Title
Optimization Frameworks for Fair Data-Driven Decision Making

Author
Olfat, Mahbod

Publication Date
2020

Peer reviewed|Thesis/dissertation

Optimization Frameworks for Fair Data-Driven Decision Making

by

Mahbod Olfat

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Industrial Engineering and Operations Research

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor Anil Aswani, Chair
Professor Rhonda Righter
Associate Professor Prasad Raghavendra
Assistant Professor Paul Grigas

Spring 2020

Optimization Frameworks for Fair Data-Driven Decision Making

To my family, for unconditional support and for teaching me to expect the most from myself

And to those close friends from whom I have learned everything. I am nothing but the sum of the experiences that you have been kind enough to grant me.

Contents

List of Figures

List of Tables

Abstract

Optimization Frameworks for Fair Data-Driven Decision Making

by

Mahbod Olfat

Doctor of Philosophy in Industrial Engineering and Operations Research

University of California, Berkeley

Associate Professor Anil Aswani, Chair

This thesis investigates the problem of fair statistical learning. We argue that critical notions of fairness can be represented by independence constraints on certain random variables, and take the approach of approximating independence by bounding moments. We propose a hierarchical Fair Optimization (FO) framework for generalized fair decision-making, prove desirable statistical properties, and extend the framework to a number of settings, ranging from supervised learning to unsupervised learning and hypothesis testing.

Algorithmic decision-making has steadily gained in prominence as more data is produced and computing resources become more abundant. However, it has been observed that these can often reflect, and perpetuate, biases apparent in the training data. To that end, we construct the FO framework as a general approach to statistical decision-making under fairness constraints. The framework revolves around bounding the moments between a "score" function underlying the decision-making process and predefined "protected" attributes. We prove that this framework is consistent and will thus asymptotically provide fair decision rules, and provide non-asymptotic bounds on how quickly the framework approaches truly fair decision-making rules. We also provide experimental results that show the efficacy of the FO hierarchy on a variety of datasets, and use it to construct fair, automated one-time and sequential dosage mechanisms for morphine and heparin.

Novel, adversarial notions of fairness are then defined for the problem of dimensionality reduction of data, and a Semidefinite Programming (SDP) relaxation of the FO hierarchy is defined that controls these notions. we provide experimental analysis, including a case study on insurance rate-setting that allows for mechanisms that are fair with respect to legally-motivated age restrictions. Similarly, we extend fairness to the problem of hypothesis testing, and make the connection between fairness and robustness in this realm. This is actuated in the form of a distributionally-robust dynamic watermarking scheme to detect attacks on dynamical systems. Finally, we extend the intuitions of data-dependent regularization

underlying the FO hierarchy to design a data-dependent regularizer that promotes robustness in classifiers in the low-data regime when data lies in a low-dimensional manifold.

Acknowledgments

This thesis is comprised of joint work with my advisor, Dr. Anil Aswani. I am hugely grateful to him for his support and guidance, and in particular for being very open and welcoming throughout these last few years. I am also grateful to the Center for Long-Term Cybersecurity, which supported the work in this thesis. I also appreciate discussions and commentary from the other faculty in the Industrial Engineering and Operations Research department, namely Drs. Rhonda Righter, Dorit Hochbaum, Paul Grigas and Javad Lavaei. I would like to thank Dr. Barbara Laraia of the Public Health department for her assistance on projects throughout my PhD. I would like to extend my thanks as well to the department staff as well, including Keith, Rebecca, Anayancy, Diana and Heather for their assistance, as well as for their friendly cheer and for some great conversations throughout the last few years.

My family were my greatest asset throughout the duration of my PhD, and I could not have completed this process without their constant support. My parents, Mahnaz and Masoud, were a party to all of my peaks and troughs; they provided a support for me to lean on at a moment's notice, which I needed at many moments. My sister, Tarra, was a refreshing and familiar voice when times were gray, and always made me a little cheerier whenever she called. I am so lucky to have you.

I've made many great friends throughout my time at Berkeley, from whom I've learned more than any number of courses and without whom my PhD experience would have been dull and unmemorable. These include, but are not limited to, Salar Fattahi, Pedro Hespanhol, George Patsakis, Dean Grosbard, Igor Molybog, Yonatan Mintz, Arman Jabbari, Erik Bertelli, Armin Askari, Alfonso Lobos, Sang Woo Park, Mahan Tajrobehkar, Han Feng, Cedric Josz, Richard Zhang and Quico Spaen, among others. Thank you for all the wonderful conversations, the intrepid arguments, the instigating questions and remarkably persistent inside jokes.

Outside of the department, there are a group of old friends that are my rock and that were always there to help me through times of difficulty or frustration. There are too many to name here, but I would like to thank Behtash Banihashemi, Syrus Razavi, Omid Tabatabaie, Soroush Etemad, Haroun Ahmed, Saiful Khan, Gautam Kanumuru and Macklin McMullen. Thank you for the conversations that never failed to lighten up my day. Finally, I would like to extend the deepest of thanks to my girlfriend, Parisa Davoodi. You bore the brunt of my daily trials and tribulations, and nothing ever seemed genuine until I had told you about it. Thank you.

Chapter 1

Introduction & Preliminaries

1.1 Introduction

The explosion of computational capacity in recent decades has opened the door for machine learning and other automated decision-making techniques to play an outsize role in day-to-day activities. Set up to learn relationships and fashion decision-rules from data, these algorithms have become ubiquitous in the most mundane and every-day uses, such as fraud detection [39], credit scoring [186] and ad targeting [201], to the most groundbreaking, such as autonomous driving and determining the entanglement of qubits in a quantum computer [117]. Their proliferation has extended even to sensitive applications with large human or societal impacts, whether immediate or delayed. Examples of these include healthcare [128,157], hiring [55,230], and criminal justice [11,253]. The requirements on algorithms to be used in such spaces differ from those meant for more aspirational or innocuous uses, as these contexts are viewed as having higher risk, and so their adoption has been significantly slower. Interestingly, decisions made by machine learning algorithms still tend to be more efficacious than those made by humans in many of these high-risk instances [242]; however, the difference lies in a human's ability to moderate, justify, and consider the ramifications of decisions. In effect, we require a retooling of machine learning and automated decision-making frameworks in order to be able to handle societal considerations of *robustness*, *interpretability* and *fairness*.

Statistical decision-making techniques rely on data produced by a true, underlying system in order to learn some optimal or nearly-optimal action, which can then yield tangible effects on the original true system. A key feature of many high-risk applications of automation from this perspective is that the true, underlying system generating data is not relatively contained, like that of a quantum computer or even an advertisement, but rather reflects the whole of society. This connection of an application with broad, societal phenomena is important from two angles: It can define the validity of the data input to algorithms, and it can define the scope of the impact of their actions. First, consider that the data produced by members of a group will likely reflect the biases of that group itself, skewing the data

produced with undesirable or misleading correlations that statistical learning techniques can mistake for true signal. For example, common predictors for creditworthiness include aspects such as the ZIP code in which a candidate lives and that candidate's income, but both can be highly correlated to a number of sensitive attributes like race, ethnicity and gender. Second, systematically-biased decisions made by automated decision-makers in economically- or societally-determinative areas can serve to perpetuate the biases that originally infected the data itself. Specifically, decisions like those involved in hiring and criminal justice have sweeping implications on the lives of those affected, both in positive and negative ways. Systematic biases in these arenas, intentional or not, can serve to exacerbate existing social inequities.

Concerns of systematic biases in machine learning algorithms have become prominent since receiving notable, though anecdotal, coverage over the last few years [11, 20, 79, 84]. A particularly impactful piece was an analysis of the COMPAS algorithm for predicting criminal recidivism, conducted by a team of investigative journalists at ProPublica, which found that it was overly-cavalier in determining black convicts to be at risk of recidivism [11]. In response, the academic community has developed a number of approaches to encourage fairness in various machine learning problems [45, 57, 77, 109, 191, 270, 276]. The problem of classification has received particular attention due to the ease of mapping class labels to positive and negative outcomes with which to understand bias, but recent work has also begun to explore bias reduction methods in the context of unsupervised learning [56, 192] and in a more general decision-analytic framework [83, 158].

In this thesis, we provide a general, hierarchical optimization framework for statistical decision-making in the presence of fairness constraints. In the most simple interpretation, we seek to find decision rules that do not contain any information about one or more "protected attributes", while continuing to perform well on accuracy metrics. The core problem is thus one of statistical estimation subject to approximate independence constraints. Aside from exogenous, yet still pertinent, societal concerns of fairness, this has a clear use in improving generalization of models: a priori knowledge of the dependence structure between relevant covariates should intuitively be able to be leveraged to increase the power of statistical tests. In this sense, our work can be seen as a step towards combining the data- and model-based frameworks for statistical decisions, and is thus similar in spirit to work on robust machine learning.

The framework introduced herein is termed the Fair Optimization (FO) hierarchy, and draws intuition from the notion of bounding moments. After defining the FO hierarchy, this thesis will prove its consistency, showing that it does guarantee fair decisions asymptotically. The hierarchy will be examined from the perspective of supervised learning first, with specific results and intuitions presented for the sub-case of classification. The framework is also extended to the context of unsupervised learning, in particular to the problem of dimensionality reduction; as this is a novel and heretofore unstudied field, it requires extensions of the notion of fairness to this domain as well. Extensive experimental results are shown in both cases, with a focus on case studies in healthcare. The notion of fairness is also extended to hypothesis testing, and this is used to construct a new, covariance-robust dynamic wa-

termarking mechanism for Linear Time-Invariant (LTI) dynamical systems. Finally, similar intuitions are employed to design a system for robust classification in the low-data regime where data lives in low-dimensional manifolds.

Background

First, we elaborate on the specific background of how systematic bias can arise, how it can be quantified, and what algorithms exist for handling this problem.

Causes of Bias Bias can arise even without malicious intent. As mentioned earlier, human biases due to previous decisions made by humans will likely be reflected in data generated, causing algorithms using that data to exhibit the same bias. In some cases, critical predictors may be unextricably correlated with sensitive attributes, and in this case trade-offs must often be made between accuracy loss and the degree of bias that is to be tolerated, as well as which predictors will be allowed to cause bias and which will not. For example, educational attainment can be a biased indicator for hiring decisions, but it may be too critical to making effective decisions to ignore. However, bias can also arise due to more avoidable phenomenon; two of the most prominent of these are measurement error and sample bias [61]. The former refers to cases where the process of data collection itself is biased. For example, it has been observed that there are gender and racial gaps in representation in clinical trials and prescriptions of pain medication [81, 107], and that crimes committed by black and whites are investigated, recorded and prosecuted at different rates. In the presence of this form of bias, the trade-off between fairness and accuracy need not exist: In fact, the imposition of exogenously-known requirements of independence can improve accuracy while mitigating bias. Subgroup validity refers to predictors that have varying levels of predictive power across different subgroups [17]. For example, a predictor that may have a strong correlation with a desired outcome for men may be almost meaningless for women; especially if combined with a lack of data availability for women, this could lead to a case where the predictor is relied upon almost entirely, and would thus lead to vastly different error rates among women and men. Even if the predictor is meaningful for all subgroups, it could occur that the proper thresholding differs between groups.

Fairness Notions The first step in dealing with issues of automated bias is to be able to quantify bias. Of existing quantitative notions of bias, a handful have gained prominence in the literature. For this quantification, it is helpful to focus on the context of a classification problem, where the output is simply a binary variable $Y \in \{\pm 1\}$ and the decision is a binary function d of a set of covariates X and a protected attribute Z, as these outcomes can then be easily mapped to desirable and undesirable social outcomes. A number of measures have been proposed to deal with this issue, but the most straightforward of these is the notion of *disparate impact*, which simply measures the difference in the likelihood of attaining a positive classification across protected classes [86]. So, zero

disparate impact would imply perfect independence between $d(X, Z)$ and Z, since it requires that $P\left(d(X, Z) = +1|Z\right) = P\left(d(X, Z) = +1\right)$. However, as mentioned above, this can be overly detrimental to accuracy when Z is highly correlated with Y. In some cases, it may be more helpful to ensure that *error rates* are similar across protected classes, i.e. $P\left(d(X, Z) = +1|Y = y, Z\right) = P\left(d(X, Z) = +1|Y = y\right)$ for $y = \pm 1$. This is referred to as *equalized odds* [109] and requires that d not be more or less aggressive in over-classifying any one protected class in any direction. Notably, the perfect classifier $d(X, Z) = Y$ satisfies this perfectly, but it has been shown that any imperfect classifier cannot satisfy equalized odds while remaining calibrated (i.e. there exists some "score" function h such that d arises from a simple thresholding of h and that $P(d(X, Z) = Y) = h(X, Z)$), except for certain degenerate cases [131]. If the majority of costs are concentrated on only one type of error (i.e. false positives or false negatives), then one may restrict the requirement to the appropriate value of y, yielding what is termed *equal opportunity*. These are all notions of *group fairness*, meaning that they rely on population statistics for each protected class. In contrast, there is the notion of *individual fairness*, which essentially requires that similar individuals be treated similarly [77]. Using the notation above, this means that $d(X, Z)$ and $d(X, Z')$ should be similar, where $Z' \neq Z$.

Fairness Algorithms To date, much of the work in the fair classification literature has centered around pre- or post-processing steps. The former generally requires a transformation of the feature space [48, 192, 271], while the latter involves intentional alteration of biased predictions [109, 124]. While both benefit from high levels of flexibility (in particular the pre-processing approach for the potential to extend to the unsupervised-learning regime), they are necessarily greedy in nature and thus profligate in sacrificing accuracy [261]. The alternative to these methods is to enforce fairness at the time of training. For many classification and decision problems, most notably those that rely on some underlying margin or score function, training requires the minimization of an empirical loss, so enforcing fairness involves designing new loss functions, regularizers or constraints that fundamentally alter the optimization problem at hand. [47, 125, 191, 270] begin down this path. In particular, [270] propose bounding the correlation between a "score function" and protected attributes as a linear proxy to bounding fairness.

Still, independence constraints are not trivial to manage: Most recent work focusing on the decision problem of whether two distributions are independent relies on estimating the Kolmogorov-Smirnov (KS) distance or Mutual Information (MI) between the two distributions via either binning or kernel-density estimation (KDE) techniques, but both of these yield highly discontinuous or nonconvex problems that are not tractable in an optimization setting [48, 182]. There have been attempts in the fairness literature to address this difficulty and to design MI-oriented constraints for statistical estimation, but these suffer from limitations in the classes of distributions that they can represent, as well as from numerical instability due to logarithmic terms [125].

Philosophical and Legal Foundations of Fairness

The concepts of fairness and justice have long been studied by members of the academic community from any number of angles and perspectives. While this thesis is a technical contribution to the literature, the base problem that it seeks to address is, at heart, a philosophical question. Furthermore, the question of how to quantify fairness, which precedes the contributions of this thesis, is itself preceded by how to define justice. As such, even though the work presented herein focuses more on computational and statistical properties of designing fair decision-making algorithms, we find it valuable to provide a brief background of the philosophical and legal foundations of fairness.

Philosophical Foundations Debate has raged about the true definition of justice for as long as philosophical debates have been recorded. For the ancient Greeks, whose metaphysics revolved around a foundational belief in a cosmic sense of order and cyclical, harmonious fluctuations between conflicting universal forces, justice necessarily required adherence to this cosmic order, most clearly stated by pre-Socratics such as Parmenides and Heraclitus [63, 130, 216]. Famously, Plato's Socrates reflects this view when he states that "justice is doing one's own work and not meddling with what isn't one's own" in *The Republic* [36]; Aristotle goes further, setting justice and equity as actively conflicting notions (though he does admit that existing social distributions may not be just) [216]. In this way, both failed to provide any positive argument for a baseline platform, however scant, for universal human rights and value.

