. We define the* $(w, \mathcal{C})$-multiaccuracy error *of* p *under the distribution* $\mathcal{D}$ *as*

$$\mathsf{MA}_{\mathcal{D}}(w, \mathcal{C}, p) = \max_{c \in \mathcal{C}} \left| \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y}) \sim \mathcal{D}} [c(\mathbf{x}) w(p(\mathbf{x}))(\mathbf{y} - p(\mathbf{x}))] \right|.$$

We say that the predictor p is $(w, \mathcal{C}, \varepsilon)$-multiaccurate *for* $\mathcal{D}$ *if* $\mathsf{MA}_{\mathcal{D}}(w, \mathcal{C}, p) \leq \epsilon$.

This is identical to the multiaccuracy definition, except that now we weight the level sets by $w(p(x))$. Beyond our applications of weighted multiaccuracy, this is a natural notion on its own, in the cases where we want to go give more importance to the errors caused on certain p-values, for example. Throughout, g denotes the input function to IHCL. Given a distribution $\bar{\mu}$, we let

$$\eta = \mathop{\mathbb{E}}_{\bar{\mu}}[g(x)].$$

We avoid using v here since we are not restricting the domain to a level set $\mathcal{X}_v$. The key quantity that we are interested in is the hardness of prediction of g under $\bar{\mu}$ with respect to $\mathcal{C}$; that is, $\Pr_{\bar{\mu}}[c(\mathbf{x}) = g(\mathbf{x})]$. As in [TTV09], we want to relate this probability to an expected value that we can later relate to the (weighted) multiaccuracy error, given that we will want to use the fact that our predictor p is a $(w, \mathcal{C})$-multiaccurate predictor. To do so, it is useful to introduce the quantity:

$$\mathrm{cor}_{\bar{\mu}}(g, \mathcal{C}) = \max_{c \in \mathcal{C}} |\mathrm{cor}_{\bar{\mu}}(g, c)| = \max_{c \in \mathcal{C}} \left| \mathop{\mathbb{E}}_{\mathbf{x} \sim \bar{\mu}}[(2g(\mathbf{x}) - 1)c(\mathbf{x})] \right|.$$

Here, unlike in Chapter 3, this quantity is in the range $[0, 1]$. Then, we relate the hardness of prediction quantity to the correlation between g and $\mathcal{C}$ as follows:

Lemma 4.16. *We have*

$$\min(\eta, 1 - \eta) + \mathrm{cor}_{\bar{\mu}}(g, \mathcal{C}) \leq \max_{c \in \mathcal{C}} \Pr_{\mathbf{x} \sim \bar{\mu}}[c(\mathbf{x}) = g(\mathbf{x})] \leq \max(\eta, 1 - \eta) + \mathrm{cor}_{\bar{\mu}}(g, \mathcal{C}). \tag{4.17}$$

Proof. We can write the event $\mathbf{1}[c(x) = g(x)]$ as the identity $\frac{1}{2}(1 + (2c(x) - 1)(2g(x) - 1))$, given that this expression evaluates to 1 if $c(x) = g(x)$, and to 0 otherwise. Thus:

$$\begin{aligned}
\Pr_{\bar{\mu}}[c(\mathbf{x}) = g(\mathbf{x})] &= \mathop{\mathbb{E}}_{\bar{\mu}}\left[\frac{1}{2}(1 + (2c(\mathbf{x}) - 1)(2g(\mathbf{x}) - 1))\right] \\
&= \mathop{\mathbb{E}}_{\bar{\mu}}[c(\mathbf{x})(2g(\mathbf{x}) - 1) + (1 - g(\mathbf{x}))] \\
&= \mathrm{cor}_{\bar{\mu}}(g, c) + 1 - \eta,
\end{aligned}$$

where we plug in the definition of correlation and use the fact that $\eta = \mathbb{E}_{\bar{\mu}}[g(x)]$. Let $\bar{c} = 1 - c$ and recall that we are assuming that $\mathcal{C}$ is closed under complementation. So we have

$$\begin{aligned}
\Pr_{\bar{\mu}}[\bar{c}(\mathbf{x}) = g(\mathbf{x})] &= 1 - \Pr_{\bar{\mu}}[c(\mathbf{x}) = g(\mathbf{x})] \\
&= \mathop{\mathbb{E}}_{\bar{\mu}}[g(\mathbf{x})] - \mathop{\mathbb{E}}_{\bar{\mu}}[c(\mathbf{x})(2g(\mathbf{x}) - 1)]] \\
&= \eta - \mathrm{cor}_{\bar{\mu}}(g, c).
\end{aligned}$$

The second equality follows from the fact that here we are interested in the complementary event $\mathbf{1}[c(x) \neq g(x)]$, which algebraically corresponds to the identity $g - c(2g - 1)$. Then, we plug in the definitions of η and $\mathrm{cor}(g, c)$.

Therefore, we get the upper bound of $\max(\eta, 1 - \eta) + |\mathrm{cor}_{\bar{\mu}}(g, c)|$ in both cases, and then we maximize over all $c \in \mathcal{C}$.

To prove the lower bound, we can predict g with accuracy $\min(\eta, 1-\eta) + |\mathrm{cor}_{\bar{\mu}}(g, c)|$ by using c if $\mathrm{cor}_{\bar{\mu}}(g, c) \geq 0$ and $\bar{c}$ otherwise, and then maximizing over $c \in \mathcal{C}$. □

We can view this lemma as related to Yao's lemma on indistinguishability (computational randomness) and unpredictability (computational hardness). Specifically:

Lemma 4.18 (Equivalence between indistinguishability and unpredictability [Yao82])**.** *A function $g : \mathcal{X} \to [0,1]$ such that $\mathbb{E}_{x\sim\mathcal{D}}[g(x)] = 1/2$ is $(\mathcal{C}, \epsilon)$-indistinguishable on $\mathcal{D}$ from the constant $1/2$ function if and only if g is $(\mathcal{C}, 1/2 - 2\epsilon)$-hard on $\mathcal{D}_{\mathcal{X}}$.*

One can prove this using the identity:

$$\Pr[c(\mathbf{x}) = g(\mathbf{x})] = 2\,\mathbb{E}[(c(\mathbf{x}) - 1/2)(g(\mathbf{x}) - 1/2)] + 1/2. \tag{4.19}$$

In Lemma 4.16, we relate indistinguishability and pseudorandomness in a similar manner, but using an arbitrary $\eta \in [0,1]$ rather than $1/2$.

Next, we relate the correlation term $\mathrm{cor}(\mathcal{C}, g)$ with the weighted multiaccuracy error:

Lemma 4.20. *Consider a distribution $\mathcal{D}_{\mathcal{X}}$ on $\mathcal{X}$, a labeling function $g : \mathcal{X} \to \{0,1\}$, a predictor p, a weight function $w \in W$ and the corresponding measure $\mu_w \in M(p, g)$. Then:*

$$\mathsf{MA}_{\mathcal{D}^g}(w, \mathcal{C}, p) = \mathrm{cor}_{\bar{\mu}_w}(\mathcal{C}, g) \cdot \mathsf{dns}(\mu_w).$$

Proof. For $y \in \{0,1\}$ and $t \in [0,1]$ we have the identity $|y - t| \cdot (2y - 1) = y - t$. Applying this with $\mathbf{y} = g(\mathbf{x})$ and $t = p(\mathbf{x})$, we can expand the weighted multiaccuracy error as follows:[2]

$$\begin{aligned}
\mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}^g}[c(\mathbf{x})w(p(\mathbf{x}))(\mathbf{y} - p(\mathbf{x}))] &= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}}[c(\mathbf{x})w(p(\mathbf{x}))(g(\mathbf{x}) - p(\mathbf{x}))] \\
&= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}}[c(\mathbf{x})w(p(\mathbf{x})) \cdot |g(\mathbf{x}) - p(\mathbf{x})| \cdot (2g(\mathbf{x}) - 1)] \\
&= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}}[\mu_w(\mathbf{x})c(\mathbf{x})(2g(\mathbf{x}) - 1)] \\
&= \frac{\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}}[\mu_w(\mathbf{x})c(\mathbf{x})(2g(\mathbf{x}) - 1)]}{\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}}[\mu_w(\mathbf{x})]} \cdot \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}}[\mu_w(\mathbf{x})] \\
&= \mathop{\mathbb{E}}_{\mathbf{x}\sim\bar{\mu}_w}[c(\mathbf{x})(2g(\mathbf{x}) - 1)] \cdot \mathsf{dns}(\mu_w).
\end{aligned} \tag{4.21}$$

To go from $\mathcal{D}_{\mathcal{X}}$ to μ_w we used the fact that $\mu_w(\mathbf{x}) = w(p(\mathbf{x})) \cdot |g(\mathbf{x}) - p(\mathbf{x})|$. Taking the absolute value on either side and maximizing over all $c \in \mathcal{C}$ gives the claimed result. □

We can now prove our main result in this section, which shows that a weighted multiaccurate predictor induces the hardcore-ness of the measure:

Lemma 4.22. *If $\mathsf{MA}_{\mathcal{D}^g}(w, \mathcal{C}, p) \leq \varepsilon$, then g is $\big(1/2 - 3\varepsilon/(2\mathsf{dns}(\mu_w))\big)$-hard for $\mathcal{C}$ under $\bar{\mu}_w$.*

Proof. In the previous lemma (Lemma 4.20), we just showed that

$$\mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}^g}[c(\mathbf{x})w(p(\mathbf{x}))(\mathbf{y} - p(\mathbf{x}))] = \mathop{\mathbb{E}}_{\mathbf{x}\sim\bar{\mu}_w}[c(\mathbf{x})(2g(\mathbf{x}) - 1)] \cdot \mathsf{dns}(\mu_w).$$

[2] This identity is similar to the one used in [TTV09, §5.2] (Equation 4.5).

Hence, by using the constant function $c = \mathbf{1}$ we get that

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}}[w(p(\mathbf{x}))(g(\mathbf{x}) - p(\mathbf{x}))] = \mathop{\mathbb{E}}_{\mathbf{x}\sim\bar{\mu}_w}[(2g(\mathbf{x}) - 1)] \cdot \mathsf{dns}(\mu_w) = (2\eta - 1) \cdot \mathsf{dns}(\mu_w).$$

The first term (LHS) corresponds to the weighted multiaccuracy error, and we are assuming this error is bounded in absolute value by ϵ. Hence,

$$|(2\eta - 1) \cdot \mathsf{dns}(\mu_w)| \leq \epsilon \implies |2\eta - 1| \leq \frac{\epsilon}{\mathsf{dns}(\mu_w)} \implies \left|\eta - \frac{1}{2}\right| \leq \frac{\varepsilon}{2\mathsf{dns}(\mu_w)}.$$

Moreover, by Lemma 4.20, we know that

$$\mathsf{MA}_{\mathcal{D}^g}(w, \mathcal{C}, p) = \mathrm{cor}_{\bar{\mu}_w}(\mathcal{C}, g) \cdot \mathsf{dns}(\mu_w) \implies \mathrm{cor}_{\bar{\mu}_w}(\mathcal{C}, g) = \frac{\mathsf{MA}_{\mathcal{D}^g}(w, \mathcal{C}, p)}{\mathsf{dns}(\mu_w)} \leq \frac{\epsilon}{\mathsf{dns}(\mu_w)}.$$

To use these bounds on our quantity of interest, namely the hardness of prediction, we use Lemma 4.16, which tells us that:

$$\begin{aligned}
\max_{c\in\mathcal{C}} \Pr_{\mathbf{x}\sim\bar{\mu}_w}[c(\mathbf{x}) = g(\mathbf{x})] &\leq \max(\eta, 1-\eta) + \mathrm{cor}_{\bar{\mu}_w}(\mathcal{C}, g) \\
&= \frac{1}{2} + |\eta - 1/2| + \mathrm{cor}_{\bar{\mu}_w}(\mathcal{C}, g) \\
&\leq \frac{1}{2} + \frac{\epsilon}{2\mathsf{dns}(\mu_w)} + \mathrm{cor}_{\bar{\mu}_w}(\mathcal{C}, g) \\
&\leq \frac{1}{2} + \frac{3\varepsilon}{2\mathsf{dns}(\mu_w)}
\end{aligned}$$

Therefore, g is $\big(1/2 - 3\varepsilon/(2\mathsf{dns}(\mu_w))\big)$-hard for $\mathcal{C}$ under $\bar{\mu}_w$, as we wanted to show. □

4.5 Optimal density analysis using calibration

To complete the proof, we need a lower bound on $\mathsf{dns}(\mu_w)$. We already showed that μ_{Max} has density 2δ assuming perfect calibration in Section 4.3. Here we do the same analysis but account for the calibration error and generalizing to the family of measures $M(p, g)$. First, in Section 4.2 we showed that if g is δ-weakly hard for $\mathcal{C}_{t,g}$, then μ_{TTV} has density at least δ. Given that every $\mu \in M(p, g)$ satisfies $\mathsf{dns}(\mu) \geq \mathsf{dns}(\mu_{\mathsf{TTV}})$, it follows that every measure in $M(p, g)$ also has density at least δ. Next, we relate the density of the measure to the quantity $2p(1 - p)$, as we did in Section 4.3, but now including the calibration error and the weight function:

Lemma 4.23. *Assume that the predictor p is τ-calibrated for $\mathcal{D}^g$. Then we have*

$$\left|\mathsf{dns}(\mu_w) - 2 \mathop{\mathbb{E}}_{\mathcal{D}_{\mathcal{X}}}[w(p(\mathbf{x}))p(\mathbf{x})(1 - p(\mathbf{x}))]\right| \leq 2\tau.$$

Proof. We will use the following identity for $v \in [0, 1]$ and $y \in \{0, 1\}$, which can be viewed as the multilinear expansion of $|y - v|$:

$$|y - v| = y(1 - 2v) + v.$$

Using this with $y = g(x)$ and $v = p(x)$ we can write

$$\begin{aligned}
\mathsf{dns}(\mu_w) &= \mathop{\mathbb{E}}_{\mathcal{D}_\mathcal{X}}[\mu_w(\mathbf{x})] \\
&= \mathop{\mathbb{E}}_{\mathcal{D}_\mathcal{X}}[w(p(\mathbf{x})) \cdot |g(\mathbf{x}) - p(\mathbf{x})|] \\
&= \mathop{\mathbb{E}}_{\mathcal{D}_\mathcal{X}}[w(p(\mathbf{x}))(g(\mathbf{x})(1 - 2p(\mathbf{x})) + p(\mathbf{x}))] \\
&= \mathop{\mathbb{E}}_{\mathcal{D}_\mathcal{X}}[w(p(\mathbf{x}))(p(\mathbf{x})(1 - 2p(\mathbf{x})) + p(\mathbf{x}))] + \mathop{\mathbb{E}}_{\mathcal{D}_\mathcal{X}}[w(p(\mathbf{x}))(1 - 2p(\mathbf{x}))(g(\mathbf{x}) - p(\mathbf{x}))]
\end{aligned}$$

Hence we have

$$\left|\mathsf{dns}(\mu_w) - 2 \mathop{\mathbb{E}}_{\mathcal{D}_\mathcal{X}}[w(p(\mathbf{x}))p(\mathbf{x})(1 - p(\mathbf{x}))]\right| = \left|\mathop{\mathbb{E}}_{\mathcal{D}_\mathcal{X}}[w(p(\mathbf{x}))(1 - 2p(\mathbf{x}))(g(\mathbf{x}) - p(\mathbf{x}))]\right|. \tag{4.24}$$

Since $w(p) \in [1, 2]$ and $1 - 2p \in [-1, 1]$, we have $|w(p)(1 - 2p)| \leq 2$. Since p is τ-calibrated, by Equation (2.11) (which expresses the expected calibration error using bounded auditing functions) applied to $v(p) = w(p)(1 - 2p)/2$, the predictor p satisfies the bound

$$\left|\mathop{\mathbb{E}}_{\mathcal{D}_\mathcal{X}}[w(p(\mathbf{x}))(1 - 2p(\mathbf{x}))(g(\mathbf{x}) - p(\mathbf{x}))]\right| \leq 2\tau.$$

Plugging this into Equation (4.24) gives the claimed bound. □

For our measure μ_{Max}, we are interested in the quantity $\min\{p, 1 - p\}$, for the reasons we explained in Section 4.3. We relate this quantity to the δ parameter using the same argument as in the perfect calibration case:

Lemma 4.25. *Assume that g is δ-hard for $\mathcal{C}_{t,q}$ under $\mathcal{D}_\mathcal{X}$, and p is τ-calibrated for $\mathcal{D}^g$. Then*

$$\mathop{\mathbb{E}}_{\mathcal{D}_\mathcal{X}}[\min(p(\mathbf{x}), 1 - p(\mathbf{x}))] \geq \delta - \tau.$$

Proof. Consider the function $f(x) = \mathbf{1}[p(x) \geq 1/2]$, which is computable in $\mathcal{C}_{t,q}$. As argued in Section XX, this is because the predictor p is discretized to precision $\Theta(\tau)$ and is a linear combination of $O(1/\tau^2)$ many hypotheses from $\mathcal{C}$, and so The function f errs when $g(x) = 1$ for points where $p(x) < 1/2$ and whenever $g(x) = 0$ otherwise. We apply the calibration condition in the typical way: we condition on each level set so that we can compute how well the predictor $\mathbf{1}[p(x) \geq 1/2]$ is doing in each.

$$\begin{aligned}
\delta &\leq \Pr_{\mathcal{D}_\mathcal{X}}[f(\mathbf{x}) \neq g(\mathbf{x})] \\
&= \Pr_{\mathcal{D}_\mathcal{X}}[g(\mathbf{x}) = 1 \wedge p(\mathbf{x}) < 1/2] + \Pr_{\mathcal{D}_\mathcal{X}}[g(\mathbf{x}) = 0 \wedge p(\mathbf{x}) \geq 1/2] \\
&= \mathop{\mathbb{E}}_{\mathcal{D}^g}[\mathbb{E}[\mathbf{y}|p(\mathbf{x})] \cdot \mathbf{1}[p(\mathbf{x}) < 1/2] + (1 - \mathbb{E}[\mathbf{y}|p(\mathbf{x})]) \cdot \mathbf{1}[p(\mathbf{x}) \geq 1/2]] \\
&= \mathop{\mathbb{E}}_{\mathcal{D}^g}[(p(\mathbf{x}) + \mathbb{E}[\mathbf{y}|p(\mathbf{x})] - p(\mathbf{x})) \cdot \mathbf{1}[p(\mathbf{x}) < 1/2] + (1 - p(\mathbf{x}) + p(\mathbf{x}) - \mathop{\mathbb{E}}_{\mathcal{D}^g}[\mathbf{y}|p(\mathbf{x})]) \cdot \mathbf{1}[p(\mathbf{x}) \geq 1/2]] \\
&\leq \mathop{\mathbb{E}}_{\mathcal{D}^g}[(p(\mathbf{x}) \cdot \mathbf{1}[p(\mathbf{x}) < 1/2] + (1 - p(\mathbf{x})) \cdot \mathbf{1}[p(\mathbf{x}) \geq 1/2])] + \mathop{\mathbb{E}}_{\mathcal{D}^g}[|p(\mathbf{x}) - \mathbb{E}[\mathbf{y}|p(\mathbf{x})]|]
\end{aligned}$$

$$= \mathop{\mathbb{E}}_{\mathcal{D}_\mathcal{X}}[\min(p(\mathbf{x}), 1-p(\mathbf{x})] + \mathsf{ECE}(p, \mathcal{D}^g),$$

where we use the identity that

$$\min(v, 1-v) = v \cdot \mathbf{1}[v < 1/2] + (1-v) \cdot \mathbf{1}[v \geq 1/2].$$

The claim now follows since $\mathsf{ECE}(p, \mathcal{D}^g) \leq \tau$. □

We finally obtain the density of μ_{Max} as a corollary to Lemmas 4.23 and 4.25, and we state it together with its hardcore-ness guarantee:

Theorem 4.26. *For $\varepsilon, \delta, \tau > 0$ where $\tau \leq \delta/2$, assume that p is $(w_{\mathsf{Max}}, \mathcal{C}, \varepsilon\delta)$-multiaccurate for $\mathcal{D}^g$, and it is $\tau/4$-calibrated. Then the measure μ_{Max} is $2\delta - \tau$ dense in $\mathcal{D}_\mathcal{X}$, and g is $(1/2 - \varepsilon)$-hard for $\mathcal{C}$ under the distribution $\bar{\mu}_{\mathsf{Max}}$.*

Proof. By Lemma 4.23 applied with w_{Max}, we have

$$\begin{aligned}
\mathsf{dns}(\mu_{\mathsf{Max}}) &\geq 2 \mathop{\mathbb{E}}_{\mathcal{D}_\mathcal{X}}\left[\frac{p(\mathbf{x})(1-p(\mathbf{x}))}{\max(p(\mathbf{x}), 1-p(\mathbf{x}))}\right] - \tau/2 \\
&= 2 \mathop{\mathbb{E}}_{\mathcal{D}_\mathcal{X}}[\min(p(\mathbf{x}), 1-p(\mathbf{x}))] - \tau/2 \\
&\geq 2\delta - \tau/2 - \tau/2 \\
&\geq 2\delta - \tau.
\end{aligned}$$

where we use Lemma 4.25.

As for the hardness of g, we apply Lemma 4.22 with weighted multiaccuracy error $\epsilon\delta$. Recall that in Lemma 4.22 we showed that $\mathsf{MA}_{\mathcal{D}^g}(w, \mathcal{C}, p) \leq \varepsilon$, then g is $\big(1/2 - 3\varepsilon/(2\mathsf{dns}(\mu_w))\big)$-hard for $\mathcal{C}$ under $\bar{\mu}_w$. Here, the weighted multiaccuracy error is $\epsilon\delta$, and we just showed that $\mathsf{dns}(\mu_w)$. Then:

$$\frac{1}{2} - \frac{3\epsilon}{2\mathsf{dns}(\mu_w)} = \frac{1}{2} - \frac{3\epsilon\delta}{2(2\delta - \tau)} \geq \frac{1}{2} - \epsilon$$

for $\tau < 0.5\delta$. □

4.6 Hardcore measures with optimal density

We conclude by putting together all of the ingredients together, stating our IHCL theorem. First, we remark that in order to obtain a weighted multiaccurate and calibrated predictor, we can readily adapt the [GHK+23] algorithm for constructing multiaccurate and calibrated predictors. As we summarize in [CGKR25, §5.3][3], because we only care about the weight function w_{Max}, the complexity of the weighted versions of multiaccuracy and calibrated multiaccuracy are the same as their unweighted counterparts. Hence we can use the guarantees from [GHK+23], which we gave in Theorem 2.16. The intuition for why, as shown in [GHK+23], calibrated multiaccuracy requires the same number of calls to the weak agnostic learner as multiaccuracy is that a recalibrating step

[3]The paper that which Part I of this thesis is based on.

also improves the squared loss of the predictor by a good amount. Then, we can use the same potential function argument, where now each iterative step reducing the potential function can be either a multiaccuracy step or a calibration step.

Theorem 4.27 (IHCL with density 2δ)**.** *Let $\mathcal{C}$ be a family of functions $c : \mathcal{X} \to \{0,1\}$, let $\mathcal{D}_{\mathcal{X}}$ be a probability distribution over $\mathcal{X}$, and let $\epsilon, \delta, \tau > 0$. There exist $t = O((1/(\epsilon^2\delta^2) + 1/\tau) \cdot \log(|\mathcal{X}|/\epsilon))$, $q = O(1/(\epsilon^2\delta^2))$ such that the following holds: If $g : \mathcal{X} \to \{0,1\}$ is $(\mathcal{C}_{t,q}, \delta)$-hard on $\mathcal{D}_{\mathcal{X}}$, then there is a measure μ satisfying:*

- *Hardness: g is $(\mathcal{C}, 1/2 - \epsilon)$-hard on $\bar{\mu}$.*
- *Optimal density: $\mathsf{dns}(\mu) = 2\delta - \tau$.*

Proof. We call the calMA algorithm from [GHK+23] with $\mathcal{C}, \mathcal{D}^g$, weight function w_{Max}, desired multiaccuracy error $\epsilon\delta$, and desired calibration error $\tau/4$. This gives us a $(w_{\mathsf{Max}}, \mathcal{C}, \epsilon\delta)$-multiaccurate predictor and $(\tau/4)$-calibrated predictor p for $\mathcal{D}^g$. Using this predictor p, the input function g, and the weight function w_{Max}, we construct the measure μ_{Max} as per Definition 4.13; namely, $\mu_{\mathsf{Max}}(x) = w_{\mathsf{Max}}(p(x)) \cdot |g(x) - p(x)|$. By Theorem 4.26, g is $(1/2 - \epsilon)$-hard for $\mathcal{C}$ under $\bar{\mu}$, and the measure μ_{Max} is $2\delta - \tau$ dense in $\mathcal{D}_{\mathcal{X}}$. This gives us the hardness and optimal density guarantees that we wanted to show. □

Our IHLC theorem thus has a circuit size loss of $q = O(1/(\epsilon^2\delta^2))$. One of our key points is that, when showing Hardcore Lemma statements from multigroup fairness notions, we knew that multiaccuracy only gave us δ density, whereas [CDV24] were able to obtain 2δ density through multicalibration, first showing the stronger IHCL++ statement from multicalibration and then recovering IHCL with 2δ density as a corollary. However, because they go through multicalibration, their IHCL corollary incurs a circuit size loss of $q = 1/(\epsilon^2\delta)^6$. Here, we have obtained the same IHCL guarantees, but from the weaker notion of calibrated multiaccuracy, and thus with much better circuit loss parameters.

Moreover, the dependence on ϵ in the number of oracle calls q that we make to the weak agnostic learner in our Theorem 4.27 is optimal. All boosting-based proofs of the hardcore lemma, starting with one of Impagliazzo's original proofs and later generalized by Kilvans and Servedio to modularly use various boosting algorithms to obtain hardcore measures, need to call a weak learner $O(1/\epsilon^2)$ many times as to obtain constant density measure on which g is $(\mathcal{C}, 1/2 - \epsilon)$-hard [KS03]. As we mentioned in the beginning of the chapter, it has been recently shown that a circuit loss size of $O(1/\epsilon^2)$ is unavoidable, regardless of proof strategy [BKST24]. In particular, Impagliazzo's boosting-based proof requires $q = O(1/(\epsilon^2\delta^2))$ oracle calls [Imp95, Thm. 1], and all subsequent hardcore constructions retain the dependence on $1/\epsilon^2$ in q [KS03; Hol05], including the uniform versions of the hardcore lemma [Tre03; BHK09; VZ13]. Meanwhile, the dependence on δ can be improved: Nisan's min-max proof in Impagliazzo's original paper yields $q = O(1/\epsilon^2 \log(1/(\epsilon\delta))$ [Imp95, Lemma 2], and it was further improved to $q = O(1/\epsilon^2 \log(1/\delta))$ by Kilvans and Servedio who showed an elegant connection to boosting [KS03, Thm. 25], which we review in Section 4.8. But these proofs do not give the optimal density.

We summarize the various IHCL statements, their proof technique, number of oracle gates q, and density of the resulting hardcore set in Table 4.1.

Paper	**Proof technique**	**Parameter q**	**Density**
[Imp95]	Boosting	$O\left(\frac{1}{\epsilon^2\delta^2}\right)$	δ
[Imp95]	Min-max theorem	$O\left(\frac{1}{\epsilon^2 \log\left(\frac{1}{\epsilon\delta}\right)}\right)$	δ
[KS03]	Boosting	$O\left(\frac{1}{\epsilon^2 \log(1/\delta)}\right)$	δ
[Hol05]	Boosting	$O\left(\frac{1}{\epsilon^2\delta^{O(1)}}\right)$	2δ
[BHK09]	Multiplicative Weights Update	$O\left(\frac{1}{\epsilon^2}\log(1/\delta)\right)$	2δ
[CDV24]	From MC + [TTV09] construction	$O\left(\frac{1}{(\epsilon^2\delta)^6}\right)$	2δ
Here	From Cal-MA + [TTV09] construction	$O\left(\frac{1}{\epsilon^2\delta^2}\right)$	2δ

Table 4.1: Comparison of the various IHCL lemmas.

4.7 Comparison to IHCL++

We conclude our discussion of the various proofs of IHCL by providing some more intuition on why the full power of multicalibration wasn't necessary to get IHCL with optimal density.

Recall the complexity-theoretic interpretation of the multicalibration theorem that we gave in Section 2.3, when explaining the connections between the Regularity Lemma and multigroup fairness notions (Theorem 2.19). Essentially, we can view multicalibration as giving us a low-complexity partition $\mathcal{P}$ of the domain, where on each (large enough) piece $P \in \mathcal{P}$, the underlying arbitrarily complex function p^* is $(\mathcal{C}, \epsilon)$-indistinguishable from the constant function v_P, where v_P is equal to the expected value of p^* on the piece P.

As shown in [CDV24], we can characterize indistinguishability from a constant function in terms of hardness of prediction.[4] Essentially, this equivalence corresponds to a generalization of Yao's equivalence between pseudorandomness and unpredictability (Lemma 4.18) from the constant function $1/2$ to any constant function v in $[0,1]$:

[4]We should be more careful when talking about the hardness of prediction of a deterministic $[0,1]$-valued function. Technically, here think of a $[0,1]$-valued function p^* as describing a randomized $\{0,1\}$-valued function $p^{*,\mathrm{rand}}$ where $\Pr_{\mathrm{coins}(p^{*,\mathrm{rand}})}[p^{*,\mathrm{rand}}(x) = 1] = p(x)$. We do not include the rand notation in this section to avoid introducing more notation; the precisely correct formulation can be found in [CDV24, §2]. We use **y** instead.

Lemma 4.28 (Subset of Lemma 2.17 in [CDV24]). *Let $\mathcal{C}$ be a class of functions $c : \mathcal{X} \to [0,1]$, let $\mathcal{D}_{\mathcal{X}}$ be a distribution over the domain $\mathcal{X}$, and let $\epsilon > 0$. Let $p^* : \mathcal{X} \to [0,1]$ be $(\mathcal{C}, \epsilon)$-indistinguishable from the constant function v on $\mathcal{D}_{\mathcal{X}}$, where $v = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}}[g(\mathbf{x})]$ and $b = \min\{v, 1-v\}$. Then, the function p^* is $(\mathcal{C}, b - 2\epsilon)$-hard on $\mathcal{D}_{\mathcal{X}}$.*

As a corollary, we can re-state the complexity-theoretic interpretation of the multicalibration theorem in terms of hardness of prediction (rather than of indistinguishability from constant functions, as we did in Lemma 4.28):

Lemma 4.29 (Multicalibration theorem in terms of hardness of prediction [CDV24]). *Given $\mathcal{X}, p^*, \mathcal{D}_{\mathcal{X}}, \mathcal{C}, \epsilon$, there exists a low-complexity partition $\mathcal{P}$ of the domain such that for all $c \in \mathcal{C}$ and all large enough $P \in \mathcal{P}$, the function p^* is $(\mathcal{C}, b_P - 2\epsilon)$-hard on $\mathcal{D}_{\mathcal{X}}|_P$.*

We defined what we mean by "low-complexity partition" in Theorem 2.19; the precise quantification of what constistutes a "large enough" piece can be found in [CDV24], where they quantify the size of each piece $P \in \mathcal{P}$ through the size parameter η_P.

Next, we can use this hardness of prediction of p^* on each $P \in \mathcal{P}$ to construct a small hardcore set on each piece. Essentially, on each $P \in \mathcal{P}$, we define a hardcore measure μ_P that assigns more weight to the minority elements in the piece, so that we "shift" the expected value of g on that piece from v_P to $1/2$, thus going from $(b_P - 2\epsilon)$-hardness of prediction to $(1/2 - \epsilon')$-hardness of prediction, which gives us the required hardcore-ness in the definition of a hardcore measure. We explained in detail in Section 4.3 why assigning higher weight to the minority elements in the piece is the right strategy as to maximize hardness. Alternatively, we can apply the IHCL theorem in each piece $P \in \mathcal{P}$, given that now we again have an assumption of hardness of prediction on each piece P. By using the function g as p^*, this gives us the following stronger and more general version of IHCL, called IHCL++ in [CDV24]:

Theorem 4.30 (IHCL++ [CDV24]). *Given $\mathcal{X}, \mathcal{D}_{\mathcal{X}}, \epsilon, \mathcal{C}$ and an arbitrary function g, there exists a low-complexity partition $\mathcal{P}$ of the domain such that, for every large enough $P \in \mathcal{P}$, there exists a measure μ_P of density $2b_P$ in $\mathcal{D}_{\mathcal{X}}|_P$ satisfying:*

- *Hardness: g is $\left(\mathcal{C}, 1/2 - \frac{\epsilon}{2b_P(1-b_P)}\right)$-hard on $\bar{\mu}_P$.*
- *Optimal density: $\mathsf{dns}(\mu_P) = 2b_P$ in $\mathcal{D}_{\mathcal{X}}|_P$.*

Crucially, unlike in IHCL, here g is not assumed to be δ-weakly hard – it is an arbitrary function. This is possible precisely because the density of each small hardcore set is expressed in terms of the b_P parameter, which can be equal to 0 if the function g is in fact easy to compute (e.g., if it is the constant **0** or constant **1** function). That is, we can view IHCL++ as saying that we can extract strong hardness from an arbitrary g by exhibiting a polynomial collection of "small" hardcore sets. In the absence of a parameter δ (given that g is no longer assumed to be δ-weakly hard), it turns out that the right abstraction for it is precisely the balance parameter b_P.

In order to recover IHCL from IHCL++, we glue each of the small hardcore sets μ_P into a global hardcore set μ. Then, [CDV24] show that:

- *Hardness:* Because g is hard on each of μ_P, g is hard on the global μ.
- *Optimal density:* First, they show that if g is δ-weakly hard, this implies that $\mathbb{E}_{\mathcal{P}(\mathcal{D}_\mathcal{X})}[b_P] \geq \delta$. Thus, since $\mathsf{dns}(\mu_P) = 2b_P$, this implies that $\mathsf{dns}(\mu) = 2\delta$.

How does this relate to our proof of IHCL with density 2δ?

- *Hardness:* For IHCL, it is overkill to require that g is hard on each of μ_P – we only need it to be hard on $\mathcal{X}$, not on each piece $P \in \mathcal{P}$. Hence, multicalibration does way more than we need, and indeed, (weighted) multiaccuracy is sufficient.
- *Optimal density:* We use the exact same argument; namely, that if g is δ-weakly hard, then $\mathbb{E}_{\mathcal{P}(\mathcal{D}_\mathcal{X})}[b_P] \geq \delta$. This is because, otherwise, the predictor that predicts the majority value on each piece (which is in $\mathcal{C}_{t,q}$) would be able to guess g too well, thus contradicting the δ-weak hardness of g. But we don't need multicalibration for this! We can get the b_P parameters just from global calibration, which allows to perform the exact same error analysis of the predictor that predicts the majority value on each piece. Indeed, for the optimal density argument, we only need to be able to work with the b_P parameters (for which calibration is necessary), but we do not need g to be indistinguishable from v_P on each piece/level set P.

4.8 The relationship between boosting and IHCL

In this chapter, we have extensively studied the relationship between (1) multigroup fairness notions and (3) hardcore set constructions. To conclude this chapter, we make some remarks about the relationship between (2) learning primitives and (3) hardcore set constructions, which tie up nicely with our discussions in the previous section on the relationship between multigroup fairness notions and hardness of prediction.

As we have explained throughout this thesis, hardness and boosting are two sides of the same coin. To make this point clearer, we can state the contrapositive verision of IHCL:

Theorem 4.31 (IHCL, contrapositive version)**.** *Suppose that a function $g : \mathcal{X} \to \{0,1\}$ is not $(\mathcal{C}, \epsilon)$-strongly hard. This means that for every measure μ over $\mathcal{X}$ with $\mathsf{dns}(\mu) = 2\delta$ in $\mathcal{D}_\mathcal{X}$, we can find a $c \in \mathcal{C}$ such that*

$$\Pr_{\tilde{\mu}}[c(\mathbf{x}) \neq g(\mathbf{x})] \leq 1/2 - \epsilon \implies \Pr_{\tilde{\mu}}[c(\mathbf{x}) = g(\mathbf{x})] > 1/2 - \epsilon.$$

That is, g is not *$(\mathcal{C}, \epsilon)$-strongly hard. Then, we can find a circuit $c' \in \mathcal{C}_{t,q}$ for $t = \mathrm{poly}(\log|\mathcal{X}|, 1/\epsilon, 1/\delta)$ and $q = \mathrm{poly}(1/\epsilon, 1/\delta)$ such that*

$$\Pr_{\mathcal{D}_\mathcal{X}}[c'(\mathbf{x}) \neq g(\mathbf{x})] \leq \delta \implies \Pr_{\mathcal{D}_\mathcal{X}}[c'(\mathbf{x}) = g(\mathbf{x})] \geq 1 - \delta.$$

That is, g is not *$(C_{t,q}, \delta)$-weakly hard.*

That is, if we can approximate g slightly better than random (according to any dense enough measure), then we can boost this "weak guessing" and build a circuit that approximate g very well (namely, better than $1 - \delta$). Following this idea, the boosting proof of IHCL (contrapositive) goes

as follows: at each iteration, we apply the assumption to the measure μ to obtain a circuit c that guesses g better than random. As in the canonical boosting algorithms, we then modify μ to give more weight to the points $x \in \mathcal{X}$ where c is not predicting g correctly. To determine the precise weight, Impagliazzo's boosting proof uses the margin by which the majority vote of the circuits that we have gathered so far correctly predicts the value of g. We continue this iterative process until the density of the measure μ drops below 2δ.