The Modern era redefined justice to be more responsive to actual human needs and interests. Thomas Hobbes defines justice as wholly artificial, a social construct in the most literal sense that it is defined solely by whatever social contract is constructed between a people and their government [112]. David Hume, a Scottish Enlightenment thinker succeeding Hobbes, does not endorse notions of a "social contract", but still claims public utility to be the main goal of justice and takes the lead of John Locke (whose work underpins much of the U.S. Constitution) in tying the application of justice to the protection of private property [118, 161]. The consequentialist view of justice is then taken to its Utilitarian extreme by Jeremy Bentham and John Stuart Mill, who claim that what is fair is solely what increases overall utility [175]. In response to the moral relativism of the empirical school of thought, Immanuel Kant proposes the "categorical imperative" as the single fundamental principle of duty, claiming that moral law "must carry with it absolute necessity" and treat all persons as "ends in themselves" [127]. Importantly, Kant creates the most extensive basis for an universal set of human rights and makes very clear that he views justice as objective, non-arbitrary and independent of intention or consequence, meaning that the pursuit of fairness can warrant marked deviations from normative observations. This principle of absolute equality is extended to the realm of socio-economic outcomes by Karl Marx [169]. Finally, John Rawls provides a modern interpretation of social contract theory, arguing that justice should be determined as if from behind a "veil of ignorance": He concludes that basic rights and civil freedoms are a primary concern and should be absolutely equally distributed, with

absolute equality of economic *opportunity* the second requirement of justice [208].

More recently, the morality of decision-making by agents without human consciousness has become a prominent topic of research among the philosophy and political science communities in its own right. A first observation from this field is that paradigmatic notions of *discrimination* lack efficaciousness when decisions are not made with a human state of intentionality [34]. Contemporary notions of discrimination largely reflect the belief that the locus of discrimination lies in the belief-structure and 'mental state' of an individual [12,152,224]. While similar arguments may be made when the ignorance or active negligence of decision-makers (algorithm designers) to possible disparities holds a similar moral weight to the case of a human making an intentionally biased decision, this paradigm of moral justice is largely exhausted by the plethora of cases where the bias resulting from automated decision-making techniques is not easily foreseen [82]. In fact, it has been argued that the true proprietors of culpability (under this framework) are those that made the distributed and biased decisions that originally yielded biased data. This argument does not help define a route of action for practitioners however.

More tangible concepts of fairness for our scenario have thus been investigated [34]. In one notion, the 'wrongness' of discrimination of an algorithm is explicitly associated with the degree to which the algorithm denies individuals their individuality; in effect, this metric states that group-based generalizations unfairly punish or reward individuals for the actions of others with whom they happen to share certain characteristics [155, 204]. To that end, it reflects the notion of individual fairness introduced above [77]. The application of this principle must be tempered since, in the extreme, it would effectively rule out any approach that relies on statistical learning as unjust. Critiques have instead argued that the critical factor to 'wrongness' is not necessarily all generalizations, but rather generalization mechanisms that are *insufficiently precise*, penalizing individuals for characteristics such as race and gender while giving credit for attributes such as job performance or educational attainment [155, 225]. This suggests, somewhat counter-intuitively, that larger models with more precise variables are inherently more fair, a notion that has developed further support in the algorithmic community [133].

Another major framework that can be more applicable to algorithmic decision-making is that of egalitarianism. In general, this is a teleological principle focused on the appropriate distribution of welfare, regardless of intentionality (and thus more amenable to notions of group fairness). The precise definitions of 'distribution' and 'welfare' have differed across interpretations of egalitarianism, with some famous examples being the 'maximin' resource distribution of John Rawls [78,208], a focus on preference-satisfaction [59], and the capability to achieve certain life goals [229]. Within this branch of moral philosophy, there still remain interesting questions. One question is that of defining "spheres of justice". Specifically, this pertains to deciding the contexts in which the goal of resource distributions should be absolute equality as opposed to harm reduction. While this debate is by no means settled, it has been argued that more blunt notions of absolute equality are more appropriate in cases of civil justice (such as the ability to vote or security check in airports) [110], while equal opportunity and related metrics, which are more responsive to differences in base rates, can

find more relevant applications in matters of 'economic justice' like job interviews [34].

Finally, questions remain in the egalitarianism literature about the role of choice and the responsibility that decision-makers have for inter-generational impacts of their decisions. The former relates to how to best design fairness metrics that do not punish individuals for the effects of luck or for choices that reflect necessity (i.e. foregoing a higher income or new house in order to take care of a loved one) [9, 12, 119], and would seem to suggest methods that limit the set of predictors that any algorithm can use. The latter reflects what is called a *deontic* sense of egalitarianism; that is, a stated concern with how an unequal state of affairs comes to be originally, as opposed to the inequality in its present form [197]. In this paradigm, any notion of fairness that relies on statistical facts is inherently incomplete, as these facts are themselves imbued with injustice and thus flawed benchmarks [155]. This spirit correlates with more aggressive use of assertive notions such as disparate impact which do not rely on existing statistical ratios and are thus not inherently flawed themselves. To that end, it would suggest that trade-offs between fairness and accuracy be handled by allowing 'wiggle-room' in fairness constraints as opposed to adopting wholly different, and purportedly flawed, fairness metrics such as equalized odds.

Legal Foundations Similar to the dominant philosophical paradigms of fairness prior to the proliferation of automation techniques, the primary legal doctrine in the constitutional law of the United States is dependent on a decision-maker's motivations. In *Washington v. Davis* (1976), the Supreme Court ruled that a written personnel test required for recruitment to the District of Columbia Police Department did not violate the Equal Protection Clause of the Fourteenth Amendment to the U.S. Constitution; it judged that the test did not have discriminatory intent, and thus did not automatically become a constitutional violation despite its racially-disproportionate impact. This original stance has adapted somewhat over the last few decades for certain types of discrimination and in certain areas. For example, it was ruled in *Fisher v. University of Texas* (2016) that race-conscious affirmative action programs for college admissions promote a governmental interest in promoting diversity and are thus permissible.

The majority of legal doctrine pertaining to statistical notions of disparate impact instead originate from federal statutes meant for specific areas of legal interest. While the legal usage of the term "disparate impact" originated in the Supreme Court case *Griggs v. Duke Power Co.* (1971), Title VII of the 1964 Civil Rights Act prohibits employers from discriminating against employees on the basis of sex, race, color, national origin and religion, while the Fair Housing Act of 1968 limits discriminatory practices related to housing [30]. In 1971, the Fair Employment Practice Commission (FEPC) of the State of California adopted the 80% rule for determining what comprised "disparate impact", and this was later adopted into the protocol for Title VII enforcement by the U.S. Equal Employment Opportunity Commission (EEOC) in 1978. This rule provides a strict technical cutoff for the level of discrimination that is permissible, and builds off of the base rates at which members of various protected groups receive beneficial treatment, similarly to the statistical concept of disparate impact.

1.2 Contributions

This thesis will propose the FO hierarchy, a hierarchical, training-time framework for fairness via bounding moments. Each level of the FO hierarchy consists of controlling a certain number of moments of data and outputs, with higher levels guaranteeing lower bias but at the cost of computational burden and possibly accuracy. This expands on some preliminary work on fair classification problems which is limited in scope and lacks rigorous analysis. These prior works make initial gestures towards, but fall far short of, suggesting how to encode full and verifiable independence of random variables into optimization structures. Thus, there is a need to extend and formalize moment-based techniques into a systematic mode of fair optimization. We take these to their logical conclusion, extending them to provably handle larger classes of protected variables, provide theoretical results and novel interpretations, and extend results to unsupervised learning and hypothesis testing regimes. The main contributions of this thesis are:

- Introduces and argues for the FO hierarchy as an approach to approximating independence in optimization and ensuring fairness in data-driven decision-making.
- Provides results on consistency and non-asymptotic rates of convergence for FO.
- Examines empirical behavior and theoretical intuitions of FO in multiple supervised learning problems, including dynamical systems and case studies on automated and fair morphine and heparin dosage.
- Defines a novel notion of fairness for unsupervised learning, and in particular dimensionality reduction problems, and extends the FO hierarchy to this setting.
- Provides the first analysis of fair hypothesis testing, and exploits these principles to design a distributionally-robust watermarking scheme for detecting attacks on dynamical systems.
- Extends the notion of data-dependent regularization to propose an average-margin regularizer for robust classification in low-data regimes for data that lives in low-dimensional manifolds.

1.3 Preliminaries

Here we define general notation for the manuscript. Further notation necessary to a single chapter will be further expounded in that specific chapter.

In this thesis, we use capital letters X, Y, Z to denote a set of data triples where X refers to a set of exogenous covariates, Y represents "target" variables that can be either directly learned or indirectly used to aid learning, and Z represents "protected" attributes with respect to which it is our goal to reduce or eliminate bias. The explicit dimensions of

these are outlined within each individual chapter, as they can alter depending on the context of their usage. We use $\mathbb{E}_n(\cdot)$ to denote expectation with respect to the empirical distribution. Recall this is the sample average of the random variable inside parenthesis. As examples, $\mathbb{E}_n(Z) = \frac{1}{n}\sum_{i=1}^n Z_i$ and $\mathbb{E}_n(ZX) = \frac{1}{n}\sum_{i=1}^n Z_i X_i$. With a few exceptions (in particular in cases concerning sequences of sets), sets are denoted using calligraphic type.

Let $M : \mathbb{R}^{dp} \to \mathbb{R}^{d\times p}$ to be the function that reshapes a vector into a matrix by placing elements into the matrix columnwise from the vector. Similarly, we define $W := M^{-1} : \mathbb{R}^{d\times p} \to \mathbb{R}^{dp}$ to be its inverse. Consider a tensor $\varphi \in \mathbb{R}^{r_1\times\cdots\times r_q}$, and let $[r] = \{1,\dots,r\}$. The norm $\|\varphi\|$ is the ℓ_∞ vector norm for the tensor considered as a vector. For two tensors $\varphi, \nu \in \mathbb{R}^{r_1\times\cdots\times r_q}$, we define their inner product $\langle\varphi,\nu\rangle$ to be the usual dot product for the tensors interpreted as vectors. We also use the symbol $\otimes$ to represent the tensor product.

For a tensor interpreted as a multilinear operator $\varphi(u_1,\dots,u_q)$, we define the two subordinate norms

$$\begin{aligned}\|\varphi\|_\circ &= \max\left\{\|\varphi(u,\dots,u)\| \;\middle|\; \|u\|_2 = 1\right\}\\ \|\varphi\|_* &= \max\left\{\|\varphi(u_1,\dots,u_q)\| \;\middle|\; \|u_k\|_2 = 1 \text{ for } k\in[q]\right\}\end{aligned} \tag{1.1}$$

where $\|\cdot\|_2$ is the Euclidean norm for vectors. These are subordinate norms since $\|\varphi(u,\dots,u)\| \le \|\varphi\|_\circ \big(\|u\|_2\big)^q$ and $\|\varphi(u_1,\dots,u_q)\| \le \|\varphi\|_* \prod_{k=1}^q \|u_k\|_2$. When $\varphi(\cdot,\dots,\cdot)$ is symmetric in its arguments, then $\|\varphi\|_\circ = \|\varphi\|_*$ [18, 37].

1.4 Outline

The FO hierarchy is outlined, and its statistical properties explored in Chapter 2. Chapter 3 then explores the use of FO specifically in supervised learning problems. This involves providing a series of visual interpretations of fairness and a number of interpretations of the constraints relevant to FO from the perspective of optimization, information theory and standard statistical learning techniques. Chapter 3 then rigorously tests the FO hierarchy applied to a number of supervised learning problems on a number of datasets, including two case studies on automated dosing algorithms. The notions of fairness, as well as some levels of the FO-hierarchy, are extended to the realm of unsupervised learning, and in particular dimensionality reduction, in Chapter 4, for which a number of empirical studies and a case study on insurance rate-setting are provided. Chapter 5 considers a slightly different problem, exploring the relationship between fairness and robustness in hypothesis testing problems, and using these intuitions to design a novel covariance-robust dynamic watermarking test for detecting attacks on Cyber-Physical Systems (CPS) and other dynamical systems. Finally, Chapter 6 extends the intuitions of data-dependent regularizers that are used throughout the previous chapters to design an Average-Margin (AM) regularization method for encouraging robustness of classifiers when dealing in low-data regimes and where data are high-dimensional in nature but are known to lie in low-dimensional manifolds. We endeavor to make each chapter self-contained to the degree possible; as a result of this, there

is inevitably some degree of repetition among chapters. Within these constraints, efforts were made to reduce repetition.

Chapter 2

Fair Optimization Framework

2.1 Introduction

There is growing concern that improperly designed data-driven approaches to decision-making may display biased or discriminatory behavior. In fact, such concerns are justified by numerous examples of unfair algorithms that have been deployed in the real world [11, 20, 79, 84]. In response, researchers have started to develop a number of approaches to encourage fairness in various statistical or machine learning problems [45, 57, 77, 109, 191, 270, 276]. The problem of classification has received particular attention due to the ease of mapping class labels to positive and negative outcomes with which to characterize fairness, but recent work has also begun to explore fair statistical methods in the context of unsupervised learning [56, 192] and in more general decision-analytic frameworks [83, 158].

Existing Approaches to Fairness

The literature on fair statistics and learning can be classified into three categories: pre-processing steps, post-processing steps, and training regularization. The general setup of these approaches is that they seek to estimate a model that predicts a dependent variable using a vector of independent variables, while trying to ensure that the model predictions are fair (we discuss quantitative measures of fairness in the next subsection) with respect to some variable that indicates a protected attribute (e.g., gender or race). Here we briefly review some of the existing approaches that have been developed for fairness.

Pre-processing approaches transform the data before estimation, to remove any protected information that could cause unfairness. For instance, [48, 271] take a nonparametric approach: They optimize over distributions to variationally transform the feature space. The nonparametric nature of their approach means that the optimization problem they design quickly becomes intractable. Alternatively, [192] take an adversarial outlook on pre-processing for fairness, and propose a semidefinite programming (SDP) formulation. Several groups have attempted to design autoencoders with a similar inspiration, although these are oriented around deep classifiers [27, 80, 163, 272]. However, pre-processing methods lead

to high generalization error when used before performing estimation, due to the theoretical difficulties associated with estimating high-dimensional densities [238].

In comparison, there is a smaller literature on post-processing for fairness. These methods take the output of a statistical technique, and process the output in order to improve fairness. A canonical example of this approach is [109], which designs a method for post-processing an arbitrary classifier in order to ensure fairness. While this method is flexible with regards to the type of classifier used, it achieves fairness by requiring different score function thresholds for different groups of protected classes. This violates a general principle called *individual fairness* [77], which says that similar individuals should be treated similarly. More significantly, [261] show that this method achieves suboptimal tradeoffs between accuracy and fairness.

Notably, both pre-processing and post-processing approaches are necessarily greedy since they unlink the process of estimation from ensuring fairness. This has motivated work on regularization approaches to fairness, which generally achieve lower generalization error while improving fairness. The regularization approaches most related to this chapter include [25, 191, 261, 270]. In particular, [270] control the correlation of a classifier score function and the protected attribute, which can be formulated as a linear constraint in the estimation problem. The method in [191] implements non-convex optimization techniques to further consider second-order deviations. However, a limitation of both is that they are applicable only when protected attributes are binary. The approach of [125,271] works for more general types of protected attributes, but it uses a heuristic approach to approximate an intractable optimization problem that includes a *mutual information* (MI) measure of fairness as a constraint. Alternatively, [100] design an iterative cutting-plane algorithm for fair support vector machine (SVM) that requires solving an SVM instance in each iteration. Moving away from classification, [46, 121] develop key concepts of fairness in the case of regression, and the work in [24] extends this to regularization techniques for ensuring different qualitative types of fairness in regression. Finally, a recent line of work has sought to generalize these ideas towards fair decision-making [83, 158].

Quantitative Measures of Fairness

We have so far casually used the terms fairness and bias without formally defining them. Part of the difficulty is a considerable lack of clarity in the existing literature as to their meaning, with different works defining different quantitative measures of fairness. We believe the underlying (and unifying) idea behind all these measures is they approximate in some way a measure of independence between the output of the statistical procedure and the variable of protected attributes. In fact, this way of thinking about fairness was first noticed by [125].

To make our discussion more concrete, we start by first discussing notions of fairness for binary classification with a binary protected attribute. Let $(X, Y, Z) \in \mathbb{R}^p \times \{\pm 1\} \times \{\pm 1\}$ be a jointly distributed random variable consisting of a vector of predictors, a binary class label, and a binary protected attribute. Let $\delta(x)$ be a score for a classifier, and suppose the classifier makes binary predictions $d(x, t) = \text{sign}(t - \delta(x))$ for a given threshold t of the score.

Since binary classifiers output a ± 1 that can be mapped to desirable/undesirable decisions, one measure of fairness is

$$KS = \max_{t \in \mathbb{R}} \left| \mathbb{P}\left[d(X,t) = +1 | Z = +1\right] - \mathbb{P}\left[d(X,t) = +1 | Z = -1\right] \right|. \tag{2.1}$$

For any fixed value of t, this is the correlation of the binary classifier and the protected attribute of interest. This quantitative measure of fairness is often called *disparate impact* [109, 191]. Effectively, disparate impact measures the total disparity in outcomes between protected classes.

This above measure of fairness can be too strict in some applications, as there may be unavoidable correlation between the classifier output and the protected label. For such cases, [109] proposes *equalized odds* as an alternative measure of fairness that instead constrains disparity in outcomes conditional on some informative variable. In the setting of binary classification, one possible informative variable is $Y \in \{\pm 1\}$ itself. This choice leads to the following quantitative measure of equalized odds fairness:

$$EO = \max_{y \in \{\pm 1\}} \max_{t \in \mathbb{R}} \left| \mathbb{P}\left[d(X,t) = +1 | Z = +1, Y = y\right] - \mathbb{P}\left[d(X,t) = +1 | Z = -1, Y = y\right] \right|. \tag{2.2}$$

Restated, the quantity (2.2) measures the disparity in *error rates* between the protected classes. An additional benefit is that a classifier with zero training error will also be fair with respect to this measure of fairness [109].

At an initial glance, the above measures of fairness do not look like manifestations of independence. Yet note the event $\{d(X,t) = +1\}$ is equivalent to the event $\{\delta(X) \leq t\}$ since $d(x,t) = \text{sign}(t - \delta(x))$. This means that (2.1) is the Kolmogorov-Smirnov (KS) distance between the distributions of $\delta(X)|Z = +1$ and $\delta(X)|Z = -1$. Since (2.2) has a very similar interpretation, we will focus our discussion on (2.1). Thus when $KS = 0$ in (2.1), we have that

$$G(t) := \mathbb{P}\left[\delta(X) \leq t | Z = +1\right] = \mathbb{P}\left[\delta(X) \leq t | Z = -1\right]. \tag{2.3}$$

This means that the joint distribution factorizes as

$$\mathbb{P}\left(\delta(X) \leq t, Z = z\right) = \mathbb{P}\left[\delta(X) \leq t | Z = z\right] \cdot \mathbb{P}(Z = z) = G(t) \cdot \mathbb{P}(Z = z), \tag{2.4}$$

which means the two random variables are independent. Summarizing, we have $KS = 0$ in (2.1) if and only if $\delta(X)$ is independent of Z. The importance of such independence in relation to fairness was first noticed by [125]. Further interpretations of these are considered in Chapter 3.

Technical Challenges with Independence

The above discussion suggests that a promising direction for generalizing fairness to a broader class of problems is to ensure independence (or rather some approximate notion of independence) between the output of a statistical technique and a random variable that measures

attributes in respect to which fairness is desired. In fact, the broader idea of quantifying independence using an empirical estimate has a long history in statistics [43,53,90,195,241]. One approach is to compute some generalized notion of correlation such as Renyi correlation or distance correlation. Another approach is to use some distance like the KS distance, total variation distance, or mutual information between the empirical probability measures of the joint and product distributions.

However, incorporating empirical independence measures into statistical procedures is not straightforward. Many statistical procedures are computed by solving an optimization problem, and so such measures must be added as constraints. However, measures like Renyi correlation, distance correlation, KS distance, total variation distance, and mutual information are all themselves the solutions of an optimization problem. (Mutual information is traditionally defined using a hard-to-compute integral, but a well-known variational characterization [41] shows that it should more properly be thought of as the solution to an optimization problem for our discussion.) This means the resulting optimization problem for a fair statistical procedure defined in this way would have another optimization problem as a constraint; these types of problems are known as bilevel programs and are very difficult to numerically solve [71,194]. The numerical difficulties are compounded for those measures defined using an empirical c.d.f., which is always discontinuous.

Contributions and Outline

This chapter develops an optimization hierarchy for fair statistical decision problems. We first generalize in Section 2.3 the framework of statistical decision problems [151] to include fairness. This provides a systematic approach for developing and studying fair versions of hypothesis testing, decision-making, estimation, regression, and classification. We use the above discussed insight relating fairness to statistical independence in order to propose in Section 2.4 an optimization hierarchy that lends itself to numerical computation. Tools from variational analysis and random set theory are used to prove in Section 2.5 that higher levels of this hierarchy lead to consistency in the sense that it asymptotically imposes independence as a constraint in corresponding statistical decision problems. Section 2.6 generalizes our framework to measure fairness using a notion of approximate independence. Specific instances of our framework, ranging from fair supervised learning to fair unsupervised learning and fair hypothesis testing, are outlined in more detail in subsequent chapters, and their efficacy proven on a number of datasets, including case studies on fair morphine dosage, heparin scheduling, and insurance rate-setting.

The distinguishing feature of our approach to ensuring independence is to use a moment-based characterization of independence that generalizes Kac's theorem [35,122] to multivariate random variables. This has the key practical benefit over other approaches to measuring independence (such as [125,271]) that all the resulting constraints in the corresponding optimization problems are smooth polynomials. This means we avoid the bilevel programming structure that arises from the use of other independence measures [125,271], and which makes numerical optimization very difficult. Because the moment constraints are smooth polyno-

mials, this further allows us to leverage advances in convex optimization [147] and related heuristics such as the constrained convex-concave procedure [235, 249, 268] for the purpose of numerically solving the resulting optimization problem. The tradeoff is that we have to include multiple (but a finite number of) constraints, one for each possible combination of moments between joint and product distributions.

Our framework also builds on preliminary work on the use of moment-based constraints for fair statistical methods [191, 192, 270]. These approaches were restricted to binary classification with binary protected classes, made use of only first- or second-order moments of only the classifier, were based on ad-hoc arguments and justifications, and lacked theoretical analysis of the resulting statistical methods. The past papers [191, 192, 270] leave open the larger question of how moment-based approaches to fairness can be generalized to continuous protected classes, multivariate protected classes, multivariate statistical decisions, and other classes of statistical problems beyond classification. This thesis unifies these past approaches into a broader theoretical framework, provides a rigorous theoretical analysis of the resulting optimization hierarchy, and successfully achieves a generalization of moment-based methods in order to handle continuous protected classes, multivariate protected classes, multivariate statistical decisions, and multiple classes of statistical decision problems, including fair versions of hypothesis testing, decision-making, estimation, regression, and classification.

Because we have to include multiple constraints, this significantly complicates the theoretical analysis of our optimization hierarchy. The limiting behavior of our framework requires a statistical analysis on the solution to an optimization problem in the limit of a countably-infinite number of random constraints involving empirical moments. Traditional results in statistics do not apply to set-valued functions [14], which are one way to interpret constraints in an optimization problem [212]. In fact, most attention in statistics on sets has been focused on estimating a single set under different measurement models [75, 106, 135, 198, 226]. The traditional theoretical argument is to use the Pompeiu-Hausdorff distance to metricize the set of sets, but this approach is intractable in our setting which has random sets defined using (in the limit) an infinite number of non-convex constraints. Instead, we build on our past work on statistics with set-valued functions [14]: We develop new theoretical arguments for statistics with random sets and set-valued functions, using variational analysis [212, 213] and random sets [173, 183].

2.2 Preliminaries

This section presents notation specific to this chapter, in addition to that provided Section 1.3. We also describe some useful (and needed) notation and definitions from variational analysis and random sets. Most of the variational analysis definitions are from [212], and the stochastic set convergence notation is originally from [14].

Variational Analysis

Let $\overline{\mathbb{R}} = [-\infty, \infty]$ denote the extended real line. We define $\Gamma(\cdot, \mathcal{S}) : E \to \overline{\mathbb{R}}$ to be the indicator function

$$\Gamma(u, \mathcal{S}) = \begin{cases} 0, & \text{if } u \in \mathcal{S} \\ +\infty, & \text{otherwise} \end{cases} \tag{2.5}$$

where E is some Euclidean space that will be clear from the context.

The outer limit of the sequence of sets C_n is defined as

$$\limsup_n C_n = \{x : \exists n_k \text{ s.t. } x_{n_k} \to x \text{ with } x_{n_k} \in C_{n_k}\}, \tag{2.6}$$

and the inner limit of the sequence of sets C_n is defined as

$$\liminf_n C_n = \{x : \exists x_n \to x \text{ with } x_n \in C_n\}. \tag{2.7}$$

The outer limit consists of all the cluster points of C_n, whereas the inner limit consists of all limit points of C_n. The limit of the sequence of sets C_n exists if the outer and inner limits are equal, and when it exists we use the notation that $\lim_n C_n := \limsup_n C_n = \liminf_n C_n$.

A sequence of extended-real-valued functions $f_n : \mathcal{X} \to \overline{\mathbb{R}}$ is said to epi-converge to f if at each $x \in \mathcal{X}$ we have

$$\begin{cases} \liminf_n f_n(x_n) \geq f(x) & \text{for every sequence } x_n \to x \\ \limsup_n f_n(x_n) \leq f(x) & \text{for some sequence } x_n \to x \end{cases} \tag{2.8}$$

Epi-convergence is so-named because it is equivalent to set convergence of the epigraphs of f_n, meaning that epi-convergence is equivalent to the condition $\lim_n \{(x, \alpha) \in \mathcal{X} \times \mathbb{R} : f_n(x) \leq \alpha\} = \{(x, \alpha) \in \mathcal{X} \times \mathbb{R} : f(x) \leq \alpha\}$. We use the notation $\text{e-lim}_n f_n = f$ to denote epi-convergence relative to $\mathcal{X}$.

A sequence of extended-real-valued functions $f_n : \mathcal{X} \to \overline{\mathbb{R}}$ is said to converge pointwise to f if at each $x \in \mathcal{X}$ we have that $\lim_n f_n(x) = f(x)$. We abbreviate pointwise convergence relative to $\mathcal{X}$ using the notation $\lim_n f_n = f$.

Random Sets

Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a complete probability space, where Ω is the sample space, $\mathfrak{F}$ is the set of events, and $\mathbb{P}$ is the probability measure. A map $S : \Omega \to \mathcal{F}$ is a random set if $\{\omega : S(\omega) \in \mathcal{X}\} \in \mathfrak{F}$ for each $\mathcal{X}$ in the Borel σ-algebra on $\mathcal{F}$ [183]. Like the usual convention for random variables, we notationally drop the argument for a random set.

When discussing stochastic convergence of random sets, we denote that a type of limit occurs almost surely by appending "as-" to the limit notation. For instance, notation $\text{as-}\limsup_n C_n \subseteq C$ denotes $\mathbb{P}(\limsup_n C_n \subseteq C) = 1$, and notation $\text{as-}\liminf_n C_n \supseteq C$ denotes $\mathbb{P}(\liminf_n C_n \supseteq C) = 1$.

2.3 Fair Statistical Decision Problems

We use the setting of statistical decision problems: Consider the random variables (X, Y, Z) that have a joint distribution $\mathcal{D}$. The interpretation is that X gives descriptive information, Y has information about some target, and Z encodes protected information which we would like to be fair with respect to. We will not explicitly use Y in this chapter, but we note that it is implicitly included within other terms that we discuss. The role (or lack of a role) of Y will be more specifically handled in subsequent chapters that separately handle supervised and unsupervised learning.

The goal is to construct a function $\delta(\cdot, \cdot)$ called a *decision rule*, which provides a decision $d = \delta(x, z)$. To evaluate the quality of a decision rule δ, we define a *risk function* $R(\delta)$. (Though it is conventional to define the risk as $R(\mathcal{D}, \delta)$, we assume without loss of generality that the risk is of the form $R(\delta)$ because when the risk is $R(\mathcal{D}, \delta)$ then the proper choice of $R(\delta)$ recovers the Bayes $R(\delta) = \mathbb{E}_{\mathcal{D}} R(\mathcal{D}, \delta)$ and minimax $R(\delta) = \max_{\mathcal{D} \in \Omega} R(\mathcal{D}, \delta)$ procedures.) In this setup, an optimal decision rule is taken to be any function from $\arg\min_{\delta(\cdot,\cdot)} R(\delta)$. However, we can define a related optimization problem that chooses an optimal fair decision rule by solving

$$\delta^*(x, z) \in \arg\min_{\delta(\cdot,\cdot)} \left\{ R(\delta) \;\middle|\; \delta(X, Z) \perp\!\!\!\perp Z \right\}, \tag{2.9}$$

where the notation $\delta(X, Z) \perp\!\!\!\perp Z$ indicates independence of $\delta(X, Z)$ and Z.

The above abstract setup is useful because it allows us to reason about fairness for a wide class of problems using a single theoretical framework. A question that may arise is why the decision function δ may be allowed to be a function of the protected attribute Z as well as the covariates X; indeed, this seems counterintuitive to the central goal that $\delta(X, Z)$ be *independent* of Z. However, there is a growing literature that posits the necessity of "disparate treatment to offset disparate impact" [62, 95, 129, 132, 134, 156]. To see why this may be the case,

Example 1. Consider the simple problem with binary X, Y, Z displayed in table 2.1. Suppose a simple classification task on these variables where we want to recover the target Y. X is slightly predictive of Y, so one possible choice of the decision function δ that does not rely on Z is the following

$$\delta_1(X, Z) = X \longrightarrow \begin{cases} P(Y = \delta_1(X, Z)) = \frac{11}{18} \\ P(\delta_1(X, Z) = 1 | Z = A) = \frac{2}{3} \\ P(\delta_1(X, Z) = 1 | Z = B) = \frac{1}{3}. \end{cases}$$

Note that the base rates of Y for each of the protected class are $P(Y = 1 | Z = A) = \frac{5}{9}$ and $P(Y = 1 | Z = B) = \frac{5}{9}$, while those of X are $P(X = 1 | Z = A) = \frac{5}{9}$ and $P(X = 1 | Z = B) = \frac{1}{3}$. While δ_1 is informative, it relies on a covariate that is a more powerful predictor for protected class A than for protected class B, and thus δ_1 perpetuates *more* bias than is truly warranted in the final decision. Now, consider a new decision function that *does* use information about the protected class.

Table 2.1: A simple example of one-dimensional, binary X, Y, Z, with the associated measure $\mathcal{D}$.