We summarize the precise boosting iterative algorithm by Impagliazzo (which is for $\mathsf{dns}(\mu) = \delta$), following the explanation given in [KS03]:

1. Set $i \leftarrow 0$, $\mu_0 := 1$.
2. While $\mathsf{dns}(\mu_i) \geq \delta$, do:
   (a) Let $c_i \in \mathcal{C}$ be such that $\Pr_{\mathcal{D}_{M_i}}[c_i(x) = g(x)] > 1/2 + \epsilon$. Such a function exists by the assumption of Theorem 4.31.
   (b) Set
   $$R_{c_i}(x) := \begin{cases} 1 & \text{if } c_i(x) = g(x), \\ -1 & \text{if } c_i(x) \neq g(x). \end{cases}$$
   (c) Set $N_i(x) := \sum_{0 \leq j \leq i} R_{c_j}(x)$. *($N_i(x)$ is the margin by which the majority vote of all of the circuits $c_0, \ldots, c_i$ that we have collected so far correctly predict $g(x)$.)*
   (d) Set
   $$\mu_{i+1}(x) := \begin{cases} 1 & \text{if } N_i(x) < 0, \\ 1 - \delta\epsilon N_i(x) & \text{if } 0 \leq N_i(x) \leq 1/(\delta\epsilon), \\ 0 & \text{if } N_i(x) > 1/(\delta\epsilon). \end{cases}$$
   (The measure μ_{i+1} assigns weight 0 to the points where the margin of correctness is large, weight 1 to points where the margin is negative, and intermediate weight to points where the margin is positive but small.)
   (e) Set $i \leftarrow i + 1$.
   (f) Return $c := \mathrm{MAJ}(c_0, c_1, \ldots, c_{i-1})$, where MAJ denotes the majority function.

Impagliazzo then shows that after at most $O(1/(\delta^2\epsilon^2))$ iterations, $\mathsf{dns}(\mu_i)$ must be less than δ, and so the algorithm terminates. Once this occurs, we can see that c (i.e., the majority circuit) agrees with g on all inputs except those that have $N_i(x) \leq 0$ and hence $\mu_i(x) = 1$. But because $\mathsf{dns}(\mu_i) < \delta$, we are guaranteed that there are not many such "bad" points, and therefore

$$\Pr_{\mathcal{D}_{\mathcal{X}}}[c(x) = g(x)] \geq 1 - \mathsf{dns}(\mu_i) > 1 - \delta.$$

Hence, we have found a circuit $c \in \mathcal{C}_{t,q}$ that contradicts the δ-weakly hardness of g, as required (this corresponds precisely to the conclusion of Theorem 4.31, IHCL contrapositive).

Following Impagliazzo's boosting proof, Kilvans and Servedio consider a general receipe for transforming boosting algorithms from learning theory into a Hardcore Lemma. They then plug into this recipe various smooth PAC-based boosting algorithms, where smooth boosting algorithms are such that all of the $\mathcal{D}_i$ distributions that are produced during the iterative algorithm never put

too much weight on any of the points $x \in \mathcal{X}$. They provide a general relationship between the smoothness parameter of the boosting algorithm and the density of the hardcore measure that it induces, namely [KS03, Thm. 14]:

Lemma 4.32. *The distributions $\mathcal{D}_i$ are $(1/d)$-smooth for all iterations i of the boosting algorithm if and only if $\mathsf{dns}(\mu) \geq d$.*

For each of the smooth boosting algorithms considered in [KS03], they obtain different circuit size and hardcore set density parameters. Specifically, they consider Freund's boost-by-majority algorithm [Fre95], Freund's $\mathrm{B}_{\mathrm{Filt}}$ algorithm [Fre95], and Freund's $\mathrm{B}_{\mathrm{Comb}}$ algorithm, which combines the previous two [Fre92]. However, none of these algorithms is able to obtain a hardcore set of density 2δ; all of their densities are suboptimal. For example, in the case of Freund's boost-by-majority algorithm [Fre95], Kilvans & Servedio show that this boosting algorithm is $O(1/\epsilon^3)$-smooth. Hence, the corresponding hardcore measure has density $\mu(M) = \Omega(\delta^3)$. Most of these boosting algorithms combine the various weak hypotheses using a majority vote. Holenstein was able to obtain 2δ density by using a more sophisticated aggregation rule in his proposed boosting algorithm for IHCL.

Some years later, Feldman revisited the relationship between boosting algorithms and hardcore set constructions [Fel09a], in his case in the agnostic setting (rather than in the realizable setting, as it was the case for [KS03]). Continuing the work of [KS03], and using a similar recipe for the transformation between boosting algorithms and hardcore set constructions, Feldman shows that hardcore set constructions that achieve the optimal hardcore set size of 2δ imply distribution-specific agnostic boosting algorithms and vice-versa. In the same paper, Feldman proposes an agnostic boosting algorithm that produces smooth $1/(2\delta - \gamma)$-smooth distributions at each iterative step of the boosting. Hence, by Lemma 4.32 and by the conversion between boosting algorithms and hardcore set constructions, it follows that his algorithm yields a hardcore set of density $2\delta - \gamma$ for any γ in time $\mathrm{poly}(1/\gamma)$.

In light of our results on the relationship between multiaccuracy, calibration, agnostic learning, and IHCL with density 2δ, a very interesting next direction is to understand exactly why the boosting algorithms considered by [KS03] only give sub-optimal δ density, whereas Feldman's agnostic boosting algorithm gives optimal 2δ density. For example, we can ask: Does calibration play a role in these different boosting algorithms?

A boosting-based perspective of IHCL++. Given the boosting interpretation of IHCL with density δ and 2δ, it is natural to try to understand the stronger IHCL++ statement from a boosting perspective. Specifically: Given that through multicalibration we now have the stronger IHCL++ hardcore theorem, what form of boosting does IHCL++ imply?

We use the characterization of multicalibration as hardness of prediction on each piece shown in [CDV24] (Lemma 4.29) to give an answer to this question. It turns out that this stronger form of learning corresponds to the recently introduced notion of *swap agnostic learning* by Gopalan, Kim, and Reingold [GKR24]. This connection is not so surprising in hindsight, given that they show an equivalence between swap agnostic learning and multicalibration, and we know that multicalibration gives us IHCL++. However, we present these connections (which lie at the top tier of our unifying picture between fairness, learning, and complexity, see Figure 1.2 in Chapter 1) through the notion

of hardness of prediction, which provides a clear interpretation of the dual relationship between hardcore set constructions and learning primitives through the notion of hardness of prediction.

In the previous section, we saw how we can view multicalibration in terms of hardness of prediction: namely, in a multicalibrated partition $\mathcal{P}$, for all $c \in \mathcal{C}$ and all large enough $P \in \mathcal{P}$, the function p^* is $(\mathcal{C}, b_P - 2\epsilon)$-hard on $\mathcal{D}_{\mathcal{X}}|_P$ (Lemma 4.29). In turn, this allowed [CDV24] to obtain IHCL++, by transforming the hardness of prediction in each $P \in \mathcal{P}$ into a small hardcore measure on each piece.

But we can also interpret hardness of prediction from a learning-theoretic perspective! The key idea is the following: if on each (large enough) $P \in \mathcal{P}$, p^* is $(\mathcal{C}, b - 2\epsilon)$-hard on $\mathcal{D}_{\mathcal{X}}$, if we now interpret $\mathcal{C}$ to be a hypothesis class rather than a collection of size-bounded circuits (returning to our various discussions in Chapter 2 about the various interpretations of $\mathcal{C}$), then this is equivalent to saying that no $c \in \mathcal{C}$ can guess p^* better than with error $b_P - 2\epsilon$. That is, in general, guessing the majority value is one natural strategy for trying to learn a function p^*. What Lemma 4.28 is showing is that, in the case where P corresponds to the piece of a multicalibrated partition, this is the *optimal* strategy. This is precisely the learning-theoretic interpretation of being hard to predict. We can also think of this as the computational analogue of irreducible noise.

Therefore, when we interpret Lemma 4.29 from a complexity-theoretic point of view, the hardness of g in each piece of an MC partition $\mathcal{P}$ yields a hardcore set within each $P \in \mathcal{P}$. When we interpret Lemma 4.28 from a learning-theoretic point of view, the hardness of g in each piece of an MC partition $\mathcal{P}$ implies that all hypothesis $c \in \mathcal{C}$ must incur error at least $b_P - 2\epsilon$. Moreover, our hypothesis h, which corresponds to a multicalibrated predictor (i.e., the MC partition $\mathcal{P}$ is obtained by taking the level sets of h), is also incurring this error on each piece, because by definition h is predicting v_P on each piece. Therefore, through multicalibration/IHCL++, we have achieved the following strong notion of learning the concept class $\mathcal{C}$: we output a hypothesis h that is piecewise constant and that which incurs optimal error on each piece of the partition, as measured in comparison to the best $c \in \mathcal{C}$ on each piece.

Moreover, given that we can prove the MC theorem using a weak agnostic learner, we can view this result as a form of boosting, where we use a weak agnostic learner to obtain agnostic learning++. This gives the answer to the question that we asked at the beginning of this section, namely: to what form of boosting does IHCL++ correspond to? Formally:

Theorem 4.33 (Multicalibration yields agnostic learning++)**.** *Let $\mathcal{C}$ be a concept class that is weak agnostically learnable. Given $p^*, \mathcal{D}, \epsilon$, there exists an efficient algorithm that produces a low-complexity partition $\mathcal{P}$ of the domain and a predictor h satisfying the following two properties over every (large enough) $P \in \mathcal{P}$:*

- *The predictor h is constant on each $P \in \mathcal{P}$.*
- $\mathrm{err}_{\mathcal{D}|_P}(\mathbf{y}, h(\mathbf{x})) \leq \mathrm{err}_{\mathcal{D}|_P}(\mathbf{y}, c_P^*) + \epsilon$, *where* $c_P^* = \arg\min_{c \in \mathcal{C}} \mathrm{err}_{\mathcal{D}|_P}(\mathbf{y}, c(\mathbf{x}))$.

Proof. We begin by calling the MC theorem from [HKRR18] with $\mathcal{X}, \mathcal{D}_{\mathcal{X}}, g$ and class of distinguishers $\mathcal{C}$ (Theorem 2.16). We obtain a $(\mathcal{C}, \epsilon)$-multicalibrated predictor p and obtain a partition $\mathcal{P}$ by taking the level sets of p, where p is equal to v_P on each $P \in \mathcal{P}$. Now we define the predictor h as the one outputting the majority value on each piece; namely, $h(x) = \mathbf{1}[v_P \geq 1/2]$ for all $x \in P$ and

for all $P \in \mathcal{P}$. (Note that this is essentially what $\mu_{\max}$ is doing in the context of IHCL.) We claim that this h and $\mathcal{P}$ satisfy the two properties of Theorem 4.33.

That the predictor is constant on each piece P (Condition 1) follows by construction. We now show that the error of h_P is optimal on each piece, where the optimal c_P^* concept in $\mathcal{C}$ depends on P (whereas h_P is single predictor). So, h_P competes with the best concept chosen tailored to each piece, realizing a strong form of agnostic learning.

By the definition of multicalibration, we know that

$$\mathop{\mathbb{E}}_{X_v \sim \mathcal{P}(\mathcal{D})} \left| \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}|_v} [c(\mathbf{x}) \cdot (p^*(\mathbf{x}) - v_P)] \right| \leq \epsilon.$$

for all $c \in \mathcal{C}$ and $P \in \mathcal{P}$. Thus, definitionally, it follows that for each $P \in \mathcal{P}$, the function p^* is $(\mathcal{C}, \epsilon)$-indistinguishable from the constant function v_P on each large enough $P \in \mathcal{P}$. By Lemma 4.29 (MC theorem as hardness of prediction), this implies that the function p^* is $(\mathcal{C}, b_P - 2\epsilon)$-hard on $\mathcal{D}|_P$, where $b_P = \min\{v_P, 1 - v_P\}$. By definition of hardness of functions, this means that

$$\Pr_{\mathcal{D}|_P}[c(\mathbf{x}) = \mathbf{y}] \leq 1 - (b_P - 2\epsilon) = 1 - b_P + 2\epsilon$$

for every $c \in \mathcal{C}$. Equivalently, stated in terms of error,

$$\mathrm{err}_{\mathcal{D}|_P}(\mathbf{y}, c) = \Pr_{\mathcal{D}|_P}[c(\mathbf{x}) \neq \mathbf{y}] > 1 - (1 - b_P + 2\epsilon) = b_P - 2\epsilon$$

for every $c \in \mathcal{C}$. Hence, in particular, the optimal classifier $c_P^* \in \mathcal{C}$ in the piece $\mathcal{D}|_P$ must also incur error at least $b_P - \epsilon$. On the other hand, our predictor h predicts the majority value on each piece $P \in \mathcal{P}$. Given that $\mathbf{y} \sim \mathrm{Bern}(p^*(\mathbf{x}))$, it follows that

$$\mathrm{err}_{\mathcal{D}|_P}(\mathbf{y}, h(\mathcal{X})) = \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}|_P} [|\mathbf{y} - h_P(\mathbf{x})|] = b_P$$

(This is indeed the same analysis that we performed for $\mu_{\max}$.) Therefore, the agnostic guarantee follows. □

It turns out that the general version of this notion of agnostic boosting++ corresponds precisely to the notion of *swap agnostic learning* that was recently introduced by Gopalan, Kim, and Reingold, which is inspired by the traditional notions of swap regret, where we switch the order of quantifiers.

Definition 4.34 (Swap agnostic learning [GKR24])**.** *For a loss function ℓ, hypothesis class $\mathcal{C}$, and error $\epsilon \geq 0$, a hypothesis h is a* $(\ell, \mathcal{C}, \epsilon)$-swap agnostic learner *if*

$$\mathbb{E}[\ell(y, h(x))] \leq \mathop{\mathbb{E}}_{v \sim \mathcal{D}_h} \left[\min_{c_v \in \mathcal{C}} \ell(y, c_v(x) \mid h(x) = v) \right] + \epsilon.$$

Here, the predictor first chooses h, and then the adversary is allowed to choose the optimal concept $c_v \in \mathcal{C}$ afterwards for each v in the range of h. Still, our single hypothesis h is able to achieve agnostic learning on each of the level sets v.

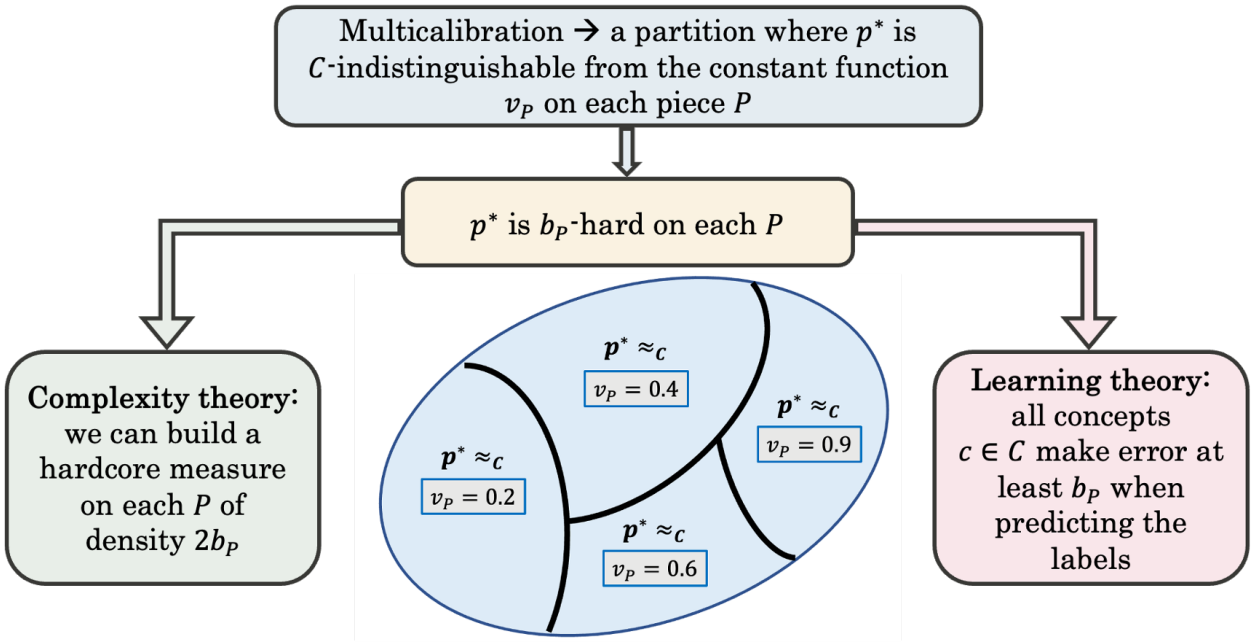


Figure 4.4: Duality of hardcore set constructions and agnostic learning from hardness of prediction.

Indeed, swap agnostic learning strengthens standard agnostic learning, where the predictor only competes against the single best hypothesis $c \in \mathcal{C}$. We can view this primitive as *agnostic learning++*, where we reproduce the usual primitive of an agnostic learner both globally and locally on every level set of h. Gopalan et al. show that we can obtain swap agnostic learning for the ℓ_2 loss from (swap) multicalibration. Note that in Theorem 4.33, we instead obtained it for the ℓ_1 loss. Therefore, given that we can obtain multicalibration from weak agnostic learning, we can view the boosting process of getting swap agnostic learning from weak agnostic learning as the boosting interpretation of IHCL++. Moreover, the partition is "maximally boosted", in the sense that we cannot further use the weak agnostic learner to extract more learnability.

We summarize this dual interpretation of hardness of prediction as yielding both hardcore set constructions and agnostic learning in Figure 4.4. Namely, we can view multicalibration as giving as a low-complexity partition such that p^* is $\mathcal{C}$-hard to predict on every piece. Then, we can take this hardness down the complexity route, where it enables us to construct a hardcore measure on every piece (yielding IHCL++), or down the learning route, where it enables us to perform agnostic learning on every piece (yielding agnostic learning++).

5
Beyond Multiaccuracy

We prove strong noise-tolerance properties of a potential-based boosting algorithm, similar to MadaBoost and SmoothBoost. Our analysis is in the agnostic framework of Kearns, Schapire and Sellie (1994), giving polynomial-time guarantees in presence of arbitrary noise.

Kalai & Kanade [KK09]

Having placed the construction of dense hardcore measures into our unified picture of learning, fairness, and complexity, we return to the relationship between multiaccuracy and weak agnostic learning that we explored in Chapter 3. We begin by recalling the main results that we presented in Chapter 3:

- A multiaccurate predictor does not necessarily yield weak agnostic learning.
- Multiaccuracy gives restricted weak agnostic learning.
- Multiaccuracy and global calibration yield strong agnostic learning.

The fact that a multiaccurate predictor itself, even when we post-process its outputs in any way, does not yield weak agnostic learning does not close the possibility of obtaining learning through more sophisticated post-processings of the multiaccurate predictor.

In this chapter, we explore various other directions for trying to obtain some form of learning from multiaccuracy. First, in Section 5.1, we explore whether being able to project onto the class $\mathcal{C}$ yields a weak learner. Second, in Section 5.2, we study the power of restricted weak agnostic learning. Third, in Section 5.3 we study the problem of learning versus auditing for multiaccuracy.

5.1 Projecting multiaccurate predictors onto the span of $\mathcal{C}$

Recall that in Chapter 3 we explained how we can decompose the correlation between a $\mathcal{C}$-multiaccurate predictor p and the labels as the correlation on the span of $\mathcal{C}$ and on the orthogonal space to $\mathsf{span}(\mathcal{C})$

(i.e., the set of all finite linear combinations of the concepts $c \in \mathcal{C}$). Namely, in Section 3.2, we showed that, if p is $(\mathcal{C}, 0)$-multiaccurate, then

$$\text{cor}(\mathbf{y}, p_{\mathcal{C}}) = \langle q^*_{\mathcal{C}}, q_{\mathcal{C}} \rangle + \langle q^*_{\perp}, q_{\mathcal{C}} \rangle = \|q^*_{\mathcal{C}}\|_2^2 = \|q_{\mathcal{C}}\|_2^2, \tag{5.1}$$

where $q^*_{\mathcal{C}}, q_{\mathcal{C}}$ are the components of $q^* = 2p^* - 1, q = 2q - 1$ onto $\mathsf{span}(\mathcal{C})$, respectively, and $q^*_{\perp}, q_{\perp}$ are the components of q^*, q orthogonal to $\mathsf{span}(\mathcal{C})$.

Suppose that given a $\mathcal{C}$-multiaccurate predictor, we could effectively project it onto $\mathsf{span}(\mathcal{C})$, obtaining $p_{\mathcal{C}}$. We can effectively see this as a more sophisticated form of post-processing of the predictor p than the ones we have considered so far. Could $p_{\mathcal{C}}$ be a weak agnostic learner for $\mathcal{C}$? The reason for why we might believe this is that, from Equation 5.1, we know that $\text{cor}(\mathbf{y}, p_{\mathcal{C}}) = \|q_{\mathcal{C}}\|_2^2$. Hence, if $\|q^*_{\mathcal{C}}\|_2^2 > 0$, then $p_{\mathcal{C}}$ would be a weak learner.

There are two potential problems that arise. First, $q_{\mathcal{C}}$ need not have range $[-1, 1]$, and it is not even bounded necessarily. Second, in practice we do not have $(\mathcal{C}, 0)$-multiaccuracy, as we assumed in Equation 5.1, but rather $(\mathcal{C}, \tau)$-multiaccuracy. Depending on the representation of $p_{\mathcal{C}}$ in terms of the vectors in $\mathsf{span}(\mathcal{C})$, the errors of the projection could accumulate. Still, we are able to show that, if the projection of p onto $\mathcal{C}$ is ℓ_1 sparse, then we do get weak agnostic learning.

First of all, what do we mean by a projection onto $\mathsf{span}(\mathcal{C})$? We mean that we can write p as a linear combination of $c \in \mathcal{C}$. In general, we cannot assume that we can achieve a "perfect" projection onto $\mathcal{C}$, and so what we want is to find an element in $\mathsf{span}(\mathcal{C})$ that is close to p. For our purposes, it will be enough to find an element $r \in \mathsf{span}(\mathcal{C})$ that satisfies the following:

Definition 5.2 (Projection onto $\mathcal{C}$)**.** *Given a predictor p and a class $\mathcal{C}$, we say that $r = \sum_s \lambda c_s$ for $c_s \in \mathcal{C}$ is a γ-*projection of p onto $\mathsf{span}(\mathcal{C})$ *if*

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}}[((2p(\mathbf{x}) - 1) - r(\mathbf{x}))^2] \leq \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}}[(2p(\mathbf{x}) - 1)^2] - \gamma.$$

Note that we can view the RHS as the squared loss obtained by a predictor that does random guessing (i.e., predicts 1/2 everywhere). In this sense, we can view Definition 5.2 as saying that the projection r is performing non-trivial squared loss minimization, as compared to the baseline of a random predictor.

Does this mean that r can give us learning? Specifically, how does weak learning relate to non-trivial squared loss minimization? As we now show, we can in fact view weak learning precisely as squared loss minimization. This provides a very elegant characterization of weak learning.

5.1.1 Weak learning as squared loss minimization

In one direction, we already know that weak learning implies non-trivial squared loss minimization. Indeed, this is exactly what is happening in the multigroup fairness algorithms! As we summarized in Chapter 2, the multigroup fairness algorithms perform the following iterative process: first, we call a weak learner for the class $\mathcal{C}$; if it succeeds, then the weak agnostic learner gives us a function $h : \mathcal{X} \to [-1, 1]$ that has correlation β with the labels. That is,

$$\text{cor}(\mathbf{y}, h(\mathbf{x})) = \mathbb{E}[h(\mathbf{x})(2\mathbf{y} - 1)] \geq \beta$$

Note that in the algorithm, we actually use the residuals $\mathbf{y} - p$ as the labels, but here we write them as $\mathbf{y}$. Then, we perform a gradient update on our predictor p using this hypothesis h, and show that this update improves the squared loss of our predictor p by at least some amount. Specifically, [HKRR18] shows that the update $p(x) = (1 + \beta c(x))/2$ satisfies

$$\begin{aligned}
\mathbb{E}[(\mathbf{y} - p(\mathbf{x}))^2] &= \mathbb{E}\left[\left(\mathbf{y} - \frac{1+\beta c(\mathbf{x})}{2}\right)^2\right] \\
&= \mathbb{E}\left[\left(\mathbf{y} - \frac{1}{2}\right)^2\right] - \beta\,\mathbb{E}[c(\mathbf{x})(2\mathbf{y}-1)] + \frac{\beta^2}{4}\,\mathbb{E}[c(\mathbf{x})^2] \\
&\leq \mathbb{E}\left[\left(\mathbf{y} - \frac{1}{2}\right)^2\right] - \frac{3\beta^2}{4}.
\end{aligned}$$

In other words, weak agnostic learning enables non-trivial squared loss error, when compared to random guessing (i.e., the constant 1/2 predictor).

Note that in Section 3.3, when showing that multiaccuracy and calibration imply strong agnostic learning, we also used the distance between p and the constant 1/2 predictor as a baseline. That is, we argued that calibrated predictor is correlated with the labels at least as well as the constant 1/2 predictor, and then showed the multiaccuracy and the existence of a $c \in \mathcal{C}$ that is correlated with the labels implies that the multiaccurate predictor p is far from the constant 1/2 predictor. In that case, we used the ℓ_1 norm; here we are using the ℓ_2 norm.

In the other direction, we also have that any function $h : \mathcal{X} \to \mathbb{R}$ which beats random guessing in squared error gives a weak agnostic learner. A similar result was shown in the work of [BFJ+94], who used a more complicated way of mapping h on to the $[-1, 1]$ interval.

Theorem 5.3. *Let $h : \mathcal{X} \to \mathbb{R}$ satisfy*

$$\mathbb{E}[(\mathbf{y} - h(\mathbf{x}))^2] \leq \mathbb{E}\left[\left(\mathbf{y} - \frac{1}{2}\right)^2\right] - \gamma.$$

Define the function $\bar{h}$ which truncates h to $[0, 1]$, and let $g(x) = 2\bar{h}(x) - 1 \in [-1, 1]$. Then $\mathrm{cor}(\mathbf{y}, g(\mathbf{x})) \geq 2\gamma$.

Proof. Since truncation to the range $[0, 1]$ only reduces the squared loss, we have

$$\mathbb{E}[(\mathbf{y} - \bar{h}(\mathbf{x}))^2] \leq \mathbb{E}\left[\left(\mathbf{y} - \frac{1}{2}\right)^2\right] - \gamma. \tag{5.4}$$

We can write

$$\begin{aligned}
\mathbb{E}[(2\mathbf{y} - 2\bar{h}(\mathbf{x}))^2] &= \mathbb{E}[(2\mathbf{y} - 1 - g(x))^2] \\
&= \mathbb{E}[(2\mathbf{y}-1)^2] - 2\,\mathbb{E}[g(\mathbf{x})(2\mathbf{y}-1)] + \mathbb{E}[g(\mathbf{x})^2].
\end{aligned}$$

Rearranging terms, we get

$$\begin{aligned}\mathbb{E}[g(\mathbf{x})(2\mathbf{y}-1)] &= \frac{1}{2}\left(\mathbb{E}[(2\mathbf{y}-1)^2] - \mathbb{E}[(2\mathbf{y}-2\bar{h}(\mathbf{x}))^2] + \mathbb{E}[g(\mathbf{x})^2]\right)\\ &\geq 2\left(\mathbb{E}\left[\left(\mathbf{y}-\frac{1}{2}\right)^2\right] - \mathbb{E}[(\mathbf{y}-\bar{h}(\mathbf{x}))^2]\right)\\ &\geq 2\gamma.\end{aligned}$$

where we use Equation (5.4). The claim follows since the LHS is equal to $\mathrm{cor}(\mathbf{y}, g(\mathbf{x}))$. □

Together, these results tell us that weak agnostic learning is equivalent to learning a predictor whose squared loss beats random guessing. This equivalence between learning a predictor versus satisfying a decision problem with respect to a random predictor is somewhat reminiscent of the result of [KL18] who show that weak agnostic learning for a class $\mathcal{C}$ is equivalent to the *refutation problem* of distinguishing between (a) labeled distributions where $\mathrm{cor}(\mathbf{y}, c(\mathbf{x})) \geq \alpha$ for some hypothesis $c \in \mathcal{C}$ from (b) the distribution where $\mathbf{y}$ consists of random bits. In the latter case, random guessing is in fact optimal. We will explore this *learning versus refutation* problem in Section 5.3.

5.1.2 Weak agnostic learning from sparse projections

Having established the equivalence between weak agnostic learning and performing non-trivial squared loss minimization, we return to the question of whether r, namely the γ-projection of p onto $\mathsf{span}(\mathcal{C})$, is a weak learner.

First, the a potential learner cannot have unbounded range. So we begin by clipping r to the range $[-1, 1]$. That is, if $r \geq 1$ then we map it to 1, if $r \leq -1$ we map it to -1, and otherwise we map r to itself. Then we use the decomposition of r into $\sum_s \lambda c_s$ with $c_s \in \mathcal{C}$ to apply the multiaccuracy condition s times to each c_s, so that we can swap p for $\mathbf{y}$ in the closeness guarantee satisfied by a γ-projection of p onto $\mathsf{span}(\mathcal{C})$. That is:

Theorem 5.5. *Let $p : \mathcal{X} \to [0,1]$ satisfy $(\mathcal{C}, \tau)$-multiaccuracy for $\mathcal{D}$. Suppose that $r = \sum_s \lambda c_s$ for $c_s \in \mathcal{C}$ is a γ-projection of p onto $\mathsf{span}(\mathcal{C})$. Let Let $h = \mathrm{clip}(r)$, which is clipped to have $h(\mathbf{x}) \in [-1, 1]$ for all $\mathbf{x} \in \mathcal{X}$. Then,*

$$\mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[((2\mathbf{y}-1) - h(\mathbf{x}))^2] \leq \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[(2\mathbf{y}-1)^2] - \gamma + 4\tau\sum_s |\lambda_s|.$$

Proof. As usual, let $q = 2p - 1$ denote the $[-1, 1]$-version of p. Then,

$$\begin{aligned}\gamma &\leq \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_\mathcal{X}}[q(\mathbf{x})^2] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_\mathcal{X}}[(q(\mathbf{x}) - r(\mathbf{x}))^2]\\ &= 2\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_\mathcal{X}}[q(\mathbf{x})r(\mathbf{x})] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_\mathcal{X}}[r(\mathbf{x})^2]\\ &= 2\mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[(2\mathbf{y}-1)r(\mathbf{x})] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_\mathcal{X}}[r(\mathbf{x})^2] + 2\mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[r(\mathbf{x})(q(\mathbf{x}) - (2\mathbf{y}-1))]\\ &= 2\mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[(2\mathbf{y}-1)r(\mathbf{x})] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_\mathcal{X}}[r(\mathbf{x})^2] + 4\mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[r(\mathbf{x})(p(\mathbf{x}) - \mathbf{y})]\\ &= 2\mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[(2\mathbf{y}-1)r(\mathbf{x})] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_\mathcal{X}}[r(\mathbf{x})^2] + 4\sum_s \lambda_s \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[c_s(\mathbf{x})(p(\mathbf{x}) - \mathbf{y})]\end{aligned}$$

$$\leq \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[(2\mathbf{y}-1)^2] - \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[((2\mathbf{y}-1)-r(\mathbf{x}))^2] + 4\tau\sum_s |\lambda_s|$$
$$\leq \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[(2\mathbf{y}-1)^2] - \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[((2\mathbf{y}-1)-h(\mathbf{x}))^2] + 4\tau\sum_s |\lambda_s|.$$

In the second last inequality, we use the fact that p is $(\mathcal{C},\tau)$-multiaccurate for $\mathcal{D}$ and that $c_s \in \mathcal{C}$ for all s, so we apply the multiaccuracy condition s times. This is why we have to be careful with the accumulation of the errors. In the last inequality we use that *clipping* r to get $h \in [-1,1]$ only reduces the squared error with respect to $2\mathbf{y}-1$. The conclusion then follows by rearranging. □

Theorem 5.5 is saying that $h = \text{clip}(r)$ is doing non-trivial squared loss minimization. This is what we want, given that in Section 5.1.1 we showed that non-trivial squared loss minimization enables weak learning. Therefore, if we did not have the error term $4\tau\sum_s |\lambda_s|$ in Theorem 5.5, by Theorem we would conclude that if r is a γ-projection of p onto $\mathsf{span}(\mathcal{C})$, then $h = \text{clip}(r)$ has correlation at least 2γ with the labels.

Due to the error term, we have to relax this result to ℓ_1 sparse projections, as to ensure that the accumulated error term $4\tau\sum_s |\lambda_s|$ remains small. This gives us the final theorem:

Theorem 5.6. *Let $p : \mathcal{X} \to [0,1]$ satisfy $(\mathcal{C},\tau)$-multiaccuracy for $\mathcal{D}$. Suppose we have $r = \sum_s \lambda_s c_s$ for $c_s \in \mathcal{C}$ such that r is a γ-projection of p onto $\mathsf{span}(\mathcal{C})$ satisfying $4\tau\sum_s |\lambda_s| \leq \gamma - \beta$ (ℓ_1 sparsity) Then, $h = \text{clip}(r)$, which clips q to the range $[-1,1]$, satisfies $\text{cor}_{\mathcal{D}}(\mathbf{y},h) \geq 2\beta$.*

Proof. By Theorem 5.5, we can apply the multiaccuracy guarantee on each c_s to obtain

$$\mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[((2\mathbf{y}-1)-h(\mathbf{x}))^2] \leq \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[(2\mathbf{y}-1)^2] - \gamma + 4\tau\sum_s |\lambda_s|.$$

By the sparsity guarantee, it follows that

$$\mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[((2\mathbf{y}-1)-h(\mathbf{x}))^2] \leq \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[(2\mathbf{y}-1)^2] - \beta.$$

That is, h performs non-trivial squared loss minimization. Then, from our theorem showing that non-trivial squared loss minimization gives a weak agnostic learner (Theorem 5.3), it follows that $\text{cor}_{\mathcal{D}}(\mathbf{y},h) \geq 2\beta$, as we wanted to show. □

5.1.3 Multiaccuracy as a query oracle

However, we do not know how to efficiently find ℓ_1 sparse projections onto an arbitrary class $\mathcal{C}$. (Otherwise, we would have the result that multiaccuracy always yields learning.) Still, can we find a concrete example where we can efficiently find such a projection? Yes! An example is in fact given by the well-known Goldreich-Levin algorithm for finding all the large Fourier coefficients of a given function [GL89]. We provide a brief summary of the Goldreich-Levin algorithm (also known as the Kushilevitz-Mansour algorithm [KM91]) in the setting of Fourier analysis.

The Goldreich-Levin algorithm. By the Fourier expansion theorem we know that every function $f : \{-1,1\} \to \mathbb{R}$ can be uniquely expressed as a multilinear polynomial:

$$f(x) = \sum_{S \subseteq [n]} \hat{f}(S)\chi_S,$$

where $\chi_S(x) = \prod_{i \in S} x_i$ for $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and $\hat{f}(S) = \mathbb{E}_{x \sim \{-1,1\}^n}[f(x)\chi_S]$ denotes the Fourier coefficient of f on S. Recall that we define the inner product $\langle f, g \rangle$ between two functions $f, g : \{-1,1\}^n \to \mathbb{R}$ as $\mathbb{E}_{\mathbf{x} \sim \{-1,1\}^n}[f(\mathbf{x})g(\mathbf{x})]$. Thus we see that the 2^n parity functions $\chi_S : \{-1,1\}^n \to \{-1,1\}$ form an orthonormal basis for the vector space of functions $\{-1,1\}^n \to \mathbb{R}$ [ODo14, Thm. 1.5].

Given *query access* to a function f (i.e., if we can query f at any chosen point), the Goldreich-Levin algorithm allows us to efficiently find its heavy Fourier coefficients. That is, for a function $f(x) = \sum_{S \subseteq [n]} \hat{f}(S)\chi_S$, it finds the Fourier coefficients $\hat{f}(S)$ that are larger than some chosen threshold γ. Specifically:

Theorem 5.7 (Goldreich-Levin [GL89], Kushilevitz-Mansour [KM91])**.** *Given query access to $f : \{-1,1\}^n \to [-1,1]$, and given $\gamma, \delta > 0$, there is a poly$(n, \frac{1}{\gamma}, \log \frac{1}{\delta})$-time algorithm that outputs a list $L = \{S_1, \ldots, S_m\}$ such that, with probability δ:*

1. *If $\hat{f}(S) \geq \gamma$, then $S \in L$, and*
2. *If $S \in L$, then $\hat{f}(S) \geq \frac{\gamma}{2}$.*

Learning parities with noise. Now let's consider the well-known problem in learning theory of *learning parities with noise* (LPN). Given $S \subseteq [n]$, a parity function is a function $\chi_S : \{0,1\}^n \to \{0,1\}$ with $\chi_S(x) = \oplus_{i \in S} x_i$. Consider the concept class $\mathcal{C}$ of all parities. Suppose that we want to learn the class of parities in the agnostic setting (i.e., with adversarial noise). It is known that under the uniform distribution, learning parities with adversarial classification noise reduces to learning parities with random classification noise [FGKP06], where the true labels are flipped independently with probability η (here η represents the noise).

A celebrated result in learning theory is that, if the learner is allowed to ask membership queries, then the Goldreich-Levin algorithm gives a polynomial time algorithm for learning parity with adversarial noise under the uniform distribution [FGKP06]. For our purposes, we can view the Goldreich-Levin algorithm as both giving us a ℓ_1 sparse projection in the sense that we required in Section 5.1.2, and it also gives a proper agnostic learning algorithm for the class of parities.