X	Y	Z	$\mathcal{D}(X,Y,Z)$
1	1	A	5/18
1	1	B	1/18
1	0	A	1/18
1	0	B	2/18
0	1	A	0/18
0	1	B	4/18
0	0	A	3/18
0	0	B	2/18

$$\delta_2(X,Z) = \max\{X, V \cdot \mathbf{1}_B\{Z\}\} \longrightarrow \begin{cases} P(Y = \delta_2(X,Z)) = \frac{2}{3} \\ P(\delta_2(X,Z) = 1 | Z = A) = \frac{2}{3} \\ P(\delta_2(X,Z) = 1 | Z = B) = \frac{2}{3}, \end{cases}$$

where V is a Bernoulli random variable with parameter 0.5, and $\mathbf{1}_B$ is the set-indicator function that takes value 1 when Z is B and 0 otherwise. In other words, there is deliberate difference in treatment between the protected classes, which members of class B are given a "second chance" to obtain classification 1. Interestingly, δ_2 here is an even more powerful predictor of the true target Y than δ_1. Furthermore, it is able to perfectly balance the per-class "opportunity" and get much closer to the true base rate of Y among each of the protected classes. The reason for this is that the covariate X is a good predictor of Y among the members of group A, but is not effective among members of group B; this exhibits one key source of bias, intentional or unintentional, in many real-world scenarios as well.

One may ask if there are any costs to including protected attribute Z in the decision function. In scenarios with discrete protected attributes such as this one, this can be described as designing distinct classification rules for each protected class. In this sense, it is possible to violate the principal of *individual fairness*, meaning that individuals from different protected classes that are nevertheless similar in all other ways will be treated differently [77]. This type of fairness may be an end in and of itself, particularly in cases dealing with humans, where setting different standards for different people may also meet with opposition.

This generality of our method and the effect of including the protected attribute in the decision rule are further demonstrated by the following (which is the first to our knowledge) example of a procedure for performing fair hypothesis testing:

Example 2. Consider a hypothesis testing setup where the null hypothesis is $H_0 : \mathbb{E}(\Xi) = 0$ for the underlying distribution

$$\begin{bmatrix}\Xi\\ \Psi\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right). \tag{2.10}$$

Suppose $X = (\Xi_1, \ldots, \Xi_n)$ and $Z = (\Psi_1, \ldots, \Psi_n)$ consist of i.i.d. samples. Let d_0 be the decision to accept the null, and let d_1 be the decision to reject the null. The traditional hypothesis test with a significance level of α corresponds to a decision rule δ that minimizes the risk function

$$R(\delta) = \mathbb{P}_{H_1}(\delta = d_0) + \Gamma(\mathbb{P}_{H_0}(\delta = d_1) - \alpha, \mathbb{R}_{\leq 0}), \tag{2.11}$$

where $H_1 = \{\mathcal{D} \in \Omega : \mathcal{D} \neq H_0\}$ [151]. An optimal decision rule for this risk is

$$\delta^* = \begin{cases} d_0, & \text{if } p \geq \alpha \\ d_1, & \text{if } p < \alpha \end{cases} \tag{2.12}$$

where p is a p-value [151]. An optimal decision rule that depends only upon X corresponds to the use of a traditional p-value

$$p = 2\Phi\Big(-\sqrt{n}\big|\tfrac{1}{n}\textstyle\sum_{i=1}^n \Xi_i\big|\Big), \tag{2.13}$$

with $\Phi(\cdot)$ being the standard normal cdf. Using the above framework, we can also compute an optimal *fair* decision rule for this risk by removing the component of Ξ that correlates with Ψ. This corresponds to

$$p = 2\Phi\Big(-\sqrt{\tfrac{n}{1-\rho^2}}\big|\tfrac{1}{n}\textstyle\sum_{i=1}^n (\Xi_i - \rho\Psi_i)\big|\Big), \tag{2.14}$$

which we can interpret as a *fair* p-value (note that $\Xi - \rho\Psi$ is independent of Ψ). An interesting observation about this setup is that using (2.14) results in a test with greater *power* than using (2.13). This is interesting because it shows that imposing fairness constraints can lead to better statistical procedures in certain contexts.

In many statistical contexts, Ω is singleton but unknown. We then instead choose the decision rule using a sample (X_i, Y_i, Z_i) for $i = 1, \ldots, n$, which is i.i.d. from the distribution $\mathcal{D}$. Towards this aim, we approximate the risk function $R(\delta)$ using an (random) approximate risk function $R_n(\delta)$ that depends upon the sample. However, computing a sample-based fair decision rule is not obvious because a statistically well-behaved, sample-based analog of the constraint $\delta(X, Z) \perp\!\!\!\perp Z$ from (2.9) has not been studied previously.

2.4 Fair Optimization Hierarchy

We next propose a framework for computing a fair decision rule by solving a sample-based analog of (2.9). We first describe our assumptions about the statistical and numerical properties of the problem. Next we present our framework and provide some intuition to justify the

structure of our formulation. We conclude by discussing some of the favorable computational properties of our framework.

Assumptions

We first make some assumptions about restrictions on our decision rule and about the random variables in question:

Assumption 1. The decision rule belongs to a parametric polynomial family and can be written as

$$\delta(x, z) = B \cdot \omega(x, z), \tag{2.15}$$

where $B \in \mathcal{B}$ is a matrix, $\mathcal{B} \subset \mathbb{R}^{d\times p}$ is a compact set, and $\omega(x, z) \in \mathbb{R}^p$ is a vector of monomials of the entries of the vectors x, z. More precisely, B parametrizes the decision rule $\delta(x, z)$, and the function $\omega(x, z)$ is assumed to be known and fixed by our design such as through feature engineering. We define the random variable $\Omega = \omega(X, Z)$, so that $\delta(X, Z) = B\Omega$.

Remark 1. In some settings, it may be desirable to have the fair decision rule depend upon only X and not Z. The above includes this case by noting $\omega(x, z)$ is free to be chosen to include only monomials of the entries of x.

Assumption 2. The entries of the random variables X, Z are almost surely bounded by $\alpha \geq 1$. Moreover, the maximal monomial degree of entries in $\omega(x, z)$ is $\rho \geq 0$, and the random variable Z has dimensions $Z \in \mathbb{R}^r$.

Assumption 3. Recall $B \in \mathcal{B}$ is a matrix that parametrizes the decision rule, for compact $\mathcal{B} \subset \mathbb{R}^{d\times p}$. We assume $\mathcal{B} \subseteq \{B \in \mathbb{R}^{d\times p} : \|W(B)\|_2 \leq \sqrt{\lambda}\}$.

Our next assumption is about statistical properties of the approximate risk function. Since our primary interest in this thesis is studying independence constraints, we directly make assumptions about the convergence of the approximate risk function. Showing that such convergence holds typically involves a separate statistical analysis specific to the problem at hand.

Assumption 4. Note the function $R_n(B \cdot \omega(x, z))$ is the approximate risk function composed with the parametric decision rule in Assumption 1. We assume that this function can be written in the form

$$h_n(B) := R_n(B \cdot \omega(x, z)) = f_n(B) + \Gamma(g_n(B), \{\mathbb{R}_{\leq 0}\}^\eta), \tag{2.16}$$

where $f_n : \mathbb{R}^{d\times p} \to \mathbb{R}$ and $g_n : \mathbb{R}^{d\times p} \to \mathbb{R}^\eta$. Moreover, define the notation $h(B) = R(B \cdot \omega(x, z))$. We assume as-e-lim h_n = as-lim $h_n = h$ relative to $\mathcal{B}$.

Remark 2. We should interpret the notation of (2.16) as simultaneously specifying an objective function $f_n(B)$ and a set of constraints $g_n(B) \leq 0$.

Remark 3. This convergence assumption may look unfamiliar, but we note that it is weaker than the convergence results that are usually shown when proving consistency of estimators. In particular, almost sure uniform convergence of h_n to h implies the above assumption.

The first four assumptions are primarily related to statistical properties, though the polynomial structure of the decision rule is also related to numerical computation. Our last assumption is about the mathematical structure of the approximate risk function, and it is related to numerical computation.

Assumption 5. In the notation of Assumption 4, we assume that the functions $f_n : \mathbb{R}^{d\times p} \to \mathbb{R}$ and $g_n : \mathbb{R}^{d\times p} \to \mathbb{R}^\eta$ are polynomials on the set $\mathcal{B}$. We also assume that h is a lower semicontinuous function on the set $\mathcal{B}$.

Remark 4. The polynomial assumption is not restrictive because the celebrated Stone-Weierstrass theorem shows that if f_n and g_n are continuous then they can be approximated to arbitrary accuracy by polynomials, since the domain of the optimization problem is within a compact set $\mathcal{B}$.

Remark 5. We note that of these, only Assumption 2 is truly outside of the control of the practitioner of our methodology, and would thus be understood as an "assumption" in the most typical sense. The rest of these assumptions can largely be interpreted as restrictions on the decision rules and risk functions used, although they still must be assumed for the purposes of the statistical analysis that we conduct.

Formulation

We are now ready to present our framework. Given the above assumptions, we study use of the following sample-based optimal fair decision rule: The level-$(\mathfrak{g}, \mathfrak{h})$ fair optimization (FO) is

$$\begin{aligned} &\min_{B\in\mathcal{B}} R_n(B\cdot\omega(x,z)) \\ &\text{s.t. } \left\|\mathbb{E}_n\big(Z^{\otimes m}\otimes(B\Omega)^{\otimes q}\big) - \mathbb{E}_n\big(Z^{\otimes m}\big)\otimes\mathbb{E}_n\big((B\Omega)^{\otimes q}\big)\right\| \le \Delta_{m,q}, \\ &\qquad\qquad \text{for } (m,q)\in[\mathfrak{g}]\times[\mathfrak{h}]. \end{aligned} \tag{2.17}$$

The hyperparameters $\mathfrak{g}$ and $\mathfrak{h}$ here reflect the moments that are to be controlled (with higher values implying more moments being controlled), and the $\Delta_{m,q}$ terms are maximal permitted deviations for the appropriate moment bounds. We will study the constraints of the above problem and show that they are statistically well-behaved analogs of the independence constraint in (2.9).

Our first result provides intuition about the constraints in the FO optimization problem (2.17). This result generalizes Kac's theorem [35, 122], which characterizes independence of random variables using moment conditions, to the setting of random vectors. This generalization is novel to the best of our knowledge, and so we include its proof below for the sake of completeness.

Theorem 1. *Suppose the multivariate random variables $U \in \mathbb{R}^p$ and $V \in \mathbb{R}^d$ are bounded. Then U and V are independent if and only if*

$$\mathbb{E}\big(U^{\otimes m} V^{\otimes q}\big) = \mathbb{E}\big(U^{\otimes m}\big) \otimes \mathbb{E}\big(V^{\otimes q}\big) \text{ for } m, q \geq 1. \tag{2.18}$$

Proof. Let the functions $M_U(s) = \mathbb{E}\exp(\langle s, U\rangle)$, $M_V(t) = \mathbb{E}\exp(\langle t, V\rangle)$, and $M_{(U,V)}(s,t) = \mathbb{E}(\exp(\langle s, U\rangle + \langle t, V\rangle)$ be the moment generating functions for U, V, and (U,V), respectively. Observe these are defined for all s, t since U, V are bounded. Our proof begins with the well-known characterization of independence using moment generating functions, that is U and V are independent if and only if $M_{(U,V)}(s,t) = M_U(s)M_V(t)$. In particular, if (2.18) holds then we have

$$\begin{aligned}
M_{(U,V)}(s,t) &= \textstyle\sum_{m=1}^{\infty}\sum_{q=1}^{\infty} \frac{1}{m!\cdot q!} \cdot \mathbb{E}\big(\langle s, U\rangle^m \langle t, V\rangle^q\big) \\
&= \textstyle\sum_{m=1}^{\infty}\sum_{q=1}^{\infty} \frac{1}{m!\cdot q!} \cdot \langle \mathbb{E}\big(U^{\otimes m} V^{\otimes q}\big), s^{\otimes m} t^{\otimes q}\rangle \\
&= \textstyle\sum_{m=1}^{\infty}\sum_{q=1}^{\infty} \frac{1}{m!\cdot q!} \cdot \langle \mathbb{E}\big(U^{\otimes m}\big) \otimes \mathbb{E}\big(V^{\otimes q}\big), s^{\otimes m} t^{\otimes q}\rangle \\
&= \textstyle\sum_{m=1}^{\infty}\sum_{q=1}^{\infty} \frac{1}{m!\cdot q!} \cdot \langle \mathbb{E}\big(U^{\otimes m}\big), s^{\otimes m}\rangle \cdot \langle \mathbb{E}\big(V^{\otimes q}\big), t^{\otimes q}\rangle \\
&= \textstyle\sum_{m=1}^{\infty}\sum_{q=1}^{\infty} \frac{1}{m!\cdot q!} \cdot \mathbb{E}\big(\langle s, U\rangle^m\big) \cdot \mathbb{E}\big(\langle t, V\rangle^q\big) \\
&= \textstyle\sum_{m=1}^{\infty} \frac{1}{m!} \cdot \mathbb{E}\big(\langle s, U\rangle^m\big) \cdot \sum_{q=1}^{\infty} \frac{1}{q!} \cdot \big(\mathbb{E}\langle t, V\rangle^q\big) \\
&= M_U(s) M_V(t)
\end{aligned} \tag{2.19}$$

This proves the reverse direction. To prove the forward direction, we note it follows immediately by applying componentwise the standard result that if U and V are independent and bounded, then $\mathbb{E}(g(U)h(V)) = \mathbb{E}(g(U)) \cdot \mathbb{E}(h(V))$ for any continuous functions $g : \mathbb{R}^p \to \mathbb{R}$ and $h : \mathbb{R}^d \to \mathbb{R}$. □

The above multivariate generalization of Kac's theorem allows us to interpret the constraints of the FO problem (2.17). We can interpret the constraints as a finite number ($\mathfrak{g} \cdot \mathfrak{h}$ many, for a level-$(\mathfrak{g}, \mathfrak{h})$ FO problem) of sample-based analogs of the corresponding moment conditions for independence (2.18).

Computational Properties

We next discuss some favorable computational properties of the FO problem (2.17). A key advantage of our framework is that the constraints are polynomials, and so we can leverage significant numerical and theoretical advances in order to solve such problems.

Theorem 2 (Theorem 5.6 and 5.7 of [147])**.** *If Assumptions 1–5 hold, then the level-$(\mathfrak{g}, \mathfrak{h})$ FO problem (2.17) can be solved to any desired accuracy by solving a convex optimization problem that can be explicitly constructed.*

Remark 6. Though the convex optimization problems resulting from the explicit construction of [147] are often large, these resulting optimization problems can be numerically solved for many interesting instances [165, 274].

We can say more about the FO problem for specific levels of the hierarchy, and we omit the proofs since they follow from the definition of the constraint:

Proposition 1. *The constraints in the FO problem (2.17) for $q = 1$ can be written as the following linear inequality constraints:*

$$\begin{aligned} B\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n \Omega_i \otimes (Z_i)^{\otimes m} - \tfrac{1}{n}\sum_{i=1}^n \Omega_i \otimes \tfrac{1}{n}\sum_{i=1}^n (Z_i)^{\otimes m}\right) \leq & \Delta_{m,1} \\ -B\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n \Omega_i \otimes (Z_i)^{\otimes m} - \tfrac{1}{n}\sum_{i=1}^n \Omega_i \otimes \tfrac{1}{n}\sum_{i=1}^n (Z_i)^{\otimes m}\right) \leq & \Delta_{m,1} \end{aligned} \tag{2.20}$$

where the inequality should be interpreted as being elementwise of the left (which is a tensor) with respect to the scalar $\Delta_{m,1}$ on the right.