Let's return to our multiaccuracy setting. Another key idea that we can explore when thinking about how much power a multiaccurate predictor p can buy us, is the fact that we have query access to p. While a priori we do not have query access to the ground truth predictor p^*, one could hope that, given that p and p^* are close in the $\mathcal{C}$-multiaccuracy sense, we could translate our query access to p to a form of query access to p^*. Indeed, we can show:

Theorem 5.8. *Suppose the class $\mathcal{C}$ is efficiently properly agnostically learnable with access to random examples from $\mathcal{D}$ and query access to the target function to accuracy ε. Moreover, suppose that we have access to a (C, τ)-multiaccurate predictor p for $\mathcal{D}$. Then, $\mathcal{C}$ is efficiently properly agnostically learnable using only access to random examples from $\mathcal{D}$ (and using the predictor p) to accuracy $\varepsilon + 3\tau$.*

Proof. Here we will assume that both the target function p^* and the concepts in $\mathcal{C}$ take values in $\{0, 1\}$. Let $\mathcal{A}$ be the algorithm for efficiently agnostically learning $\mathcal{C}$ using the random examples from $\mathcal{D}$ and query access to the target function, and let p be a $(\mathcal{C}, \tau)$-multiaccurate predictor for $\mathcal{D}$.

To show the desired reduction, our goal is to show that we can simulate the query access to the target function by using query access to our multiaccurate predictor p instead, to which we actually have access to (unlike to p^*). Whenever the algorithm asks for a random example, we replace $(\mathbf{x}, \mathbf{y})$ by $(\mathbf{x}, \mathrm{Bern}(p(\mathbf{x}))$, where we take a fresh independent draw from a Bernoulli random variable with parameter $p(\mathbf{x})$ at each time. Whenever the algorithm asks a membership query, we simulate them by returning $\mathrm{Bern}(p(x))$ when queried at point x.

At the end of the learning process, by the assumption that $\mathcal{A}$ knows how to agnostically learn $\mathcal{C}$ using the actual random examples from $\mathcal{D}$ and query access to p^*, with our simulated queries it returns a hypothesis $c_{\mathcal{A}} \in \mathcal{C}$ that with high probability satisfies

$$\mathrm{err}(c_{\mathcal{A}}; \mathcal{D}) \leq \min_{c \in \mathcal{C}} \mathbb{E}[|c(\mathbf{x}) - p(\mathbf{x})|] + \varepsilon.$$

Next, we want to relate this error guarantee on $c_{\mathcal{A}}$ with the usual agnostic learning guarantee, to conclude that we have agnostically learnt $\mathcal{C}$. To do so, we use the fact that p is $(\mathcal{C}, \tau)$-multiaccurate.

Specifically, for the target function $p^* : \mathcal{X} \to \{0, 1\}$ we have that

$$\begin{aligned}
\Pr[c(\mathbf{x}) \neq p^*(\mathbf{x})] &= \mathbb{E}[c(\mathbf{x})(1 - p^*(\mathbf{x})) + (1 - c(\mathbf{x}))p^*(\mathbf{x})] \\
&= \mathbb{E}[c(\mathbf{x}) + p^*(\mathbf{x})] - 2\,\mathbb{E}[c(\mathbf{x})p^*(\mathbf{x})] \\
&\leq \mathbb{E}[c(\mathbf{x}) + p(\mathbf{x})] - 2\,\mathbb{E}[c(\mathbf{x})p(\mathbf{x})] + 3\tau \\
&= \mathbb{E}[c(\mathbf{x})(1 - p(\mathbf{x})) + (1 - c(\mathbf{x}))p(\mathbf{x})] + 3\tau \\
&= \mathbb{E}[|c(\mathbf{x}) - p(\mathbf{x})|] + 3\tau.
\end{aligned}$$

The first equality follows by a simple algebraic identity (similar to the ones we used in Chapter 4). The inequality follows by applying the multiaccuracy guarantee on both expectations; for the one in the left term we use the fact that $\mathbf{1} \in \mathcal{C}$, and so we can write $p^*(\mathbf{x})$ as $\mathbf{1} \cdot p^*(\mathbf{x})$ and then apply the multiaccuracy guarantee on it. Symmetrically, we can show that

$$\Pr[c(\mathbf{x}) \neq p^*(\mathbf{x})] \geq \mathbb{E}[|c(\mathbf{x}) - p(\mathbf{x})|] - 3\tau.$$

Thus, since this holds for all $c \in \mathcal{C}$, by minimizing over $\mathcal{C}$ and using the fact that $\mathrm{err}_{\mathcal{D}}(c) = \Pr_{\mathcal{D}}[c(\mathbf{x}) \neq p^*(\mathbf{x})]$ it follows that

$$\mathrm{err}_{\mathcal{D}}(c_{\mathcal{A}}) \leq \min_{c \in C} \mathrm{err}_{\mathcal{D}}(c) + \varepsilon + 3\tau.$$

Therefore, $\mathcal{A}$ has efficiently agnostically properly learnt $\mathcal{C}$ using only access to random examples

from $\mathcal{D}$ (while we used the multiaccurate predictor p to simulate the two types of queries), as we wanted to show. □

This result, together with the fact that the Goldreich-Levin algorithm gives a proper agnostic learning algorithm for the class of parities if membership query access is allowed, implies the following corollary:

Lemma 5.9. *Let $\mathcal{C}$ be the class of parities, and let $\mathcal{D}$ be such that such that the marginal distribution over $\mathcal{X}$ is uniform. Then, finding a $(\mathcal{C}, \tau)$-multiaccurate predictor p for $\mathcal{D}$ and $\tau \leq 0.1$ is at least as hard as properly agnostically learning $\mathcal{C}$ with random classification noise with noise rate* 0.1.

Proof. Suppose that we have access to a $(\mathcal{C}, \tau)$-multiaccurate predictor p for $\mathcal{D}$ and $\tau \leq 0.1$. As we have discussed, the class $\mathcal{C}$ of parities is efficiently properly agnostically lernable with access to random examples from $\mathcal{D}$ and query access to the target function by using the Goldreich-Levin algorithm. Because of this, and since we have access to a multiaccurate predictor, Theorem 5.8 applies, and thus we are able to learn $\mathcal{C}$ to accuracy $\min_{c \in \mathcal{C}} \mathrm{err}_{\mathcal{D}}(c) + \epsilon + 3\tau$.

Let χ_S be the target parity. In the random classification noise model, the labels are generated as follows: with probability η, we keep the true label χ_S, and with probability $1 - \eta$ we flip it. Then, the target parity has correlation $\mathbb{E}[\mathbf{y}\chi_S] = 1 - 2\eta$ with the labels. In this case, because the noise rate is $\eta = 0.1$, this means that the optimal error is at most 0.1.

Then, by Theorem 5.8, using our multiaccurate predictor we are able to find a parity c that has error at most

$$\min_{c \in \mathcal{C}} \mathrm{err}_{\mathcal{D}}(c) + \epsilon + 3\tau \leq 0.1 + \epsilon + 3 \cdot 0.1 = 0.4 + \epsilon < 1/2.$$

Moreover, in the random classification noise we have that all parities, except for the target parity χ_S, are uncorrelated with the labels. Indeed, given subsets $S, T \subseteq [n]$, we have that

$$\mathbb{E}[\chi_S \cdot \eta \cdot \chi_T] = \begin{cases} 1 - 2\eta & \text{if } S = T, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, the fact that we have found a parity that has error bounded away from $1/2$ means that the parity that we have found is *not* uncorrelated with the labels, and so we have in fact found the target parity χ_S. Therefore, we have learnt the class of parities in the random classification noise model. □

Since LPN is widely believed to be hard, we can conclude that the existence of an efficient algorithm that produces a p that is $(\mathcal{C}, \tau)$-multiaccurate predictor for small values of τ for the class $\mathcal{C}$ of parities is unlikely to exist, at least when the marginal distribution over $\mathcal{X}$ is uniform.

5.2 Restricted weak agnostic learning

So far in this chapter, we have tried to obtain learning from multiaccuracy by trying to use a multiaccurate predictor in various other ways, such as by projecting it onto $\mathcal{C}$ or by using it as a query oracle.

Recall that we showed that a multiaccurate predictor does yield a more restricted notion of weak learning, which we called *restricted weak agnostic learning*, where the learning promise is only required to hold if α is at least some value γ_0. Specifically, in Theorem 3.4 in Section 3.2 we showed that a $\mathcal{C}$-multiaccurate predictor p can be viewed as a $(2\alpha - 1)$-weak learner when $\alpha > 1/2$. Hence, this gives non-trivial learning only when $\alpha > 1/2$, where α is the maximum correlation between the labels and the concepts in $\mathcal{C}$. More generally, note that the task of (α, β)-weak agnostic learning gets harder as:

- α decreases, since this corresponds to detecting lower correlation.
- β increases, since this corresponds to learning a better hypothesis.

The typical notion of boosting refers to the β parameter: indeed, these kind of statements use a weak agnostic learner for $\mathcal{C}$ to construct a strong agnostic learner for $\mathcal{C}$. We know that this type of boosting is possible in the agnostic setting: indeed, [KK09; Fel09a] showed that an (α, β) weak learner implies a $(\gamma, \gamma - \alpha)$ weak learner for all $\gamma > \alpha$. This result is typically interpreted to say that if weak agnostic learning is possible for α approaching 0, then strong agnostic learning is possible.

In our notion of restricted weak agnostic learning, however, the parameter that we would like to improve is α. This is because, starting from multiaccuracy, we can only obtain a weak learner for $\alpha > 1/2$. Therefore, if we could reduce the value of α (which we can view as a form of boosting, given that reducing the value of α makes the learning problem harder), then we could hope to get a stronger learning guarantee from the assumption that we can efficiently construct a multiaccurate predictor for a class $\mathcal{C}$. Unfortunately, as we will see in this section, this type of boosting is not possible in general by providing a concrete counter-example.

We begin by explaining how we can view restricted weak agnostic learning as a form of approximate agnostic learning. This will allow us to already give an example (namely, using the class of halfspaces) where doing restricted weak agnostic learning is feasible for some values of α, but it becomes computationally unfeasible as $\alpha \to 0$. Then, in Section 5.2.2 we will give a tight example.

5.2.1 Approximation algorithms

Even though we come up with the notion of *restricted weak agnostic learning*, we give some more justification for why this is a natural concept in learning. In particular, it is closely related to the notion of *approximate agnostic learning*, which has been extensively studied in the learning theoretic literature [FGKP06; DKK+21; Dan15]. We remark that the results in this section are only meant to provide a greater understanding of the concept of a restricted weak agnostic learning by drawing a connection to a previously-studied notion in learning theory, but the approximation parameters stated in our lemmas follow directly from Daniely's results on halfspaces [Dan15].

Specifically, approximation algorithms in the agnostic setting are defined as follows:

Definition 5.10. *Given a concept class $\mathcal{C}$, distribution $\mathcal{D}$, and parameter $\mu \geq 0$, we say that an algorithm $\mathcal{A}$ is a* $(1+\mu, \epsilon)$-approximation algorithm *for $\mathcal{C}, \mathcal{D}$ if it outputs a predictor $h : \mathcal{X} \to [-1, 1]$ such that* $\operatorname{err}(h, \mathcal{D}) \leq (1 + \mu) \cdot \left(\min_{c \in \mathcal{C}} \operatorname{err}(c, \mathcal{D}) \right) + \epsilon$ *for any $\epsilon > 0$. We call such an h a* $(1 + \mu)$-approximation predictor.

Any such approximate agnostic learning algorithm can be viewed as a restricted weak agnostic learner. In particular, a guarantee of the form where the error is bounded by $(1 + \mu)\mathsf{OPT}$ is only

meaningful when $\mathsf{OPT} < \frac{1}{2(1+\mu)}$; in such cases, these algorithms yield restricted weak agnostic learners, as the lemma below shows.

Lemma 5.11. *Any $(1+\mu,\epsilon)$-approximation algorithm for a concept class $\mathcal{C}$ and distribution $\mathcal{D}$ yields an $(1-\frac{1}{1+\mu}+\gamma,(1+\mu)\gamma-2\epsilon)$-weak agnostic learner for $\mathcal{C}$ for any parameters $\gamma,\epsilon,\mu\geq 0$.*

Proof. Let $c^* \in \mathcal{C}$ be a concept in $\mathcal{C}$ achieving the minimal error with respect to $\mathcal{D}$. Given that the error of a predictor is always trivially upper bounded by $1/2$, it follows that for the $(1+\mu)$-approximation guarantee to be non-trivial, we must have $\mathrm{err}(c^*;\mathcal{D}) < \frac{1}{2(1+\mu)}$. By the definition of $\mathrm{err}(c^*;\mathcal{D})$, it follows that $\mathrm{cor}_{\mathcal{D}}(\mathbf{y},c^*) = 1-2\mathrm{err}(c^*;\mathcal{D})$, and hence $\mathrm{cor}_{\mathcal{D}}(\mathbf{y},c^*) \geq 1-\frac{1}{1+\mu}$. Suppose further that $\mathrm{cor}_{\mathcal{D}}(\mathbf{y},c^*) \geq 1-\frac{1}{1+\mu}+\gamma$ for any choice of $\gamma\geq 0$; equivalently,

$$\mathrm{err}(c^*;\mathcal{D}) = \frac{1-\mathrm{cor}_{\mathcal{D}}(\mathbf{y},c^*)}{2} \leq \frac{1-\left(1-\frac{1}{1+\mu}+\gamma\right)}{2} = \frac{\frac{1}{1+\mu}-\gamma}{2} = \frac{1}{2(1+\mu)} - \gamma/2.$$

Let h be the output predictor of a $(1+\mu)$-approximation algorithm for $\mathcal{C},\mathcal{D}$. Then, by the approximation guarantee on h, for any choice of $\epsilon\geq 0$ it follows that

$$\begin{aligned}
\mathrm{cor}_{\mathcal{D}}(\mathbf{y},h) = 1-2\mathrm{err}(h;\mathcal{D}) &\geq 1-2\big((1+\mu)\cdot\mathrm{err}(c^*;\mathcal{D})+\epsilon\big)\\
&\geq 1-2\left((1+\mu)\cdot\left(\frac{1}{2(1+\mu)}-\gamma/2\right)+\epsilon\right)\\
&= 1-2\left(\frac{1}{2}-\frac{(1+\mu)\gamma}{2}+\epsilon\right)\\
&= (1+\mu)\gamma-2\epsilon.
\end{aligned}$$

□

Given our characterization of restricted weak agnostic learning as approximate agnostic algorithms, we can now used known results about approximate agnostic learning to see whether the α parameter in an (α,β)-weak agnostic learner can be reduced.

Amit Daniely has has established the following approximate agnostic learning algorithm (in fact a PTAS) for learning halfspaces. Specifically:

Theorem 5.12 ([Dan15])**.** *For every $\mu>0$, there exists an algorithm for agnostically learning halfspaces under the uniform distribution on the d-dimensional sphere with an approximation ratio of $(1+\mu,\epsilon)$ that runs in time* $\mathrm{poly}\left(d^{\frac{\log^3(1/\mu)}{\mu^2}},1/\epsilon\right)$.

Together with Claim 5.11, this implies that

Corollary 5.13. *We can construct a $(\alpha,\frac{\alpha}{2-\alpha}-2\epsilon)$-weak agnostic learner for the class of halfspaces under the uniform distribution on the d-dimensional sphere in time* $\mathrm{poly}\left(d^{\frac{(2-\alpha)^2\log^3\left(\frac{2-\alpha}{\alpha}\right)}{\alpha^2}},1/\epsilon\right)$.

Proof. As in the proof of Claim 5.11, we write α as $\alpha = 1-\frac{1}{1+\mu}+\gamma$ and set $\gamma=\alpha/2$. Re-arranging, we obtain that $1+\mu = \frac{1}{1-\alpha/2}$ and $\mu = \frac{\alpha}{2-\alpha}$. By Claim 5.11, $\mathrm{cor}(h)\geq(1+\mu)\gamma-2\epsilon$. By plugging

in the values of μ and γ, we obtain that $\text{cor}(h) \geq \left(\frac{1}{1-\alpha/2}\right) \cdot (\alpha/2) - 2\epsilon = \frac{\alpha}{2-\alpha} - 2\epsilon$. Lastly, the claimed runtime follows by plugging $\mu = \frac{\alpha}{2-\alpha}$, into Theorem 5.12. □

Given the dependence on α in the runtime shown in Corollary 5.13, we conclude that we can construct restricted weak agnostic learners for the class of halfspaces under the uniform distribution in polynomial time for constant values of α, but the algorithm is no longer efficient as $\alpha \to 0$.

This result already provides evidence that, in general, we cannot reduce the α parameter arbitrarily, given that in this case the task of constructing an (α, β)-restricted agnostic learner for halfspaces goes from being computationally efficient for constant values of α to unfeasible as $\alpha \to 0$. However, it does not rule out the possibility that it could be efficient for small values of α, given that this separation only applies to Daniely's algorithm for learning halfspaces [Dan15].

In the next section, we provide a tight counter-example that provides a particular case where we can show that it is impossible to boost the α parameter.

5.2.2 Boosting a restricted weak agnostic learner

Here, we show that in general we cannot boost the α parameter of a (α, β)-weak agnostic learner. To do so, we need to choose some problem that is widely believe to be hard; we choose to make the standard cryptographic assumption that pseudorandom functions cannot be distinguished from random functions with non-negligible advantage by *efficient* algorithms. This, for example, follows from the existence of one-way functions [GGM86]. Specifically, we make the following assumption:

Assumption on pseudorandom functions. For any function $T : \mathbb{N} \to \mathbb{N}$, there exists a family of pseudorandom functions, denoted $\mathcal{F} = \{f_r : \{0,1\}^n \to \{0,1\} \mid r \in \{0,1\}^n\}$, such that:

- f_r is easy to compute given r as input.
- For $r \leftarrow \{0,1\}^n$ chosen uniformly at random, any (potentially randomized) algorithm that runs in time $T(n)$ and has query access to $f_r : \{0,1\}^n \to \{0,1\}$ cannot distinguish it from a truly random function with advantage greater than $1/T(n)$ over the uniform distribution on $x \in \{0,1\}^n$.

Our goal is to use pseudorandom functions to construct a domain $\mathcal{X}$ and a concept class $\mathcal{C}$ such that, for every $\alpha \in (0,1)$ there exists a marginal distribution $\mathcal{D}_{\mathcal{X}}$ so that,

- $\mathcal{C}$ is *efficiently* $(\alpha + \epsilon, \alpha)$-weak agnostically learnable under any $\mathcal{D}$ with marginal $\mathcal{D}_{\mathcal{X}}$.
- $\mathcal{C}$ is *not efficiently* (α, β)-weak agnostically learnable for distributions $\mathcal{D}$ with marginal $\mathcal{D}_{\mathcal{X}}$ for any $\beta \geq 2\alpha/T(n)$.

Our construction is inspired by Feldman's work on the power of membership queries in agnostic learning [Fel09b]. In the realizable setting, making the same assumption on pseudorandom functions, it is known that the PAC model with membership queries is strictly stronger than the PAC model without membership queries. This is done by encoding the target concept in the domain such that we can "decode" it with membership queries, but with high probability is not observable through only random examples.

In the agnostic setting, however, this same idea doesn't work, because the target function can be anything, and so it could be random on the part of the domain that contains the encoding [Fel09b].

By providing a much more ingenious construction, Feldman shows that membership queries can in fact give us more power than just random examples when learning with respect to the uniform distribution. Here are some key observations of his construction, which we will use in ours:

- In order for this approach to work in the agnostic setting, we have to encode the secret over a very large fraction over the domain.
- The encoding of the secret has to be resilient to almost any amount of noise, and so we need to be able to decode it with very high amounts of noise. Feldman uses the concatenation of the Reed-Solomon code with the binary Hadamard code by Guruswami and Sudan [GS00], which we will also use.
- The encoding should not be readable from random examples. A way to ensure this is to use pseudorandom functions as part of the concepts themselves in our concept class $\mathcal{C}$.

With these ideas in mind, we proceed to give our construction. Before we make it formal, we give some intuition. We will divide the domain $\mathcal{X}$ into two disjoint parts E and F, where E contains $1-\alpha$ of the probability mass and F only an α-fraction of the mass. Each concept in $\mathcal{C}$ is defined as a pseudorandom function f_r and its key r. We use E and F as follows:

- In E: we encode the key r.
- In F: use the pseudorandom function f_r.

If there is not too much noise, then we can learn $\mathcal{C}$ by constructing the key r from random examples in $\mathbb{E}$. Once we have the key r, we can identify f_r exactly. However, when there is enough noise as to make the labels on E be completely random, then the learning problem becomes equivalent to distinguishing pseudorandom functions from random, which is a cryptographically hard problem.

In order to obtain the tightest gap possible in the boosting of the α parameter, we use list-decodable error-correcting codes that can tolerate error up to slightly less than $1/2$. We remark that unique decoding can only correct errors up to relative distance $1/4$. The RSH code by [GS00] (where the abbreviation comes from the fact that the code consists of a concatenation of the Reed-Solomon code with the binary Hadamard code) satisfies such a guarantee. Specifically, the code $\mathrm{RSH} : \{0,1\}^n \to \{0,1\}^m$ where $m = O(n^2/\varepsilon^4)$ satisfies the following property: Upon receiving a word $w \in \{0,1\}^m$ such that there exists $x \in \{0,1\}^n$ satisfying $d(w, \mathrm{RSH}(x)) \leq (1/2 - \epsilon)m$, where d corresponds to the Hamming distance, there is an algorithm that runs in time $\mathrm{poly}(n)$ and returns a list containing the true x with high probability.[1]

Using the RSH code, we now prove the following:

[1] For details regarding the properties of the RSH code, see [CGKR25] (the paper that Part I of this thesis is based on).

Theorem 5.14. *For every $\alpha \in (0,1)$, there exists $\mathcal{X} = \cup_n \mathcal{X}_n$, $(\mathcal{D}_{\mathcal{X}_n})_{n\geq 1}$, $\mathcal{C} = \cup_n \mathcal{C}_n$ such that:*

(i) There exists a learning algorithm L, such that for every $0 < \varepsilon < 1 - \alpha$, $n \geq 1$, $\mathcal{C}_n$ is $(\alpha + \varepsilon, \alpha)$-weak agnostically learnable by L in time $poly(n, 1/\varepsilon)$ for any distribution $\mathcal{D}$ with marginal $\mathcal{D}_{\mathcal{X}_n}$. The learner is proper and only requires random examples.

(ii) Assuming that pseudorandom functions cannot be distinguished from random functions with non-negigible advantage by efficient algorithms, there does not exist any learning algorithm L that runs in time $T(n)/poly(n)$ that is an (α, β)-weak agnostic learner for $\mathcal{C}_n$ for any $\beta \geq 2\alpha/T(n)$ for distributions with marginal $\mathcal{D}_{\mathcal{X}_n}$ over $\mathcal{X}_n$, and succeeds with probability at least $1/T(n)$, even with access to membership queries.

Note that for Part (i), the learner only requires random samples and is a proper learner, but for Part (ii), the weak learner is allowed query access, and is also allowed to be improper. For simplicitly, we drop the subscript n in the proof.

Proof. Let $m = O(n^2/\epsilon^2)$ (so that we can use the guarantees of the RSH code) and let $E = [m]$ and $F = \{0,1\}^n$. We take $\mathcal{X} = E \cup F$ and define the following marginal distribution given any distribution $\mathcal{D}$:[2] We let $\Pr_{\mathcal{D}_{\mathcal{X}}}[\mathbf{x} \in E] = 1 - \alpha$, $\Pr_{\mathcal{D}_{\mathcal{X}}}[\mathbf{x} \in F] = \alpha$, and make the conditional distribution on each of E and F be the uniform. Given a string $r \in \{0,1\}^n$, we denote its encoding by the RSH code as $\mathrm{RSH}(r)$. Recall that $\mathrm{RSH}(r) \in \{0,1\}^m$; we will index the coordinates of RSH using E and $\mathcal{F} = \{f_r : \{0,1\}^n \to \{0,1\} \mid r \in \{0,1\}^n\}$ denote the $\mathbf{x}^{th}$ coordinate of the encoding of r by RSH as $\mathrm{RSH}(r)_{\mathbf{x}}$ for each $\mathbf{x} \in E$.

Let $\mathcal{F} = \{f_r : \{0,1\}^n \to \{0,1\} \mid r \in \{0,1\}^n\}$ be a family of pseudorandom functions. For each random seed $r \in \{0,1\}^n$, we use the E region to encode the key r to the PRF f_r, and then put the PRF f_r on the F region. Specifically, for each $r \in \{0,1\}^n$, we define a Boolean function $g_r : \mathcal{X} \to \{0,1\}$ as follows:

$$g_r(\mathbf{x}) = \begin{cases} \mathrm{RSH}(r)_{\mathbf{x}} \text{ for } \mathbf{x} \in E \\ f_r(\mathbf{x}) \text{ for } \mathbf{x} \in F \end{cases}$$

We then define the $\{\pm 1\}$ version of g_r as $c_r(\mathbf{x}) = 2g_r(\mathbf{x}) - 1$ and define our concept class as the collection of all such g_r; that is, $\mathcal{C} = \{c_r \mid r \in \{0,1\}^n\}$. We depict our construction in Figure 5.1.

[2]Note that the distribution $\mathcal{D}$ is arbitrary, but we do get to choose the marginal distribution.

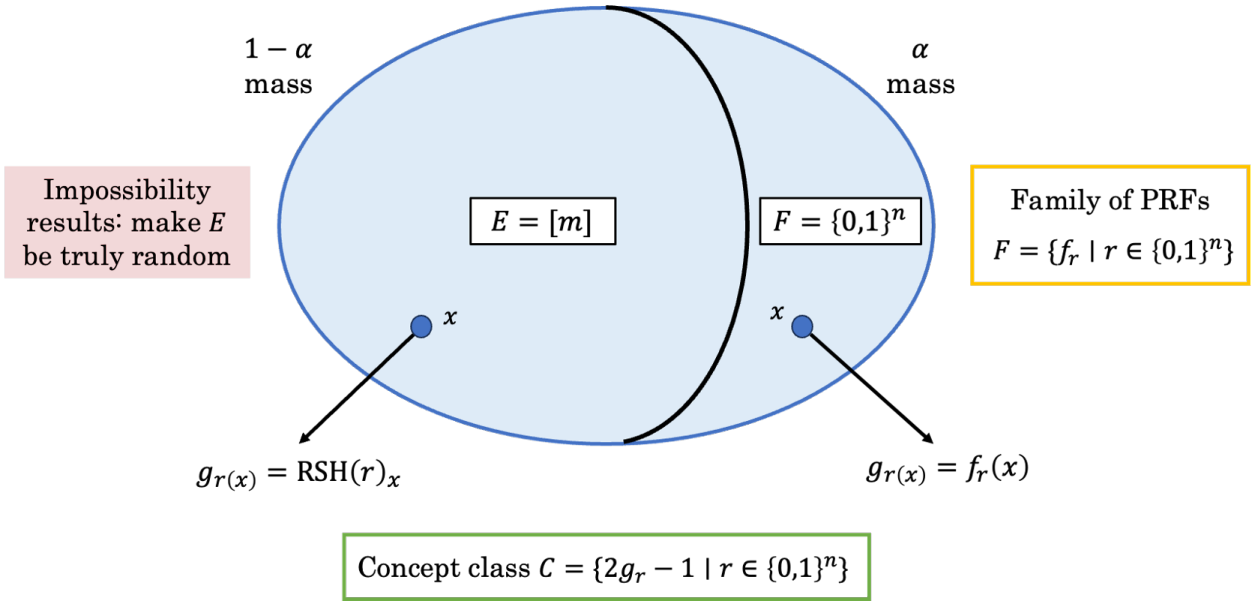


Figure 5.1: Construction of $\mathcal{X}, \mathcal{D}_{\mathcal{X}}$ and $\mathcal{C}$.

Part (i): For the positive result in Part (i), consider any distribution $\mathcal{D}$ on $(\mathbf{x}, \mathbf{y})$ where $\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}$ and there exists $c_r \in \mathcal{C}$ such that $\mathrm{cor}(\mathbf{y}, c_r(\mathbf{x})) \geq \alpha + \varepsilon$.

Because the region E occupies an $(1-\alpha)$ fraction of the domain, it must be that c_r has some correlation with the labels on E. Otherwise, given that the total correlation is equal to the sum of the weighted correlations on each of E and F, this would contradict the fact that c_r satisfies $\mathrm{cor}(\mathbf{y}, c_r(\mathbf{x})) \geq \alpha + \varepsilon$. Formally, it must be that $\mathrm{cor}(\mathbf{y}, c_r(\mathbf{x}) | \mathbf{x} \in E) \geq \varepsilon/(1-\alpha)$, because otherwise we reach a contradiction:

$$\begin{aligned}
\mathrm{cor}(\mathbf{y}, c_r) &= (1-\alpha) \cdot \mathrm{cor}(\mathbf{y}, c_r | \mathbf{x} \in E) + \alpha \cdot \mathrm{cor}(\mathbf{y}, c_r | \mathbf{x} \in F) \\
&< (1-\alpha)\frac{\varepsilon}{1-\alpha} + \alpha \cdot 1 \\
&= \alpha + \varepsilon.
\end{aligned}$$

We can re-write $\mathrm{cor}(\mathbf{y}, c_r)$ using the standard translation from correlation to Hamming distance for variables in $\{\pm 1\}$ as follows:

$$\begin{aligned}
\mathrm{cor}(\mathbf{y}, c_r(\mathbf{x}) | \mathbf{x} \in E) &= \mathbb{E}[c_r(\mathbf{x})(2\mathbf{y} - 1) | \mathbf{x} \in E] \\
&= \mathbb{E}[(2g_r(\mathbf{x}) - 1)(2\mathbf{y} - 1) | \mathbf{x} \in E] \\
&= 1 - 2\Pr[\mathbf{y} \neq g_r(\mathbf{x}) | \mathbf{x} \in E].
\end{aligned}$$

As we showed, the LHS is at least $\varepsilon/(1-\alpha)$. Moreover, by definition, $g_r(x) = \mathrm{RSH}(r)_x$ for $x \in E$. Hence, from the above chain of equalities it follows that

$$\frac{\varepsilon}{1-\alpha} \leq \mathrm{cor}(\mathbf{y}, c_r(\mathbf{x}) | \mathbf{x} \in E) = 1 - 2\Pr[\mathbf{y} \neq \mathrm{RSH}(r)_{\mathbf{x}} | \mathbf{x} \in E].$$

Therefore,

$$\Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\mathbf{y} \neq \mathrm{RSH}(r)_{\mathbf{x}} | \mathbf{x} \in E] \leq \frac{1}{2}\left(1 - \frac{\varepsilon}{1-\alpha}\right) \leq \frac{(1-\varepsilon-\varepsilon\alpha)}{2}. \tag{5.15}$$

This is a high probability, which is what we want. Namely, this is telling us that, because we have given a lot of mass to the E region, the initial assumption that $\mathrm{cor}(\mathbf{y}, c_r(\mathbf{x})) \geq \alpha + \epsilon$ ensures that the noise is not too bad, in the sense that we get the correct label of $\mathbf{y} = \mathrm{RSH}(r)_{\mathbf{x}}$ for each $\mathbf{x} \in E$.

This bound holds independently for each $\mathbf{x} \in E$. Because each $\mathbf{x} \in E$ indices one bit of the code $\mathrm{RSH}(r)$, in order to obtain our candidate word $w \in \{0,1\}^m$ for $\mathrm{RSH}(r)$ we need to observe at least one sample $(\mathbf{x}, \mathbf{y})$ for each $\mathbf{x} \in E$. Because $E = [m]$, this requires $O(m\log(m)/(1-\alpha))$ samples. Once we have them, we concatenate all of the $\mathrm{RSH}(r)_x$ in order to form our word $\mathbf{w}$.

What is the Hamming distance between $\mathbf{w}$ and $\mathrm{RSH}(r)$? By Equation 5.15, we are doing quite well on each x-coordinate. Therefore, adding up across the m coordinates we get that

$$\begin{aligned}
\mathbb{E}[d(\mathbf{w}, \mathrm{RSH}(r))] &= \mathbb{E}\left[\sum_{x\in[m]} \mathbf{1}[\mathbf{w}(x) \neq \mathrm{RSH}(r)_x]\right] \\
&= \sum_{x\in[m]} \Pr[\mathbf{w}(x) \neq \mathrm{RSH}(r)_x] \\
&= \sum_{x\in[m]} \Pr[\mathbf{y} \neq \mathrm{RSH}(r)_x | \mathbf{x} = x] \\
&= m \Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\mathbf{y} \neq \mathrm{RSH}(r)_{\mathbf{x}} | \mathbf{x} \in E] \\
&\leq \frac{(1-\varepsilon-\varepsilon\alpha)}{2} m.
\end{aligned}$$

Hence by a Chernoff bound, with probability $1 - \exp(-\varepsilon^2\alpha^2 m)$ we have $d(\mathbf{w}, \mathrm{RSH}(r)) \leq (1-\varepsilon)m/2$.

This is precisely the guarantee that we need for the RSH code to be able to decode. Then, we use the list-decoder from random samples, which gives us a list containing the seed r with high probability. To find it, we enumerate over all the candidates r' in the list, and check whether $\mathrm{cor}(\mathbf{y}, c_{r'}) \geq \alpha$ using random samples. We can do this given that f_r is easy to compute given r. We output the best hypothesis that we find, which will have correlation at least α with the labels with high probability. Note that we get α rather than $\alpha + \epsilon$ because we lose an ϵ in the sampling error when estimating the correlations.

Part (ii): For the negative result in Part (ii), we do get to choose the distribution. We define a distribution $\mathcal{D}_r$ where we draw $r \in \{0,1\}^n$ at random. This makes it so that $\mathbf{y}|\mathbf{x}$ is a uniformly random bit for $x \in E$ and $\mathbf{y}|\mathbf{x} = f_r(\mathbf{x})$ for $\mathbf{x} \in F$. The idea here is that we have reduced learning to distinguishing between truly random and a pseudorandom function, which is not possible per our cryptographic assumption.

Given that our hypotheses $c_r(\mathbf{x}) = 2g_r(\mathbf{x}) - 1$ are defined such that $g_r(\mathbf{x}) = \mathrm{RSH}(r)_{\mathbf{x}}$ for $\mathbf{x} \in E$ and $g_r(\mathbf{x}) = f_r(\mathbf{x})$ for $\mathbf{x} \in F$, this means that any $c_r \in \mathcal{C}$ gets no correlation on E (because the labels are random) and perfect correlation on F. Hence:

$$\mathrm{cor}(\mathbf{y}, c_r) = (1-\alpha)\cdot\mathrm{cor}(\mathbf{y}, h(\mathbf{x})|\mathbf{x}\in E) + \alpha\cdot\mathrm{cor}(\mathbf{y}, h(\mathbf{x})|\mathbf{x}\in F) = 0 + \alpha\cdot 1 = \alpha.$$

Let's assume by contradiction that there exists an algorithm which runs in time $T(n)$, which is

allowed query access to $\mathcal{D}$, and outputs a hypothesis $h : \mathcal{X} \to [-1, 1]$ that achieves correlation β. Then,

$$\begin{aligned}\beta &= (1-\alpha) \cdot \mathrm{cor}(\mathbf{y}, h(\mathbf{x}) | \mathbf{x} \in E) + \alpha \cdot \mathrm{cor}(\mathbf{y}, h(\mathbf{x}) | \mathbf{x} \in F) \\ &= \alpha \cdot \mathrm{cor}(f_r(\mathbf{x}), h(\mathbf{x}) | \mathbf{x} \in F),\end{aligned}$$

given that no hypothesis can achieve better than 0 correlation on E, because the labels there are random.

But this implies that $\mathrm{cor}(f_r(\mathbf{x}), h(\mathbf{x}) | \mathbf{x} \in F) \geq \beta/\alpha$, so h (restricted to F) has distinguishing advantage $\beta/(2\alpha)$ for f_r versus a random function. This is a contradiction if $\beta \geq 2\alpha/T(n)$. □

5.2.3 Multiaccuracy with better error parameter

We conclude our study of restricted weak agnostic learning by trying a different type of boosting. By the original multiaccuracy algorithm [HKRR18], we know that given any (restricted) (α, β)-weak agnostic learner for $\mathcal{C}$, we can efficiently construct a $(\mathcal{C}, \alpha + \varepsilon)$-multiaccurate predictor. As we have summarized at multiple points in this thesis, this is because every call to the weak agnostic learner allowes to perform a gradient update that improves the squared loss of our predictor. We can ask: is it possible that with the same primitive of an (α, β)-weak agnostic learner we could obtain a multiaccurate predictor with better error parameter? That is, can we get an $(\mathcal{C}, \alpha')$-multiaccurate predictor for $\alpha' < \alpha$?

Unfortunately, the answer is also no. Formally:

Theorem 5.16. *For every $\alpha \in (0, 1)$, there exists $\mathcal{X} = \cup_n \mathcal{X}_n$, $(\mathcal{D}_{\mathcal{X}_n})_{n \geq 1}$, $\mathcal{C} = \cup_n \mathcal{C}_n$ such that:*

(i) There exists a learning algorithm L, such that for every $0 < \varepsilon < 1 - \alpha$, $n \geq 1$, it can output a $(\mathcal{C}, \alpha + \epsilon)$-multiaccurate predictor for any distribution $\mathcal{D}$ with marginal $\mathcal{D}_{\mathcal{X}_n}$ in time poly$(n, 1/\varepsilon)$.

(ii) Assuming that pseudorandom functions cannot be distinguished from random functions with non-negigible advantage by efficient algorithms, for any $0 < \varepsilon < \min\{\alpha, 1 - \alpha\}$, there does not exist any algorithm $\mathcal{A}$ that runs in time $T(n)/$poly$(n, 1/\varepsilon)$ and outputs with probability at least $1/T(n)$, a predictor $p : \mathcal{X}_n \to [0, 1]$ that is $(\mathcal{C}_n, \alpha/2 - \epsilon)$-multiaccurate for distributions with marginal $\mathcal{D}_{\mathcal{X}_n}$ over $\mathcal{X}_n$, even with access to membership queries.

To prove this, we use essentially the same construction as in Theorem 5.14 from the previous section. Indeed, the [HKRR18] algorithm allows us to go from feasible agnostic learning to feasible multiaccuracy.

Proof. We use the same $E, F, \mathcal{C}$ as in Theorem 5.14.

Part (i): Part (i) of Theorem 5.16 follows directly from our proof of Part (i) in Theorem 5.14. Namely, we showed that $\mathcal{C}_n$ can be efficiently $(\alpha + \epsilon, \alpha)$-weak agnosticaly learnt in time (poly$(n, 1/\epsilon)$ for any distribution $\mathcal{D}$ with marginal $\mathcal{D}_{\mathcal{X}_n}$. By the [HKRR18] algorithm, it follows that we can

construct a $(\mathcal{C}, \alpha + \epsilon)$-multiaccurate predictor for the same distribution in time (poly$(n, 1/\epsilon)$ as well.

Part (ii): We again use the same $E, F, \mathcal{C}$ as in Theorem 5.14, and here we also use the distribution $\mathcal{D}_r$ where we draw $r \in \{0,1\}^n$ at random, as in Part (ii) of Theorem 5.14. Then, because the labels in E are random, any $c_r \in \mathcal{C}$ again gets 0 correlation on E and perfect correlation on F, given that $g_r(\mathbf{x}) = f_r(\mathbf{x})$ for $\mathbf{x} \in F$. Hence, $\text{cor}(\mathbf{y}, c_r) = \alpha$.

Suppose p is a $(\mathcal{C}, \tau)$-multiaccurate predictor for $\mathcal{D}_r$ for $\tau \leq \alpha/2 - \varepsilon$. Let $\mathbf{y}_p \sim \text{Bern}(p(\mathbf{x}))$. Then we observe that since p is $(\mathcal{C}, \tau)$-multiaccurate,

$$\begin{aligned}
|\text{cor}(\mathbf{y}, c_r) - \text{cor}(\mathbf{y}_p, c_r)| &\leq |\mathbb{E}[c_r(\mathbf{x})(2\mathbf{y} - 1) - c_r(\mathbf{x})(2\mathbf{y}_p - 1)]| \\
&\leq 2\,|\mathbb{E}[c_r(\mathbf{x})(\mathbf{y} - \mathbf{y}_p)]| \\
&\leq 2\,|\mathbb{E}[c_r(\mathbf{x})(\mathbf{y} - p(\mathbf{x}))]| \\
&\leq \alpha - 2\varepsilon,
\end{aligned}$$

where the last inequality follows by $(\mathcal{C}, \tau)$-multiaccuracy. Since $\text{cor}(\mathbf{y}, c_r) \geq \alpha$, this implies that $\text{cor}(\mathbf{y}_p, c_r) \geq 2\varepsilon$.

We will show that such a predictor p can be used to predict f_r with non-trivial probability, which contradicts the hardness assumption. Since

$$\text{cor}(\mathbf{y}_p, c_r) = (1 - \alpha) \cdot \text{cor}(\mathbf{y}_p, c_r | \mathbf{x} \in E) + \alpha \cdot \text{cor}(\mathbf{y}_p, c_r | \mathbf{x} \in F),$$

either $\text{cor}(\mathbf{y}_p, c_r | \mathbf{x} \in E) \geq \varepsilon/(1 - \alpha)$, or $\text{cor}(\mathbf{y}_p, c_r | \mathbf{x} \in F) \geq \varepsilon/\alpha$. In the former case, we can use list-decoding to obtain r and thence f_r as in Part (i). For the latter case, we observe that $\mathbf{y}_p$ lets us predict f_r with non-trivial advantage, since

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{x} \sim F}[f_r(\mathbf{x}) \neq \mathbf{y}_p] &= \frac{1}{2} \mathop{\mathbb{E}}_{\mathbf{x} \sim F}[1 - (2\mathbf{y}_p - 1)(2f_r(\mathbf{x}) - 1)] \\
&= \frac{1}{2} - \frac{1}{2} \mathop{\mathbb{E}}_{\mathbf{x} \sim F}[(2\mathbf{y}_p - 1)c_r(\mathbf{x})] \\
&= \frac{1}{2} - \frac{1}{2}\text{cor}(\mathbf{y}_p, c_r | \mathbf{x} \in F)
\end{aligned}$$

where we use $c_r(x) = 2f_r(x) - 1$ for $x \in F$. Thus, in either case we can predict the pseudorandom function f_r with non-trivial advantage with non-negligible probability. Hence we have reached a contradiction. □

We remark that for the proof of Theorem 5.16 we could have used a similar construction with any hard problem (e.g., learning parities with noise), whereas in Theorem 5.14 we crucially rely on being able to decode the key to the PRF.

The spectrum of weak agnostic learning. We put together our various results on the spectrum of the feasibility of weak agnostic learning in Figure 5.2, which we reproduced in Chapter 1. The region of interest is $\beta \leq \alpha$. Going right or down, the problem becomes easier, whereas going up or left is harder. Green indicates regions that are known to be easy in general (under the right learning assumptions), while Red indicates regions that are hard (in the worst case over all $\mathcal{C}$).

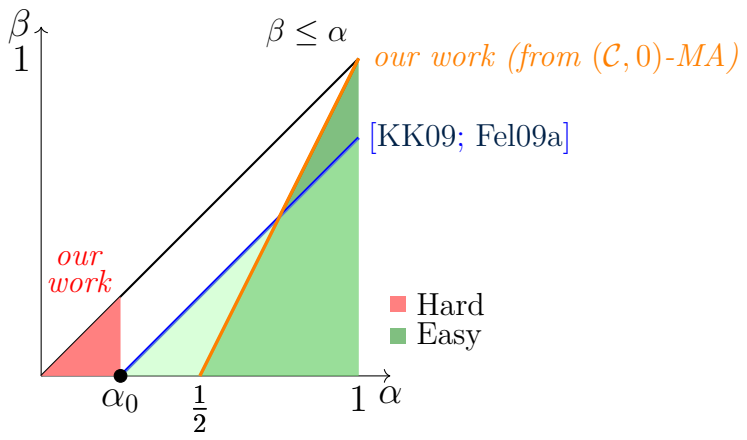


Figure 5.2: The spectrum of weak agnostic learning.

- The Orange line bounds the region that is feasible if we have access to a $(\mathcal{C}, 0)$-multiaccurate predictor. This follows from our Theorem 3.4. If we also assume calibration, the entire region $\beta \leq \alpha$ is feasible.
- The Blue line with slope 1 follows from the boosting results of [KK09; Fel09a]. The value of α_0 indicates an arbitrary small correlation α; the precise boosting statement using α_0 can be found in [KK09].
- The Red region indicates that there is no black-box reduction from (α, β) weak agnostic learning to $(\alpha - \varepsilon, \beta')$ weak agnostic learning, assuming the existence of one-way functions. This follows from our Theorem 5.14.

5.3 Auditing versus learning for multiaccuracy

To conclude our investigations on the relationship between multiaccuracy and weak agnostic learning with a different approach inspired by the works of Vadhan [Vad17] and Kothari and Livni [KL18] on the *learning versus refutation* problem in agnostic learning. We know that in the original algorithm for constructing multiaccurate predictors [HKRR18], we audit for $\mathcal{C}$-multiaccuracy in each step of the algorithm using a $\mathcal{C}$-weak agnostic learner. Given that auditing for C-multiaccuracy in this search-based way is exactly equivalent to proper weak agnostic learner for $\mathcal{C}$, our results in Section 3.1 showing that one cannot post-process a multiaccurate predictor to get a weak learner can appear surprising at first. So we can ask: what exactly is the relationship between the problem of *auditing* for multiaccuracy versus that of *learning* a multiaccurate predictor?

Definition 5.17. *Given a concept class $\mathcal{C}$, distribution $\mathcal{D}$, predictor p, and parameters $\delta, \epsilon > 0$, the* audit-MA-decision *problem for $\mathcal{C}, \delta$ returns "yes" if p is a $(\mathcal{C}, \delta + \epsilon)$-multiaccurate predictor for $\mathcal{D}$, and "no" if p is not a $(\mathcal{C}, \delta)$-multiaccurate predictor for $\mathcal{D}$.*

Definition 5.18. *Given a concept class $\mathcal{C}$, distribution $\mathcal{D}$, predictor p, and parameters $\delta, \epsilon > 0$, the* audit-MA-search *for $\mathcal{C}, \delta$ problem returns "yes" if p is a $(\mathcal{C}, \delta + \epsilon)$-multiaccurate predictor for $\mathcal{D}$, and a concept g such that $|\mathbb{E}[g(\mathbf{x})(\mathbf{y} - p(\mathbf{x}))]| > \delta$ if p is not a $(\mathcal{C}, \delta)$-multiaccurate predictor for $\mathcal{D}$. If we require g to belong to $\mathcal{C}$, then we call this algorithm* proper*. Otherwise, we call it* improper*.*

Clearly, if we can solve the AUDIT-MA-SEARCH problem for $\mathcal{C}, \delta, \epsilon$, we can solve the AUDIT-MA-DECISION problem for $\mathcal{C}, \delta, \epsilon$ as well.

The following equivalence follows directly from Definition 5.18:

Claim 5.19. *We can efficiently solve the proper* AUDIT-MA-SEARCH *problem for $\mathcal{C}, \mathcal{D}, \delta, \epsilon$ if and only if we can efficiently properly $(\delta + \epsilon, \delta)$-weak agnostically learn $\mathcal{C}$ under $\mathcal{D}$.*

Proof. Given a proper weak agnostic learner $\mathcal{A}$ for $\mathcal{C}$, we solve the AUDIT-MA-SEARCH for a given predictor p as follows. We give samples $\{(\mathbf{x}, \mathbf{y} - p(\mathbf{x}))\}$ to $\mathcal{A}$, where $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}$, and weak agnostic learning parameters $(\delta + \epsilon, \delta)$. If $\mathcal{A}$ returns a hypothesis $c \in \mathcal{C}$, then we return c. Otherwise, we return "yes". In the other direction, given an AUDIT-MA-SEARCH algorithm $\mathcal{B}$, we can properly weak agnostically learn $\mathcal{C}$ as follows. We call algorithm $\mathcal{B}$ with parameter $\delta + \epsilon$ and with the constant $\mathbf{0}$ predictor p. If $\mathcal{B}$ returns "yes", then we do not return anything; otherwise if $\mathcal{B}$ returns some concept $c \in \mathcal{C}$, then we return this same concept c. □

The work of [HKRR18] shows that we can reduce the problem of learning a multiaccurate predictor to the AUDIT-MA-SEARCH problem, which in turn can be solved with a weak agnostic learner. However, for the purpose of learning a multiaccurate predictor in this way, it is not necessary to require the AUDIT-MA-SEARCH algorithm (and in turn, the weak agnostic learning algorithm) to be proper. Namely, to make enough progress in the squared loss, we can do with any hypothesis g that has high enough correlation with $\mathbf{y} - p(\mathbf{x})$, regardless of whether $g \in \mathcal{C}$ or not.

Thus, when we use an *improper* weak agnostic learner to solve the AUDIT-MA-SEARCH problem, the equivalence only holds in one direction: if the weak agnostic learner fails, then we know that our predictor p is multiaccurate. But if it returns a hypothesis g that is not necessarily in $\mathcal{C}$, then we do not know whether or not p is multiaccurate.

Given the equivalence between the AUDIT-MA-SEARCH problem and proper weak agnostic learning (Claim 5.19), it is clear that the distinction between the search and decision versions of the problem of auditing for multiaccuracy is closely related to the equivalent question but for the problem of weak agnostic learning instead. The answer to the latter question was given in the works of Kothari-Livni and Vadhan, who studied the *learning versus refutation problem*. Namely, consider the following version of the decision-version of the problem of weak agnostic learning:

Definition 5.20 (Refutation algorithm, informal [KL18; Vad17])**.** *Given a concept class $\mathcal{C}$, distribution $\mathcal{D}_{\mathcal{X}}$, and parameter δ, the* REFUTATION *problem for $\mathcal{C}, \mathcal{D}, \delta$ consists of distinguishing between the following two cases:*

1. Structure. *The labels $\mathbf{y}$ have correlation at least δ with some concept $c \in \mathcal{C}$.*
2. Noise. *The labels $\mathbf{y}$ are uniform and independent Rademacher random variables.*

We remark that the precise definition of a refutation algorithm is slightly different, in that the refutation algorithm is given an m-tuple of points and an m-tuple of labels. The *refutation complexity* of a concept class is then defined as the smallest m such that there exists an efficient refutation algorithm with m samples [KL18, Def. 4].

It is clear that if a class $\mathcal{C}$ is weak agnostically learnable, then we can solve the refutation problem for $\mathcal{C}$. But what about the other direction? Perhaps surprisingly, Kothari-Livni and Vadhan show that the REFUTATION problem is equivalent to standard weak agnostic learning:

Theorem 5.21 (Learning and Refutation are equivalent, informal [KL18; Vad17])**.** *A class $\mathcal{C}$ can be efficiently weak agnostically learnt (not necessarily properly) if and only if we can efficiently solve the* REFUTATION *problem for $\mathcal{C}$.*

In our case, the problem that we are interested in is the DECISION-WAL problem, where the task is to decide whether there is some concept in $\mathcal{C}$ that has some correlation with the labels, or whether the labels are truly random. This problem is equivalent to the AUDIT-MA-DECISION problem, where we just have to swap between the labels $\mathbf{y}$ and the residuals $\mathbf{y} - p$. Using the result by [KL18; Vad17], we would hope to then show that the AUDIT-MA-DECISION is equivalent to the problem of learning a weak agnostic learner for $\mathcal{C}$ (with some loss of parameters in both directions, particularly in the backwards direction, where we need a larger change of parameters in the representation class of the weak agnostic learner due to its improperness).

However, the issue is that the DECISION-WAL problem is not *exactly* equivalent to the REFUTATION problem considered in [KL18] due to their m-finite sample formulation, so we cannot directly use their result. An interesting next direction is to understand how these two problems relate precisely, so that potentially we can show that if we have an efficient algorithm for *auditing* for multiaccuracy for a concept class $\mathcal{C}$, then $\mathcal{C}$ is agnostically learnable.

II

Learning to Abstain Optimally and Fairly

6
Learning with Abstentions

It is well-known that in many applications erroneous predictions of one type or another must be avoided. In some applications, like spam detection, false positive errors are serious problems. In other applications, like medical diagnosis, abstaining from making a prediction may be more desirable than making an incorrect prediction.

Kalai, Kanade, and Mansour [KKM12]

In Part II of this thesis, we study the problem of learning with abstentions from the perspective of the multigroup fairness framework. As discussed in Chapter 1, we consider *selective classifiers* which output values in the range $[0,1] \cup ?$, where ? corresponds to "I don't know." The key idea of selective classification is that we want to achieve higher accuracy at the cost of abstaining from making predictions on certain points [Cho57; KKM12; KK21a; KK21b; GKKM20].

As summarized in the introduction, there are several lines of work that have studied the problem of learning with abstentions. From an algorithmic fairness point of view, adding abstentions to the usual learning framework is desirable as to avoid making arbitrary decisions and unnecessary false positives and false negatives [ALG21; BAZ+21; KKR22; LHAC23; KHS23; LHAC24; CLC+24]. Still, most of the works in the algorithmic fairness space still focus heavily on the prediction themselves (e.g., as in the multigroup fairness framework), and so there is a lack of a formal study of the role of uncertainty in predictions, particularly when applied to societal settings. Moreover, certain forms of selective classification can magnify disparities across groups [JSK+20; LBR+21; SC21; YTG+], and so we have to be careful in how we decide to abstain.

Some works that study the reliability of predictions focus on the *model multiplicity problem*, which is concerned with the following fact: for a given fixed dataset, there are multiple ways in which we can train a predictor on the dataset such that it achieves high accuracy, but where these various potential and equally good predictors can then disagree on individual predictions [BRB22; MCU20]. If we train a whole class of models $\mathcal{M}$, then a natural way to compute the uncertainty empirically at each point x is to compute $m(x)$ for each $m \in \mathcal{M}$ and then compute how much variance/disagreement there is among these various predictions. Indeed, some works use this method as a way to add abstentions, and find that we can obtain close-to-fair predictions simply by

abstaining on the individuals with high variance [CLC+24]. This once again illustrates the fact that it is misguided to focus so heavily on de-biasing the predicted probabilities if these turn out to be arbitrary. Hence, it is important for us to develop formal methods for quantifying the uncertainty of predictors. Several other works study the relationship between the variance within the class $\mathcal{M}$ and common group fairness metrics [LHAC23; LHAC24; ALG21; KHS23; JRLT23].

Relatedly, still within the problem of predictive multiplicity, various algorithms have been proposed for ensembling the various competing models in different ways [BLF21; RTW23; DNW24; BCDT25]. An algorithm that is particularly interesting here is the *Reconcile* algorithm, which uses a multiaccuracy-type algorithm for reconciling two predictors such that the two agree by the end and we have only improved their respective squared losses during the reconciliation process [RTW23]. Indeed, as explained in [BCDT25], instead of using a weak agnostic learner as a distinguisher, we use the disagreement area between the two predictors as the witness. Given that these ensembling and reconciliation methods aim to reduce the disagreement within the model class, we can view them as helping reduce the abstention rate of the final predictor. Naturally, a drawback of all of these variance-based methods is that they require fitting an entire class of models.

A popular frequentist uncertainty quantification method (as opposed to Bayesian-based approaches, some of which have been applied in the setting of fair classification [BAZ+21; KKR22; HCM+23; TCL23]) is that of *conformal prediction*, first proposed by Gammerman, Vovk, and Vapnik [VGS05; AB21]. Conformal prediction is a technique for determining precise levels of confidence which can be applied to any method that has already been trained on the data [SV08]. In the classification case, instead of a point-prediction (e.g., 0.8 for individual x), a conformal prediction method returns a *prediction set* containing the possible labels for that point, such that the true label is in this set with high probability. Thus, the bigger the set, the lower the confidence and the higher the uncertainty. Some recent works have used multicalibration-like algorithms to extend the conformal prediction setting to provide conditional guarantees instead of only marginal guarantees, where we condition on groups in a collection [JLP+21; JNRR23].

Lastly, abstentions in learning have also been used to provide bounds on efficient learning in the presence of arbitrary covariate shifts [KK21a; KK21b; GKKM20] and in the introduction of *partial* hypotheses classes [Lon01; HP23].

6.1 Reliable agnostic learning

We want to take a rigorous and formal learning-theoretic approach to the problem of learning with abstentions. In the setting of learning theory, the study of selective classification was started in 2009 by Kalai, Kanade, and Mansour, who coined it "reliable agnostic learning" [KKM12]. By adapting the traditional agnostic learning framework, their model answers the questions: How can we come up with a formal model of classifiers that abstain, what does it mean to "abstain optimally", and how can we learn such classifiers? Here, the goal is to output a selective classifier whose accuracy nearly matches the accuracy of the best selective classifier from a pre-specified concept class. For all statements and definitions regarding reliable agnostic learning, the concepts in $\mathcal{C}$ are always Boolean.

Given its formal learning-theoretic guarantees, our starting point into the problem of learning with abstentions is precisely the model of reliable agnostic learning introduced in [KKM12]. In

order to formally introduce the model, we need to set up some notation, which we will be using throughout Part II of this thesis.

Loss functions. In addition to the notation that we used in Part I of this thesis, we need to introduce new definitions for dealing with loss functions, which play a central role in this second part. In Part II of the thesis, we only consider Boolean concepts classes $\mathcal{C}$. A loss function ℓ takes a label $y \in \mathcal{Y}$ and an action $t \in \mathbb{R}$ and returns a loss value $\ell(y,t)$. Examples include the logistic loss $\ell(y,t) = \log(1+\exp(-yt))$ or binary classification with different false-positive/negative costs $\ell(y,t) = c_y|y-t|$. We let $\mathcal{L} = \{\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}\}$ denote a collection of loss functions.

As usual, the goal is to find a hypothesis $h : \mathcal{X} \to \mathbb{R}$ that minimizes the expected loss $\ell_{\mathcal{D}} := \mathbb{E}_{\mathcal{D}}[\ell(y,h(x))]$. Specifically, we continue to work in the agnostic setting, where we want the expected loss of our hypothesis h to be at most epsilon higher than the loss incurred by the best concept in $\mathcal{C}$. We formalized this notion of (strong) agnostic learning in Chapter 2. A key observation, which is precisely what motivates the definition of an omnipredictor, is that the best concept in $\mathcal{C}$ depends on the chosen loss function [GKR+22]. For example, if $\mathbf{y} \sim \text{Bern}(0.2)$, then for the ℓ_2 loss it is optimal to predict 0.2, whereas for the ℓ_1 loss we should predict 0.

We focus on the setting $\mathcal{Y} = \{0,1\}$ of binary classification. In our setting, we allow predictors to output an abstention ?, and so we consider triplets of loss functions $(\ell_+, \ell_-, \ell_?)$. If a predictor is allowed to abstain and thus ? is in its support, we denote it with an abstention sign in the subscript of the predictor.

Losses $(\ell_+, \ell_-, \ell_?)$. We further specify loss functions depending on the value of $y \in \{0,1\}$. Given a loss function $\ell : \mathcal{Y} \times \{\mathbb{R} \cup \{?\}\} \to \mathbb{R}$, we decompose $\ell = (\ell_+, \ell_-, \ell_?)$ as follows:

1. **Negative labels.** For inputs (y,t) where $t \neq ?$ and $y = 0$, we write $\ell_+(t)$ for $\ell_{\mathcal{D}}(0,t)$.
2. **Positive labels.** For inputs (y,t) where $t \neq ?$ and $y = 1$, we write $\ell_-(t)$ for $\ell_{\mathcal{D}}(1,t)$.
3. **Abstentions.** For inputs (y,t) where $t = ?$, we write $\ell_?(y)$ for $\ell(y,t)$. In turn to separate the cases $y = 1$ and $y = 0$, we write $\ell_?(1) = \alpha_+$ and $\ell_?(0) = \alpha_-$. Whenever the predictor cannot be uniquely inferred from the context, we still write $\ell_?(y,t)$.

We drop $\mathcal{D}$ from the subscript if it can be directly inferred.

For example, in the specific case of the 0-1 loss, $\ell_+(0,t) = |0-t|$ and $\ell_-(1,t) = |1-t|$, and so the expected loss $\mathbb{E}[\ell_+(\mathbf{y}, h(\mathbf{x}))]$ is equal to the rate of false positives and $\mathbb{E}[\ell_-(\mathbf{y}, h(\mathbf{x}))]$ to the rate of false negatives. The sum of losses $\ell_+ + \ell_-$ corresponds to the usual definition of *error* of the predictor. In turn, $\ell_?$ generalizes the definition of the abstention rate of a predictor $\mathbb{E}_{\mathcal{D}}[\mathbb{1}[h(\mathbf{x}) = ?]]$. The case where $\ell_?(y) = \alpha$ for a fixed value of $\alpha > 0$ corresponds to the traditional Chow model [Cho57; KK21a]. Note that we allow different abstention costs depending on whether the corresponding label is $y = 0$ or $y = 1$.

We are now ready to formally introduce the model of reliable agnostic learning. So far, we have been using the general term "reliable agnostic learning", which is in fact an umbrella term for three different notions introduced by [KKM12]:

- *Positive reliable learning* (PRL), where the goal is to almost never produce false positives.
- *Negative reliable learning* (NRL), where the goal is to almost never produce false negatives.
- *Fully reliable learning* (FRL), where the goal is to almost never make any errors, which is doable because we add abstentions to the range of the predictor.

Thus we can view PRL and NRL as a model of learning for one-sided noise. This provides some intuition for why Kanade and Thaler showed that reliable agnostic learning is easier than (weak) agnostic learning [KT14]. Note that we only add abstentions in the case of FRL; indeed, PRL and NRL are *not* selective classifiers.

To define PRL and NRL, we need to define the following two concept classes derived from the input concept class $\mathcal{C}$. Specifically, given a concept class $\mathcal{C}$ of Boolean concepts $c : \mathcal{X} \to \{0,1\}$ and a distribution $\mathcal{D}$, we further define the following concept classes derived from $\mathcal{C}$ [KKM12]:

$$\mathcal{C}^+ = \{c \in \mathcal{C} \mid \ell_+(c) = 0\}, \qquad \mathcal{C}^- = \{c \in \mathcal{C} \mid \ell_-(c) = 0\},$$

where $\mathcal{D}$ is always implicit in the loss functions.

Definition 6.1 (PRL for a family of loss functions $\mathcal{L}$ [KKM12; KT14])**.** *A concept class $\mathcal{C}$ of Boolean concepts is* $\mathcal{L}$-positively reliably learnable *if there exists a learning algorithm that for any distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, any $\ell_+, \ell_- \in \mathcal{L}$, and any $\epsilon, \delta > 0$, when given access to the example oracle $\mathsf{EX}(\mathcal{D})$ and loss functions ℓ_+, ℓ_-, outputs a hypothesis $h : \mathcal{X} \to [0,1]$ that satisfies the following with probability at least $1 - \delta$:*

1. $\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\ell_+(h(\mathbf{x}))] \leq \epsilon$,
2. $\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\ell_-(h(\mathbf{x}))] \leq \min_{c_+\in\mathcal{C}^+} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\ell_-(c^+(\mathbf{x}))] + \epsilon$.

The notion of $\mathcal{L}$-*negative reliable learning* (NRL) is defined analogously, by switching the positives and the negatives:

Definition 6.2 (NRL for a family of loss functions $\mathcal{L}$ [KKM12; KT14])**.** *A concept class $\mathcal{C}$ of Boolean concepts is* $\mathcal{L}$-negatively reliably learnable *if there exists a learning algorithm that for any distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, any $\ell_+, \ell_- \in \mathcal{L}$, and any $\epsilon, \delta > 0$, when given access to the example oracle $\mathsf{EX}(\mathcal{D})$ and loss functions ℓ_+, ℓ_-, outputs a hypothesis $h : \mathcal{X} \to [0,1]$ that satisfies the following with probability at least $1 - \delta$:*

1. $\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\ell_-(h(\mathbf{x}))] \leq \epsilon$,
2. $\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\ell_+(h(\mathbf{x}))] \leq \min_{c^-\in\mathcal{C}^-} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\ell_+(c^-(\mathbf{x}))] + \epsilon$.

Hence, in the case of the 0-1 loss, a positive reliable classifier is one that almost never produces false positives while simultaneously minimizing false negative errors, attaining a rate comparable to the false negative error rate of the best classifier $c_+ \in \mathcal{C}^+$ [KT14]. Symmetrically, a negative reliable classifier is one that almost never produces false negatives while simultaneously minimizing false positive errors, attaining a rate comparable to the false positive error rate of the best $c_- \in \mathcal{C}^-$.

Note that, by the definitions of $\mathcal{C}^+$ and $\mathcal{C}^-$, the concepts in $\mathcal{C}^+$ and $\mathcal{C}^-$ also satisfy Condition 1 in the respective definitions (i.e., almost no error; in their case, it is exactly 0 error). We remark that

the original definitions by [KKM12] were introduced for the case of the 0-1 loss; we are proposing the generalized versions of $\mathcal{L}$-PRL and $\mathcal{L}$-NRL. In the case of the 0-1 loss, ℓ_+ corresponds to the rate of false positives, and ℓ_- corresponds to the rate of false negatives (which is why we chose the subscripts this way). It can be more intuitive to think about the definitions in the case of the 0-1 loss, so we also include the original [KKM12] definitions for completeness:

$$\mathrm{err}(h, \mathcal{D}) = \Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[h(\mathbf{x}) \neq \mathbf{y}],$$

$$\mathrm{false}_+(h, \mathcal{D}) = \Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[c(\mathbf{x}) = 1 \wedge \mathbf{y} = 0],$$

$$\mathrm{false}_-(h, \mathcal{D}) = \Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[c(\mathbf{x}) = 0 \wedge \mathbf{y} = 1].$$

Note that $\mathrm{err}(h, \mathcal{D}) = \mathrm{false}_+(h, \mathcal{D}) + \mathrm{false}_-(h, \mathcal{D})$.

Definition 6.3 (PRL for the 0-1 loss [KKM12])**.** *A concept class $\mathcal{C}$ of Boolean concepts is* positive reliably learnable *if there exists a learning algorithm that for any distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, any $\epsilon, \delta > 0$, with access to the example oracle $\mathsf{EX}(\mathcal{D})$, outputs a hypothesis $h : \mathcal{X} \to \{0,1\}$ that satisfies the following with probability at least $1 - \delta$,*

1. $\mathrm{false}_+(h, D) \leq \epsilon$,
2. $\mathrm{false}_-(h, D) \leq \min_{c\in\mathcal{C}^+} \mathrm{false}_-(c, D) + \epsilon$, *where* $\mathcal{C}^+ = \{c \in \mathcal{C} \mid \mathrm{false}_+(c, \mathcal{D}) = 0\}$.

Symmetrically, we have that:

Definition 6.4 (NRL for the 0-1 loss [KKM12])**.** *A concept class $\mathcal{C}$ of Boolean concepts is* negative reliably learnable *if there exists a learning algorithm that for any distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, any $\epsilon, \delta > 0$, with access to the example oracle $EX(\mathcal{D})$, outputs a hypothesis $h : \mathcal{X} \to \{0,1\}$ that satisfies the following with probability at least $1 - \delta$,*

1. $\mathrm{false}_-(h, \mathcal{D}) \leq \epsilon$, *and*
2. $\mathrm{false}_+(h, \mathcal{D}) \leq \min_{c\in\mathcal{C}^-(\mathcal{D})} \mathrm{false}_+(c, \mathcal{D}) + \epsilon$, *where* $\mathcal{C}^- = \{c \in \mathcal{C} \mid \mathrm{false}_-(c, \mathcal{D}) = 0\}$.

Adding abstentions. Both PRL and NRL are non-selective classifiers; indeed, we still map to the range $[0, 1]$ rather than $[0, 1] \cup ?$. But if we want *both* the positive and the negative rates to be low (i.e., for the total error to be low), then this is not possible unless we allow for ? to be in the range of h as well. We need to include abstentions to the input concept class $\mathcal{C}$, so that we can compare the abstention rate incurred by our predictor to the best abstention rate in the class. To do so, [KKM12] define the following class of selective classifiers from $\mathcal{C}$, which we call $\mathtt{SC}(\mathcal{C})$ (for "selective classifiers"). Given any $c_+ \in \mathcal{C}^+$ and $c_- \in \mathcal{C}^-$, we ensemble them to construct a selective classifier $c_? = (c_+, c_-)$ as follows:

$$c_?(x) = \begin{cases} 1 & \text{if } c_+(x) = c_-(x) = 1, \\ 0 & \text{if } c_+(x) = c_-(x) = 0, \\ ? & \text{if } c_+(x) \neq c_-(x). \end{cases}$$

We then let $\mathtt{SC}(\mathcal{C}) = \{(c_+, c_-) \mid c_+ \in \mathcal{C}^+, c- \in \mathcal{C}^-\}$ (for "selective classifiers") be another concept class derived from $\mathcal{C}$. As usual, the distribution $\mathcal{D}$ is implicitly in the construction of $\mathtt{SC}(\mathcal{C})$, given that it is in turn implicit in the construction of $\mathcal{C}^-$ and $\mathcal{C}^+$. Importantly, given the definitions of $\mathcal{C}^+$ and $\mathcal{C}^-$, note that all concepts in $\mathtt{SC}$ have exactly 0 error over the non-abstaining region.

We can now define *fully reliable learning* using the base class $\mathtt{SC}(\mathcal{C})$:

Definition 6.5 (FRL for a family of loss functions $\mathcal{L}$ [KKM12; KT14])**.** *A concept class $\mathcal{C}$ is* $\mathcal{L}$-fully reliably learnable *if there exists a learning algorithm that for any distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, any $\ell_+, \ell_-, \ell_? \in \mathcal{L}$, and any $\epsilon, \delta > 0$, when given access to the example oracle $\mathsf{EX}(\mathcal{D})$ and loss functions $\ell_+, \ell_-, \ell_?$, outputs a hypothesis $h_? : \mathcal{X} \to [0,1] \cup \{?\}$ that satisfies the following with probability at least $1 - \delta$:*

1. $\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\ell_+(h_?(\mathbf{x})) + \ell_-(h_?(\mathbf{x}))] \leq \epsilon$,
2. $\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\ell_?(\mathbf{y}, h_?(\mathbf{x}))] \leq \min_{c_?\in\mathtt{SC}(\mathcal{C})} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\ell_?(\mathbf{y}, c_?(\mathbf{x}))] + \epsilon$.

We similarly generalized $\mathcal{L}$-FRL from the original notion of FRL for the 0-1 loss [KKM12]. In that case, the uncertainty $?(h_?, \mathcal{D})$ of a Boolean selective classifier $h_?$ is defined as:

$$?(h_?, \mathcal{D}) = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[h(\mathbf{x}) = ?].$$

Then, the original FRL notion is defined as:

Definition 6.6 (FRL for the 0-1 loss [KKM12])**.** *A concept class $\mathcal{C}$ is* fully reliably learnable *if there exists a learning algorithm that for any distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, any $\epsilon, \delta > 0$, with access to the example oracle $\mathsf{EX}(\mathcal{D})$, outputs a selective classifier $h : \mathcal{X} \to \{0,1,?\}$, that satisfies the following with probability at least $1 - \delta$,*

1. $\mathrm{err}(h_?, \mathcal{D}) = \mathrm{false}_+(h_?, \mathcal{D}) + \mathrm{false}_-(h_?, \mathcal{D}) \leq \epsilon$,
2. $?(h_?, \mathcal{D}) \leq \min_{c_?\in\mathtt{SC}(\mathcal{C})} ?(c_?, \mathcal{D}) + \epsilon$.

Constructing PRL, NRL, and FRL predictors. For the specific case where $\mathcal{L}$ corresponds only to the 0-1 loss, Kalai, Kanade, and Mansour showed that if $\mathcal{C}$ is efficiently agnostically learnable, then $\mathcal{C}$ is also efficiently $\mathcal{L}$-reliable learnable, all for PRL, NRL, and FRL [KKM12]:

Theorem 6.7 ([KKM12])**.** *Let $\mathcal{L}$ contain only the 0-1 loss. If a concept class $\mathcal{C}$ is agnostically learnable under distribution $\mathcal{D}$ in time $T(\epsilon, \delta)$, then $\mathcal{C}$ is $\mathcal{L}$-positively reliably learnable and $\mathcal{L}$-negative reliably learnable, both in time $T(\epsilon^2/2, \delta)$. Then, $\mathcal{C}$ is also $\mathcal{L}$-fully reliably learnable, in time $2T(\epsilon^2/8, \delta/2)$.*

To prove Theorem 6.7, they first construct PRL and NRL predictors, and then construct FRL predictors by ensembling them. Specifically, to construct a PRL predictor from agnostic learning, [KKM12] modify the target function (while leaving the underlying distribution unchanged) such that false positives (with respect to the original function) are penalized much more than false negatives. By choosing the parameters adequately, the output of the black-box agnostic learner is close to the best positive-reliable classifier. The same argument holds symmetrically for the case of NRL.

To construct FRL predictors, we call our now existing oracles for PRL and NRL to get a PRL predictor h_+ and a NRL predictor h_- with the appropriate parameters. Then, we can construct a selective classifier $h_? : \mathcal{X} \to \{0, 1, ?\}$ exactly in the way that we defined $\mathtt{SC}(\mathcal{C})$. Namely,

$$h_?(x) = \begin{cases} 1 & \text{if } h_+(x) = h_-(x) = 1, \\ 0 & \text{if } h_+(x) = h_-(x) = 0, \\ ? & \text{if } h_+(x) \neq h_-(x). \end{cases}$$

The key idea here is that, whenever $h_?$ abstains, one of h_+, h_- is making an error. The same is true for the class $\mathtt{SC}(\mathcal{C})$: whenever any $c_? = (c_+, c_-) \in \mathtt{SC}(\mathcal{C})$ abstains, one of c_+ or c_- is making an error. In the case of c_+, c_- however, since by definition of $\mathcal{C}^+, \mathcal{C}^-$ these concepts only make errors of one type (for $\mathbf{y} = 0$ or $\mathbf{y} = 1$), then the abstention rate of $c_?$ is exactly equal to the sum of the errors of c_+ and c_-.

Throughout the subsequent statements and proofs, by "efficiently" we mean that the algorithm runs in polynomial time in the appropriate parameters. We drop the failure probability δ from the statements.

6.2 Generalized Chow model

The PQ-learning and Chow models. The reliable agnostic framework is closely related to two other well-studied theoretical models for learning with abstentions in the learning theory literature. The first is the PQ-learning model introduced by Goldwasser, Kalai, Kalai, and Montasser [GKKM20]. In their work, adding abstentions to learning is motivated by the covariate shift problem, in which the training data is distributed according to P and the test data according to Q, where P and Q can be arbitrary distributions over the domain $\mathcal{X}$. This generic form of learning is not possible to do in general, given that P and Q might not even overlap. To make this problem tractable, Goldwasser et al. introduce the model of *PQ-learning*, where the learner has access to unlabeled test examples from Q and the option to abstain on any point $x \in \mathcal{X}$. They consider the rejection rate of the algorithm (i.e., the fraction of $\mathcal{X}$ over which the classifier abstains) and the misclassification rate, which quantifies the error of the classifier only over the subset of the domain on which the classifier does not abstain. Goldwasser et al. give algorithms for building selective classifiers in the PQ-learning model which guarantee low test error rate and low rejection rate with respect to P for concept classes of bounded VC dimension [GKKM20]. Their algorithm is efficient if we have access to an Empirical Risk Minimizer (ERM) for $\mathcal{C}$. Note that classes $\mathcal{C}$ of bounded VC dimension, being able to do ERM efficiently is equivalent to efficient proper agnostic learning [KK21a].