This results says constraints with $q = 1$ are always convex. This means that the FO problem (2.17) with $\mathfrak{h} = 1$ is a convex optimization problem whenever R_n is convex in B. Such convexity of R_n occurs in many interesting problems, including linear regression and support vector machines.

Proposition 2. *The constraints in the FO problem (2.17) for $q = 2$ are inequalities that each involve a difference of two convex quadratic functions.*

This results says constraints with $q = 2$ are always a difference of convex functions. This means that stationary points of the FO problem (2.17) with $\mathfrak{h} = 2$ can be found using the effective constrained convex-concave procedure [235, 249, 268] whenever R_n is convex in B. Recall that R_n is convex in many interesting problems like linear regression and support vector machines.

Proposition 3. *If Z is a binary random variable, which is coded as either $Z \in \{0, 1\}$ or $Z \in \{\pm 1\}$, then the constraints in the FO problem (2.17) for $m \geq 2$ are redundant with the corresponding constraint for $m = 1$.*

This result says that when Z is binary, then the hierarchy simplifies and we only need to consider applying the level-$(1, \mathfrak{h})$ FO problems. We will use this simplification when conducting numerical experiments in Chapters 3 and 4.

2.5 Statistical Consistency of FO Hierarchy

We prove in this section that the sample-based constraints of the FO problem (2.17) are in fact statistically well-behaved analogs of the independence constraint in (2.9).

Concentration of Tensor Moment Estimates

We begin by defining several multilinear operators. We define the empirical operators

$$\begin{aligned}\widehat{\varphi}_{m,q}(B_1,\dots,B_q) &= \mathbb{E}_n\big(Z^{\otimes m}\bigotimes_{k=1}^q (B_k\Omega)\big)\\ \widehat{\nu}_{m,q}(B_1,\dots,B_q) &= \mathbb{E}_n\big(Z^{\otimes m}\big)\otimes \mathbb{E}_n\big(\bigotimes_{k=1}^q (B_k\Omega)\big)\end{aligned} \tag{2.21}$$

and the expected operators

$$\begin{aligned}\varphi_{m,q}(B_1,\dots,B_q) &= \mathbb{E}\big(Z^{\otimes m}\bigotimes_{k=1}^q (B_k\Omega)\big)\\ \nu_{m,q}(B_1,\dots,B_q) &= \mathbb{E}\big(Z^{\otimes m}\big)\otimes \mathbb{E}\big(\bigotimes_{k=1}^q (B_k\Omega)\big)\end{aligned} \tag{2.22}$$

As a slight simplification of notation, when the argument of these multilinear operators is (B) we take that to mean the argument is $(B,\dots,B)$. We can thus identify these operators with terms in the FO problem (2.17): The $\widehat{\varphi}_{m,q}(B)$ and $\widehat{\nu}_{m,q}(B)$ are precisely the terms appearing in the constraints.

Proposition 4. *If Assumptions 1 and 2 hold, then we have*

$$\mathbb{P}\big(\|\widehat{\varphi}_{m,q}-\varphi_{m,q}\|_\circ > \mathcal{R}_{m,q}[n]+\gamma\big) \le 2\exp\big(-\tfrac{n\gamma^2}{64p^q\alpha^{2m+2\rho q}}\big) \tag{2.23}$$

for $\mathcal{R}_{m,q}[n] = 8\alpha^{m+\rho q}p^{q/2}\sqrt{\frac{dp\log(1+4q)+m\log r+q\log d}{n}}$.

Proof. We use a chaining argument. Suppose $\{t_i\}_{i=1}^N$ is a $\frac{1}{2q}$ covering of $\mathbb{S}^{dp-1}$, and note $N\le(1+4q)^{dp}$ by the volume ratio bound [254]. Define $T_i = M(t_i)\in\mathbb{R}^{d\times p}$. Let $\mathcal{P}_q$ be the set of all permutations of $[q]$, and let

$$\Phi(B_1,\dots,B_q) = \tfrac{1}{q!}\textstyle\sum_{\pi\in\mathcal{P}_q}\big(\widehat{\varphi}_{m,q}(B_{\pi_1},\dots,B_{\pi_q})-\varphi_{m,q}(B_{\pi_1},\dots,B_{\pi_q})\big). \tag{2.24}$$

Observe that by construction: $\Phi(\cdot,\dots,\cdot)$ is symmetric, and it satisfies the identity $\Phi(B)=\widehat{\varphi}_{m,q}(B)-\varphi_{m,q}(B)$. Now consider the telescoping sum

$$\Phi(B) = \Phi(T_i) + \textstyle\sum_{k=1}^q \Phi(\overbrace{B,\dots,B}^{q-k}, B-T_i, \overbrace{T_i,\dots,T_i}^{k-1}). \tag{2.25}$$

Recall $\|W(T_i)\|_2=1$ and $\|W(B-T_i)\|_2\le\frac{1}{2q}$ for $W(B)\in\mathbb{S}^{dp-1}$. Since $\|\cdot\|_*$ is a subordinate norm, we have $\|\Phi\|_\circ \le \|\Phi(T_i)\| + \sum_{k=1}^q \frac{1}{2q}\|\Phi\|_*$. But note that $\Phi(\cdot,\dots,\cdot)$ is symmetric, and so $\|\Phi\|_\circ = \|\Phi\|_*$ [18, 37]. Thus we have $\|\Phi\|_\circ \le 2\|\Phi(T_i)\|$. But by definition of the tensor norm $\|\cdot\|$ we have

$$\|\Phi(T_i)\| = \max_{u_k,v_k}\big|\big\langle\Phi(T_i), \textstyle\bigotimes_{k=1}^m u_k\bigotimes_{k=1}^q v_k\big\rangle\big| \tag{2.26}$$

for $u_k\in E_r, v_k\in E_d$; where $E_d=\{x\in\{0,1\}^d : \|x\|_1=1\}$. So it holds that

$$\|\Phi\|_\circ \le 2\max_{i,u_k,v_k}\big|\big\langle\Phi(T_i), \textstyle\bigotimes_{k=1}^m u_k\bigotimes_{k=1}^q v_k\big\rangle\big| \tag{2.27}$$

for $i \in [N], u_k \in E_r, v_k \in E_d$. Next consider any $s \in \mathbb{R}$, and observe that

$$\begin{aligned}\mathbb{E}\exp\left(s\|\Phi\|_\circ\right) &\leq \mathbb{E}\exp\left(2s\max_{i,u_k,v_k}\left|\left\langle\Phi(T_i), \textstyle\bigotimes_{k=1}^m u_k \bigotimes_{k=1}^q v_k\right\rangle\right|\right)\\ &\leq 2Nr^m d^q \max_{i,u_k,v_k}\mathbb{E}\exp\left(2s\left\langle\Phi(T_i), \textstyle\bigotimes_{k=1}^m u_k \bigotimes_{k=1}^q v_k\right\rangle\right)\end{aligned} \tag{2.28}$$

We seek to bound the term on the right-hand side. Towards this end, note $\|B\Omega_i\| \leq \sqrt{p}\|W(B)\|_2\|\Omega_i\| \leq \sqrt{p}\alpha^\rho$ by the Cauchy-Schwarz inequality and Assumption 2. This means that for $S_i = \left\langle Z^{\otimes m}(T_i\Omega)^{\otimes q}, \bigotimes_{k=1}^m u_k \bigotimes_{k=1}^q v_k\right\rangle$ we have $\left|S_i\right| \leq \alpha^{m+\rho q}p^{q/2}$. Next observe that

$$\begin{aligned}\mathbb{E}\exp\left(2s\left\langle\Phi(T_i), \textstyle\bigotimes_{k=1}^m u_k \bigotimes_{k=1}^q v_k\right\rangle\right) &\leq \left(\mathbb{E}\exp\left(\tfrac{4\epsilon s S_i}{n}\right)\right)^n\\ &= \left(\mathbb{E}\textstyle\sum_{k=0}^\infty \frac{1}{k!}\left(\frac{4\epsilon s S_i}{n}\right)^k\right)^n\\ &= \left(\mathbb{E}\textstyle\sum_{k=0}^\infty \frac{1}{(2k)!}\left(\frac{4 s S_i}{n}\right)^{2k}\right)^n\\ &\leq \left(\textstyle\sum_{k=0}^\infty \frac{1}{k!}\left(\frac{16 s^2 p^q \alpha^{2m+2\rho q}}{n^2}\right)^k\right)^n\\ &= \exp\left(\tfrac{16 s^2 p^q \alpha^{2m+2\rho q}}{n}\right)\end{aligned} \tag{2.29}$$

where the first line follows by a stochastic symmetrization step (i.e., Jensen's inequality, followed by multiplication with i.i.d. Rademacher random variables ϵ having distribution $\mathbb{P}(\epsilon = \pm 1) = \frac{1}{2}$, and concluded by using the triangle inequality), the third line follows since ϵ is a symmetric random variable, and the fourth line follows by replacing $(2k!)$ with $k!$ and substituting the absolute bound on $|S_i|$. Combining the above with (2.28) gives

$$\mathbb{E}\exp\left(s\|\Phi\|_\circ\right) \leq 2(1+4q)^{dp} r^m d^q \exp\left(\tfrac{16 s^2 p^q \alpha^{2m+2\rho q}}{n}\right). \tag{2.30}$$

Using the Chernoff bound gives

$$\begin{aligned}\mathbb{P}\left(\|\Phi\|_\circ > t\right) &\leq 2(1+4q)^{dp} r^m d^q \inf_{s\in\mathbb{R}} \exp\left(\tfrac{16 s^2 p^q \alpha^{2m+2\rho q}}{n} - st\right)\\ &= 2(1+4q)^{dp} r^m d^q \exp\left(-\tfrac{nt^2}{64 p^q \alpha^{2m+2\rho q}}\right)\end{aligned} \tag{2.31}$$

The result now follows by choosing

$$t = \sqrt{\tfrac{64 p^q \alpha^{2m+2\rho q}}{n}\left(dp\log(1+4q) + m\log r + q\log d\right) + \gamma^2} \tag{2.32}$$

and accordingly simplifying the resulting expression. □

Remark 7. Though a similar proof was used in [254] for random matrices and in [244] for random tensors, we use a stronger argument that is adapted to our setup and results in a faster convergence rate where some terms are logarithmic that would otherwise be polynomial with a weaker argument. We use a stronger chaining argument than [244, 254] by using a telescoping sum (2.25) that reduces cross terms. We use a tensor symmetrization construction (2.24) that allows us to exploit Banach's theorem [18, 37]. We achieve better constants than [254] by more carefully bounding our moment series expansion.

Proposition 5. *If Assumptions 1 and 2 hold, then we have*

$$\mathbb{P}\big(\|\widehat{\nu}_{m,q} - \nu_{m,q}\|_\circ > 2\mathcal{R}_{m,q}[n] + 2\gamma\big) \leq 4\exp\big(-\tfrac{n\gamma^2}{64p^q\alpha^{2m+2\rho q}}\big). \tag{2.33}$$

for $\mathcal{R}_{m,q}[n] = 8\alpha^{m+\rho q}p^{q/2}\sqrt{\frac{dp\log(1+4q)+m\log r+q\log d}{n}}$.

Proof. We cannot prove the result directly as in Proposition 4 because $\mathbb{E}\widehat{\nu}_{m,q}(B) \neq \nu_{m,q}(B)$, whereas the proof of Proposition 4 used the fact that $\mathbb{E}\widehat{\varphi}_{m,q}(B) = \varphi_{m,q}(B)$ in the symmetrization step of (2.29). We instead have to use an indirect approach to prove this result. We begin by noting $\widehat{\varphi}_{m,0}(B) = \mathbb{E}_n(Z^{\otimes m})$, $\varphi_{m,0}(B) = \mathbb{E}(Z^{\otimes m})$, $\widehat{\varphi}_{0,q}(B) = \mathbb{E}_n((B\Omega)^{\otimes q})$, and $\varphi_{0,q}(B) = \mathbb{E}((B\Omega)^{\otimes q})$. For any $W(B) \in \mathbb{S}^{dp-1}$ we have that $\|B\Omega_i\| \leq \sqrt{p}\|W(B)\|_2\|\Omega_i\| \leq \sqrt{p}\alpha^\rho$ by the Cauchy-Schwarz inequality and Assumption 2. This means that $\|\widehat{\varphi}_{m,0}\|_\circ \leq \alpha^m$ and $\|\varphi_{0,q}\|_\circ \leq \alpha^{\rho q}p^{q/2}$. Now consider

$$\begin{aligned}\|\widehat{\nu}_{m,q} - \nu_{m,q}\|_\circ &= \|\widehat{\varphi}_{m,0}\otimes\widehat{\varphi}_{0,q} - \varphi_{m,0}\otimes\varphi_{0,q}\|_\circ \\ &\leq \|\widehat{\varphi}_{m,0}\|_\circ\cdot\|\widehat{\varphi}_{0,q} - \varphi_{0,q}\|_\circ + \|\varphi_{0,q}\|_\circ\cdot\|\widehat{\varphi}_{m,0} - \varphi_{m,0}\|_\circ \\ &\leq \alpha^m\|\widehat{\varphi}_{0,q} - \varphi_{0,q}\|_\circ + \alpha^{\rho q}p^{q/2}\|\widehat{\varphi}_{m,0} - \varphi_{m,0}\|_\circ\end{aligned} \tag{2.34}$$

Then the union bound implies

$$\begin{aligned}\mathbb{P}\big(\|\widehat{\nu}_{m,q} - \nu_{m,q}\|_\circ \leq 2\mathcal{R}_{m,q}[n] + 2\gamma\big) \leq \\ 1 - \mathbb{P}\big(\alpha^m\|\widehat{\varphi}_{0,q} - \varphi_{0,q}\|_\circ > \mathcal{R}_{m,q}[n] + \gamma\big) + \\ - \mathbb{P}\big(\alpha^{\rho q}p^{q/2}\|\widehat{\varphi}_{m,0} - \varphi_{m,0}\|_\circ > \mathcal{R}_{m,q}[n] + \gamma\big)\end{aligned} \tag{2.35}$$

for $\mathcal{R}_{m,q}[n] = 8\alpha^{m+\rho q}p^{q/2}\sqrt{\frac{dp\log(1+4q)+m\log r+q\log d}{n}}$, which upon using (2.23) from Proposition 4 gives (2.33), which is the desired result. □

Feasible Set Consistency

We are now in a position to study the constraints of the FO problem (2.17). Towards this goal, we first define

$$\mathcal{S} = \big\{B \in \mathcal{B} : B\Omega \perp\!\!\!\perp Z\big\}. \tag{2.36}$$

This is the feasible set of (2.9), which chooses an optimal fair decision rule when the underlying distributions are exactly known, for a decision rule that satisfies Assumption 1. We next define the family of random sets

$$\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} = \big\{B \in \mathcal{B} : \big\|\widehat{\varphi}_{m,q}(B) - \widehat{\nu}_{m,q}(B)\big\| \leq \Delta_{m,q}, \text{for } (m,q) \in [\mathfrak{g}]\times[\mathfrak{h}]\big\}. \tag{2.37}$$

This is simply the feasible set of the level-$(\mathfrak{g},\mathfrak{h})$ FO problem (2.17).

Proposition 6. *The sets* $\mathcal{S}$ *and* $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ *are closed, under Assumption 1.*

Proof. We first prove the result for $\mathcal{S}$. Consider any convergent sequence $B_k \in \mathbb{R}^{d\times p}$ with $B_k \in \mathcal{S}$ and $\lim_k B_k = B_0$. Theorem 1 says for all k we have

$$\varphi_{m,q}(B_k) = \nu_{m,q}(B_k), \text{for } m, q \geq 1. \tag{2.38}$$

But the φ and ν are continuous since they are multlinear operators on Euclidean space. This means $\lim_k \varphi_{m,q}(B_k) = \varphi_{m,q}(B_0)$ and $\lim_k \nu_{m,q}(B_k) = \nu_{m,q}(B_0)$ for $m, q \geq 1$. As a result we have

$$\varphi_{m,q}(B_0) = \nu_{m,q}(B_0), \text{for } m, q \geq 1, \tag{2.39}$$

which by Theorem 1 implies $B_0 \in \mathcal{S}$. This proves that $\mathcal{S}$ is closed.

The proof for $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ is a simple modification of the above argument. Consider any convergent sequence $B_k \in \mathbb{R}^{d\times p}$ with $B_k \in \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ and $\lim_k B_k = B_0$. By definition of $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ we have for all k that

$$\left\|\widehat{\varphi}_{m,q}(B_k) - \widehat{\nu}_{m,q}(B_k)\right\| \leq \Delta_{m,q}, \text{for } (m, q) \in [\mathfrak{g}] \times [\mathfrak{h}]. \tag{2.40}$$

But the $\widehat{\varphi}$ and $\widehat{\nu}$ are continuous since they are multlinear operators on Euclidean space, and so the normed function $\left\|\widehat{\varphi}_{m,q}(B) - \widehat{\nu}_{m,q}(B)\right\|$ is also continuous. As a result we have

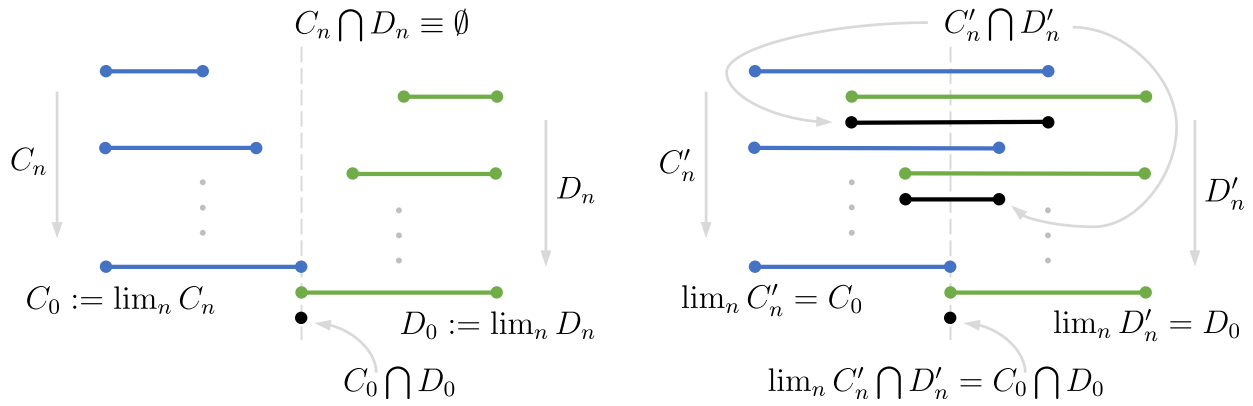
$$\left\|\widehat{\varphi}_{m,q}(B_0) - \widehat{\nu}_{m,q}(B_0)\right\| = \lim_k \left\|\widehat{\varphi}_{m,q}(B_k) - \widehat{\nu}_{m,q}(B_k)\right\| \leq \Delta_{m,q}, \quad \text{for } m, q \geq 1. \tag{2.41}$$

This means $B_0 \in \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ by definition. This proves that $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ is closed. □

The sequence of random sets $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ is technically difficult to study because each random set is defined by the intersection of many random constraint inequalities, with the number of these random constraints increasing towards infinity. There is a more subtle technical difficulty that needs to be addressed. The issue is that when intersecting a sequence of sets, the intersection of the sequence terms generally does not converge to the intersection of the limiting sets [14, 173]. The next example demonstrates this phenomenon in a deterministic setting, and it provides some insight into how the situation can be addressed through a carefully designed regularization approach.

Example 3. Fig. 2.1 provides a visualization of this example. Let us first define $C_n = [-1, -\frac{1}{n}]$ and $D_n = [\frac{1}{n}, 1]$, which each specify a deterministic sequence of compact sets. Then we have that $\lim_n C_n = [-1, 0] =: C_0$ and that $\lim_n D_n = [0, 1] =: D_0$. However, note that $C_n \bigcap D_n = \emptyset$. This means $\lim_n C_n \bigcap D_n = \emptyset \neq C_0 \bigcap D_0 = \{0\}$. Now suppose we carefully regularize these sequences of sets. Specifically consider the regularized sequence of deterministic, compact sets $C'_n = [-1, -\frac{1}{n} + \Delta_n]$ and $D'_n = [\frac{1}{n} - \Delta_n, 1]$ for $\Delta_n = \frac{2}{n}$, where we think of the Δ_n as regularizing by inflating the sets. Clearly this choice of regularization goes to zero since $\lim_n \Delta_n = 0$. More importantly, we now have $C'_n \bigcap D'_n = [-\frac{1}{n}, \frac{1}{n}]$. This means we have $\lim_n C'_n = C_0$ and $\lim_n D'_n = D_0$ with $\lim_n C'_n \bigcap D'_n = \{0\} = C_0 \bigcap D_0$.

The above example was deterministic, and it may not initially be clear whether such behavior is an issue for our random setting. The next example demonstrates a situation where this non-convergence occurs for $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$.



(a) Unregularized Set Intersections (b) Regularized Set Intersections

Figure 2.1: The left shows how the intersection of a sequence of sets may not converge to the intersection of the limiting sets. The right shows how regularization of the sequence of sets can help to ensure that the intersection of the regularized sets converges to the intersection of the limiting sets.

Example 4. Consider a setting where $B \in \mathbb{R}$ and the distributions are $X \sim \text{Ber}(x)$ and $Z \sim \text{Ber}(z)$ with $X \perp\!\!\!\perp Z$. We assume that $x \in (0,1)$ and $z \in (0,1)$ to prevent degeneracies in this example. In this setup $\mathcal{S} = \mathcal{B}$. Now observe that $(Z_i)^m = Z_i$ and $(X_i)^q = X_i$ for $(m,q) \geq 1$ since $X_i, Z_i \in \{0,1\}$. This means the $(m,q) \geq 1$ constraints in $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ for $\Delta_{m,q} = 0$ are

$$\begin{aligned}\Big|\big(\tfrac{1}{n}\textstyle\sum_{i=1}^n (Z_i)^m (X_i)^q - \tfrac{1}{n}\sum_{i=1}^n (Z_i)^q \cdot \tfrac{1}{n}\sum_{i=1}^n (X_i)^q\big) B^q\Big| = \\ \Big|\big(\tfrac{1}{n}\textstyle\sum_{i=1}^n Z_i X_i - \tfrac{1}{n}\sum_{i=1}^n Z_i \cdot \tfrac{1}{n}\sum_{i=1}^n X_i\big) B^q\Big| = 0. \quad (2.42)\end{aligned}$$

This means $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} = \mathcal{B}$ whenever $\mathcal{E}_n = \{\frac{1}{n}\sum_{i=1}^n Z_i X_i = \frac{1}{n}\sum_{i=1}^n Z_i \cdot \frac{1}{n}\sum_{i=1}^n X_i\}$ occurs, and that $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} = \emptyset$ otherwise. And so trivially by the definition of $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ we have as-lim sup$_n$ $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} \subseteq \mathcal{B}$. If we recall the classical setting of a 2×2 contingency table, this event $\mathcal{E}_n$ is equivalent to having exact equality between a marginal and cross-term in the contingency table. As a result, we consider a test statistic inspired by the Pearson test for independence

$$T_n = n \cdot \big(\mathbb{E}_n(ZX) - \mathbb{E}_n(Z)\mathbb{E}_n(X)\big)^2. \quad (2.43)$$

Clearly by its definition, we have that $T_n = 0$ if and only if $\mathcal{E}_n$ holds. Also, a straightforward calculation gives

$$\mathbb{E}(T_n) = (\tfrac{n-1}{n})(zx)(1-z-x-zx). \quad (2.44)$$

Note that $\mathbb{E}(T_n) > 0$ since we assumed $x, z \in (0,1)$, and note that $\mathbb{E}(T_n)$ is monotonically increasing towards $\lim_n \mathbb{E}(T_n) = (zx)(1-z-x-zx) > 0$. Now using McDiarmid's inequality

we get for any $t > 0$ that

$$\mathbb{P}(\mathcal{E}_n) \leq \mathbb{P}(T_n \leq \mathbb{E}(T_n) - t) \leq \exp(-nt^2/8). \tag{2.45}$$

Choosing $t = (zx)(1 - z - x - zx)/2$, the Borel-Cantelli lemma implies $\mathcal{E}_n$ cannot occur infinitely often. Hence we must have as-lim inf$_n \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} = \emptyset \not\supseteq \mathcal{S}$.

Example 3 provides the key intuition for how potential non-convergence of $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$, as demonstrated in Example 4, can be resolved. If we can regularize the sets $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ by sufficiently inflating them in such a way that the amount of inflation decreases with n, then we may be able to ensure the almost sure stochastic convergence of $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ to $\mathcal{S}$. In fact, the notation of Example 3 was chosen to be suggestive of how we will perform this regularization: We will purposefully keep the $\Delta_{m,q} > 0$ while allowing them to shrink towards zero.

More broadly, the FO problem (2.17) has two types of tuning parameters, namely the $(\mathfrak{g}, \mathfrak{h})$ that controls the number of moment constraints and the $\Delta_{m,q}$ that controls the strictness of the moment constraint. This gives us considerable flexibility when studying asymptotic properties. A choice of faster rates requires knowledge of an appropriate value of α from Assumption 2. An alternative approach is to choose slower rates that have the benefit of working for any value of α. Here, we will take the latter approach.

Theorem 3. *Suppose that* $\Delta_{m,q} = O(n^{-1/4})$ *and* $\mathfrak{g} = \mathfrak{h} = O(\log \log n)$. *If Assumptions 1, 2, and 3 hold, then* as-lim$_n \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} = \mathcal{S}$.

Proof. For the first part of the proof we will show as-lim inf$_n \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} \supseteq \mathcal{S}$. Indeed, suppose this is not true. Then there exists $B_0 \in \mathcal{S}$ and an open neighborhood $\mathcal{N} \subseteq \mathcal{B}$ of B_0 such that $\mathcal{N} \bigcap \mathcal{S}_{\mathfrak{g},\mathfrak{h}} = \emptyset$ infinitely often (Theorem 4.5 of [212]). We can rewrite one of these events as

$$\{\mathcal{N} \bigcap \mathcal{S}_{\mathfrak{g},\mathfrak{h}} = \emptyset\} = \bigcup_{m\in[\mathfrak{g}]} \bigcup_{q\in[\mathfrak{h}]} \{ \inf_{B\in\mathcal{N}} \|\widehat{\Xi}_{m,q}(B)\| > \Delta_{m,q}\}, \tag{2.46}$$

where for convenience we define the multilinear operators $\Xi_{m,q} = \varphi_{m,q} - \nu_{m,q}$, $\widehat{\Xi}_{m,q} = \widehat{\varphi}_{m,q} - \widehat{\nu}_{m,q}$, $\Phi_{m,q} = \widehat{\varphi}_{m,q} - \varphi_{m,q}$, and $\Psi_{m,q} = \widehat{\nu}_{m,q} - \nu_{m,q}$. Because Theorem 1 can be rewritten under the assumptions of this theorem as

$$\sup_{B\in\mathcal{S}} \|\varphi_{m,q}(B) - \nu_{m,q}(B)\| = 0 \text{ for } m, q \geq 1, \tag{2.47}$$

application of the triangle inequality yields

$$\begin{aligned} \|\widehat{\Xi}_{m,q}(B_0)\| &\leq \|\Xi_{m,q}(B_0)\| + \|\Phi_{m,q}(B_0)\| + \|\Psi_{m,q}(B_0)\| \\ &\leq \lambda^{q/2}\|\Phi_{m,q}\|_\circ + \lambda^{q/2}\|\Psi_{m,q}\|_\circ \end{aligned} \tag{2.48}$$

Let $\mathcal{G}_{m,q}[n] = (1 + \log n)\lambda^{q/2}\mathcal{R}_{m,q}[n]$, and note that $\mathcal{G}_{m,q}[n] = o(n^{-1/4})$ for $(m, q) \in [\mathfrak{g}] \times [\mathfrak{h}]$ under the hypothesis of this theorem. This means that for all n sufficiently large, the union

bound gives us that

$$\begin{aligned}\mathbb{P}\big(\mathcal{N}\bigcap\mathcal{S}_{\mathfrak{g},\mathfrak{h}}=\emptyset\big) &\leq \textstyle\sum_{m\in[\mathfrak{g}]}\sum_{q\in[\mathfrak{h}]}\mathbb{P}\big(\lambda^{q/2}\|\Phi_{m,q}\|_\circ > \mathcal{G}_{m,q}[n]\big)+\\ &\qquad \textstyle\sum_{m\in[\mathfrak{g}]}\sum_{q\in[\mathfrak{h}]}\mathbb{P}\big(\lambda^{q/2}\|\Psi_{m,q}\|_\circ > 2\mathcal{G}_{m,q}[n]\big) \\ &\leq O((\log\log n/n)^2)\end{aligned} \tag{2.49}$$

where the last line used Propositions 4 and 5, along with the relation that $\exp(-\frac{n\gamma^2}{64p^q\alpha^{2m+2\rho q}}) = O(1/n^2)$ for $\gamma = \log n\cdot\mathcal{R}_{m,q}[n]$. Thus the Borel-Cantelli lemma says $\mathcal{N}\bigcap\mathcal{S}_{\mathfrak{g},\mathfrak{h}}=\emptyset$ only finitely many times, which is a contradiction. This proves as-$\liminf_n \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} \supseteq \mathcal{S}$.