In a follow-up work, Kalai and Kanade showed that PQ-learning is in fact equivalent to reliable learning. Moreover, they provide further evidence that the computational hardness of PQ-learning and reliable learning lies in-between PAC and agnostic learning (under the usual hardness assumptions), by showing that the class of parities is PQ-learnable. [KK21b]. This separation was already shown by Kanade and Thaler, who gave an algorithm for reliably learning majorities over $\{0,1\}^d$ in time $2^{\tilde{O}(\sqrt{d})}$, whereas there are no known agnostic learning algorithms for this problem that run in time less than $2^{\Omega(d)}$ [KT14].

In another follow-up work, Kalai and Kanade consider a different formulation of the selective classification problem considered by Goldwasser et al., which is based on Chow's abstention model [Cho57]. Here, instead of finding a trade-off between two error rates (namely the rejection rate and the misclassification rate), we have a fixed parameter $\alpha > 0$ which corresponds to the abstention cost. I.e., for each $x \in \mathcal{X}$, we either make a prediction and suffer a potential prediction loss, or we abstain and pay a price of α. Importantly, Kalai and Kanade show that this is a stronger learning model than that of PQ-learning/reliable learning, given their guarantees obtained in the Chow model imply the PQ-bounds from [GKKM20], but the but the reverse does not hold. Specifically, [KK21b] are able to by-pass the lower bounds shown in [GKKM20]. To optimize over the total loss (i.e., the sum of the prediction loss and the abstention loss), [KK21b] also use an ERM oracle.

The generalized Chow model. In light of the results obtained in [GKKM20; KK21a; KK21b], we formulate our selective omnipredictors following the Chow model, given that it is the strongest learning-theoretic model. We use a generalized version of the model, which we call the *generalized Chow loss function*:

Definition 6.8 (Generalized Chow loss)**.** *Given a triplet of loss functions $(\ell_+, \ell_-, \ell_?)$ induced by $\ell : \mathcal{Y} \times \{\mathbb{R} \cup \{?\}\} \to \mathbb{R}$ with associated weights (λ, μ, ν) and a selective classifier $h_? : \mathcal{X} \to \mathbb{R} \cup \{?\}$, the* generalized Chow loss *incurred by $h_?$ is equal to*

$$\ell_{\mathrm{GC},\mathcal{D}}(h_?; \lambda, \mu, \nu) = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\lambda\ell_+(h(\mathbf{x})) + \mu\ell_-(h(\mathbf{x})) + \nu\ell_?(\mathbf{y})].$$

One can include randomized predictors $h_?$ to this definition (where $h_?$ assigns a *probability* of abstaining to each $x \in \mathcal{X}$); however, as we show later, randomization does not help in minimizing the generalized Chow loss function.

6.3 Selective omniprediction

Recall the notion of *omniprediction* that we described in Chapter 2.4. Our goal is to introduce the notion of abstentions to the omnipredictors framework. This goal is useful and beneficial in both directions: for one, it expands the multigroup fairness framework from being entirely prediction-based to being able to incorporate abstentions. The multigroup fairness framework originated as an algorithmic fairness metric, and as we discussed in the introduction, it is natural and important to be able to include abstentions in any automated decision-making process. In the other direction, as we will show, the omniprediction framework allows us efficiently build predictors that obtain optimal generalized Chow loss for any such loss in a very rich class of loss functions. Moreover, by virtue of omniprediction, we can change the generalized Chow loss function at any point after training, and our predictor (post-processed accordingly) will always remain optimal with respect to the base concept class. This learning paradigm is extremely useful in practice: for example, this allows us to change the abstention costs over time, or to change the cost of false positives or false negatives over time. One can envision many practical settings in which this flexibility is highly desirable; e.g., if not catching patients with a specific illness becomes increasingly more dangerous. Note that our framework also allows us to separate the costs of both prediction and abstention for the cases of $y = 1$ and $y = 0$ respectively.

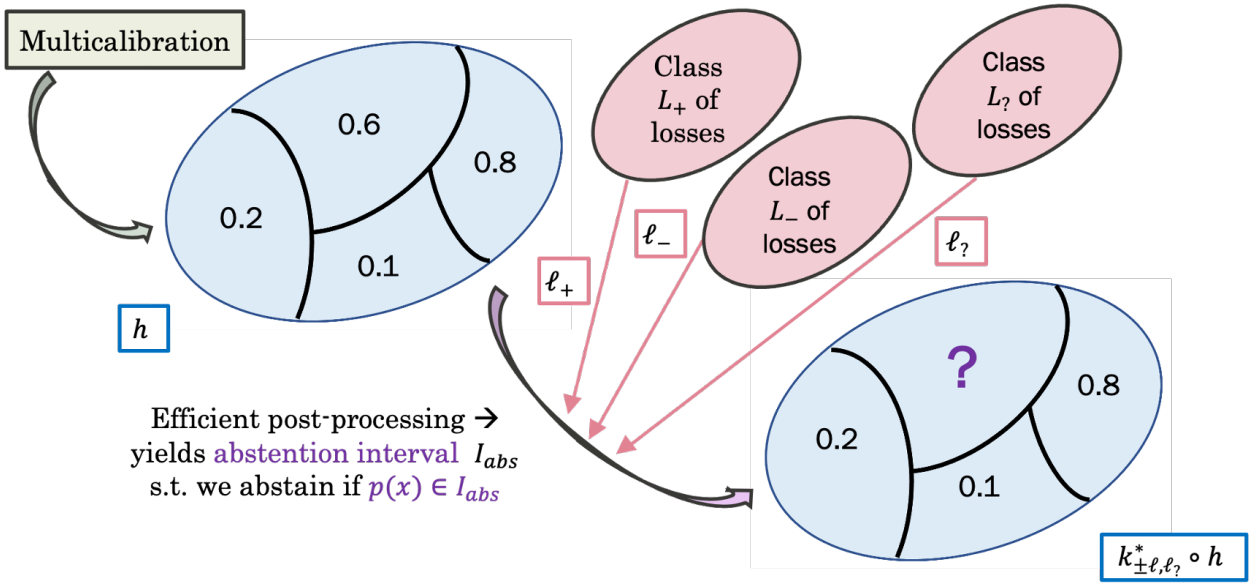


Figure 6.1: Selective omniprediction.

Following the original notion of omnipredictors, and using our generalized Chow loss function, we define the notion of a selective omnipredictor as follows:

Definition 6.9 (Selective omniprediction)**.** *Given a concept class $\mathcal{C}$ of concepts $c : \mathcal{X} \to \mathbb{R}$, distribution $\mathcal{D}$, $\epsilon > 0$, and a class of loss functions $\mathcal{L}$, we say that a predictor $h : \mathcal{X} \to [0,1]$ is a* $(\mathcal{L}, \mathcal{C}, \epsilon)$-selective omnipredictor *if for every $(\ell_+, \ell_-, \ell_?) \in \mathcal{L}$ and any associated weights (λ, μ, ν), there exists a function $k^*_{\ell_\pm,\ell_?} : [0,1] \to \mathbb{R} \cup \{?\}$ such that for any post-processing function $k : [0,1] \to \mathbb{R} \cup \{?\}$,*

$$\ell_{\mathrm{GC},\mathcal{D}}(k^*_{\ell_\pm,\ell_?} \circ h; \lambda, \mu, \nu) \leq \min_{c_? \in k \circ \mathcal{C}} \ell_{\mathrm{GC},\mathcal{D}}(c_?; \lambda, \mu, \nu) + \epsilon.$$

We remark that for all statements and definitions concerning omnipredictors and selective omnipredictors, the concepts in $\mathcal{C}$ always have range $\mathbb{R}$.

Importantly, as in the original omniprediction framework, the optimal classifier $c_?$ in $k \circ \mathcal{C}$ is tailored to the specific triplet of loss functions, whereas h is a *single* classifier for all of $\mathcal{L}$ and (λ, μ, ν). In the next chapter, we show that we can efficiently construct a $(\mathcal{L}, \mathcal{C})$-selective classifier from a $\mathcal{C}$-multicalibrated predictor. Moreover, we will show that it is an optimal strategy to map each level set of the multicalibrated predictor to ? or not (if not, we might still modify the p-values in the post-processing step, but we always maintain the same level sets). Moreover, all of the p-values on which we optimally abstain form a continguous interval, which we can compute directly from the generalized Chow function *independently of the data*. We summarize the notion of a $(\mathcal{L}, \mathcal{C})$-selective omnipredictor in Figure 6.1.

7

Selective Omniprediction & Reliable Learning

Loss minimization is a dominant paradigm in machine learning, where a predictor is trained to minimize some loss function that depends on an uncertain event (e.g., "will it rain tomorrow?"). Different loss functions imply different learning algorithms and, at times, very different predictors. While widespread and appealing, a clear drawback of this approach is that the loss function may not be known at the time of learning, requiring the algorithm to use a best-guess loss function.

Gopalan, Kalai, Reingold, Sharan, and Wieder [GKR+22]

In this chapter, we show how to construct selective omnipredictors efficiently for a rich class of loss functions. We show how we can determine the provably optimal interval of abstention directly from the chosen generalized Chow loss, independently from the data. We give concrete examples with typical examples of loss functions as to provide more intuition, and we also provide experiments on synthetic data to demonstrate the feasibility and efficiency of selective omniprediction in practice. Lastly, we show how we can use selective omnipredictors to obtain $\mathcal{L}$-fully reliable learners for any loss function in a rich class of loss functions $\mathcal{L}$, thus generalizing the results of [KKM12] for the 0-1 loss.

7.1 Constructing selective omnipredictors efficiently

As we have explained throughout this thesis, the key idea in the omniprediction framework is to first learn a model p that is $(\mathcal{C}, \epsilon)$-computationally indistinguishable from the ground truth optimal predictor p^*, which is accomplished through the technique of multicalibration, and then apply a post-processing function k^*_ℓ to p once a loss function $\ell \in \mathcal{L}$ has been fixed. In our setting of selective classification, similar to the original paper on omniprediction [GKR+22], we use the following post-processing function that minimizes expected loss under the Bernoulli distribution, which we show yields an optimal final loss (in the agnostic sense):

Definition 7.1. *Given loss functions $(\ell_+, \ell_-, \ell_?)$ and corresponding weights (λ, μ, ν), let the function $k^*_{\ell_\pm,\ell_?} : [0,1] \to \mathbb{R} \cup \{?\}$ be defined as*

$$\begin{aligned} k^*_{\ell_\pm,\ell_?}(p) &= \arg\min_{t \in \mathbb{R}\cup\{?\}} \mathop{\mathbb{E}}_{\mathbf{y}\sim\mathrm{Bern}(p)} [\lambda\ell_+(t) + \mu\ell_-(t) + \nu\ell_?(\mathbf{y})] \\ &= \arg\min_{t \in \mathbb{R}\cup\{?\}} p \cdot \Big(\mu\ell_-(t) + \nu\alpha_+\Big) + (1-p) \cdot \Big(\lambda\ell_+(t) + \nu\alpha_-\Big). \end{aligned}$$

Intuitively, this is the right post-processing function because it is the optimal best-response for the true predictor p^*, given that the true labels are distributed as $\mathrm{Bern}(p^*(\mathbf{x}))$. Given that p is $(\mathcal{C}, \epsilon)$-multicalibrated for p^*, one can show that this is also the best-response for p.

Our main result in this section is the feasibility of efficiently constructing selective classifiers. We remark that recent work has shown how we can construct omnipredictors from the weaker primitive of (proper) calibrated multiaccuracy [OKK25]. However, for our construction, we seem to require the full power of multicalibration. It is a very interesting future direction to understand whether we can construct selective omnipredictors from a weaker primitive than multicalibration and to determine how that relates to the learning primitives required to construct an omnipredictor. For example, it is unclear whether we can have a general reduction from an omnipredictor to a selective omnipredictor.

Theorem 7.2 (Constructing selective omnipredictors)**.** *Let $\mathcal{C}$ be a concept class of concepts $c : \mathcal{X} \to [-M, M]$, $\mathcal{D}$ a distribution on $\mathcal{X} \times \{0,1\}$, $\epsilon > 0$, and $\mathcal{L}$ a family loss functions with associated weights (λ, μ, ν) with $\lambda + \mu + \nu \le 1$, such that all $\ell \in \mathcal{L}$ are B-Lipschitz. Then, a $(\mathcal{C}, \epsilon)$-multicalibrated predictor is a $(\mathcal{L}, \mathcal{C}, 4\epsilon\beta + \epsilon B)$-selective omnipredictor, where β is an absolute bound on $\ell_+, \ell_-, \ell_?$.*

Remark 7.3. *The condition $\lambda + \mu + \nu \le 1$ is to ensure that the additive error term does not scale up; equivalently we could just get $2B(\lambda + \mu + \nu)\epsilon$ as the additive error.*

Proof. Given the concept class $\mathcal{C}$ and parameter ϵ, we discretize each $c \in \mathcal{C}$ to precision ϵ (i.e., into $\lceil 1/\epsilon \rceil$ many buckets). We denote these discretized concepts by $\hat{c}$ and the corresponding concept class by $\hat{\mathcal{C}}$. Because all loss functions are B-Lipschitz, discretizing the concepts to precision ϵ incurs an additive error of at most ϵB.

We begin by calling the multicalibration theorem of [HKRR18; GHK+23] with $\mathcal{X}, \mathcal{D}, \epsilon$, and $\hat{C}$ to obtain a $(\hat{\mathcal{C}}, \epsilon)$-multicalibrated predictor h. For any fixed loss function $\ell = (\ell_+, \ell_-, \ell_?) \in \mathcal{L}$, we claim that $k^*_{\ell_\pm,\ell_?} \circ h$, where $k^*_{\ell_\pm,\ell_?}$ is the post-processing function defined in Definition 7.1, is a selective omnipredictor.

By the definition of a selective omnipredictor, we want to show that generalized Chow loss incurred by $k^*_{\ell_\pm,\ell_?} \circ h$ is upper-bounded by the generalized Chow loss incurred by $k \circ c$ for every $c \in \mathcal{C}$, where $k : [0,1] \to \mathbb{R} \cup \{?\}$ is an arbitrary post-processing function. By definition of $k^*_{\ell_\pm,\ell_?}$, recall that

$$k^*_{\ell_\pm,\ell_?}(p) = \mathrm{argmin}_{t \in \mathbb{R}\cup\{?\}} p \cdot \Big(\mu\ell_-(t) + \nu\ell_?(1)\Big) + (1-p) \cdot \Big(\lambda\ell_+(t) + \nu\ell_?(0)\Big).$$

Following this generalized Chow loss expression (Definition 6.8), we decompose it into the expected cost of predicting, which we denote by κ_{pred}, and the expected cost of abstaining, which we denote

by κ_{abs}, both under the Bernoulli distribution (i.e., for $\mathbf{y} \sim \text{Bern}(p(\mathbf{x}))$ for each $p \in \text{range}(h)$):

$$\kappa_{\text{pred}}(p) = \min_t p \cdot \mu\ell_-(t) + (1-p) \cdot \lambda\ell_+(t),$$

$$\kappa_{\text{abs}}(p) = \nu\big(p \cdot \ell_?(1) + (1-p) \cdot \ell_?(0)\big) = \nu\big(p \cdot \alpha_+ + (1-p) \cdot \alpha_-\big).$$

Note that the prediction cost depends on the value of t, whereas the abstention cost is independent of t. Then, for each point p, in order to minimize the expected generalized Chow loss under the Bernoulli distribution $\text{Bern}(p(\mathbf{x}))$, we proceed in two steps:

- Find the value $t^*(p) \in [-M, M]$ that minimizes the value of $\kappa_{\text{pred}}(p)$ given a fixed value of p. This depends only on the choice of $\mu, \ell_-, \lambda, \ell_+$ and not on the underlying data. This corresponds to finding the $t^*(p)$ value that minimizes the value of $k^*_\ell(p)$.
- For each predicted value $h(x) = p$, we map it to either $t^*(p)$ or ? depending on whether it is cheaper to predict or to abstain; that is, depending on the value of $\min\{\kappa_{\text{pred}}(t^*(p)), \kappa_{\text{abs}}\}$.

In other words, we can re-write $k^*_{\ell_\pm, \ell_?}$ as $k_{\text{abs}} \circ k^*_\ell$, where

$$k_{\text{abs}} = \begin{cases} t^* & \text{if } \kappa_{\text{pred}} \leq \kappa_{\text{abs}} \\ ? & \text{if } \kappa_{\text{pred}} > \kappa_{\text{abs}} \end{cases}$$

The key idea is the following: the true labels are distributed as $\mathbf{y} \sim \text{Bern}(p^*(\mathbf{x}))$. So if we had access to the true $p^*(x)$ value for each x, then the optimal prediction/abstention decision rule (i.e., the post-processing function applied to the p^* value that yields the minimum total generalized Chow loss) is precisely k_{abs}.

We can write the cost function incurred by the post-processing function $k^*_{\ell_\pm, \ell_?} = k_{\text{abs}} \circ k^*_\ell$ as:

$$\kappa(t, p) = \min\{p \cdot \mu\ell_-(t) + (1-p) \cdot \lambda\ell_+(t), p \cdot \ell_?(1) + (1-p) \cdot \ell_?(0)\}.$$

An important point to remark is that, while the prediction/abstention decision rule is decided with the p-values (to which we have access to, since they correspond to the predictions of the multicalibrated predictor h), the actual cost that we incur is computed with the true values p^*. Multicalibration precisely allows us to bridge this gap: the predictor h believes that the labels are distributed according to $\text{Bern}(p(\mathbf{x}))$, and so it uses the function $k_{\text{abs}} \circ k^*_\ell$ as post-processing, which yields the optimal cost under the distribution $\mathbf{y} \sim \text{Bern}(p(\mathbf{x}))$. In order to bridge the "simulated" labels $\mathbf{y} \sim \text{Bern}(p(\mathbf{x}))$, which are used in our decision rule, and the "true" labels $\mathbf{y} \sim \text{Bern}(p^*(\mathbf{x}))$, which yield the actual cost that we pay, we use the fact that h is $\mathcal{C}$-multicalibrated, which ensures that

$$\mathbb{E}[\mathbf{y} \mid h = p, \hat{c} = \gamma] \approx \mathbb{E}[p \mid h = p, c = \gamma]$$

for each γ in the range of $\hat{c}$. The RHS can equivalently be written as $\mathbb{E}[h \mid h = p, \hat{c} = \gamma]$. This "bridging" enabled by the multicalibration property satisfied by the predictor h is what allows us to show that our selective omnipredictor (namely, the predictor $k^*_{\ell_\pm, \ell_?}$) incurs optimal loss with respect to the class $\hat{C}$.

We have discussed how the prediction/abstention decision rule for the multicalibrated predictor is given by $\kappa(t^*(p), p)$. As per the definition of a selective omnipredictor, we need to show that the

generalized Chow loss incurred by $k^*_{\ell_\pm,\ell_?}$ is upper-bounded by the generalized Chow loss incurred by any of the selective concepts $c_?$, where $c \in \mathcal{C}$ and $k : [0,1] \to \mathbb{R} \cup \{?\}$ is *any* post-processing function that adds abstentions to the concepts $c \in \mathcal{C}$. The only natural restriction on k is that it is a function of the values $\hat{c}(\cdot)$, and cannot be a function of x. E.g., if $\hat{c}(x_1) = \gamma = \hat{c}(x_2)$, then it must be that $\hat{c}_?(x_1) = \hat{c}_?(x_2)$. Hence, in the case of the concepts $\hat{c} \in \hat{C}$, we cannot directly assume that the prediction/abstention decision rule corresponds to $k^*_{\ell_\pm,\ell_?}$ as well.

However, we can use the multicalibration condition to reason about the loss incurred by the concepts $\hat{c}$. Specifically, for each concept $\hat{c} \in \hat{\mathcal{C}}$, we define the sets $\mathcal{X}_{(p,\gamma)} = \{x \in \mathcal{X} \mid h(x) = p, \hat{c}(x) = \gamma\}$ for each $p \in \text{range}(h)$ and each $\gamma \in \text{range}(\hat{c})$. Moreover, we let

$$\phi_{(p,\gamma)} = \mathbb{E}[\mathbf{y} | \mathbf{x} \in \mathcal{X}_{(p,\gamma)}].$$

The fact that h is $(\hat{C}, \epsilon)$-multicalibrated implies that

$$\phi_{(p,\gamma)} = \mathbb{E}[\mathbf{y} \mid h(\mathbf{x}) = p, \hat{c}(\mathbf{x}) = \gamma] \approx_\epsilon \mathbb{E}[h(\mathbf{x}) \mid h(\mathbf{x}) = p, \hat{c}(\mathbf{x}) = \gamma] = p \implies \phi_{(p,\gamma)} \approx_\epsilon p.$$

In practice, the multicalibration condition applies on expectation over the level sets $\mathcal{X}_{(p,\gamma)}$ for all p, γ. We fix a level set $h(x) = p$ of h and a concept $\hat{c} \in \hat{C}$. We want to compare the loss incurred by $k^*_{\ell_\pm,\ell_?}(h)$ with the loss incurred by $k \circ c$, where $k : [0,1] \to \mathbb{R} \cup \{?\}$ is any post-processing function. Within the level set $h(x) = p$, all of the values $k^*_{\ell_\pm,\ell_?}(h)$ are the same, and so $k^*_{\ell_\pm,\ell_?}(h)$ is either predicting the value $t^*(p)$ on all of the points in the level set $h(x) = p$, or abstaining in all of the points in the level set, as determined by the cost function $\kappa(t^*(p), p)$. Within the level set $h(x) = p$, we further partition it according to the level sets of $\hat{c}$. That is, we consider the partition of $\mathcal{X}_p$ into the sets $\mathcal{X}_{(p,\gamma)}$ for each $\gamma \in \text{range}(\hat{c})$. For each $x \in \mathcal{X}_{(p,\gamma)}$, $\hat{c}$ either predicts or abstains, using a decision rule k that is allowed to be arbitrary. For each set $\mathcal{X}_{(p,\gamma)}$, we split the proof into 4 cases, depending on whether h and $\hat{c}$ decide to abstain or predict as per their respective decision rules.

Throughout, we let β denote a bound on the absolute values of $\ell_+, \ell_-, \ell_?$. Moreover, we can write the loss function as

$$\ell_{\text{GC},\mathcal{D}}(k^*_{\ell_\pm,\ell_?} \circ h; \lambda, \mu, \nu) = \mathop{\mathbb{E}}_{p \in \text{range}(h)} \mathop{\mathbb{E}}_{\gamma \in \text{range}(\hat{c})} \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}|_{\mathcal{X}_{(p,\gamma)}}} [\ell_{\text{GC},\mathcal{D}}(\mathbf{y}, k^*_{\ell_\pm,\ell_?}(h(\mathbf{x})))],$$

and similarly for $k \circ \hat{c}$.

Having fixed the values $h = p$ and $\hat{c} = \gamma$, we argue about the expected loss incurred by the post-processed multicalibrated predictor versus the expected loss incurred by the concept on the level set $\mathcal{X}_{(p,\gamma)}$.

1. $k^*_{\ell_\pm,\ell_?} \circ h$ **predicts &** $k \circ \hat{c}$ **predicts.** We begin by swapping $\phi_{(p,\gamma)}$ for p in the following expression:

$$\begin{aligned}\phi_{(p,\gamma)} \cdot \mu\ell_-(t^*(p)) + (1 - \phi_{(p,\gamma)}) \cdot \lambda\ell_+(t^*(p)) = p \cdot \mu\ell_-(t^*(p)) + (1-p) \cdot \lambda\ell_+(t^*(p)) \\ + (\phi_{(p,\gamma)} - p)(\mu\ell_-(t^*(p)) - \lambda\ell_+(t^*(p)).\end{aligned}$$

By definition of $t^*(p)$, it follows that $t^*(p)$ is the minimizer of $\kappa_{\text{pred}}(p)$ in $[-M, M]$ given a fixed

value of p. Hence,

$$p \cdot \mu\ell_-(t^*(p)) + (1-p) \cdot \lambda\ell_+(t^*(p)) \leq p \cdot \mu\ell_-(\gamma) + (1-p) \cdot \lambda\ell_+(\gamma).$$

Swapping p for $\phi_{(p,\gamma)}$ again, we get that

$$p \cdot \mu\ell_-(\gamma) + (1-p) \cdot \lambda\ell_+(\gamma) + (\phi_{(p,\gamma)} - p)(\mu\ell_-(t^*(p)) - \lambda\ell_+(t^*(p))$$

$$= \phi_{(p,\gamma)} \cdot \mu\ell_-(\gamma) + (1-\phi_{(p,\gamma)}) \cdot \lambda\ell_+(\gamma) + (\phi_{(p,\gamma)} - p)\big[\big(\mu\ell_-(t^*(p)) - \lambda\ell_+(t^*(p))\big) - \big(\mu\ell_-(\gamma) - \lambda\ell_+(\gamma)\big)\big].$$

Putting everything together, we get that

$$\phi_{(p,\gamma)} \cdot \mu\ell_-(t^*(p)) + (1-\phi_{(p,\gamma)}) \cdot \lambda\ell_+(t^*(p))$$

$$= \phi_{(p,\gamma)} \cdot \mu\ell_-(\gamma) + (1-\phi_{(p,\gamma)}) \cdot \lambda\ell_+(\gamma) + (\phi_{(p,\gamma)} - p)\big[\big(\mu\ell_-(t^*(p)) - \ell\lambda\ell_+(t^*(p)\big) - \big(\mu\ell_-(\gamma) - \lambda\ell_+(\gamma)\big)\big].$$

By the β-bound on the loss functions, and given that $\lambda + \mu + \nu \leq 1$, it follows that

$$(\phi_{(p,\gamma)} - p)\big[\big(\mu\ell_-(t^*(p)) - \ell\lambda\ell_+(t^*(p)\big) - \big(\mu\ell_-(\gamma) - \lambda\ell_+(\gamma)\big)\big] \leq |\phi_{(p,\gamma)} - p| \cdot 4\beta.$$

Therefore, over $\mathcal{X}_{(p,\gamma)}$, where $k^*_{\ell_\pm,\ell_?} \circ h$ is predicting p and $k \circ \hat{c}$ is predicting γ, the expected generalized Chow losses compare as follows:

$$\ell_{\mathrm{GC},\mathcal{D}}(k^*_{\ell_\pm,\ell_?} \circ h; \lambda, \mu, \nu \mid \mathbf{x} \in \mathcal{X}_{(p,\gamma)}) \leq \ell_{\mathrm{GC},\mathcal{D}}(k \circ \hat{c}; \lambda, \mu, \nu \mid \mathbf{x} \in \mathcal{X}_{(p,\gamma)}) + |\phi_{(p,\gamma)} - p| \cdot 4\beta.$$

2. $k^*_{\ell_\pm,\ell_?} \circ h$ **predicts &** $k \circ \hat{c}$ **abstains.** As in the previous case, we swap $\rho_{(p,\gamma)}$ by p:

$$\begin{aligned}\phi_{(p,\gamma)} \cdot \mu\ell_-(t^*(p)) + (1-\phi_{(p,\gamma)}) \cdot \lambda\ell_+(t^*(p)) = p \cdot \mu\ell_-(t^*(p)) + (1-p) \cdot \lambda\ell_+(t^*(p)) \\ + (\phi_{(p,\gamma)} - p)(\mu\ell_-(t^*(p)) - \lambda\ell_+(t^*(p)).\end{aligned}$$

By the decision rule for h determined by the value of $\kappa(t^*(p), p)$, if $k^*_{\ell_\pm,\ell_?} \circ h$ predicts on $X_{(p,\gamma)}$ this implies that

$$p \cdot \mu\ell_-(t^*(p)) + (1-p) \cdot \lambda\ell_+(t^*(p)) \leq p \cdot \nu\ell_?(1) + (1-p) \cdot \nu\ell_?(0).$$

Again swapping p for $\phi_{(p,\gamma)}$, we obtain:

$$p \cdot \nu\ell_?(1) + (1-p) \cdot \nu\ell_?(0) + (\phi_{(p,\gamma)} - p)(\mu\ell_-(t^*(p)) - \lambda\ell_+(t^*(p))$$

$$= \phi_{(p,\gamma)} \cdot \nu\ell_?(1) + (1-\phi_{(p,\gamma)}) \cdot \nu\ell_?(0) + (\phi_{(p,\gamma)} - p)\Big[\big(\mu\ell_-(t^*(p)) - \lambda\ell_+(t^*(p))\big) - \big(\nu\ell_?(1) - \nu\ell_?(0)\big)\Big].$$

By the β-bound on the loss functions, and given that $\lambda + \mu + \nu \leq 1$, it follows that

$$(\phi_{(p,\gamma)} - p)\Big[\big(\mu\ell_-(t^*(p)) - \lambda\ell_+(t^*(p))\big) - \big(\nu\ell_?(1) - \nu\ell_?(0)\big)\Big] \leq |\phi_{(p,\gamma)} - p| \cdot 4\beta.$$

Therefore, putting everything together, we obtain that

$$\ell_{\mathrm{GC},\mathcal{D}}(k^*_{\ell_\pm,\ell_?} \circ h; \lambda, \mu, \nu \mid \mathbf{x} \in \mathcal{X}_{(p,\gamma)}) \leq \ell_{\mathrm{GC},\mathcal{D}}(k \circ \hat{c}; \lambda, \mu, \nu \mid \mathbf{x} \in \mathcal{X}_{(p,\gamma)}) + |\phi_{(p,\gamma)} - p| \cdot 4\beta.$$

3. $k^*_{\ell_\pm,\ell_?} \circ h$ **abstains &** $k \circ \hat{c}$ **predicts.** We start by swapping ρ for p as in the previous two cases:

$$\phi_{(p,\gamma)} \cdot \nu\ell_?(1) + (1 - \phi_{(p,\gamma)}) \cdot \nu\ell_?(0) = p \cdot \nu\ell_?(1) + (1-p) \cdot \nu\ell_?(0) + (\phi_{(p,\gamma)} - p)(\nu\ell_?(1) - \nu\ell_?(0)).$$

Because $k^*_{\ell_\pm,\ell_?} \circ h$ abstains on this level set, it must be that

$$p \cdot \nu\ell_?(1) + (1-p) \cdot \nu\ell_?(0) \leq p \cdot \mu\ell_-(t^*(p)) + (1-p) \cdot \lambda\ell_+(t^*(p)).$$

By definition of $t^*(p)$ as the minimizer of κ_{pred}, for any value of γ we have that

$$p \cdot \mu\ell_-(t^*(p)) + (1-p) \cdot \lambda\ell_+(t^*(p)) \leq p \cdot \mu\ell_-(\gamma) + (1-p) \cdot \lambda\ell_+(\gamma).$$

We again switch p back to $\phi_{(p,\gamma)}$:

$$p \cdot \mu\ell_-(\gamma) + (1-p) \cdot \lambda\ell_+(\gamma) + (\phi_{(p,\gamma)} - p)(\nu\ell_?(1) - \nu\ell_?(0))$$

$$= \phi_{(p,\gamma)} \cdot \mu\ell_-(\gamma) + (1 - \phi_{(p,\gamma)})) \cdot \lambda\ell_+(\gamma) + (\phi_{(p,\gamma)} - p)\big[\big(\nu\ell_?(1) - \nu\ell_?(0)\big) - \big(\mu\ell_-(\gamma) + \lambda\ell_+(\gamma)\big)\big]$$

By the β-bound on the loss functions, and given that $\lambda + \mu + \nu \leq 1$, it follows that

$$(\phi_{(p,\gamma)} - p)\big[\big(\nu\ell_?(1) - \nu\ell_?(0)\big) - \big(\mu\ell_-(\gamma) + \lambda\ell_+(\gamma)\big)\big] \leq |\phi_{(p,\gamma)} - p| \cdot 4\beta.$$

Putting everything together, we obtain that

$$\ell_{\text{GC},\mathcal{D}}(k^*_{\ell_\pm,\ell_?} \circ h; \lambda, \mu, \nu \mid \mathbf{x} \in \mathcal{X}_{(p,\gamma)}) \leq \ell_{\text{GC},\mathcal{D}}(k \circ \hat{c}; \lambda, \mu, \nu \mid \mathbf{x} \in \mathcal{X}_{(p,\gamma)}) + |\phi_{(p,\gamma)} - p| \cdot 4\beta.$$

4. $k^*_{\ell_\pm,\ell_?} \circ h$ **abstains &** $k \circ \hat{c}$ **abstains.** In this case, given that the values of $\ell_?(1)$ and $\ell_?(0)$ are independent of t, it directly follows that both $k^*_{\ell_\pm,\ell_?} \circ h$ and $k \circ \hat{c}$ incur the exact same generalized Chow loss on $\mathcal{X}_{(p,\gamma)}$.

Conclusion. Putting these four cases together, and by taking the expected value over all level sets $\mathcal{X}_{(p,\gamma)}$, for all p in the range of h and γ in the range of $\hat{c}$, and by thus applying the multi-calibration guarantee on $\mathbb{E}[|\phi_{(p,\gamma)} - p|]$ (i.e., which guarantees that $\mathbb{E}[|\phi_{(p,\gamma)} - p|] \leq \epsilon$), we obtain that

$$\ell_{\text{GC},\mathcal{D}}(k^*_{\ell_\pm,\ell_?} \circ h; \lambda, \mu, \nu) \leq \ell_{\text{GC},\mathcal{D}}(k \circ c; \lambda, \mu, \nu) + 4\epsilon\beta.$$

Because these four cases are exhaustive and hold for all values of p and γ, we conclude that

$$\begin{aligned}
\ell_{\text{GC},\mathcal{D}}(k^*_{\ell_\pm,\ell_?} \circ h; \lambda, \mu, \nu) &= \mathop{\mathbb{E}}_{p \in \text{range}(h)} \mathop{\mathbb{E}}_{\gamma \in \text{range}(\hat{c})} \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}|_{\mathcal{X}_{(\phi,\gamma)}}} [\ell_{\text{GC},\mathcal{D}}(\mathbf{y}, k^*_{\ell_\pm,\ell_?}(h(\mathbf{x})))] \\
&\leq \mathop{\mathbb{E}}_{p \in \text{range}(h)} \mathop{\mathbb{E}}_{\gamma \in \text{range}(\hat{c})} \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}|_{\mathcal{X}_{(\phi,\gamma)}}} [\ell_{\text{GC},\mathcal{D}}(\mathbf{y}, k(c(\mathbf{x})))] + 4\epsilon\beta \\
&= \ell_{\text{GC},\mathcal{D}}(k \circ c; \lambda, \mu, \nu) + 4\epsilon\beta.
\end{aligned}$$

Therefore, for the non-discretized concept class $\mathcal{C}$, and accounting for ϵB loss incurred in the clippings of each of c and $t^*(p)$, we conclude that

$$\ell_{\mathrm{GC},\mathcal{D}}(k^*_{\ell_\pm,\ell_?} \circ h; \lambda, \mu, \nu) \leq \ell_{\mathrm{GC},\mathcal{D}}(k \circ c; \lambda, \mu, \nu) + 4\epsilon\beta + \epsilon B$$

for all loss functions in $\mathcal{L}$, and hence h is a selective omnipredictor, as we wanted to show.

Then, the value of $k^*_{\ell_\pm,\ell_?}$ can be computed efficiently because (1) computing the value of $\kappa_{\mathrm{pred}}(p)$ corresponds to solving a one-dimensional minimization problem, and (2) the value of $k^*_{\ell_\pm,\ell_?}$ is then fully determined from the values $\kappa_{\mathrm{pred}}(p)$ with the optimal $t = t^*(p)$ and of κ_{abs}, independent from the data.

Lastly, we show that allowing for randomized selective predictors does not help in minimizing the generalized Chow loss function. Suppose that the selective predictor was randomized, such that for each point $x \in \mathcal{X}$ it would predict a value $a \in [0, 1]$ indicating the probability of abstention on x. Then, for a fixed t we can write our post-processing function as

$$k^*_{\ell_\pm,\ell_?}(p) = \mathrm{argmin}_{t,a}(1-a)\kappa_{\mathrm{pred}} + a\kappa_{\mathrm{abs}} = (\kappa_{\mathrm{abs}} - \kappa_{\mathrm{pred}}) \cdot a + \kappa_{\mathrm{pred}}.$$

For a fixed $t \in \mathbb{R} \cup \{?\}$ (and once the loss functions have been fixed), the total loss only depends on κ_{pred} and κ_{abs}, which in turn only depend on p. Note that after fixing t, $k^*_{\ell,a}$ is a linear function on $\kappa_{\mathrm{pred}}, \kappa_{\mathrm{abs}}$. This implies that the total generalized Chow loss is minimized at either $a = 0$ or $a = 1$, so no fractional abstention is required. □

Interval of abstention. Having shown that we can efficiently build selective omnipredictors, we further show that all points p that are set to ? by $k^*_{\ell_\pm,\ell_?}$ are in a contiguous interval. We formalize this in the following lemma:

Lemma 7.4. *Given any triplet of loss functions $(\ell_+, \ell_-, \ell_?) \in \mathcal{L}$, the points $x \in \mathbb{R}$ such that $k^*_{\ell_\pm,\ell_?}(x) = ?$ form a contiguous interval, which we denote by I_{abs}.*

To prove Lemma 7.4, we first show the following intermediate lemma:

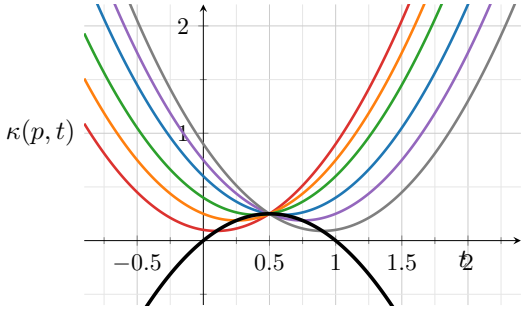
Lemma 7.5. *The function $\kappa_{\mathrm{pred}}(p, t^*)$ is concave as a function of p.*

Proof. Recall that $\kappa_{\mathrm{pred}}(p, t^*) = \min_t p\cdot\mu\ell_-(t)+(1-p)\cdot\lambda\ell_+(t)$, where $\kappa_{\mathrm{pred}}(p, t^*) = \min_{t\in\mathbb{R}} \kappa_{\mathrm{pred}}(p, t)$.[1] For every fixed $t_0 \in \mathbb{R}$, the function $\kappa_{\mathrm{pred}}(p, t_0)$ is affine in p:

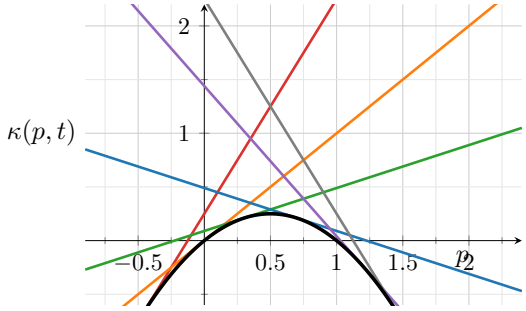
$$\kappa_{\mathrm{pred}}(p, t_0) = \big(\mu\ell_-(t_0) - \lambda\ell_+(t_0)\big) \cdot p + \lambda\ell_+(t_0).$$

Affine functions are convex and concave, and so $\kappa_{\mathrm{pred}}(p) = \min_{t_0\in\mathbb{R}} \kappa_{\mathrm{pred}}(p, t_0)$ is equal to the pointwise infimum of a family of affine functions in p. It is a known fact in analysis that the pointwise infimum of affine functions is concave, and so $\kappa_{\mathrm{pred}}(p, t^*)$ is indeed concave as a function of p. □

[1] In the proof of Theorem 7.2 we used $\kappa_{\mathrm{pred}}(p)$. Here we write t^* to remind that the function has been minimized with respect to t.



(a) Example with the ℓ_2 function; $\kappa_{\text{pred}}(p,t)$ with fixed values of p. For each fixed value of p, the resulting function is convex in t.



(b) Example with the ℓ_2 function; $\kappa_{\text{pred}}(p,t)$ with fixed values of t. For each fixed value of t, the resulting function is affine in p.

Proof of Lemma 7.4. Recall that $k^*_{\ell_\pm,\ell_?}(p) = \min\{\kappa_{\text{pred}}(p,t^*), \kappa_{\text{abs}}(p)\}$.[2] Per Lemma 7.5, the function $\kappa_{\text{pred}}(p,t^*)$ is concave in p. By definition, note that $\kappa_{\text{abs}}(p)$ is affine in p:

$$\kappa_{\text{abs}}(p) = \big(\nu\alpha_+ - \alpha_-\big)p + \alpha_-.$$

Therefore, the function $\kappa_{\text{pred}}(p,t^*) - \kappa_{\text{abs}}(p)$ is still concave in p. Hence this function has at most two roots, and hence the set of points p where $\kappa_{\text{pred}}(p,t^*) \geq \kappa_{\text{abs}}(p)$ forms a contiguous interval (which can be empty, in the case where it is always better to predict than to abstain). By the definition of our post-processing function $k^*_{\ell_\pm,\ell_?}(p)$, this interval corresponds precisely to the set of points where $k^*_{\ell_\pm,\ell_?}(p) = ?$, and hence this interval is equal to I_{abs}. □

Figure 7.1b illustrates the concavity of $\kappa_{\text{pred}}(p,t)$ as a function of t (left) and the affinity of $\kappa_{\text{pred}}(p,t)$ as a function of p (right), which prove that the abstentions as allocated by the selective omnipredictor occur in one contiguous (and possibly empty) interval I_{abs}.

Importantly, note that we can determine I_{abs} directly from the chosen triplet of functions $(\ell_+, \ell_-, \ell_?)$, without any dependence on the underlying data. This is why our method is highly efficient: once we run the multicalibration algorithm, we directly apply our off-the-shelf post-processing function $k^*_{\ell_\pm,\ell_?}$.

Omniprediction with constraints. Recent work by Hu, Livni-Navon, Reingold, and Yang extends the line of work on omniprediction to constrained optimization problems [HNRY23]. This allows the learner to train agnostic to the final choice of loss function as well as of constraints that will be later imposed (as long as these satisfy certain conditions). By viewing Condition 1 in the definitions of PRL, NRL, and FRL as a constraint, one could potentially adapt their results in order to construct $\mathcal{L}$-PRL and $\mathcal{L}$-NRL predictors, and then ensemble them in the usual way to obtain $\mathcal{L}$-FRL predictors. It is unclear how this approach could be used to obtain optimality in the more general framework of selective omniprediction with generalized Chow losses (Definition 6.9), rather than restricted to the concept class $\texttt{SC}(\mathcal{C})$.

Lastly, one could relax the B-Lipschitzness requirement by using the formulation of (B,ϵ)-nice loss functions introduced in [GKR+22].

[2] In the case of ties, we decide to predict.

7.1.1 Selective omnipredictors in action

While our results are theoretical, we provide experiments with synthetic data to demonstrate the feasibility of selective omniprediction in practice and to provide some concrete examples of the function $k^*_{\ell_\pm,\ell_?}$ for specific choices of triplets of loss functions $(\ell_+, \ell_-, \ell_?)$.[3]

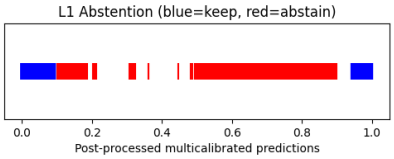
Experimental details. First, we generate $10,000$ samples and 20 features as our data using `sk-learn`'s function `make_classification` and train a random forest to obtain baseline predictions. We then implement the multicalibration algorithm from scratch using the concept class $\mathcal{C}$ of decision trees of depth 3. At each step, we check for correlation between any concept in $\mathcal{C}$ and the residuals computed with the current predictions. For the multicalibration algorithm, we use a discretization parameter of 0.1, a learning rate of 0.01, and 200 maximum iterations. The multicalibration algorithm is run on the validation set (20% of the data) and we then report all of our statistics on the test set (20% of the data). This gives us a predictor h; note that so far we have not used any loss functions.

Next, we choose specific triplets of loss functions $(\ell_+, \ell_-, \ell_?)$ and for each we apply our post-processing function $k^*_{\ell_\pm,\ell_?}$. We use the triplets are shown in Table 7.1. This is a pre-computed function that we derive mathematically, so it is extremely efficient to post-process our h in this way. For each triplet, we compute the total coverage, the total loss over the non-abstaining region (i.e., $\ell_+ + \ell_-$), and the abstention loss (i.e., $\ell_?$). Separately, we perform the usual loss-specific minimization to compare with. In order to use the same concept class $\mathcal{C}$, we implement decision trees that minimize the Chow losses for the chosen triplet $(\ell_+, \ell_-, \ell_?)$. That is, we train a different decision tree (that adds abstentions) optimized specifically for minimizing the Chow loss corresponding to each of the different triplets in Table 7.1. We similarly report the total coverage, the total loss over the non-abstaining region, and the abstention loss (i.e., $\ell_?$).

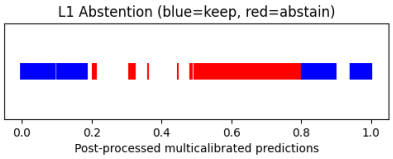
The bar plots show the points $p \in [0,1]$ where $k^*_{\ell_\pm,\ell_?} \circ h$ abstains for different triplets of loss functions, indicated in the caption of each subfigure. We compute I_{abs} theoretically for each of these triplets, independently of the data (we provide the calculations below after the table); in Figure 7.2 one can see that these are indeed the regions where our algorithm abstains in practice. Table 7.1 shows the final loss and coverage incurred by our single selective omnipredictor (post-processed accordingly) compared to those obtained by each of the decision trees, which are trained separately optimizing for each of the triplets of loss functions.

We repeat all of this process for different initializations of the synthetic data and report one such run in Table 7.1. All runs demonstrate the same pattern: our post-processed predictor h, even though it is a single predictor for all of the loss triplets, achieves better coverage and better loss than the generalized Chow loss-specific abstaining decision tree. Intuitively, this occurs because the multicalibration post-processing helps calibrate the predictions by pushing them towards 0 and 1. In contrast, the decision tree abstains significantly more and tends to obtain lower loss over the non-abstention region). The experiments demonstrate the utility of using selective omniprediction in practice.

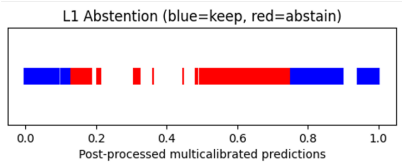
[3]The code can be found at `https://github.com/silviacasac/learning-to-abstain`.



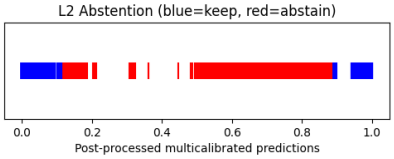
$\ell_+ = \ell_- = \ell_1, \ell_? = 0.1,$
$I_{\text{abs}} = [0.1, 0.9].$



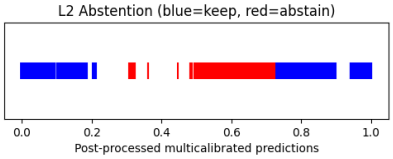
$\ell_+ = \ell_- = \ell_1, \ell_? = 0.2,$
$I_{\text{abs}} = [0.2, 0.8].$



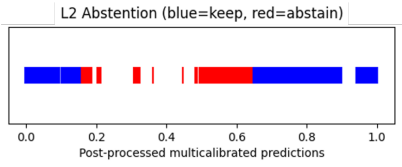
$\ell_1, \ell_?(1) = 0.3,\ \ell_?(0) = 0.1,$
$I_{\text{abs}} = [0.125, 0.75].$



$\ell_2, \ell_? = 0.1,$
$I_{\text{abs}} = \left[\frac{1-\sqrt{0.6}}{2}, \frac{1+\sqrt{0.6}}{2}\right].$



$\ell_2, \ell_? = 0.2,$
$I_{\text{abs}} = \left[\frac{1-\sqrt{0.2}}{2}, \frac{1+\sqrt{0.2}}{2}\right].$



$\ell_2, \ell_?(1) = 0.3,\ \ell_?(0) = 0.1,$
$I_{\text{abs}} = \left[\frac{4-\sqrt{6}}{10}, \frac{4+\sqrt{6}}{10}\right].$

Figure 7.2: Final loss and coverage incurred by our single selective omnipredictor, post-processed with the function $k^*_{\ell_\pm,\ell_?}$ for each of the specific triplets indicated in the subfigure caption. Red: abstain. Blue: predict.

Table 7.1: Comparison of coverage and losses. "Cov" stands for *coverage*, "Pred" for *predicted loss* (i.e., over the non-abstaining region), and "Total" for the total generalized Chow loss.

	$\ell_? = 0.1$		$\ell_? = 0.2$		$\ell_?(1) = 0.3,\ \ell_?(0) = 0.1$	
	$k^*_{\ell_\pm,\ell_?} \circ h$	Abst. DT	$k^*_{\ell_\pm,\ell_?} \circ h$	Abst. DT	$k^*_{\ell_\pm,\ell_?} \circ h$	Abst. DT
ℓ_1	Cov: 22.00% Pred: 0.045 Total: 0.088	Cov: 2.30% Pred: 0.001 Total: 0.098	Cov: 52.70% Pred: 0.083 Total: 0.139	Cov: 3.80% Pred: 0.002 Total: 0.194	Cov: 42.90% Pred: 0.058 Total: 0.144	Cov: 2.50% Pred: 0.002 Total: 0.198
ℓ_2	Cov: 24.15% Pred: 0.046 Total: 0.087	Cov: 3.80% Pred: 0.001 Total: 0.098	Cov: 61.15% Pred: 0.078 Total: 0.125	Cov: 13.20% Pred: 0.010 Total: 0.183	Cov: 64.60% Pred: 0.082 Total: 0.122	Cov: 13.00% Pred: 0.002 Total: 0.193

Examples of selective omnipredictors. To give some more concrete intuition, we provide concrete examples of the post-processing function $k^*_{\ell_\pm,\ell_?}$ for specific choices of loss function triplets $(\ell_+, \ell_-, \ell_?)$. We do so for the triplets of loss functions used in Figure 7.2, and for each triplet we theoretically derive the abstention interval I_{abs}. Recall that I_{abs} only depends on the chosen generalized Chow loss, and not on the underlying data, which is what makes our selective omniprediction framework extremely efficient to adapt to many different loss functions.

ℓ_1 LOSS WITH DIFFERENT ABSTENTION COSTS. Consider letting ℓ_-, ℓ_+ correspond to the 0-1 loss, and let the abstention cost be fixed at some value $\alpha > 0$. That is:

$$\ell_-(t) = |1 - t|, \quad \ell_+(t) = |t|, \quad \ell_?(y) = \alpha.$$

Then,

$$\kappa_{\text{pred}}(p) = \min_t p \cdot |1 - t| + (1 - p) \cdot |t|, \quad \kappa_{\text{abs}} = \alpha.$$

The value $t^*(p)$ that minimizes $\kappa_{\text{pred}}(p)$ for each p corresponds to $k^*_\ell(p) = t^*(p) = \mathbb{1}[p \geq 1/2]$. Then, for each predicted value $h(x) = p$, $k^*_{\ell_\pm,\ell_?}$ will compare the expected prediction cost under the Bernoulli distribution $\mathbf{y} \sim \text{Bern}(p)$, namely

$$\kappa_{\text{pred}}(p) = p \cdot |1 - t^*(p)| + (1 - p) \cdot |t^*(p)| = p \cdot \big|1 - \mathbb{1}[p \geq 1/2]\big| + (1 - p) \cdot \big|\mathbb{1}[p \geq 1/2]\big|, \tag{7.6}$$

with the expected cost of abstention $\kappa_{\text{abs}} = \alpha$. If $\kappa_{\text{pred}}(p) \leq \kappa_{\text{abs}}$, then $k^*_{\ell_\pm,\ell_?}(p) = t^*(p)$; otherwise, $k^*_{\ell_\pm,\ell_?}(p) = ?$.

We compute the interval I_{abs} for different choices of α. For an arbitrary value of $\kappa_{\text{abs}} = \alpha$, I_{abs} corresponds to the interval contained between the roots of the polynomial $\kappa_{\text{pred}}(p) - \kappa_{\text{abs}}$, which is a function in p. In the case of $\alpha = 0.1$, the roots of $\kappa_{\text{pred}}(p) - 0.1$ (where κ_{pred} is given in Equation 7.6) yield $I_{\text{abs}} = [0.1, 0.9]$. For $\alpha = 0.2$, the roots of $\kappa_{\text{pred}}(p) - 0.2$ yield $I_{\text{abs}} = [0.2, 0.8]$.

Lastly, we consider the case where $\ell_?(1) = \alpha_+ = 0.3$ and $\ell_?(0) = \alpha_- = 0.1$. Here, we have that

$$\kappa_{\text{abs}}(p) = p \cdot \alpha_+ + (1 - p) \cdot \alpha_-.$$

Then, the roots of the polynomial $\kappa_{\text{pred}}(p) - p \cdot \alpha_+ + (1 - p) \cdot \alpha_-$ for $\alpha_+ = 0.3$ and $\alpha_- = 0.1$ yield $I_{\text{abs}} = [0.125, 0.75]$.

ℓ_2 LOSS WITH DIFFERENT ABSTENTION COSTS. In the case of the ℓ_2 loss, and still with fixed abstention cost $\ell_?(y) = \alpha$, we have that

$$k^*_{\ell_\pm,\ell_?}(p, t) = p \cdot (1 - t)^2 + (1 - p) \cdot (0 - t)^2 + a\alpha.$$

Hence, if we decide to predict, we pay an expected cost of $\kappa_{\text{pred}}(p) = \min_t p \cdot (1-t)^2 + (1-p) \cdot (0-t)^2$. If we abstain, we pay an expected cost $\kappa_{\text{abs}}(p) = \alpha$. Both of these expected costs are under the Bernoulli distribution $\mathbf{y} \sim \text{Bern}(p)$. For a fixed value of p, the function $\kappa_{\text{pred}}(p, t)$ is minimized at $t^*(p) = p$, and hence

$$\kappa_{\text{pred}}(p, t^*) = p \cdot (1 - p)^2 + (1 - p) \cdot p^2.$$

Hence, the value of $k^*_{\ell_\pm,\ell_?}(p)$ is fully determined by the quantity $\min\{p \cdot (1 - p)^2 + (1 - p) \cdot p^2, \alpha\}$.

We can similarly compute the interval I_{abs} for different choices of α by computing the roots of the polynomial $\kappa_{\text{pred}} - \kappa_{\text{abs}}$. In the case of $\alpha = 0.1$, the roots of $\kappa_{\text{pred}}(p) - 0.1$ yield $I_{\text{abs}} = [\frac{1-\sqrt{0.6}}{2}, \frac{1+\sqrt{0.6}}{2}]$. For $\alpha = 0.2$, the roots of $\kappa_{\text{pred}}(p) - 0.2$ yield $I_{\text{abs}} = [\frac{1-\sqrt{0.2}}{2}, \frac{1+\sqrt{0.2}}{2}]$.

Lastly, we again consider the case where $\ell_?(1) = \alpha_+ = 0.3$ and $\ell_?(0) = \alpha_- = 0.1$. The roots of the polynomial $\kappa_{\text{pred}}(p) - p \cdot \alpha_+ + (1-p) \cdot \alpha_-$ for $\alpha_+ = 0.3$ and $\alpha_- = 0.1$ yield $I_{\text{abs}} = [\frac{4-\sqrt{6}}{10}, \frac{4+\sqrt{6}}{10}]$.

All of these theoretically-derived abstention intervals I_{abs} can be visualized in our experiments, as summarized in Figure 7.2. We further provide an illustrative example in Figure 7.3 to show how the abstention interval I_{abs} widens when we decrease the cost of abstention from 0.2 to 0.1.

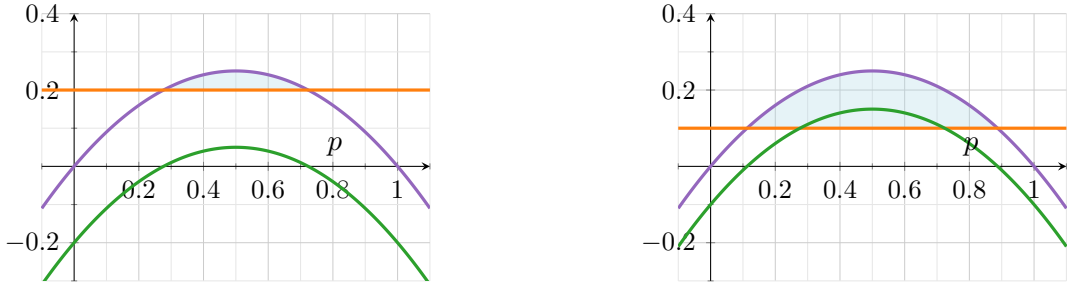


Figure 7.3: Example with the ℓ_2 loss and $\ell_?(y) = 0.2$ for all y (i.e., the traditional Chow abstention model) in the left subfigure and $\ell_?(y) = 0.1$ for all y in the right subfigure. The purple line represents $\kappa_{\text{pred}}(p, t^*)$ and the orange line represents $\kappa_{\text{abs}}(p)$. The green line represents $\kappa_{\text{pred}}(p, t^*) - \kappa_{\text{abs}}(p)$, so its roots determine the interval I_{abs} where we abstain optimally (shaded in blue). In the left example, $I_{\text{abs}} = \big[(1 - \sqrt{0.2})/2, (1 + \sqrt{0.2})/2\big]$, independent of the data. As we decrease the cost of abstention from $\alpha = 0.2$ to $\alpha = 0.1$ (and so the orange line moves down), our selective omnipredictor obtains an increasingly wider abstention interval I_{abs} (e.g., in the right subfigure), as one would expect.

An indistinguishability-based view on abstention. The fact that we can determine the abstention of interval I_{abs} of a multicalibrated predictor directly from the triplet $(\ell_+, \ell_-, \ell_?)$ independently from the data is coherent with the indistinguishability-based view on learning that we have explained at various points in this thesis. For example, in the case of the ℓ_1, ℓ_2 losses with symmetric $\alpha_+ = \alpha_-$ abstention costs, if we had access to the actual ground truth predictor p^*, we would abstain in the same I_{abs} interval that we have computed in this section, and we would expect I_{abs} to be centered around $1/2$. Intuitively, in the omniprediction paradigm we always act "as if" we were dealing with the p^* probabilities rather than those of p. This approach yields the optimal loss because p is $\mathcal{C}$-multiaccurate, and hence, as we explained in Chapter 2, the action we take with p looks like the optimal action we would have taken with p^*, as far as the concepts in $\mathcal{C}$ are concerned. Similar for us, it yields the optimal generalized Chow loss. This works because both the omniprediction and selective omniprediction definitions operate within the agnostic setting, where the losses are always considered optimal with respect to the base concept class $\mathcal{C}$.

Selective classification and the model multiplicity problem. As we discussed in the beginning of Chapter 6, recent work has studied the reliability of predictions through the model multiplicity problem. In this context, we train a model class of multiple competing predictors, and then we decide to abstain on a point if there is too much disagreement in the class concerning its prediction. In a sense, we can view the model multiplicity approach as computing uncertainty by using a model class to compare to, whereas in our selective omniprediction method the predictor is able to measure its own reliability, without needing to train any other predictor.

Recently, an approach based on the multigroup fairness literature called *Reconcile* has been proposed for reconciling two predictors that disagree substantially in their predictions [RTW23; BCDT25]. Essentially, we use a multiaccuracy-type algorithm to use their area of disagreement as a witness, and then we perform a gradient update on the predictor using the witness, which we know can only improve the squared loss. It would be interesting to couple this reconciliation procedure with our selective omniprediction method, especially since both approaches reside within the multigroup fairness framework. That is, it is very natural to instead try to reconcile a pair of *selective* classifiers. By the nature of the Reconcile algorithm, we would expect such a process to yield an ensembled classifier that has brought the disagreement region to disappear, and such that in the process (1) we have only improved the accuracy of the two predictors and (2) reduced their abstention rates. This is coherent with our interpretation of the I_{abs} interval, since in the reconciliation procedure we polarize the values towards 0 or 1. Thus, even though I_{abs} is fixed independent of the data, the reconciliation/boosting procedure would push the values closer to 0 and 1, and thus points would "leave" the abstention interval I_{abs}, causing both the accuracy to increase and the abstention rate to decrease.

7.2 Building general reliable agnostic learners

We show how we can use our efficient construction of selective omnipredictors to recover and generalize the results on reliable agnostic learning obtained by [KKM12] from 0-1 loss to an entire family of loss functions (as defined in Section 6.1). In this sense, we observe a phenomenon similar to the recently discovered relationship between multicalibration and the Regularity Lemma [CDV24; DLLT23]: by starting from the multigroup fairness framework, we can cast these notions back to the fields of complexity and learning theory and recover and generalize previously-known and much older "classical" results. Here, the omniprediction framework has allowed us to revisit the work by Kalai, Kanade, and Mansour from 2009 and generalize their constructions to a rich class of loss functions. Formally:

Theorem 7.7. *Let $\mathcal{C}$ be a concept class on $\mathcal{X}$, $\mathcal{L}$ any class of B-Lipschitz loss functions, and $\mathcal{D}$ a distribution on $\mathcal{X} \times \mathcal{Y}$. If $\mathcal{C}$ is agnostically learnable under $\mathcal{D}$ in time $T(\epsilon, \delta)$, then $\mathcal{C}$ is $\mathcal{L}$-fully reliably learnable under $\mathcal{D}$ in time $T(\epsilon^2/6, \delta)$.*

Indeed, note that here we use our proposed notion of $\mathcal{L}$-FRL (Definition 6.5), which is a generalization of the 0-1 version proposed in [KKM12].

Proof. Let ϵ be the target error parameter for fully reliable learning. We consider the parameters $\lambda = \mu = 1/3$ and $\nu = \epsilon/6$. Let $c^*_?$ be an optimal abstaining classifier in $\texttt{SC}(\mathcal{C})$. By definition of $\texttt{SC}(\mathcal{C})$, it follows that all concepts in $\texttt{SC}(\mathcal{C})$ incur 0 error over the non-abstaining region, and hence

$$\ell_{\text{GC},\mathcal{D}}(c^*_?; \lambda, \mu, \nu) = \nu \ell_?(c^*_?).$$

Let h be a $(\mathcal{L}, \mathcal{C}, \gamma)$-selective omnipredictor (which we can obtain for $\mathcal{C}$ given that $\mathcal{C}$ is agnostically learnable, as shown in Theorem 7.2). By the selective omniprediction guarantee, it follows that for

any post-processing function $k : [0,1] \to \mathbb{R} \cup \{?\}$,

$$\ell_{\mathrm{GC},\mathcal{D}}(h;\lambda,\mu,\nu) \leq \min_{c_? \in k \circ \mathcal{C}} \ell_{\mathrm{GC},\mathcal{D}}(c_?;\lambda,\mu,\nu) + \gamma.$$

In particular, this implies that

$$\ell_{\mathrm{GC},\mathcal{D}}(h;\lambda,\mu,\nu) \leq \ell_{\mathrm{GC},\mathcal{D}}(c^*_?;\lambda,\mu,\nu) + \gamma = \nu\ell_?(c^*_?) + \gamma.$$

By setting $\gamma = \epsilon^2/6$, and using the fact that $\ell_?(c^*_?) \leq 1$, since $\lambda = 1/3$, $\nu = \epsilon/6$ we get that

$$\frac{1}{3}\ell_+(h) \leq \epsilon/6 + \epsilon^2/6,$$

and so $\ell_+(h) \leq \epsilon$. Similarly, since $\mu = 1/3$, we have that $\ell_-(h) \leq \epsilon$. Finally,

$$\nu\ell_?(h) \leq \nu\ell_?(c^*_?) + \epsilon^2/6,$$

and since $\nu = \epsilon/6$ it follows that $\ell_?(h) \leq \ell_?(c^*_?) + \epsilon$. This completes the proof, given that h satisfies Conditions 1 and 2 in the definition of $\mathcal{L}$-fully reliable learning. □

It is interesting to remark that both here and in the original proof in [KKM12], we have an ϵ^2 dependence. Namely, when calling the agnostic learning oracle, [KKM12] need to call it with error ϵ^2. Likewise, here we need to call our selective omnipredictor construction with error ϵ^2. An intriguing future direction would be to determine whether or not this ϵ^2 dependence is inherent to reliable agnostic learning, or whether we can relax it ϵ.

8

Learning Abstentions Fairly

Multigroup fairness is a class of definitions of fairness proposed in the last five years. Instead of providing aggregate statistical guarantees for a few protected categories (defined by race, ethnicity, gender, age, and so on), these definitions suggest giving such guarantees to a large (often exponential) number of sets. The intuition is that providing meaningful fairness guarantees to a group requires extending these guarantees to subgroups that are relevant in the setting we consider.

Simons Institute workshop on *Multigroup Fairness and the Validity of Statistical Judgment*, April 2023.

So far, we have considered the question of how to learn abstentions optimally, where we measure optimality in the agnostic sense, with respect to a base concept class $\mathcal{C}$ and a generalized Chow loss function ℓ_{GC}. The motivation for abstaining is to be able to make almost no errors when predicting, which we can accomplish by mapping the uncertain points $x \in \mathcal{X}$ to '?' instead.

However, as we are coming from the multigroup fairness framework, we might additionally have fairness concerns, which should be very natural for us to think about at this point of the thesis: suppose that we have a collection $\mathcal{G} = \{g : \mathcal{X} \to \{0,1\}\}$ of subgroups of interest of the domain, which can intersect arbitrarily. Besides achieving high accuracy over the points where we do predict a numerical value, we would also like to abstain fairly (i.e., optimally) on *each* of the groups $g \in \mathcal{G}$, so that we avoid achieving high global accuracy at the expense of overly abstaining on some subgroups. That is, we want to bring the "multi" perspective into the notion of reliable agnostic learning. All throughout this chapter, the concepts in $\mathcal{C}$ are Boolean.

8.1 Multigroup reliable learning

How can we measure how an optimal abstention rate looks like within each of the groups $g \in \mathcal{G}$? Motivated by the reliable agnostic learning framework [KKM12], we do so by requiring our predictor to abstain no more than the optimal selective classifier $c_? \in \mathtt{SC}(\mathcal{C})$ (with an ϵ slack) on *each* group $g \in \mathcal{G}$, where the optimal $c_? \in \mathtt{SC}(\mathcal{C})$ can naturally be different for each group. That is, similar

to our notion of a selective omnipredictor, we want to construct a single classifier simultaneously for all groups, but its abstention rate competes with that of a $c_?$ that is chosen optimally in each group. This is an idea that we have repeatedly seen in the thesis; for example, in the notion of swap agnostic learning.

Formally, we introduce the following definition, which can be viewed as the "multigroup" version of the FRL definition.

Definition 8.1 ($(\mathcal{C},\mathcal{G})$-multigroup selective classifier)**.** *Given a collection of subgroups $\mathcal{G}$ of $\mathcal{X}$, a concept class $\mathcal{C}$, distribution $\mathcal{D}$, and $\epsilon > 0$, we say that a predictor $h_? : \mathcal{X} \to [0,1] \cup \{?\}$ is a* $(\mathcal{C},\mathcal{G},\epsilon)$-multigroup selective classifier *if the following two conditions are satisfied:*

1. ***Global accuracy.*** $\mathrm{err}_{\mathcal{D}}(h_?) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[|\mathbf{y} - h_?(\mathbf{x})| \cdot \mathbb{1}[h_?(\mathbf{x}) \neq ?]] \leq \epsilon$.
2. ***Optimal local abstention rate.*** *For every $g \in \mathcal{G}$,*

$$\Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[g(\mathbf{x}) \cdot \mathbb{1}[h_?(\mathbf{x}) = ?]\right] \leq \min_{c_? \in \mathtt{SC}(\mathcal{C})} \Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[g(\mathbf{x}) \cdot \mathbb{1}[c_?(\mathbf{x}) = ?]\right] + \epsilon.$$

Note that Condition 2 automatically implies an optimal global abstention rate as well, by applying the local condition with the constant function $g = \mathbf{1}$ (which we assume is always contained in $\mathcal{G}$ without loss of generality).

Note that so far in this thesis we had always (implicitly or explicitly) taken $\mathcal{C} = \mathcal{G}$. While this can be advantageous in some situations, this notion demonstrates why it is useful to separate the base concept class $\mathcal{C}$ from the collection of groups $\mathcal{G}$. At the end of the day, they are different objects.

While the notion of $(\mathcal{C},\mathcal{G})$-multigroup selective classification is clearly desirable, we need to be able to construct if efficiently. Can we do that? Perhaps surprisingly, we now have the primitive of calibrated multiaccuracy, which played a central role in Part I of this thesis, make a comeback! Specifically, it turns out that we can efficiently construct a $(\mathcal{C},\mathcal{G},\epsilon)$-multigroup selective classifier for any $\mathcal{C},\mathcal{G}$ from a multiaccurate predictor for the class $\mathcal{C}\cdot\mathcal{G} = \{cg \mid c \in \mathcal{C}, g \in \mathcal{G}\}$ that is also globally calibrated:

Theorem 8.2. *Given access to a $(\mathcal{C}\cdot\mathcal{G}, \epsilon^2/8)$-multiaccurate and calibrated predictor that runs in time $T(\epsilon,\delta)$, we can construct a $(\mathcal{C},\mathcal{G},\epsilon)$-multigroup selective classifier in time $T(\epsilon,\delta)$.*

Because from Part I we know that we need the primitive of a weak agnostic learner in order to construct a calibrated and multiaccurate predictor, this means that here we require access to a $\mathcal{C}\cdot\mathcal{G}$-calibrated and multiaccurate predictor. A priori, it does not seem like we can achieve such learning from having a separate weak agnostic learner for each of $\mathcal{C}$ and $\mathcal{G}$.

The key idea in our proof is to convert a $(\mathcal{C}\cdot\mathcal{G},\epsilon)$-multiaccurate and calibrated predictor $h : \mathcal{X} \to [0,1]$ into a multigroup selective classifier by mapping all x such that $h(x) \in (\epsilon, 1-\epsilon)$ to ?. Intuitively, this thresholding will ensure that we can achieve global accuracy while not over-abstaining. For each group g, let $c^g_? = (c^g_+, c^g_-)$ be an optimal selective classifier in $\mathtt{SC}(\mathcal{C})$ within g. Since $c^g_+ \in \mathcal{C}^+$, it follows that whenever $c^g_+ = 1$, the true label y on that point is also 1. By the multiaccuracy guarantee for c^g_+ and g (which is in $\mathcal{C}\cdot\mathcal{G}$), we obtain that $\mathbb{E}_{\mathcal{D}}[\mathbf{y}] \approx \mathbb{E}_{\mathcal{D}}[h(\mathbf{x})] \approx 1$ in the region where $c^g_+(x) = 1, g(x) = 1$. A symmetric argument holds with c^g_- and g. Lastly, we use

the global calibration condition to ensure that our thresholded $h_?$ remains accurate in the entire domain.

A key property to realize about the concepts $c_+ \in C^+$, $c_- \in \mathcal{C}^-$ is that, by definition of $\mathcal{C}^+, C^-$, it follows that

$$c_+ = 1 \implies y = 1 \qquad y = 0 \implies c_+ = 0$$
$$c_- = 0 \implies y = 0 \qquad y = 1 \implies c_- = 1$$

Proof. Let $g \in \mathcal{G}$ be some group and let us focus on the part of the domain $\mathcal{X}_g = \{x \mid g(x) = 1\}$. Consider the classes $\mathcal{C}_g^+, \mathcal{C}_g^-$ derived from the Boolean concept class $\mathcal{C}$, where $\mathcal{C}_g^+(\mathcal{D}) = \{c \in \mathcal{C} \mid \Pr[c(\mathbf{x}) = 1, g(\mathbf{x}) = 1, \mathbf{y} = 0] = 0\}$ and $\mathcal{C}_g^-(\mathcal{D}) = \{c \in \mathcal{C} \mid \Pr[c(\mathbf{x}) = 0, g(\mathbf{x}) = 1, \mathbf{y} = 1] = 0\}$.