For the second part of the proof we will show as-$\limsup_n \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} \subseteq \mathcal{S}$. Indeed, suppose this is not true. Then there exists $B_0 \in \limsup_n \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ and a closed neighborhood $\mathcal{N} \subseteq \mathcal{B}$ of B_0 such that $\mathcal{N}\bigcap\mathcal{S}=\emptyset$ and $\mathcal{N}\bigcap\mathcal{S}_{\mathfrak{g},\mathfrak{h}}\neq\emptyset$ infinitely often (Theorem 4.5 of [212]). But Theorem 1 implies there exists some $m,q\geq 1$ such that we have

$$\zeta := \inf_{B\in\mathcal{N}}\|\varphi_{m,q}(B)-\nu_{m,q}(B)\| > 0. \tag{2.50}$$

We will keep m,q fixed at these values for the remainder of the proof. Now note that for one of the events $\mathcal{N}\bigcap\mathcal{S}_{\mathfrak{g},\mathfrak{h}}\neq\emptyset$ we have

$$\big\{\mathcal{N}\bigcap\mathcal{S}_{\mathfrak{g},\mathfrak{h}}\neq\emptyset\big\} \subseteq \big\{\inf_{B\in\mathcal{N}}\|\widehat{\Xi}_{m,q}(B)\| \leq \Delta_{m,q}\big\}. \tag{2.51}$$

Application of the triangle inequality yields

$$\begin{aligned}\zeta = \inf_{B\in\mathcal{N}}\|\Xi_{m,q}(B)\| \leq& \\ \inf_{B\in\mathcal{N}}\|\widehat{\Xi}_{m,q}(B)\| + \sup_{B\in\mathcal{N}}\|\Phi_{m,q}(B)\| + \sup_{B\in\mathcal{N}}\|\Psi_{m,q}(B)\| \leq& \\ \inf_{B\in\mathcal{N}}\|\widehat{\Xi}_{m,q}(B)\| + \lambda^{q/2}\|\Phi_{m,q}\|_\circ + \lambda^{q/2}\|\Psi_{m,q}\|_\circ.&\end{aligned} \tag{2.52}$$

Let $\mathcal{G}_{m,q}[n] = (1+\log n)\lambda^{q/2}\mathcal{R}_{m,q}[n]$, and note that $\mathcal{G}_{m,q}[n] = o(n^{-1/4})$ and that $\Delta_{m,q} = o(1)$ under the hypothesis of this theorem. For all n sufficiently large, we have $\zeta - \Delta_{m,q} \geq \zeta/2 \geq 3\mathcal{G}_{m,q}[n]$. Hence the union bound gives

$$\begin{aligned}\mathbb{P}\big(\mathcal{N}\bigcap\mathcal{S}_{\mathfrak{g},\mathfrak{h}}\neq\emptyset\big) &\leq \mathbb{P}\big(\lambda^{q/2}\|\Phi_{m,q}\|_\circ > \mathcal{G}_{m,q}[n]\big)+\\ &\qquad \mathbb{P}\big(\lambda^{q/2}\|\Psi_{m,q}\|_\circ > 2\mathcal{G}_{m,q}[n]\big) \\ &\leq O(1/n^2)\end{aligned} \tag{2.53}$$

where the last line used Propositions 4 and 5, along with the relation that $\exp(-\frac{n\gamma^2}{64p^q\alpha^{2m+2\rho q}}) = O(1/n^2)$ for $\gamma = \log n\cdot\mathcal{R}_{m,q}[n]$. Thus the Borel-Cantelli lemma says $\mathcal{N}\bigcap\mathcal{S}_{\mathfrak{g},\mathfrak{h}}\neq\emptyset$ only finitely many times, which is a contradiction. This proves as-$\limsup_n \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} \subseteq \mathcal{S}$. □

Solution Set Consistency

Next consider the solution set

$$\widehat{\mathcal{O}}_{\mathfrak{g},\mathfrak{h}} = \arg\min_{B} \left\{ R_n(B \cdot \omega(x,z)) \;\middle|\; B \in \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} \right\} \tag{2.54}$$

for the level-$(\mathfrak{g},\mathfrak{h})$ FO problem (2.17). Similarly, consider the solution set

$$\mathcal{O} = \arg\min_{B} \left\{ R(B \cdot \omega(x,z)) \;\middle|\; B \in \mathcal{S} \right\} \tag{2.55}$$

for the optimization problem (2.9), which chooses an optimal fair decision rule when the underlying distributions are exactly known.

Our next result shows that solving the FO problem (2.17) provides a statistically consistent approximation to solving the optimization problem (2.9), and we state the result using the solutions sets $\widehat{\mathcal{O}}_{\mathfrak{g},\mathfrak{h}}$ and $\mathcal{O}$ defined above.

Theorem 4. *Suppose that* $\Delta_{m,q} = O(n^{-1/4})$ *and* $\mathfrak{g} = \mathfrak{h} = O(\log\log n)$. *If Assumptions 1–4 hold, then* $\text{as-lim sup}_n\, \widehat{\mathcal{O}}_{\mathfrak{g},\mathfrak{h}} \subseteq \mathcal{O}$.

Proof. First consider the indicator function $\Gamma(B, \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}})$. Combining our Theorem 3 with Proposition 7.4 of [212] gives $\text{as-e-lim}\, \Gamma(\cdot, \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}) = \Gamma(\cdot, \mathcal{S})$ relative to $\mathbb{R}^{d\times p}$. Next we claim $\text{as-lim}\, \Gamma(\cdot, \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}) = \Gamma(\cdot, \mathcal{S})$ relative to $\mathbb{R}^{d\times p}$. Since Proposition 6 says the $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ are closed, the remark after Theorem 7.10 of [212] implies it is sufficient to show that for every $B_0 \in \mathcal{S}$ we have $B_0 \notin \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ only a finite number of times. A similar argument to the first part of the proof for Theorem 3 can be used to show this, and so we omit the details.

Next we note that the level-$(\mathfrak{g},\mathfrak{h})$ FO problem (2.17) can be written as $\min_B h_n(B) + \Gamma(B, \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}})$, and the optimization problem (2.9) can be written as $\min_B h(B) + \Gamma(B, \mathcal{S})$. Now using Theorem 7.46 of [212] gives us that

$$\text{as-e-lim}\left(h_n(\cdot) + \Gamma(\cdot, \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}})\right) = h(\cdot) + \Gamma(\cdot, \mathcal{S}). \tag{2.56}$$

The result now follows by direct application of Proposition 7.30 of [212]. □

Remark 8. If the optimization problem (2.9) is infeasible, then we will have $\mathcal{O} = \emptyset$ and $\text{as-lim sup}_n\, \widehat{\mathcal{O}}_{\mathfrak{g},\mathfrak{h}} = \emptyset$, with $\widehat{\mathcal{O}}_{\mathfrak{g},\mathfrak{h}} \neq \emptyset$ only finitely many times.

Remark 9. We can guarantee under the case of additional assumptions that $\text{as-lim sup}_n\, \widehat{\mathcal{O}}_{\mathfrak{g},\mathfrak{h}} \neq \emptyset$, with $\widehat{\mathcal{O}}_{\mathfrak{g},\mathfrak{h}} = \emptyset$ only finitely many times. In particular, it can be shown that this occurs when Assumption 5 holds and $\mathcal{O} \neq \emptyset$. If $\mathcal{O}$ consists of a single point, then it can also be shown that $\text{as-lim}_n\, \widehat{\mathcal{O}}_{\mathfrak{g},\mathfrak{h}} = \mathcal{O}$.

The conclusion "$\text{as-lim sup}_n\, \widehat{\mathcal{O}}_{\mathfrak{g},\mathfrak{h}} \subseteq \mathcal{O}$" of the above theorem says all cluster points (i.e., convergent subsequences) as n increases of optimal solutions to the sample-based FO problem (2.17) belong to the set of optimal solutions to the problem (2.9) that we initially set out to solve using a sample-based approach. A stronger result is generally not true [212]; however, as mentioned above it can be shown that if $\mathcal{O}$ is singleton then we have $\text{as-lim}_n\, \widehat{\mathcal{O}}_{\mathfrak{g},\mathfrak{h}} = \mathcal{O}$.

2.6 Approximate Independence

Let $U \in \mathbb{R}^p$ and $V \in \mathbb{R}^d$ be random vectors, and consider the quantity

$$\mathbb{M}(U;V) = \sup_{m,q\geq 1} \frac{\left\|\mathbb{E}\left(U^{\otimes m}V^{\otimes q}\right) - \mathbb{E}\left(U^{\otimes m}\right)\otimes\mathbb{E}\left(V^{\otimes q}\right)\right\|}{(m+q)!}. \tag{2.57}$$

We call the quantity $\mathbb{M}(U;V)$ the *mutual majorization* of U and V, and the choice of this name is meant to draw a direct analogy to mutual information.

Proposition 7. *The mutual majorization is nonnegative* $\mathbb{M}(U;V) \geq 0$*, symmetric* $\mathbb{M}(U;V) = \mathbb{M}(V;U)$*, and satisfies* $\mathbb{M}(U;V) \leq \epsilon$ *if and only if*

$$\left\|\mathbb{E}\left(U^{\otimes m}V^{\otimes q}\right) - \mathbb{E}\left(U^{\otimes m}\right)\otimes\mathbb{E}\left(V^{\otimes q}\right)\right\| \leq \epsilon\cdot(m+q)! \text{ for } m,q\geq 1. \tag{2.58}$$

The mutual majorization also characterizes independence in the sense that for bounded multivariate random variables U *and* V*, we have* $\mathbb{M}(U;V) = 0$ *if and only if* U *and* V *are independent.*

Proof. The first three claims are obvious from the definition of mutual majorization, and so we focus on the fourth claim. If U and V are independent, then $\mathbb{M}(U;V) = 0$ by Theorem 1. To show the converse, we prove its contrapositive: If U and V are dependent, then $\mathbb{M}(U;V) > 0$ since Theorem 1 implies $\left\|\mathbb{E}\left(U^{\otimes m}V^{\otimes q}\right) - \mathbb{E}\left(U^{\otimes m}\right)\otimes\mathbb{E}\left(V^{\otimes q}\right)\right\| > 0$ for some $m,q\geq 1$. □

The implication of this result is that we can use mutual majorization to quantify approximate independence. We thus define an optimization problem that chooses an optimal ϵ-approximately-fair decision rule by solving

$$\delta^*(x,z) \in \arg\min_{\delta(\cdot,\cdot)} \left\{R(\delta) \;\middle|\; \mathbb{M}(\delta(X,Z);Z) \leq \epsilon\right\}. \tag{2.59}$$

The level-$(\mathfrak{g},\mathfrak{h})$ FO problem (2.17) with appropriate choice of $\Delta_{m,q}$ is a statistically well-behaved, sample-based approximation of the above problem. In order to be able to discuss this, we first define the set

$$\mathcal{S}(\epsilon) = \left\{B\in\mathcal{B} : \mathbb{M}(B\Omega;Z) \leq \epsilon\right\} \tag{2.60}$$

and the solution set

$$\mathcal{O}(\epsilon) = \arg\min_{B} \left\{R(B\cdot\omega(x,z)) \;\middle|\; B\in\mathcal{S}(\epsilon)\right\}. \tag{2.61}$$

These are respectively the feasible set and solution set of the optimization problem (2.59), which chooses an optimal ϵ-approximately-fair decision rule when the underlying distributions are exactly known.

Theorem 5. *Let* $\Delta_{m,q} = \epsilon \cdot (m+q)! + O(n^{-1/4})$ *and* $\mathfrak{g} = \mathfrak{h} = O(\log\log n)$. *If Assumption 1 holds, then* $\mathcal{S}(\epsilon)$ *is closed. If Assumptions 2 and 3 also hold,* $\text{as-lim}_n \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} = \mathcal{S}(\epsilon)$. *If Assumption 4 also holds,* $\text{as-lim sup}_n \widehat{\mathcal{O}}_{\mathfrak{g},\mathfrak{h}} \subseteq \mathcal{O}(\epsilon)$.

Remark 10. The proof is omitted because it is a straightforward modification of the proofs for Proposition 6 and Theorems 3 and 4. The main difference in the modified proofs is the use of (2.58) from Proposition 7.

Remark 11. Recall we already proved $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ is closed in Proposition 6.

2.7 Conclusion

We proposed an optimization hierarchy for fair statistical decision problems, which provides a systematic approach to fair versions of hypothesis testing, decision-making, estimation, regression, and classification. We showed that the FO hierarchy is general to many different notions of fairness as well as many different forms of decisions, and has a high level of flexibility in terms of the number of constraints that need to be added. We proved that higher levels of this hierarchy asymptotically impose independence between the output of the decision rule and the protected variable as a constraint in corresponding statistical decision problems. A version of this hierarchy was also proposed that involved tuning fewer hyperparameters. An important question that remains to be answered is how to tune the hyperparameters in our hierarchy. Our theoretical results provide some guidance on how to choose the level of the hierarchy and how to reduce the number of tuning parameters to just one. However, further theoretical and empirical study is needed to better understand the tuning process.

Chapter 3

Fairness in Supervised Learning

3.1 Introduction

Supervised learning describes the process of learning a mapping between input-output pairs. The development of this field reaches back to the pioneering works of Legendre in 1805 [150] and Gauss in 1809 [10] in developing the method of least squares for tracking the motion of cosmic bodies, and subsequent progression and formalizations by Galton [97], Yule [269], Pearson [199] and Fisher [91]. When the output is no longer continuous but rather takes values in a countable set, this problem is referred to as classification, and has been a key problem in statistical learning theory since Fisher's original use of linear discriminant functions [92]. Today, convex margin-based classification techniques such as logistic regression and support vector machines (SVMs), which depend on thresholding a linear "score" function, are workhorse algorithms that can quickly and easily be applied to wide variety of use-cases. Indeed, such models benefit from small generalization error due to the structural simplicity of the set of decision functions [22,250,251], and from computational efficiency due to the existence of statistically-innocuous, convex surrogate loss functions that can be easily optimized over [21].

The last decade has seen a resurgence of the use of Artificial Neural Networks (ANN), also referred to as "deep-learning", for various classification and prediction tasks, such as image processing [140, 215], natural language processing [101, 200], automatic featurization [137, 168], and reinforcement learning [153, 177]. Due to this, deep-learning models have even begun gaining traction in more sensitive, societal contexts such as healthcare [128,157], hiring [55, 230], and criminal justice [11, 253]. While these do not share in the beneficial properties that arise from convex classifiers, they have been shown to outperform on many key tasks, and so have inspired a contemporary push towards better understanding the behavior of deep-learning models.

Problems of bias and unfairness are easy to conceptualize in the setting of supervised learning (and particularly that of binary classification), as the "output" in each input-output pair can map to a more or less socially-desirable outcome, thus imbuing the predictions of

the model with significance beyond the limited scope of its statistical task. Accordingly, the earliest studies that exposed bias in automated decision-making frameworks focused on biased classifiers [11, 20]. Similarly, the first work in this fair machine learning also focused on designing fair classifiers [45, 48, 109, 270, 271].

This chapter focuses on applications of the Fair Optimization (FO) hierarchy introduced in Chapter 2 for supervised learning. In this, it makes three main contributions. First, we reinterpret two fairness notions using receiver operating characteristic (ROC) curves, which leads to a new visualization for classifier fairness. Second, we provide a number of interpretations of the FO hierarchy for supervised learners. Third, we conduct numerical experiments on real data to evaluate the efficacy of FO in a variety of supervised learning settings, including fair SVM, fair linear regression and fair quantile regression applied to a morphine-dosage case study.

Fairness Notions for Classifiers

Ensuring classifiers are fair requires quantifying their fairness. However, [94] and [134] showed that no single metric can capture all intuitive aspects of fairness, and so any metric must choose a specific aspect of fairness to quantify. Here, we consider arguably the two most popular notions: disparate impact [45, 270, 276] and equal opportunity [77, 109]. Precise definitions of these are given in Section 3.3. Interpretations are given that build on the original definitions in Section 2.1. In most cases, the notions that we consider are defined for only the output value of a margin classifier, meaning that the metrics are only considered at one value of the threshold implicit in the process. On the other hand, we argue that it is necessary to bound these metrics at all possible thresholding levels. We believe this is more in-line with malicious usage of classifiers in which strategic choice of thresholds can be used to practice discrimination, and will increase the robustness of fairness measures.

Algorithms to Compute Fair Classifiers

Several approaches have been developed to construct fair classifiers. Some [2, 48, 162, 221, 271] compute transformations of the data to make it independent of the protected class. However, these kinds of preprocessing approaches are necessarily blind to the ultimate usage of the transformed data, and so share in the disadvantages of greedy approaches: They can be too conservative and reduce predictive accuracy more than desired. Alternatively, post-processing methods [109] modify classifiers post-hoc to reduce its accuracy with respect to protected classes until fairness is achieved. Note that these methods are explicitly focused on the output of the classifier at one thresholding level by design. Several techniques compute a fair classifier via regularization methods [45, 66]; however, these also only apply at single thresholding levels. The only method for convex classifiers we are aware of that tries to compute a fair classifier for all thresholds is that of [270], which is a specific instance of the FO hierarchy.