Let h be a $(\mathcal{C} \cdot \mathcal{G}, \epsilon^2/8)$-multiaccurate and calibrated predictor for $\mathcal{X}, \mathcal{D}$, which we can build using Theorem 2.16. Consider any $c_+ \in \mathcal{C}_g^+ \subseteq \mathcal{C}$. The multiaccuracy condition is ensured for all $c \cdot g \in \mathcal{C} \cdot \mathcal{G}$, and so for this particular c_+ and fixed group g we have that

$$\left| \mathop{\mathbb{E}}_{\mathcal{D}} \left[(g(\mathbf{x}) \cdot c_+(\mathbf{x}))(\mathbf{y} - h(\mathbf{x})) \right] \right| \leq \epsilon^2/8.$$

Given that whenever $c_+(x) = 1$ and $g(x) = 1$, we have that $y = 1$ by definition of $\mathcal{C}_g^+$, it follows that

$$\mathop{\mathbb{E}}_{\mathcal{D}}[\mathbf{y} \mid g(\mathbf{x}) = 1, c_+(\mathbf{x}) = 1] = 1.$$

That is, within the region in $\mathcal{X}_g$ where $c_+(x) = 1$, we have that $\mathbb{E}_{\mathcal{D}}[\mathbf{y}] = 1$. Then, by the multiaccuracy guarantee on $c_+ \cdot g$, it follows that the expected value of h over the same region (i.e., where $c_+(\mathbf{x}) = 1$ inside of $\mathcal{X}_g$) is also close to 1:

$$\left| \mathop{\mathbb{E}}_{\mathcal{D}} \left[(g(\mathbf{x}) \cdot c_+(\mathbf{x}))(\mathbf{y} - h(\mathbf{x})) \right] \right| = \left| \mathop{\mathbb{E}}_{\mathcal{D}} \left[(g(\mathbf{x}) \cdot c_+(\mathbf{x}))(1 - h(\mathbf{x})) \right] \right| \leq \epsilon^2/8,$$

and so

$$\mathop{\mathbb{E}}_{\mathcal{D}}[h(\mathbf{x}) \mid g(\mathbf{x}) = 1, c_+(\mathbf{x}) = 1] \geq 1 - \frac{\epsilon^2}{8\,\mathbb{E}_{\mathcal{D}}[g(\mathbf{x})c_+(\mathbf{x})]}. \tag{8.3}$$

Now, we have two cases, either $\Pr[c_+(\mathbf{x}) = 1, g(x) = 1] = \mathbb{E}_{\mathcal{D}}[c_+(\mathbf{x})g(\mathbf{x})] \leq \epsilon/2$, in which case we trivially have,

$$\mathbb{E}\left[\mathbb{1}[h(\mathbf{x}) \geq 1 - \epsilon] \cdot g(\mathbf{x})\right] \geq \mathop{\mathbb{E}}_{\mathcal{D}}[c_+(\mathbf{x})g(\mathbf{x})] - \epsilon/2,$$

or, from Equation 8.3, we have that

$$\mathop{\mathbb{E}}_{\mathcal{D}}\left[\mathbb{1}[h(x) \leq 1 - \epsilon] \mid g(\mathbf{x}) = 1, c_+(\mathbf{x}) = 1\right] \leq \frac{\epsilon}{4\,\mathbb{E}_{\mathcal{D}}[g(\mathbf{x})c_+(\mathbf{x})]}.$$

In this case, by taking the complement and multiplying both sides by $\mathbb{E}_{\mathcal{D}}[g(\mathbf{x})c_+(\mathbf{x})]$, we get,

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{D}}\left[\mathbb{1}[h(\mathbf{x}) \geq 1 - \epsilon] \cdot g(\mathbf{x})\right] &\geq \mathop{\mathbb{E}}_{\mathcal{D}}\left[\mathbb{1}[h(x) \geq 1 - \epsilon] \cdot g(\mathbf{x})c_+(\mathbf{x})\right] \\ &\geq \mathop{\mathbb{E}}_{\mathcal{D}}[c_+(\mathbf{x})g(\mathbf{x})] - \epsilon/4.\end{aligned}$$

Thus, in either case we have,

$$\mathop{\mathbb{E}}_{\mathcal{D}}\left[\mathbb{1}[h(\mathbf{x}) \geq 1-\epsilon] \cdot g(\mathbf{x})\right] \geq \mathop{\mathbb{E}}_{\mathcal{D}}[c_+(\mathbf{x})g(\mathbf{x})] - \epsilon/2.$$

A symmetric argument holds in the case of $c_- \in \mathcal{C}^-$ with the same group g. By the definition of $\mathcal{C}_g^-(\mathcal{D})$, it holds that whenever $c_-(\mathbf{x}) = 0$ and $g(\mathbf{x}) = 1$, $\mathbf{y} = 0$, and so within the region in $\mathcal{X}_g$ where $c_-(\mathbf{x}) = 0$, $\mathbb{E}_{\mathcal{D}}[\mathbf{y}] = 0$. Let $\bar{c}_-$ be the complement of c_- (which we can take since $\mathcal{C}$ is closed under complement). Then, it follows that:

$$\mathop{\mathbb{E}}_{\mathcal{D}}[\mathbf{y} \mid g(\mathbf{x}) = 1, \bar{c}_-(\mathbf{x}) = 1] = 0.$$

By the multiaccuracy guarantee on $\bar{c}_- \cdot g$, we can use an identical argument above (by a case distinction on whether $\Pr[c_-(\mathbf{x}) = 0, g(\mathbf{x}) = 1] \geq \epsilon/2$) to show that

$$\mathop{\mathbb{E}}_{\mathcal{D}}\left[\mathbb{1}[h(\mathbf{x}) \leq \epsilon] \cdot g(\mathbf{x})\right] \geq \mathop{\mathbb{E}}_{\mathcal{D}}[\bar{c}_-(\mathbf{x})g(\mathbf{x})] - \epsilon/2,$$

From the $(\mathcal{C}\cdot\mathcal{G}, \epsilon^2/8)$-multiaccurate and calibrated h, we construct our $(\mathcal{C}, \mathcal{G}, \epsilon)$-multigroup selective classifier $h_?$ by post-processing h as follows:

$$h_?(x) = \begin{cases} h(x) & \text{if } h(x) \in [0, \epsilon] \cup [1-\epsilon, 1], \\ ? & \text{if } h(x) \in (\epsilon, 1-\epsilon). \end{cases}$$

Observe that

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{D}}\left[\mathbb{1}[h_?(\mathbf{x}) \neq ?] \cdot g(\mathbf{x})\right] &= \mathop{\mathbb{E}}_{\mathcal{D}}\left[\mathbb{1}[h(\mathbf{x}) \geq 1-\epsilon] \cdot g(\mathbf{x})\right] + \mathop{\mathbb{E}}_{\mathcal{D}}\left[\mathbb{1}[h(\mathbf{x}) \leq \epsilon] \cdot g(\mathbf{x})\right] \\ &\geq \mathop{\mathbb{E}}_{\mathcal{D}}[c_+(\mathbf{x})g(\mathbf{x})] + \mathop{\mathbb{E}}_{\mathcal{D}}[\bar{c}_-(\mathbf{x})g(\mathbf{x})] - \epsilon.\end{aligned}$$

Recall that any $c_? \in \mathtt{SC}(\mathcal{C})$ is of the form (c_+, c_-). By the construction of $\mathtt{SC}(\mathcal{C})$ (see Section 6.1) and given that $\bar{c}_-$ is the complement of c_- it follows that

$$\mathop{\mathbb{E}}_{\mathcal{D}}\left[\mathbb{1}[c_?(\mathbf{x}) \neq ?] \cdot g(\mathbf{x})\right] \leq \mathop{\mathbb{E}}_{\mathcal{D}}[(c_+(\mathbf{x})) + \bar{c}_-(\mathbf{x}))g(\mathbf{x})]$$

Therefore, the two above equation show that

$$\mathop{\mathbb{E}}_{\mathcal{D}}\left[\mathbb{1}[h_?(\mathbf{x}) \neq ?] \cdot g(\mathbf{x})\right] \geq \mathop{\mathbb{E}}_{\mathcal{D}}\left[\mathbb{1}[c_?(\mathbf{x}) \neq ?] \cdot g(\mathbf{x})\right] - \epsilon.$$

Hence, on each $g \in \mathcal{G}$, $h_?$ does not abstain significantly more often than the optimal $c_?$ for each group.

Note that the global calibration property of $h_?$ and the fact that the predicted values of $h_?(\mathbf{x})$ are in $[0, \epsilon)$ or $(1-\epsilon, 1]$ implies that both the false positive and negative rates of $h_?$ are bounded by ϵ thus satisfying Condition 1 in Definition 8.1. □

Note that to achieve the global accuracy guarantee, we only used the global calibration condition and the fact that we abstain in the region where $h(x) \in (\epsilon, 1-\epsilon)$ (per our definition of $h_?$ from h). In

the other hand, to achieve the local abstention guarantee, we used the multiaccuracy guarantee and the definitions of $\mathcal{C}^+$ and $\mathcal{C}^-$. Hence, our proof clearly delineates the complementary roles played by each of the multiaccuracy and calibration. This complementarity fits very nicely with our results on calibrated multiaccuracy in Part I of this thesis. Recall that for both of our results showing that (1) multiaccuracy and calibration give strong agnostic learning (Section 3.3) and (2) (weighted) multiaccuracy and calibration give a hardcore measure of optimal density, the roles played by each component are similarly modular. Moreover, $(\mathcal{C}, \mathcal{G})$-multigroup selective classification now provides a new example on the growing family of useful implications from multiaccuracy with global calibration.

8.2 Multigroup fairness primitives

It is natural to ask whether we can efficiently construct $(\mathcal{C}, \mathcal{G})$-multigroup selective classifiers starting from a weaker learning primitive than that of a calibrated multiaccurate predictor for the class $\mathcal{C} \cdot \mathcal{G}$. What is the weakest learning primitive that we can build it from? In Theorem 8.4 below, we answer this question in the positive in the case where $|\mathcal{G}|$ is small. Specifically, we show that in this case, we can construct it from fully reliable learning:

Lemma 8.4. *If the class $\mathcal{C}$ is fully reliably learnable for the ℓ_1 loss and $\epsilon > 0$, we can construct a $(\mathcal{C}, \mathcal{G}, \epsilon)$-multigroup selective classifier in time poly$(|\mathcal{G}|, 1/\epsilon)$ given oracle access to the full reliable learner for $\mathcal{C}$.*

Recall that in Part I we showed that calibrated multiaccuracy yields strong agnostic learning (Section 3.3). Therefore, Theorem 8.2 showing that we can construct $(\mathcal{C}, \mathcal{G})$-multigroup selective classification from calibrated multiaccuracy implies that we need the class $\mathcal{C} \cdot \mathcal{G}$ to be agnostically learnable. As we have seen, however, reliable agnostic learning is believed to be easier than agnostic learning [KT14; KK21b]. Therefore, it is desirable to be able to construct $(\mathcal{C}, \mathcal{G})$-multigroup selective classification from FRL, given that then we can get it from a weaker learning primitive.

Proof. For each $g \in \mathcal{G}$, we call the FRL oracle on the restricted domain $\{x \mid g(x) = 1\}$, with error parameter $\epsilon/|\mathcal{G}|$, class $\mathcal{C}$, and the 0-1 loss function. Let $h^g_? : \{x \mid g(x) = 1\} \to \{0, 1, ?\}$ denote the selective classifier that we obtain with an FRL call with the domain $\mathcal{X}_g = \{x \mid g(x) = 1\}$. We construct our final global classifier $h_? : \mathcal{X} \to \{0, 1, ?\}$ as follows: for every point $x \in \mathcal{X}$, consider the set $\mathcal{K} \subseteq \mathcal{G}$ of all groups $g \in \mathcal{G}$ such that $g(x) = 1$. Then, $h_?$ is equal to the majority vote of the $|\mathcal{K}|$ total classifiers $h^g_?$.

By the FRL guarantee, each $h^g_?$ makes a wrong prediction on at most $\epsilon/|\mathcal{G}|$ points in $\mathcal{X}$. In the worst case, all the $|\mathcal{G}|$ error regions for each of the $h^g_?$ predictors are disjoint and cause all the majority votes taken on the points x in these regions to cause the wrong prediction. Hence,

$$\mathrm{err}_{\mathcal{D}}(h_?) = \mathop{\mathbb{E}}_{\mathcal{D}}\left[|\mathbf{y} - h_?(\mathbf{x})| \cdot \mathbb{1}[h_?(\mathbf{x}) \neq ?]\right] \leq |\mathcal{G}| \cdot \frac{\epsilon}{|\mathcal{G}|} = \epsilon,$$

and thus we satisfy Condition 1 in the definition of $(\mathcal{C}, \mathcal{G})$-multigroup selective classification (global accuracy).

Secondly, we have not added further abstentions in our ensembling of the $h^g_?$ predictors, and so we continue to satisfy optimal local abstention rate within each $g \in \mathcal{G}$. Namely, the FRL guarantee on each $g \in \mathcal{G}$ ensures that

$$\Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[g(\mathbf{x}) \cdot \mathbb{1}[h_?(\mathbf{x}) = ?]\right] \leq \min_{c_? \in \mathtt{SC}(\mathcal{C})} \Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[g(\mathbf{x}) \cdot \mathbb{1}[c_?(\mathbf{x}) = ?]\right] + \epsilon.$$

In fact, we are getting a better $\epsilon/|\mathcal{G}|$ error. Hence, we satisfy Condition 2 in the definition in the definition of $(\mathcal{C}, \mathcal{G})$-multigroup selective classification. Thus we have shown that $h_?$ is a $(\mathcal{C}, \mathcal{G}, \epsilon)$-multigroup selective classifier, as desired. □

Answering this question in generality (i.e., where $\mathcal{G}$ can be arbitrarily large) appears to be a very interesting open question.

Calibrated multiaccuracy and multicalibration. We make a couple of further remarks about how the various multigroup fairness definitions relate to the problem of $(\mathcal{C}, \mathcal{G})$-multigroup selective classification. First, if we have the stronger primitive of a $(\mathcal{C} \cdot \mathcal{G})$-multicalibrated predictor (which implies a $(\mathcal{C} \cdot \mathcal{G})$-multiaccurate calibrated predictor), then we can have a non-selective predictor with local agnostic guarantees. Formally:

Lemma 8.5. *Given access to a $(\mathcal{C} \cdot \mathcal{G}, \epsilon)$-multicalibrated predictor that runs in time $T(\epsilon, \delta)$, we can construct a classifier $h : \mathcal{X} \to [0, 1]$ in time $T(\epsilon, \delta)$ satisfying the following local accuracy property: for every $g \in \mathcal{G}$,*

$$\mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[|\mathbf{y} - h(\mathbf{x})| \mid g(\mathbf{x}) = 1] \leq \min_{c\in\mathcal{C}} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[|\mathbf{y} - c(\mathbf{x})| \mid g(\mathbf{x}) = 1] + \epsilon.$$

Note that a $(\mathcal{C}, \mathcal{G}, \epsilon)$-multigroup selective classifier achieves a local accuracy property within each $g \in \mathcal{G}$, where the error term is weighted by $1/\Pr_{\mathcal{D}}[g(\mathbf{x}) = 1]$. In Lemma 8.5, however, we do so without requiring abstentions, which is why we need the stronger notion of multicalibration rather than of calibrated multiaccuracy. Note that accuracy notion in Lemma 8.5 is of the agnostic form, rather than an absolute error guarantee.

Proof. This follows directly from our Theorem 4.33 in Chapter 4 in Part I of this thesis. There we showed that from multicalibration we can obtain agnostic learning++ for the ℓ_1 loss, which is exactly the guarantee claimed in the statement of Lemma 8.5. Given that here we have access to a $(\mathcal{C} \cdot \mathcal{G})$-multicalibrated predictor, rather than just a $\mathcal{C}$-multicalibrated predictor, we are able to further condition on g and still obtain the same agnostic ℓ_1 error guarantee. □

Second, we remark that given a $(\mathcal{C} \cdot \mathcal{G}, \epsilon)$-multiaccurate and calibrated predictor, we can directly obtain FRL predictors for the class $\mathcal{C}$ for the ℓ_1 loss by thresholding the predictor as we do in the proof of Theorem 8.2. Given that calibrated multiaccuracy implies agnostic learning, as we showed in Chapter 3 of this thesis, it is already implied by Theorem 8.2 that we can achieve reliable agnostic learning from calibrated multiaccuracy. However, the approach below in Lemma 8.5 gives a direct reduction; i.e., we just need to post-process the outputs of the predictor. Formally:

Lemma 8.6. *Given access to a $(\mathcal{C} \cdot \mathcal{G}, \epsilon)$-multiaccurate and calibrated predictor h, we can apply a post-processing function to h to obtain a $\mathcal{L}$-FRL predictor for the ℓ_1 loss.*

Proof. Let h be a $(\mathcal{C} \cdot \mathcal{G}, \epsilon)$-multiaccurate and calibrated predictor. We first construct a $\mathcal{L}$-FRL predictor $h_?$ from h by applying the same post-processing function as in the proof of Theorem 8.2. Namely, we let

$$h_?(x) = \begin{cases} h(x) & \text{if } h(x) \in [0, \epsilon] \cup [1-\epsilon, 1], \\ ? & \text{if } h(x) \in (\epsilon, 1-\epsilon). \end{cases}$$

By the definition of FRL, we need to show that $h_?$ satisfies 1) low global error, and 2) optimal abstention. This follows directly from our Theorem 8.2 by using the group $g = \mathbf{1}$. □

8.3 Applications to conformal prediction

We conclude Part II of this thesis by demonstrating another useful application of (1) fully reliable learners and of (2) $(\mathcal{C}, \mathcal{G})$-multigroup selective classifiers. Specifically, both methods can be seen as conformal prediction algorithms, with FRL achieving marginal guarantees and $(\mathcal{C}, \mathcal{G})$-multigroup selective classifiers achieving group guarantees. Specifically, to apply these notions as conformal prediction methods in the binary classification case, we view the $\{0, 1\}$ prediction set as ?.

8.3.1 Conformal prediction from FRL

In the classification setting, the goal of conformal prediction is to construct a *prediction set* of the possible labels $\mathcal{S}(\mathbf{x}) \subseteq \mathcal{Y}$ for each point $x \in \mathcal{X}$ such that $\Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\mathbf{y} \in \mathcal{S}(\mathbf{x})] \geq 1 - \epsilon$ for a chosen error rate $\epsilon \in (0, 1)$ [VGS05; AB21]. This is known as the *ϵ-marginal coverage guarantee*. In the case of binary classification, the possible prediction sets can only be $\{0\}, \{1\}, \{0, 1\}$.

We show that an FRL predictor does indeed satisfy the ϵ-marginal coverage guarantee:

Lemma 8.7. *Let $h_? : \mathcal{X} \to \{0, 1, ?\}$ be an FRL predictor for $\mathcal{C}, \mathcal{D}, \epsilon$ and the 0-1 loss. Then, the prediction sets induced by the level sets of h, where we map $0 \mapsto \{0\}$, $1 \mapsto \{1\}$, and $? \mapsto \{0, 1\}$, satisfy the ϵ-marginal coverage guarantee.*

Proof. Let $\mathrm{err}(h_?, \mathcal{D}) = \Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[h_?(\mathbf{x}) \neq \mathbf{y}]$ denote the error of the selective predictor $h_?$ and recall that

$$\mathrm{false}_+(h_?, \mathcal{D}) = \Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[h_?(\mathbf{x}) = 1 \wedge \mathbf{y} = 0] \quad \mathrm{false}_-(h_?, \mathcal{D}) = \Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[h_?(\mathbf{x}) = 0 \wedge \mathbf{y} = 1].$$

Since $\mathrm{err}(h_?, \mathcal{D}) \leq \epsilon$ by the FRL guarantee of $h_?$ and $\mathrm{err}(h_?, \mathcal{D}) = \mathrm{false}_+(h_?, \mathcal{D}) + \mathrm{false}_-(h_?, \mathcal{D})$, it follows that

$$\mathrm{false}_+(h_?, \mathcal{D}) \leq \epsilon, \qquad \mathrm{false}_-(h_?, \mathcal{D}) \leq \epsilon.$$

This implies that for all $\mathbf{y} = 1$,

$$\Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\mathbf{y} \notin \{1\}] \leq \epsilon,$$

and symmetrically for all $\mathbf{y} = 0$. Therefore, we satisfy the marginal coverage guarantee for the prediction sets $\{0\}, \{1\}$. We also trivially satisfy it for the prediction set $\{0, 1\}$, since $\mathbf{y}$ is Boolean and so $\Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\mathbf{y} \in \{0, 1\}] = 1$. □

Remark 8.8. *While any conformal prediction method must definitionally satisfy the ϵ-marginal coverage guarantee, typical algorithms offer no theoretical bounds on the size of the prediction sets. In the case of binary classification, this means that we could have an arbitrarily large number of points mapped to $\{0,1\}$ (even the entire domain, which is a trivial way of satisfying the marginal coverage guarantee). However, FRL does provide provable abstention guarantees with respect to a base concept class $\mathcal{C}$ of our choice. Let $h_? : \mathcal{X} \to \{0,1,?\}$ be an FRL predictor for the class $\mathcal{C}$, 0-1 loss, and $\epsilon > 0$, and let $\{0\},\{1\},\{0,1\}$ be its induced prediction sets as a conformal prediction method. Then:*

Lemma 8.9. *Let $h_? : \mathcal{X} \to \{0,1,?\}$ be an FRL predictor for $\mathcal{C}, \mathcal{D}, \epsilon$, and the 0-1 loss. Then, the prediction sets induced by the level sets of $h_?$ satisfy:*

$$\Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[h_?(\mathbf{x}) = \{0,1\}] \leq \min_{c_? \in \mathtt{SC}(\mathcal{C})} \Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[c_?(x) = ?] + \epsilon.$$

Proof. This follows directly from Condition 2 in the definition of FRL, which ensures that

$$\Pr_{\mathcal{D}}[h_? = ?] \leq \min_{c_? \in \mathtt{SC}(\mathcal{C})} \Pr_{\mathcal{D}}[c_? = ?] + \epsilon.$$

□

Hence, FRL viewed as a conformal prediction method, besides satisfying the ϵ-marginal coverage guarantee, also provides a provable upper bound on the size of the set $\{0,1\}$, which constitutes a measure of uncertainty in the case of binary classification. Moreover, we can choose any base class $\mathcal{C}$ we want for the provable guarantee on the size of the prediction sets, which is a very useful guarantee.

Experiments of FRL as conformal prediction. As shown in Lemma 8.7 and Remark 8.8 (formalized in Lemma 8.9), we can view FRL as a conformal prediction method. Here we illustrate the feasibility of implementing FRL in practice and compare its coverage and ℓ_1 error with that of standard conformal prediction algorithms.

We generate $5,000$ data samples synthetically using `sk-learn`'s `make_blobs` function, which generates isotropic Gaussian blobs for clustering. To create an FRL predictor, we do it by training a PRL and an NRL predictor separately for the same data, and then ensembling them in the usual way (i.e., if the two agree on a prediction we keep it, otherwise we abstain). To create the PRL and NRL predictors, we use random forests where we give extra weight w to the label that we wish to penalize most for the predictions (so false positives in the case of PRL, and false negatives in the case of NRL). We try weights $w = [1, 5, 10, 25, 50, 75, 150, 200]$. This ensures that our trained predictors achieve the required one-sided guarantees. For the ensembled FRL, we compute its global coverage and ℓ_1 error.

Next, we use a popular conformal prediction algorithm for binary classification to compare to. We choose the widely-used MAPIE library [TBM+22] and use the standard split conformal prediction method using the LAC method to compute the conformity score. We note that this is the only allowed conformity score method in MAPIE for binary classification, which is why

Table 8.1: Match on ℓ_1: for each weight w, choose α minimizing the difference in ℓ_1 error between the FRL and the conformal prediction methods.

Weight	α	FRL MAE	CP MAE	FRL coverage	CP coverage
1.0	0.25	0.250	0.260	100.0%	97.0%
5.0	0.05	0.139	0.130	54.5%	42.3%
10.0	0.05	0.113	0.130	40.0%	42.3%
25.0	0.01	0.078	0.027	25.5%	11.0%
50.0	0.01	0.045	0.027	17.7%	11.0%
75.0	0.01	0.062	0.027	16.0%	11.0%
150.0	0.01	0.015	0.027	13.1%	11.0%
200.0	0.01	0.000	0.027	9.4%	11.0%

we did not add other conformity score methods. We use a random forest as the base class, and train different models for different confidence levels of $\alpha = [0.01, 0.05, 0.10, 0.20, 0.25]$ (i.e., this is the parameter for the marginal coverage guarantee). When tested on the test set, the conformal prediction method returns prediction sets $\{0\}$, $\{1\}$, or $\{0, 1\}$. We view the set $\{0, 1\}$ as equivalent to an abstention ?, and then compute the global coverage and ℓ_1 error for the conformal prediction method.

To compare both methods for all of the weights $w = [1, 5, 10, 25, 50, 75, 150, 200]$ and coverage parameter $\alpha = [0.01, 0.05, 0.10, 0.20, 0.25]$ we match them based on a) matched ℓ_1 error, and b) matched coverage. We report the results pairs of tables as follows: a) for each weight w, we take the α value that corresponds to the closest ℓ_1 error of the FRL predictor for this w, and we report the errors and coverages for this pair. We then also report the results by matching coverages instead: b) for each weight w, we instead take the coverage value that corresponds to the closest coverage to that of the FRL predictor for this w, and we report the errors and coverages for this pair.

We do several repeats of our experiment for different initializations of the synthetic data and report the pair of Tables 8.1 and 8.2 as one such example. In all of our runs, we observe how FRL provides similar coverage and error guarantees to that of the standard conformal prediction method, with FRL sometimes having a higher coverage and lower ℓ_1 error (without a clear generalizable trend).

8.3.2 Conformal prediction from $(\mathcal{C}, \mathcal{G})$-multigroup selective classification

Similarly, we show that we can view $(\mathcal{C}, \mathcal{G})$-multigroup selective classification as a conformal prediction method in the case of binary classification. Besides the marginal coverage guarantee, now that we have a collection $\mathcal{G}$ of groups one can also hope to satisfy a conditional version of coverage. Namely, for every $g \in \mathcal{G}$, we want to satisfy $\Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\mathbf{y} \in \mathcal{S}(\mathbf{x}) \mid g(\mathbf{x}) = 1] \geq 1 - \epsilon$. This property is known as the $(\mathcal{G}, \epsilon)$-*group conditional coverage guarantee* [JNRR23]. We show that $(\mathcal{C}, \mathcal{G})$-multigroup selective classifiers do indeed satisfy this conditional guarantee:

Table 8.2: Match on coverage: for each weight w, choose α minimizing the difference in coverage between the FRL and the conformal prediction methods.

Weight	α	FRL ℓ_1	CP ℓ_1	FRL coverage	CP coverage
1.0	0.25	0.250	0.260	100.0%	97.0%
5.0	0.10	0.139	0.172	54.5%	62.7%
10.0	0.05	0.113	0.130	40.0%	42.3%
25.0	0.01	0.078	0.027	25.5%	11.0%
50.0	0.01	0.045	0.027	17.7%	11.0%
75.0	0.01	0.062	0.027	16.0%	11.0%
150.0	0.01	0.015	0.027	13.1%	11.0%
200.0	0.01	0.000	0.027	9.4%	11.0%

Lemma 8.10. *Let $h_? : \mathcal{X} \to [0,1] \cup ?$ be a $(C, \mathcal{G}, \epsilon)$-multigroup selective classifier. Then, the prediction sets induced by the level sets of h, where we map $[0, \epsilon] \mapsto \{0\}, [1-\epsilon, 1] \mapsto \{1\}$, and $(\epsilon, 1-\epsilon) \mapsto \{0,1\}$ satisfy $\Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\mathbf{y} \in \mathcal{S}(x) \mid g(\mathbf{x}) = 1] \geq 1 - \frac{\epsilon}{\Pr_{\mathcal{D}}[g(\mathbf{x})=1]}$.*

Proof. By the definition of a $(\mathcal{C}, \mathcal{G}, \epsilon)$-multigroup selective classifier, it follows that $h_?$ satisfies the global accuracy guarantee with an ϵ error parameter. This directly implies a local accuracy guarantee on each $g \in \mathcal{G}$ if we weight the error parameter by the probability mass assigned by $\mathcal{D}$ to g; that is:

$$\mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\big[|\mathbf{y} - h_?(\mathbf{x})| \cdot \mathbb{1}[h_?(\mathbf{x}) \neq ?] \mid g(\mathbf{x}) = 1\big] \leq \frac{\epsilon}{\Pr_{\mathcal{D}}[g(\mathbf{x}) = 1]}.$$

Then, the result follows from the fact that for all $\mathbf{y} = 1$,

$$\Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\mathbf{y} \in \{1\} \mid g(\mathbf{x}) = 1] \leq \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\big[|\mathbf{y} - h_?(\mathbf{x})| \cdot \mathbb{1}[h_?(\mathbf{x}) \neq ?] \mid g(\mathbf{x}) = 1\big],$$

and symmetrically for all $\mathbf{y} = 0$ it follows that

$$\Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\mathbf{y} \in \{0\} \mid g(\mathbf{x}) = 1] \leq \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\big[|\mathbf{y} - h_?(\mathbf{x})| \cdot \mathbb{1}[h_?(\mathbf{x}) \neq ?] \mid g(\mathbf{x}) = 1\big].$$

□

Remark 8.11. *As pointed out in Section 8.3.1, typical conformal prediction methods offer no theoretical bounds on the size of the prediction sets. Through our framework of $(\mathcal{C}, \mathcal{G})$-multigroup selective classification, however, we do obtain provable abstention guarantees for each $g \in \mathcal{G}$ with respect to a base concept class $\mathcal{C}$ of our choice. Specifically, for each $g \in \mathcal{G}$, the prediction sets of*

our $(\mathcal{C}, \mathcal{G})$-multigroup selective classifier $h_?$ as specified in Lemma 8.10 satisfy

$$\Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[g(\mathbf{x}) \cdot \mathbb{1}[h_?(\mathbf{x}) = \{0,1\}]\right] \leq \min_{c_? \in \mathcal{SC}(\mathcal{C})} \Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[g(\mathbf{x}) \cdot \mathbb{1}[c_?(\mathbf{x}) = ?]\right] + \epsilon.$$

Importantly, $(\mathcal{C}, \mathcal{G})$-multigroup selective classification as a conformal prediction method ensures group conditional coverage and a provable abstention bound on each group even when these intersect.

Formally:

Lemma 8.12. *Let $h_? : \mathcal{X} \to [0,1] \cup ?$ be a $(\mathcal{C}, \mathcal{G}, \epsilon)$-multigroup selective classifier. Then, the prediction sets induced by the level sets of $h_?$ satisfy:*

$$\Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[g(\mathbf{x}) \cdot \mathbb{1}[h_?(\mathbf{x}) = \{0,1\}]\right] \leq \min_{c_? \in \mathtt{SC}(\mathcal{C})} \Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[g(\mathbf{x}) \cdot \mathbb{1}[c_?(\mathbf{x}) = ?]\right] + \epsilon$$

for every group $g \in \mathcal{G}$.

Proof. This follows directly from the optimal local abstention rate guarantee of a $(\mathcal{C}, \mathcal{G}, \epsilon)$-multigroup selective classifier $h_?$, which ensures that for every $g \in \mathcal{G}$,

$$\Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[g(\mathbf{x}) \cdot \mathbb{1}[h_?(\mathbf{x}) = ?]\right] \leq \min_{c_? \in \mathtt{SC}(\mathcal{C})} \Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[g(\mathbf{x}) \cdot \mathbb{1}[c_?(\mathbf{x}) = ?]\right] + \epsilon.$$

□

We emphasize that the groups in the collection $\mathcal{G}$ can intersect arbitrarily, and hence it is a very useful property to be able to satisfy group conditional coverage guarantees.

Experiments with $(\mathcal{C}, \mathcal{G})$-multigroup selective classifiers. Lastly, we implement our $(\mathcal{C}, \mathcal{G})$-multigroup selective clasifiers in practice. To do so, we modify the construction our proof of Theorem 8.2 to adapt the multicalibration algorithm that we used in the experiments for Section 7.1.1.

Again for our proof of concept, we generate $10,000$ samples synthetically using `sk-learn`'s `make_classification` function. For the concept class $\mathcal{C}$, we use the same class of decision trees of depth 3. For the class of groups $\mathcal{G}$, we generate them from the data using randomness by allowing the groups to intersect and by ensuring some correlation with the labels within each group. In these experiments, we use 10 groups, and each ends up having size of between 300 and 400 samples.

We use the same discretization parameter 0.1 and learning rate 0.01 in the update step as in the experiments for Section 7.1.1. We perform the multicalibration algorithm across all groups in $\mathcal{G}$. In each, we use the same concept class $\mathcal{C}$ of decision trees of depth 3 to find correlation with the residuals and we cap the number of iterations at 150. We see that the $(\mathcal{C}, \mathcal{G})$-multicalibration greatly reduces the ℓ_2 (i.e., Brier score) and expected calibration error (ECE) for each of the groups in $\mathcal{G}$, as shown in Table 8.3. We emphasize that, as far as we are aware, our work is the first to separate the roles of the concept class $\mathcal{G}$ and the group collection $\mathcal{G}$ (in the multigroup fairness literature, one usually sets $\mathcal{C} = \mathcal{G}$).

Next, to turn our $(\mathcal{C}, \mathcal{G})$-multicalibrated predictor into a $(\mathcal{C}, \mathcal{G})$-multigroup selective classifier, we apply the thresholding function that we have repeatedly used in our proofs: namely, for a chosen

Table 8.3: Group-wise metrics before and after $(\mathcal{C}, \mathcal{G})$-multicalibration.

G	N	$\text{Brier}_{\text{pre}}$	$\text{Brier}_{\text{post}}$	ECE_{pre}	ECE_{post}
0	414	0.106	0.004	0.162	0.036
1	389	0.122	0.003	0.139	0.036
2	419	0.120	0.005	0.158	0.046
3	395	0.119	0.006	0.164	0.047
4	416	0.111	0.005	0.170	0.047
5	387	0.111	0.005	0.178	0.051
6	410	0.116	0.009	0.118	0.054
7	400	0.120	0.008	0.152	0.060
8	388	0.120	0.011	0.168	0.063
9	380	0.116	0.011	0.156	0.060

value of ϵ, we map the predicted value to 0 if it is $\leq \epsilon$, to 1 if it is $\geq \epsilon$, and to ? otherwise. We then compute the coverage and ℓ_1 error (i.e., mean absolute error) of our $(\mathcal{C}, \mathcal{G})$-multigroup selective classifier within each of the groups $g \in \mathcal{G}$ for different values of ϵ.

Separately, we again use the MAPIE library to train a conformal prediction method *separately within each of the groups* $g \in \mathcal{G}$. Notably, this methodology does not technically allow the groups to intersect, but we report the statistics independently on every group. In contrast, our method yields one global predictor, instead of a predictor for each group that does not allow for intersections. This is a significant advantage of using $(\mathcal{C}, \mathcal{G})$-multigroup selective classification as a conformal prediction method. We remark that two recent works use the multicalibration algorithm to obtain group conditional coverage guarantees for an intersecting collection of groups and perform extensive evaluations [JLP+21; JNRR23].

For the MAPIE training, we use random forests as the base class and $\alpha = 0.1$ as the parameter for the coverage guarantee. After training the predictor, we obtain the prediction sets $\{0\}, \{1\}$, and $\{0, 1\}$ for each of the points in the test set. Viewing $\{0, 1\}$ as equivalent to ?, we compute the coverage and ℓ_1 error of the conformal prediction method within each group (where we remark that the predictor is trained separately for every group, unlike our selective classifier). We report the coverage and ℓ_1 errors of the selective classifier for different values of ϵ (specifically, $\epsilon = 0.2, 0.3, 0.4$) and compare it to the coverage and ℓ_1 errors of the per-group conformal prediction method.

All runs show a similar pattern: the selective classifier is competitive with the conformal prediction method within each group $g \in \mathcal{G}$, even though we have a single predictor for all groups and the conformal prediction method is trained separately on each group.

Table 8.4: Pergroup Coverage and ℓ_1 error for $(\mathcal{C}, \mathcal{G})$-multigroup selective classification thresholding at $\epsilon = 0.20$, $\epsilon = 0.30$, $\epsilon = 0.40$, compared with a separate conformal prediction predictor trained separately for each group.

Group	N	$\text{Cov}_{0.20}$	$\ell_{1,0.20}$	$\text{Cov}_{0.30}$	$\ell_{1,0.30}$	$\text{Cov}_{0.40}$	$\ell_{1,0.40}$	$\text{Cov}^{\mathcal{G}}$	$\ell_1^{\mathcal{G}}$
0	421	0.983	0.104	0.988	0.106	1.000	0.107	0.988	0.108
1	405	0.983	0.118	0.985	0.118	0.990	0.120	0.938	0.100
2	408	0.975	0.131	0.983	0.135	0.993	0.141	0.963	0.115
3	384	0.977	0.120	0.979	0.122	0.992	0.123	0.958	0.114
4	410	0.949	0.118	0.956	0.117	0.973	0.128	0.956	0.128
5	401	0.960	0.114	0.970	0.116	0.980	0.120	0.963	0.098
6	388	0.961	0.123	0.977	0.129	0.985	0.134	0.951	0.108
7	399	0.952	0.124	0.967	0.122	0.987	0.124	0.962	0.109
8	400	0.932	0.123	0.948	0.127	0.970	0.139	0.963	0.109
9	409	0.914	0.131	0.917	0.133	0.961	0.155	0.941	0.132

9

Conclusions & Future Work

Now what is science? (...) It is before all a classification, a manner of bringing together facts which appearances separate, though they are bound together by some natural and hidden kinship.

Henri Poincaré, The Value of Science (1907)

In this thesis, we have thoroughly studied the connections between the multigroup fairness framework and the fields of learning theory and complexity theory. In particular, we have focused on the relationship between multiaccuracy and agnostic learning and between multiaccuracy and hardcore set constructions. In both cases, we have shown how the addition of calibration provides much stronger constructions.

Through the framework of omnipredictors, we have studied the problem of learning with abstentions from the multigroup fairness perspective, proposing the notions of selective omniprediction and multigroup selective classification We have shown that we can construct selective omnipredictors from multicalibration and multigroup selective classifiers from calibrated multiaccuracy. Naturally, our work leaves several open questions.

1. Closing gaps. There are some gaps in our results that could be potentially improved. For example, in Theorem 5.14 we showed that an (α, β)-weak agnostic learner does not yield anything better than $(\mathcal{C}, \alpha/2)$-multiaccuracy; however, Theorem 2.16 only shows that it is possible to achieve $(\mathcal{C}, \alpha)$-multiaccuracy using access to an (α, β)-weak agnostic learner.

2. Multiaccuracy and weak agnostic learning. The question of whether being able to construct a multiaccurate predictor for a class $\mathcal{C}$ implies that $\mathcal{C}$ is agnostically learnable remains open. The result in Theorem 3.1 only rules out learners obtained by post-processing a predictor satisfying certain multigroup guarantees. We do not know whether more general reductions to *multiaccuracy* can yield a weak agnostic learner. A stronger model, for example, would be an algorithm that has oracle access to a learner for multiaccurate predictors, which it can call multiple times.

3. The power of calibrated multiaccuracy. Lastly, by the equivalence between the Regularity Lemma by [TTV09] and the multiaccuracy theorem, multicalibration implies stronger and more general versions of other well-known theorems in complexity theory besides the hardcore lemma, such as the Dense Model Theorem and characterizations of pseudo-average min-entropy [RTTV08; VZ12; VZ13; CDV24; HV25], and Yao's XOR lemma [MPV25]. Similar to our result for Impagliazzo's Hardcore Lemma, it is natural to ask if versions of these theorems can be proved under weaker assumptions like calibrated multiaccuracy.

4. Relaxing the notion of calibration. As we have discussed, recent works have proposed relaxed notions of calibration, such as smooth calibration, U-calibration, or proper calibration [KLST23; OKK25]. A natural question is whether we can further relax our positive results for the case of calibrated multiaccuracy to weaker notions of calibration.

5. Learning versus refutation. As we summarized in Section 5.3, an intriguing next direction is to try to use the learning versus refutation framework for weak agnostic learning [KL18; Vad17] in our setting. Specifically, it seems that being able to audit for multiaccuracy would allow us to have weak agnostic learning.

6. The complexity of reliable agnostic learning. Recall the dependence on ϵ when obtaining a fully reliable learner from agnostic learning. In the case of [KK09], they are able to learn PRL, NRL, and FRL predictors for a concept class $\mathcal{C}$ in time $T(O(\epsilon^2))$ using an agnostic learner for $\mathcal{C}$ that runs in time $T(\epsilon)$. In our case, when we obtain fully reliable learners from selective omnipredictors in Theorem 7.7, in order to obtain a $\mathcal{L}$-fully reliable learner with error ϵ we require a selective omnipredictor with error ϵ^2. It appears that the nature of the two constraints in the definition of reliable agnostic learning (unlike the case of the generalized Chow loss formulation) induces this overhead, but it is unclear whether it is unavoidable.

7. Weak agnostic learner for $\mathcal{C} \cdot \mathcal{G}$. Our construction of a $(\mathcal{C}, \mathcal{G})$-multigroup selective classifier requires access to a $(\mathcal{C} \cdot \mathcal{G})$-multiaccurate and calibrated predictor. From the works on multigroup fairness [GHK+23; CGKR25], this in turn requires access to a weak agnostic learner for the class $\mathcal{C} \cdot \mathcal{G}$. We do not know whether it is possible to construct $(\mathcal{C}, \mathcal{G})$-multigroup selective classifiers having only access to separate weak agnostic learners for $\mathcal{C}$ and $\mathcal{G}$, without requiring a weak agnostic learner for their intersection. This can be seen as a broader question about the learnability of intersections of concept classes.

8. Building selective classifiers. In Lemma 8.4 we show that we can construct a $(\mathcal{C}, \mathcal{G})$-multigroup selective classifier from a fully reliable learner, which is believed to be a weaker primitive than (weak) agnostic learning [KT14]. However, we are only able to show this for classes $\mathcal{G}$ that are small in size, given that we need to call the FRL oracle $|\mathcal{G}|$ times. Hence the question of whether we can build selective classifiers from a weaker primitive than agnostic learning for a general class $\mathcal{G}$ remains open.

9. Conformal prediction & model multiplicity. Lastly, in light of our connections between reliable agnostic learning and $(\mathcal{C}, \mathcal{G})$-multigroup selective classification with conformal prediction, it would be interesting to develop this connection further, particularly focusing on the ability of these methods to provide provable guarantees on the sizes of the prediction sets beyond the setting of binary classification. It also appears fruitful to study how our framework of learning with abstentions, where a predictor is able to measure its own reliability, relates to the recent works on model multiplicity, as we have preliminarily discussed. Another interesting connection would be between our framework of selective classification and newly proposed methods for using higher order calibration as an uncertainty decomposition method [AGG+24].

Glossary

Notations

$\mathcal{X}$	Domain
$\mathcal{Y}$	Set of labels
x	Point in the domain
y	Label
$\mathcal{P}$	Partition of the domain
P	Piece in the partition $\mathcal{P}$
$\mathbb{E}$	Expectation
Pr	Probability distribution
$\mathcal{D}$	Distribution on $\mathcal{X} \times \mathcal{Y}$
$\mathcal{D}_{\mathcal{X}}$	Marginal distribution on $\mathcal{X}$
$\mathbb{1}$	Indicator random variable
$\mathcal{L}$	Class of loss functions
ℓ	Loss function
p^*	Ground-truth predictor
p	Our predictor
$\mathcal{C}$	Set of distinguishers/subgroups/concepts
c	Distinguisher/subgroup/concept
$\mathcal{F}$	Family of pseudorandom functions
f_r	Pseudorandom function with seed r
h	Hypothesis
$\text{Bern}(v)$	Bernoulli random variable of parameter v
$\mathcal{G}$	Collection of subgroups
g	Group in the collection $\mathcal{G}$
cor	Correlation
err	Error
ϵ	Error parameter
δ	Density parameter in IHCL
$\mathcal{C}_{t,q}$	Class with complexity at most (t, q) rel. to $\mathcal{C}$
$\langle \cdot, \cdot \rangle$	Inner product
d_H	Normalized Hamming distance

v_P	Expected value of $\mathbf{y}$ on piece $P \in \mathcal{P}$
b_P	Balance of $\mathbf{y}$ on piece $P \in \mathcal{P}$
$\mathcal{X}_v$	Set of points in $\mathcal{X}$ where $p(x) = v$
μ	Measure
$\bar{\mu}$	Normalized measure
$\mathsf{dns}(\mu)$	Density of measure μ
(α, β)	Input parameters to a WAL
μ_{TTV}	Measure constructed in [TTV09]
μ_{Max}	Our maximal measure
w	Weight function
w_{Max}	Maximum weight function
k^*_ℓ	Best-response function for loss function ℓ
$(\ell_+, \ell_-, \ell_?)$	Loss function triplets
(λ, μ, ν)	Associated weights to a triplet of loss functions
ℓ_{GC}	Generalized Chow loss
$\mathcal{C}^+$	Concepts c in $\mathcal{C}$ satisfying $\ell_+(c) = 0$
$\mathcal{C}^-$	Concepts c in $\mathcal{C}$ satisfying $\ell_-(c) = 0$
$\mathsf{SC}(\mathcal{C})$	Post-processed class $\mathcal{C}$ with selective classifiers $c_?$
I_{abs}	Abstention interval
κ_{pred}	Cost of predicting
κ_{abs}	Cost of abstaining
α	Abstention cost under the Chow model

Abbreviations

MA	Multiaccuracy
MC	Multicalibration
Cal-MA	Calibrated multiaccuracy
IHCL	Impagliazzo's Hardcore Lemma
IHCL++	Generalized & stronger Impagliazzo's Hardcore Lemma
OI	Outcome Indistinguishability
ECE	Expected calibration error
MAJ	Majority function
WAL	Weak agnostic learner
PRL	Positive reliable learning
NRL	Negative reliable learning
FRL	Fully reliable learning

Bibliography

[AB21] Anastasios N Angelopoulos and Stephen Bates. "A gentle introduction to conformal prediction and distribution-free uncertainty quantification". In: *arXiv preprint arXiv:2107.07511* (2021).

[AGG+24] Gustaf Ahdritz, Aravind Gollakota, Parikshit Gopalan, Charlotte Peale, and Udi Wieder. "Provable Uncertainty Decomposition via Higher-Order Calibration". In: *The Thirteenth International Conference on Learning Representations*. 2024.

[AGHM21] Noga Alon, Alon Gonen, Elad Hazan, and Shay Moran. "Boosting simple learners". In: *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. 2021, pp. 481–489.

[ALG21] Junaid Ali, Preethi Lahoti, and Krishna P Gummadi. "Accounting for model uncertainty in algorithmic discrimination". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021, pp. 336–345.

[ARS24] Rohan Alur, Manish Raghavan, and Devavrat Shah. "Human expertise in algorithmic prediction". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 138088–138129.

[BAZ+21] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. "Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021, pp. 401–413.

[BCDT25] Tina Behzad, Sílvia Casacuberta, Emily Ruth Diana, and Alexander Williams Tolbert. "Reconciling Predictive Multiplicity in Practice". In: *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 2025, pp. 3350–3369.

[Ben23] Ruha Benjamin. "Race after technology". In: *Social Theory Re-Wired*. Routledge, 2023, pp. 405–415.

[BFJ+94] Avrim Blum, Merrick L. Furst, Jeffrey C. Jackson, Michael J. Kearns, Yishay Mansour, and Steven Rudich. "Weakly learning DNF and characterizing statistical query learning using Fourier analysis". In: *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing*. ACM, 1994, pp. 253–262.

[BG18] Joy Buolamwini and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification". In: *Conference on fairness, accountability and transparency*. PMLR. 2018, pp. 77–91.

[BGHN23] Jarosaw Basiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. "A unifying theory of distance from calibration". In: *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*. 2023, pp. 1727–1740.

[BGJ+22] Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. "Practical adversarial multivalid conformal prediction". In: *Advances in neural information processing systems* 35 (2022), pp. 29362–29373.

[BHK09] Boaz Barak, Moritz Hardt, and Satyen Kale. "The uniform hardcore lemma via approximate bregman projections". In: *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*. SIAM. 2009, pp. 1193–1200.

[BKST24] Guy Blanc, Caleb Koch, Carmen Strassle, and Li-Yang Tan. "The sample complexity of smooth boosting and the tightness of the hardcore theorem". In: *FOCS* (2024).

[BLF21] Emily Black, Klas Leino, and Matt Fredrikson. "Selective ensembles for consistent predictions". In: *arXiv preprint arXiv:2111.08230* (2021).

[BRA+20] Noam Barda, Dan Riesel, Amichay Akriv, Joseph Levy, Uriah Finkel, Gal Yona, Daniel Greenfeld, Shimon Sheiba, Jonathan Somer, Eitan Bachmat, et al. "Developing a COVID-19 mortality risk prediction model when individual-level data are not available". In: *Nature communications* 11.1 (2020), p. 4439.

[BRB22] Emily Black, Manish Raghavan, and Solon Barocas. "Model multiplicity: Opportunities, concerns, and solutions". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022, pp. 850–863.

[Bro18] Meredith Broussard. *Artificial unintelligence: How computers misunderstand the world*. MIT Press, 2018.

[Cas23] Sílvia Casacuberta Puig. "Finding Simple Models of Complex Objects: From Regularity Lemmas to Algorithmic Fairness". Bachelor's Thesis. Harvard University, 2023.

[CDV24] Sílvia Casacuberta, Cynthia Dwork, and Salil Vadhan. "Complexity-Theoretic Implications of Multicalibration". In: *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*. 2024, pp. 1071–1082.

[CGKR25] Sílvia Casacuberta, Parikshit Gopalan, Varun Kanade, and Omer Reingold. "How global calibration strengthens multiaccuracy". In: *Proceedings of the 66th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*. 2025.

[Cho57] Chi-Keung Chow. "An optimum character recognition system using decision functions". In: *IRE Transactions on Electronic Computers* 4 (1957), pp. 247–254.

[CK25] Sílvia Casacuberta and Varun Kanade. "Selective Omniprediction and Fair Abstention". In: *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 2025.

[CLC+24] A Feder Cooper, Katherine Lee, Madiha Zahrah Choksi, Solon Barocas, Christopher De Sa, James Grimmelmann, Jon Kleinberg, Siddhartha Sen, and Baobao Zhang. "Arbitrariness and Social Prediction: The Confounding Role of Variance in Fair Classification". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 20. 2024, pp. 22004–22012.

[CLP15] Kai-Min Chung, Edward Lui, and Rafael Pass. "From Weak to Strong Zero-Knowledge and Applications". In: *Theory of Cryptography - 12th Theory of Cryptography Conference, TCC 2015, Warsaw, Poland, March 23-25, 2015, Proceedings, Part I*. Ed. by Yevgeniy Dodis and Jesper Buus Nielsen. Vol. 9014. Lecture Notes in Computer Science. Springer, 2015, pp. 66–92.

[CS24] Mohammad-Amin Charusaie and Samira Samadi. "A Unifying Post-Processing Framework for Multi-Objective Learn-to-Defer Problems". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2024.

[Dan15] Amit Daniely. "A PTAS for agnostically learning halfspaces". In: *Conference on Learning Theory*. PMLR. 2015, pp. 484–502.

[Daw82] A Philip Dawid. "The well-calibrated Bayesian". In: *Journal of the American statistical Association* 77.379 (1982), pp. 605–610.

[Daw85] A Philip Dawid. "Calibration-based empirical probability". In: *The Annals of Statistics* 13.4 (1985), pp. 1251–1274.

[DBFR24] Gianluca Detommaso, Martin Andres Bertran, Riccardo Fogliato, and Aaron Roth. "Multicalibration for Confidence Scoring in LLMs". In: *International Conference on Machine Learning*. PMLR. 2024, pp. 10624–10641.

[DDF+25] Yuval Dagan, Constantinos Daskalakis, Maxwell Fishelson, Noah Golowich, Robert Kleinberg, and Princewill Okoroafor. "Breaking the T^(2/3) Barrier for Sequential Calibration". In: *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*. 2025, pp. 2007–2018.

[DHP+12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. "Fairness through awareness". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science conference*. 2012, pp. 214–226.

[DK23] Catherine D'ignazio and Lauren F Klein. *Data feminism*. MIT press, 2023.

[DKK+21] Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. "Agnostic proper learning of halfspaces under Gaussian marginals". In: *Conference on Learning Theory*. PMLR. 2021, pp. 1522–1551.

[DKKS24] Siddartha Devic, Aleksandra Korolova, David Kempe, and Vatsal Sharan. "Stability and Multigroup Fairness in Ranking with Uncertain Predictions". In: *International Conference on Machine Learning*. PMLR. 2024, pp. 10661–10686.

[DKR+21] Cynthia Dwork, Michael P Kim, Omer Reingold, Guy N Rothblum, and Gal Yona. "Outcome indistinguishability". In: *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. 2021, pp. 1095–1108.

[DKR+22] Cynthia Dwork, Michael P Kim, Omer Reingold, Guy N Rothblum, and Gal Yona. "Beyond bernoulli: Generating random outcomes that cannot be distinguished from nature". In: *International Conference on Algorithmic Learning Theory*. PMLR. 2022, pp. 342–380.

[DLLT23] Cynthia Dwork, Daniel Lee, Huijia Lin, and Pranay Tankala. "From Pseudorandomness to Multi-Group Fairness and Back". In: *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023, 12-15 July 2023, Bangalore, India*. Vol. 195. Proceedings of Machine Learning Research. PMLR, 2023, pp. 3566–3614.

[DNW24] Ally Yalei Du, Dung Daniel Ngo, and Zhiwei Steven Wu. "Reconciling Model Multiplicity for Downstream Decision Making". In: *arXiv preprint arXiv:2405.19667* (2024).

[DP08] Stefan Dziembowski and Krzysztof Pietrzak. "Leakage-resilient cryptography". In: *2008 49th Annual IEEE Symposium on Foundations of Computer Science*. IEEE. 2008, pp. 293–302.

[DRR23] Cynthia Dwork, Omer Reingold, and Guy N Rothblum. "From the Real Towards the Ideal: Risk Prediction in a Better World". In: *4th Symposium on Foundations of Responsible Computing (FORC 2023)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik. 2023.

[Eub18] Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

[Fel09a] Vitaly Feldman. "Distribution-specific agnostic boosting". In: *arXiv preprint arXiv:0909.2927* (2009).

[Fel09b] Vitaly Feldman. "On the power of membership queries in agnostic learning". In: *The Journal of Machine Learning Research* 10 (2009), pp. 163–182.

[FGKP06] Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. "New results for learning noisy parities and halfspaces". In: *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*. IEEE. 2006, pp. 563–574.

[FGMS25] Maxwell Fishelson, Noah Golowich, Mehryar Mohri, and Jon Schneider. "High-Dimensional Calibration from Swap Regret". In: *arXiv preprint arXiv:2505.21460* (2025).

[FK99] Alan Frieze and Ravi Kannan. "Quick approximation to matrices and applications". In: *Combinatorica* 19.2 (1999), pp. 175–220.

[FKP25] Unai Fischer-Abaigar, Christoph Kern, and Juan Carlos Perdomo. "The Value of Prediction in Identifying the Worst-Off". In: *Forty-second International Conference on Machine Learning*. 2025.

[FN24] Yizirui Fang and Eric Nalisnick. "Learning to Defer with an Uncertain Rejector via Conformal Prediction". In: *NeurIPS 2024 Workshop on Bayesian Decision-making and Uncertainty*. 2024.

[Fre92] Yoav Freund. "An improved boosting algorithm and its implications on learning complexity". In: *Proceedings of the Fifth Annual workshop on Computational Learning Theory*. 1992, pp. 391–398.

[Fre95] Yoav Freund. "Boosting a weak learning algorithm by majority". In: *Information and Computation* 121.2 (1995), pp. 256–285.

[FS97] Yoav Freund and Robert E Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". In: *Journal of Computer and System Sciences* 55.1 (1997), pp. 119–139.

[FSA99] Yoav Freund, Robert Schapire, and Naoki Abe. "A short introduction to boosting". In: *Journal-Japanese Society For Artificial Intelligence* 14.771-780 (1999), p. 1612.

[FT25] Yiding Feng and Wei Tang. "Persuasive Calibration". In: *arXiv preprint arXiv:2504.03211* (2025).

[FV97] Dean P Foster and Rakesh V Vohra. "Calibrated learning and correlated equilibrium". In: *Games and Economic Behavior* 21.589 (1997), pp. 40–55.

[FV98] Dean P Foster and Rakesh V Vohra. "Asymptotic calibration". In: *Biometrika* 85.2 (1998), pp. 379–390.

[GGKS23] Aravind Gollakota, Parikshit Gopalan, Adam Klivans, and Konstantinos Stavropoulos. "Agnostically learning single-index models using omnipredictors". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 14685–14704.

[GGM86] Oded Goldreich, Shafi Goldwasser, and Silvio Micali. "How to construct random functions". In: *J. ACM* 33.4 (1986), pp. 792–807.

[GHK+23] Parikshit Gopalan, Lunjia Hu, Michael P Kim, Omer Reingold, and Udi Wieder. "Loss Minimization Through the Lens Of Outcome Indistinguishability". In: *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik. 2023, pp. 60–1.

[GHR24] Parikshit Gopalan, Lunjia Hu, and Guy N Rothblum. "On computationally efficient multiclass calibration". In: *The Thirty Seventh Annual Conference on Learning Theory*. PMLR. 2024, pp. 1983–2026.

[GJN+22] Varun Gupta, Christopher Jung, Georgy Noarov, Mallesh M Pai, and Aaron Roth. "Online Multivalid Learning: Means, Moments, and Prediction Intervals". In: *Innovations in Theoretical Computer Science (ITCS)* (2022).

[GJRR24] Sumegha Garg, Christopher Jung, Omer Reingold, and Aaron Roth. "Oracle efficient online multicalibration and omniprediction". In: *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM. 2024, pp. 2725–2792.

[GKKM20] Shafi Goldwasser, Adam Tauman Kalai, Yael Kalai, and Omar Montasser. "Beyond perturbations: Learning guarantees with arbitrary adversarial test examples". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 15859–15870.

[GKR+22] Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. "Omnipredictors". In: *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik. 2022, pp. 79–1.

[GKR23] Parikshit Gopalan, Michael Kim, and Omer Reingold. "Swap agnostic learning, or characterizing omniprediction via multicalibration". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 39936–39956.

[GKR24] Parikshit Gopalan, Michael Kim, and Omer Reingold. "Swap agnostic learning, or characterizing omniprediction via multicalibration". In: *Advances in Neural Information Processing Systems* 36 (2024).

[GKSZ22] Parikshit Gopalan, Michael P Kim, Mihir A Singhal, and Shengjia Zhao. "Low-degree multicalibration". In: *Conference on Learning Theory*. PMLR. 2022, pp. 3193–3234.

[GL89] Oded Goldreich and Leonid A Levin. "A hard-core predicate for all one-way functions". In: *Proceedings of the twenty-first annual ACM Symposium on Theory of Computing*. 1989, pp. 25–32.

[GNW11] Oded Goldreich, Noam Nisan, and Avi Wigderson. "On Yao's XOR-Lemma". In: *Studies in Complexity and Cryptography: Miscellanea on the Interplay Between Randomness and Computation* 6650 (2011), p. 273.

[GOR+24] Parikshit Gopalan, Princewill Okoroafor, Prasad Raghavendra, Abhishek Sherry, and Mihir Singhal. "Omnipredictors for regression and the approximate rank of convex functions". In: *The Thirty Seventh Annual Conference on Learning Theory*. PMLR. 2024, pp. 2027–2070.

[GRSW22] Parikshit Gopalan, Omer Reingold, Vatsal Sharan, and Udi Wieder. "Multicalibrated partitions for importance weights". In: *International Conference on Algorithmic Learning Theory*. PMLR. 2022, pp. 408–435.

[GS00] Venkatesan Guruswami and Madhu Sudan. "List decoding algorithms for certain concatenated codes". In: *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*. ACM, 2000, pp. 181–190.

[GT08] Ben Green and Terence Tao. "The primes contain arbitrarily long arithmetic progressions". In: *Annals of Mathematics* (2008), pp. 481–547.

[GW11] Craig Gentry and Daniel Wichs. "Separating succinct non-interactive arguments from all falsifiable assumptions". In: *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*. Ed. by Lance Fortnow and Salil P. Vadhan. ACM, 2011, pp. 99–108.

[Hau92] David Haussler. "Decision theoretic generalizations of the PAC model for neural net and other learning applications". In: *Information and Computation* 100.1 (1992), pp. 78–150.

[HCM+23] Maria Heuss, Daniel Cohen, Masoud Mansoury, Maarten de Rijke, and Carsten Eickhoff. "Predictive uncertainty-based bias mitigation in ranking". In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 2023, pp. 762–772.

[HDNS24] Dutch Hansen, Siddartha Devic, Preetum Nakkiran, and Vatsal Sharan. "When is multicalibration post-processing necessary?" In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 38383–38455.

[HJZ24] Nika Haghtalab, Michael Jordan, and Eric Zhao. "A unifying perspective on multi-calibration: Game dynamics for multi-objective learning". In: *Advances in Neural Information Processing Systems* 36 (2024).

[HKRR18] Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. "Multicalibration: Calibration for the (computationally-identifiable) masses". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 1939–1948.

[HNRY23] Lunjia Hu, Inbal Rachel Livni Navon, Omer Reingold, and Chutong Yang. "Omnipredictors for constrained optimization". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 13497–13527.

[Hol05] Thomas Holenstein. "Key agreement from weak bit agreement". In: *Proceedings of the Thirty-Seventh Annual ACM Symposium on Theory of Computing*. 2005, pp. 664–673.

[HP23] Lunjia Hu and Charlotte Peale. "Comparative Learning: A Sample Complexity Theory for Two Hypothesis Classes". In: *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*. 2023.

[HPR22] Lunjia Hu, Charlotte Peale, and Omer Reingold. "Metric entropy duality and the sample complexity of outcome indistinguishability". In: *International Conference on Algorithmic Learning Theory*. PMLR. 2022, pp. 515–552.

[HPS16] Moritz Hardt, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning". In: *Advances in Neural Information Processing Systems* 29 (2016).

[HQYZ24] Nika Haghtalab, Mingda Qiao, Kunhe Yang, and Eric Zhao. "Truthfulness of Calibration Measures". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2024.

[HTGG22] Isabelle Hupont, Songül Tolan, Hatice Gunes, and Emilia Gómez. "The landscape of facial processing applications in the context of the European AI Act and the development of trustworthy systems". In: *Scientific Reports* 12.1 (2022), p. 10688.

[HV25] Lunjia Hu and Salil Vadhan. "Generalized and unified equivalences between hardness and pseudoentropy". In: *Theory of Cryptography Conference*. Springer. 2025, pp. 258–288.

[Imp95] Russell Impagliazzo. "Hard-core distributions for somewhat hard problems". In: *Proceedings of IEEE 36th Annual Foundations of Computer Science*. IEEE. 1995, pp. 538–545.

[IMR14] Russell Impagliazzo, Cristopher Moore, and Alexander Russell. "An Entropic Proof of Chang's Inequality". In: *SIAM J. Discret. Math.* 28.1 (2014), pp. 173–176.

[JLP+21] Christopher Jung, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra. "Moment multicalibration for uncertainty estimation". In: *Conference on Learning Theory*. PMLR. 2021, pp. 2634–2678.

[JNRR23] Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. "Batch Multivalid Conformal Prediction". In: *International Conference on Learning Representations (ICLR)*. 2023.

[JP14] Dimitar Jetchev and Krzysztof Pietrzak. "How to fake auxiliary input". In: *Theory of Cryptography Conference*. Springer. 2014, pp. 566–590.

[JRLT23] Junqi Jiang, Antonio Rago, Francesco Leofante, and Francesca Toni. "Recourse under model multiplicity via argumentative ensembling". In: *arXiv preprint arXiv:2312.15097* (2023).

[JSK+20] Erik Jones, Shiori Sagawa, Pang Wei Koh, Ananya Kumar, and Percy Liang. "Selective Classification Can Magnify Disparities Across Groups". In: *International Conference on Learning Representations*. 2020.

[Kab02] Valentine Kabanets. "Derandomization: A brief overview". In: *Current Trends in Theoretical Computer Science* 1 (2002), pp. 165–188.

[KCGK23] Selim Kuzucu, Jiaee Cheong, Hatice Gunes, and Sinan Kalkan. "Uncertainty-based fairness measures". In: *arXiv preprint arXiv:2312.11299* (2023).

[KF04] Sham M Kakade and Dean P Foster. "Deterministic calibration and Nash equilibrium". In: *International Conference on Computational Learning Theory*. Springer. 2004, pp. 33–48.

[KGZ19] Michael P Kim, Amirata Ghorbani, and James Zou. "Multiaccuracy: Black-box post-processing for fairness in classification". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019, pp. 247–254.

[KHS23] Falaah Arif Khan, Denys Herasymuk, and Julia Stoyanovich. "On Fairness and Stability: Is Estimator Variance a Friend or a Foe?" In: *arXiv preprint arXiv:2302.04525* (2023).

[KK09] Varun Kanade and Adam Kalai. "Potential-based agnostic boosting". In: *Advances in Neural Information Processing Systems* 22 (2009).

[KK21a] Adam Kalai and Varun Kanade. "Towards optimally abstaining from prediction with OOD test examples". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 12774–12785.

[KK21b] Adam Tauman Kalai and Varun Kanade. "Efficient learning with arbitrary covariate shift". In: *Algorithmic Learning Theory*. PMLR. 2021, pp. 850–864.

[KKG+22] Michael P Kim, Christoph Kern, Shafi Goldwasser, Frauke Kreuter, and Omer Reingold. "Universal adaptability: Target-independent inference that competes with propensity scoring". In: *Proceedings of the National Academy of Sciences* 119.4 (2022), e2108097119.

[KKM12] Adam Tauman Kalai, Varun Kanade, and Yishay Mansour. "Reliable agnostic learning". In: *Journal of Computer and System Sciences* 78.5 (2012), pp. 1481–1495.

[KKR22] Patrick Kaiser, Christoph Kern, and David Rügamer. "Uncertainty-aware predictive modeling for fair data-driven decisions". In: *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*. 2022.

[KKZ24] Christoph Kern, Michael Kim, and Angela Zhou. "Multi-CATE: Multi-Accurate Conditional Average Treatment Effect Estimation Robust to Unknown Covariate Shifts". In: *arXiv preprint arXiv:2405.18206* (2024).

[KL18] Pravesh K Kothari and Roi Livni. "Agnostic learning by refuting". In: *9th Innovations in Theoretical Computer Science, ITCS 2018*. Schloss Dagstuhl-Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing. 2018, p. 55.

[KLST23] Bobby Kleinberg, Renato Paes Leme, Jon Schneider, and Yifeng Teng. "U-calibration: Forecasting for an unknown agent". In: *The Thirty Sixth Annual Conference on Learning Theory*. PMLR. 2023, pp. 5143–5145.

[KM91] Eyal Kushilevitz and Yishay Mansour. "Learning decision trees using the Fourier spectrum". In: *Proceedings of the 23rd Annual ACM Symposium on Theory of Computing*. 1991, pp. 455–464.

[KMV08] Adam Tauman Kalai, Yishay Mansour, and Elad Verbin. "On agnostic boosting and parity learning". In: *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*. ACM, 2008, pp. 629–638.

[KNRW18] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 2564–2572.

[KP23] Michael P Kim and Juan C Perdomo. "Making Decisions Under Outcome Performativity". In: *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik. 2023, pp. 79–1.

[KS03] Adam R Klivans and Rocco A Servedio. "Boosting and hard-core set construction". In: *Machine Learning* 51 (2003), pp. 217–238.

[KSS92] Michael J Kearns, Robert E Schapire, and Linda M Sellie. "Toward efficient agnostic learning". In: *Proceedings of the fifth annual workshop on Computational learning theory*. 1992, pp. 341–352.

[KT14] Varun Kanade and Justin Thaler. "Distribution-independent reliable learning". In: *Conference on Learning Theory*. PMLR. 2014, pp. 3–24.

[LBR+21] Joshua K Lee, Yuheng Bu, Deepta Rajan, Prasanna Sattigeri, Rameswar Panda, Subhro Das, and Gregory W Wornell. "Fair selective classification via sufficiency". In: *International conference on machine learning*. PMLR. 2021, pp. 6076–6086.

[LHAC23] Carol Xuan Long, Hsiang Hsu, Wael Alghamdi, and Flavio P Calmon. "Arbitrariness lies beyond the fairness-accuracy frontier". In: *arXiv preprint arXiv:2306.09425* (2023).

[LHAC24] Carol Long, Hsiang Hsu, Wael Alghamdi, and Flavio Calmon. "Individual Arbitrariness and Group Fairness". In: *Advances in Neural Information Processing Systems* 36 (2024).

[LMKA16] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. "How we analyzed the COMPAS recidivism algorithm". In: *ProPublica (5 2016)* 9.1 (2016).

[LNPR22] Daniel Lee, Georgy Noarov, Mallesh Pai, and Aaron Roth. "Online minimax multiobjective optimization: Multicalibeating and other applications". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 29051–29063.

[Lon01] Philip M Long. "On Agnostic Learning with {0,*, 1}-Valued and Real-Valued Hypotheses". In: *14th Annual Conference on Computational Learning Theory*. Vol. 14. Springer Science & Business Media. 2001, p. 289.

[LSH19] Lydia T Liu, Max Simchowitz, and Moritz Hardt. "The implicit fairness criterion of unconstrained learning". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 4051–4060.

[LW24] Terrance Liu and Steven Wu. "Multi-group Uncertainty Quantification for Long-form Text Generation". In: *The 41st Conference on Uncertainty in Artificial Intelligence*. 2024.

[MCU20] Charles Marx, Flavio Calmon, and Berk Ustun. "Predictive multiplicity in classification". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 6765–6774.

[MPV25] Cassandra Marcussen, Aaron Putterman, and Salil Vadhan. "Characterizing the Distinguishability of Product Distributions Through Multicalibration". In: *40th Computational Complexity Conference (CCC 2025)*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik. 2025, pp. 19–1.

[MPZ18] David Madras, Toniann Pitassi, and Richard Zemel. "Predict responsibly: improving fairness and accuracy by learning to defer". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 2018, pp. 6150–6160.

[NR23] Georgy Noarov and Aaron Roth. "The statistical scope of multicalibration". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 26283–26310.

[NRRX23] Georgy Noarov, Ramya Ramalingam, Aaron Roth, and Stephan Xie. "High-Dimensional Prediction for Sequential Decision Making". In: *Forty-second International Conference on Machine Learning*. 2023.

[ODo14] Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

[OKK25] Princewill Okoroafor, Robert Kleinberg, and Michael P. Kim. "Near-Optimal Algorithms for Omniprediction". In: *Proceedings of the 66th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*. 2025.

[ONe17] Cathy O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2017.

[Per24] Juan Carlos Perdomo. "The Relative Value of Prediction in Algorithmic Decision Making". In: *International Conference on Machine Learning*. PMLR. 2024, pp. 40439–40460.

[PKD+21] Florian Pfisterer, Christoph Kern, Susanne Dandl, Matthew Sun, Michael P Kim, and Bernd Bischl. "mcboost: Multi-calibration boosting for R". In: *Journal of Open Source Software* 6.64 (2021), p. 3453.

[PRR24] Charlotte Peale, Vinod Raman, and Omer Reingold. "Representative Language Generation". In: *Forty-second International Conference on Machine Learning*. 2024.

[QZ25] Mingda Qiao and Eric Zhao. "Truthfulness of Decision-Theoretic Calibration Measures". In: *Proceedings of Machine Learning Research vol* 291 (2025), pp. 1–54.

[Rot23] Guy N Rothblum. "Indistinguishable Predictions and Multi-group Fair Learning". In: *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer. 2023, pp. 3–21.

[RS24] Aaron Roth and Mirah Shi. "Forecasting for swap regret for all downstream agents". In: *Proceedings of the 25th ACM Conference on Economics and Computation*. 2024, pp. 466–488.

[RTTV08] Omer Reingold, Luca Trevisan, Madhur Tulsiani, and Salil Vadhan. "Dense subsets of pseudorandom sets". In: *2008 49th Annual IEEE Symposium on Foundations of Computer Science*. IEEE. 2008, pp. 76–85.

[RTW23] Aaron Roth, Alexander Tolbert, and Scott Weinstein. "Reconciling Individual Probability Forecasts". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 2023, pp. 101–110.

[SAH24] Ali Shirali, Rediet Abebe, and Moritz Hardt. "Allocation requires prediction only if inequality is low". In: *Proceedings of the 41st International Conference on Machine Learning*. 2024, pp. 45114–45153.

[San03] Alvaro Sandroni. "The reproducible properties of correct forecasts". In: *International Journal of Game Theory* 32.1 (2003), pp. 151–159.

[SC21] Nicolas Schreuder and Evgenii Chzhen. "Classification with abstention but without disparities". In: *Uncertainty in Artificial Intelligence*. PMLR. 2021, pp. 1227–1236.

[Sch90] Robert E Schapire. "The strength of weak learnability". In: *Machine learning* 5.2 (1990), pp. 197–227.

[Skó17] Maciej Skórski. "A cryptographic view of regularity lemmas: Simpler unified proofs and refined bounds". In: *International Conference on Theory and Applications of Models of Computation*. Springer. 2017, pp. 586–599.

[SSV03] Alvaro Sandroni, Rann Smorodinsky, and Rakesh V Vohra. "Calibration with many checking rules". In: *Mathematics of operations Research* 28.1 (2003), pp. 141–153.

[SV08] Glenn Shafer and Vladimir Vovk. "A tutorial on conformal prediction." In: *Journal of Machine Learning Research* 9.3 (2008).

[TBM+22] Vianney Taquet, Vincent Blot, Thomas Morzadec, Louis Lacombe, and Nicolas Brunel. "MAPIE: an open-source library for distribution-free uncertainty quantification (2022)". In: *arXiv preprint arXiv:2207.12274* (2022).

[TCL23] Anique Tahir, Lu Cheng, and Huan Liu. "Fairness through aleatoric uncertainty". In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 2023, pp. 2372–2381.

[Tre03] Luca Trevisan. "List-decoding using the XOR lemma". In: *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.* IEEE. 2003, pp. 126–135.

[TTV09] Luca Trevisan, Madhur Tulsiani, and Salil Vadhan. "Regularity, boosting, and efficiently simulating every high-entropy distribution". In: *2009 24th Annual IEEE Conference on Computational Complexity*. IEEE. 2009, pp. 126–136.

[TZ08] Terence Tao and Tamar Ziegler. "The primes contain arbitrarily long polynomial progressions". In: *Acta Math.* 201.2 (2008), pp. 213–305. ISSN: 0001-5962.

[Vad17] Salil Vadhan. "On learning vs. refutation". In: *Conference on Learning Theory*. PMLR. 2017, pp. 1835–1848.

[Val84] Leslie G Valiant. "A theory of the learnable". In: *Communications of the ACM* 27.11 (1984), pp. 1134–1142.

[VGS05] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Vol. 29. Springer, 2005.

[Vin18] James Vincent. "Amazon reportedly scraps internal AI recruiting tool that was biased against women". In: *The Verge* 10 (2018).

[VZ12] Salil P. Vadhan and Colin Jia Zheng. "Characterizing pseudoentropy and simplifying pseudorandom generator constructions". In: *Proceedings of the 44th Symposium on Theory of Computing Conference, STOC 2012, New York, NY, USA, May 19 - 22, 2012*. Ed. by Howard J. Karloff and Toniann Pitassi. ACM, 2012, pp. 817–836.

[VZ13] Salil Vadhan and Colin Jia Zheng. "A uniform min-max theorem with applications in cryptography". In: *Annual Cryptology Conference*. Springer. 2013, pp. 93–110.

[WLCW24] Jiayun Wu, Jiashuo Liu, Peng Cui, and Steven Wu. "Bridging Multicalibration and Out-of-distribution Generalization Beyond Covariate Shift". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2024.

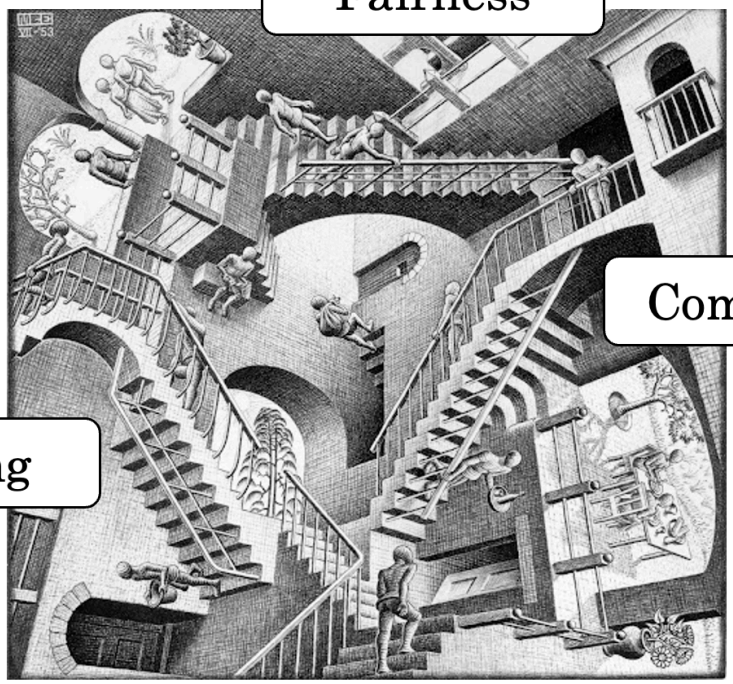
[Yao82] Andrew C. Yao. "Theory and Applications of Trapdoor Functions (Extended Abstract)". In: *23rd Annual Symposium on Foundations of Computer Science, Chicago, Illinois, USA, 3-5 November 1982*. IEEE Computer Society, 1982, pp. 80–91.

[YTG+] Tongxin Yin, Jean-Francois Ton, Ruocheng Guo, Yuanshun Yao, Mingyan Liu, and Yang Liu. "Fair Classifiers that Abstain without Harm". In: *The Twelfth International Conference on Learning Representations*.

[Zhe14] Jia Zheng. *A uniform min-max theorem and characterizations of computational randomness*. Harvard University, 2014.

[ZKS+21] Shengjia Zhao, Michael Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. "Calibrating predictions to decisions: A novel approach to multi-class calibration". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 22313–22324.

Fairness



Complexity

Learning