Similar to the case for convex classifiers, fairness in deep-learning models has recently attracted much attention. In some cases, regularization methods originally designed for convex classification techniques may be extended to the deep-learning setting [125]. However, there have been several techniques designed specifically for deep-learning applications. While some of these take pre-processing approaches similar in spirit to those for convex classifiers [40], a large number take an adversarial approach [27,80,163,272]. In general, these approaches use a generative adversarial structure where a primary ANN is designed for the desired prediction task, an 'adversarial' ANN is used to predict the protected class using the predictor's output, and a loss function is used to train these models simultaneously, such that the predictor obtains accurate predictions that cannot be used to reconstruct the protected data using the adversary. The most similar work to ours is that of [44], which use an alternative form of regularization for image labeling tasks. However, their method is tailored to their considered task, while our method is much more general.

Outline

After describing the data and our notation in Section 3.2, we next define two fairness notions and provide a new ROC visualization of fairness, along with further intuition, in Section 3.3. Section 3.4 presents specific instances of the FO hierarchy developed in Chapter 2 to specific supervised learning problems of interest in this chapter, namely SVM, regression and deep-learning classification. Section 3.5 then develops intuitions and interpretations largely specific to these models, and Section 3.6 presents kernel versions of the models. Computational properties are discussed in Section 3.7. Penultimately, Section 3.8 conducts numerical experiments using both synthetic and real datasets to demonstrate the efficacy of our approach in promoting fairness while preserving accuracy. These involve SVM, regression and deep classification problems. This section is finished with two case studies in automated dosage: one for one-time Morphine injections, and one for schedule of Heparin dosage using a "fair LSTM". Finally, Section 3.9 concludes.

3.2 Preliminaries

Our data consists of 3-tuples (X, Y, Z) where $X \in \mathbb{R}^p$ are predictors, $Y \in \mathcal{Y} \in \mathbb{R}$ are labels, and $Z \in \mathcal{Z}$ label a protected class. In many cases in this chapter, we will have either $\mathcal{Y} = \{\pm 1\}$ or $\mathcal{Z} = \{\pm 1\}$, to denote the simplified context of binary classification or a situation with only two protected classes. Let $(x)_+ = \max\{x, 0\}$. When it is necessary to denote several observations of Y or X concatenated into a vector or matrix, we use the notation $\vec{Y}$ or $\vec{X}$; this means that $\vec{Y} \in \mathbb{R}^n$ is a vector such that the i'th element is the realization Y_i, and $\vec{X} \in \mathbb{R}^{n \times p}$ is a matrix such that the i'th row is the realization X_i. Furthermore, the i'th element of vector $\vec{Y}$ is denoted $\vec{Y}_i$, and the i'th row of matrix $\vec{X}$ is denoted $\vec{X}_i$.

Next let $K(x, x') : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ be a kernel function, and consider the notation

$$K(\vec{X}, \vec{X}') = \begin{bmatrix} K(X_1, X_1') & K(X_1, X_2') & \cdots \\ K(X_2, X_1') & K(X_2, X_2') & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \tag{3.1}$$

Recall that the essence of the *kernel trick* is to replace $X_i^\mathsf{T} X_j$ with $K(X_i, X_j)$, and so the benefit of the matrix notation given in (3.1) is that it allows us to replace $\vec{X}(\vec{X}')^\mathsf{T}$ with $K(\vec{X}, \vec{X}')$ as part of the kernel trick.

Last, we define some additional notation. Let $[n] = \{1, \ldots, n\}$, and note $\mathbf{1}(u)$ is the indicator function. A positive semidefinite matrix U is denoted $U \succeq 0$. If U, V are vectors of equal dimension, then the notation $U \circ V$ refers to their element-wise product: $(U \circ V)_i = U_i \cdot V_i$. Also, $\mathbf{e}$ is the vector whose entries are all 1.

3.3 Visualization of Fairness

In this section, we discuss popular quantifications of fairness. For expository reasons, we focus initially on a binary classifier.

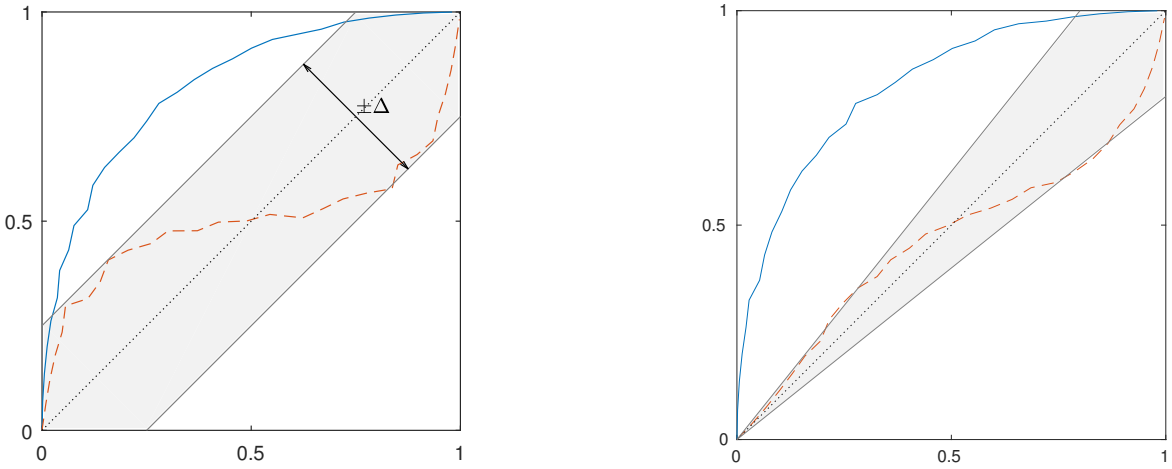
Disparate Impact

One popular notion of fairness is that predictions of the label Y are independent of the protected class Z. This definition is typically stated [45, 270, 276] in terms of a single threshold, though it can be generalized to multiple thresholds. We say that a classifier $d(x, t)$, which arises from thresholding a score function $\delta(x)$ at value t, has disparate impact (also referred to as demographic parity) Δ if

$$\left| \mathbb{P}\left[d(X,t) = +1 \middle| Z = +1 \right] - \mathbb{P}\left[d(X,t) = +1 \middle| Z = -1 \right] \right| \leq \Delta, \ \forall t \in \mathbb{R}. \tag{3.2}$$

To understand this, note $\mathbb{P}\left[d(X,t) = +1 \middle| Z = +1 \right]$ is the true positive rate when predicting the protected class at threshold t, while $\mathbb{P}\left[d(X,t) = +1 \middle| Z = -1 \right]$ is the false positive rate when predicting the protected class at threshold t. So the intuition is that a classifier has disparate impact level Δ if its true positive and false positive rates with respect to its ability to predict the protected class are approximately (up to Δ deviation) equal at all threshold levels.

Reinterpreted, reducing disparate impact, or imposing demographic parity, requires that predictions of the classifier cannot reveal information about the protected class any better (up to Δ deviation) than random guessing. In this sense, it is equivalent to independence between the output of $d(X, t)$ and the protected variable Z at any value of t, and can be interpreted as the Kolmogorov-Smirnov (KS) distance between the conditional distributions of the score function underlying d, conditional on Z. Equivalently, having a disparate impact level of Δ is in fact equivalent to requiring that the ROC curve for the classifier $d(X, t)$ in



(a) Disparate impact at level Δ.

(b) 80% Rule.

Figure 3.1: A visual representation of our notion of fairness. Here, the solid blue line is the ROC curve for the Y label and the dotted red line the ROC curve for the protected Z label. Δ refers to the maximum distance of the latter from the diagonal, which represents a perfect lack of predictability.

predicting Z is within Δ of the *line of no-discrimination*, which is the line that is achievable by biased random guessing. Figure 3.1a visualizes how disparate impact can be understood using an ROC curve.

Relation to the 80% Rule

The Fair Employment Practice Commission (FEPC) of the State of California announced a new regulation in 1971 termed the "80% rule", which was designed to provide determination guidelines for whether corporate selection systems led to disparate impact. This rule was later instituted as a Title VII enforcement mechanism by the Department of Labor and Department of Justice in 1978 [30]. The rule states that, if members of any group are less than 80% as likely as those of any other group to be hired or otherwise achieve some desirable outcome, then that outcome is sufficiently different from a random allocation to be labeled as "disparate impact" [51].

Clearly, this amounts to requiring that

$$0.8 \cdot \mathbb{P}\left[d(X,t) = +1 \middle| Z = +1\right] \leq \mathbb{P}\left[d(X,t) = +1 \middle| Z = -1\right] \leq 1.25 \cdot \mathbb{P}\left[d(X,t) = +1 \middle| Z = +1\right], \tag{3.3}$$

which cannot be easily stated through the lens of disparate impact as defined above, but is closely related. It is further visualized in ROC form in fig. 3.1b. However, this can be overly restrictive, particularly when dealing with lower-probability events. In fact, it could be routinely violated due to simple statistical error. If we consider $\mathbb{P}\left[d(X,t) = +1 \middle| Z = z\right]$ as a function of z and t, we can consider the case where it has exponentially-shrinking tails in t for both $z = +1$ and $z = -1$. In this case, minor empirical errors can cause the 80% rule to be violated for small or large values of t.

Equalized Odds

Since disparate impact measures have been criticized as too strict [77,109], another notion of fairness has been proposed in which predictions of the label Y are required to be independent of the protected class Z *conditional on the protected class* (i.e. this holds for cases where $Y = +1$ and for cases where $Y = -1$, separately). In this definition, we must interpret $Y = +1$ as a better label than $Y = -1$; for instance, $Y = -1$ may be a loan default, while $Y = +1$ is full repayment of a loan. This definition is typically stated [109] in terms of a single threshold, though it can be generalized to multiple thresholds. We say that a classifier $d(X,t)$ has equalized odds (EO) with level Δ if

$$\left|\mathbb{P}\left[d(X,t) = +1 \middle| Z = +1, Y = y\right] - \mathbb{P}\left[d(X,t) = +1 \middle| Z = -1, Y = y\right]\right| \leq \Delta, \quad \forall t \in \mathbb{R}, y \in \{\pm 1\}. \quad (3.4)$$

The definition of equalized odds takes a different approach to fairness by requiring that the error rates of any classifier be similar among protected classes. To see this, note that $\mathbb{P}\left[d(X,t) = +1 \middle| Z = z, Y = +1\right]$ is the true-positive rate among members of protected class z, and $\mathbb{P}\left[d(X,t) = +1 \middle| Z = z, Y = -1\right]$ is the false-positive rate among members of protected class z. This could make more sense in a case where the protected label Z is inextricably linked to the target label, Y, or in cases where distorted base rates of one or more protected class amplify statistical issues with disparate impact.

Reinterpreted, a form of intuition similar to that provided for disparate impact is that equalized odds requires the false positive rates and true positive rates for a classifier to be approximately (up to Δ deviation) equal at all threshold levels for the protected class, when conditioned on Y. Thus, equal opportunity requires that predictions of the classifier cannot reveal information about the protected class any better (up to Δ deviation) than random guessing, conditioned on the true label. In cases where one specific true label is more desirable than others, or where the costs to one type of error outweigh other types, a relaxation of equalized odds is applicable. This is referred to as equal opportunity, and consists of imposing only one of the constraints necessary for equalized odds (i.e. only for the case when $Y = +1$).

A substantial difference between the notions of equalized odds/equal opportunity and disparate impact is the directional impact that they have on a classifier's accuracy. The classifier is largely limited in its accuracy by the relationship between Z and Y in the case of disparate impact; if these two are highly correlated, then the restriction visualized in fig. 3.1a will keep the accuracy of δ in predicting Y low as well. In the extreme case where they are perfectly correlated, the only feasible classifier will essentially be a completely uninformative one. However, equalized odds only requires that the ROC curves of d in predicting Y conditional on Z be approximately similar (this is apparent from the interpretation of eq. (3.4) above). Notably, they could both be perfect classifiers without violating equalized odds (or equal opportunity). In contrast to the notion of disparate impact, a randomized classifier will not necessarily satisfy the notion of equalized odds [131]. Consider an example where $\mathbb{P}[Y = +1|Z = +1] = 0.4$ and $\mathbb{P}[Y = +1|Z = -1] = 0.6$. Here, a random classifier would have a true-positive rate of 0.4 for protected class $Z = +1$, but only 0.6 for protected class $Z = -1$ (and the opposite for false-negatives). Thus, a highly-restrictive constraint on disparate impact can push a classifier towards being uninformative, while a constraint on equalized odds instead could be attained by randomizing to reduce informativeness for all protected classes down to that of the protected class with the least informative classifier (in a point-wise manner across all thresholds).

Extention to General Supervised Learning

In the above, we have used the simplistic case wherein Y and Z are both binary and univariate. This may be easily generalized by extending the definitions give for disparate impact, equalized odds, and equal opportunity to account for the expanded sets of possible values. For example, this would mean extending disparate impact to requiring that

$$\left|\mathbb{P}\left[d(X,t) = y \middle| Z = z_1\right] - \mathbb{P}\left[d(X,t) = y \middle| Z = z_2\right]\right| \leq \Delta, \ \forall t \in \mathbb{R}; y \in \mathcal{Y}; z_1, z_2 \in \mathcal{Z}. \quad (3.5)$$

In the case of continuous $\mathcal{Y}$ or $\mathcal{Z}$, this can be computed by a simple discretization. This is particularly helpful for situations like fair regression. Note that all of the main intuitions developed throughout the previous subsections of this section still apply.

In important note should be made about the granularity of the set $\mathcal{Z}$. If there are multiple dimensions along which protected classes may be defined (i.e. race, gender, ethnicity, age, etc.), it is important that any fairness metric be measured with respect to all combinations of protected classes [129]. For example, ensuring a certain level of disparate impact with regards to women vs. men and minorities vs. non-minorities does not enforce that same level of fairness for minority women with respect to other sub-classes. In the FO formulation, this is why we must consider the tensor product $Z^{\otimes m}$ instead of simply using an entry-wise product. This is exemplified in the following example.

Example 5. Consider a situation where there are two dimensions of protections: protections based on ethnicity (A or B) and protections based on sex (M or F). Assume that the population breakdown is half of sex M and half of sex F, and 70% of ethnicity A and 30% of

		Ethnicity		
		A	B	Total
Gender	M	0.2	0.3	0.5
	F	0.5	0.0	0.5
	Total	0.7	0.3	1.0

Table 3.1: A simple example showcasing the possibility of "fairness gerrymandering"

ethnicity B, with no correlation between between ethnicity and sex. Now, consider a classifier that decides some desirable outcome (i.e. receiving a loan, being hired, etc.), with the proportions of those receiving the positive outcome that come from each protected class outlined in table 3.1. In this case, the classification satisfies a disparate impact level of zero for both ethnicity and sex, but does clearly not do so when considering the interaction of these two.

3.4 Models Considered

Throughout this chapter, we consider the FO-formulation of the standard soft-margin SVM, with a single protected attribute with respect to which we want to be fair. Considering the formulation introduced in Chapter 2, this amounts to a problem of the following form

$$\begin{aligned} \min\ & \frac{\lambda}{2}\|\beta\|_2^2 + E_n\left(\left(Y\cdot\beta^\mathsf{T}X\right)_+\right) \\ \text{s.t. }& \left|\mathbb{E}_n\left(Z^m\cdot(\beta^\mathsf{T}X)^q\right) - \mathbb{E}_n\left(Z^m\right)\cdot\mathbb{E}_n\left((\beta^\mathsf{T}X)^q\right)\right| \le \Delta_{m,q}, \\ & \qquad \text{for } (m,q)\in[\mathfrak{g}]\times[\mathfrak{h}]. \end{aligned} \tag{3.6}$$

Here, $\beta\in\mathbb{R}^p$ defines the score function $\beta^\mathsf{T}X$ (we assume for simplicity that one element of X is a contant value of 1, associated with a shift term in the score function), and λ is a standard regularization parameter. Note that this follows simply from the formulation in Chapter 2 by setting $B=\beta$, w to the identity for inputs X and to the zero function for inputs Z, and $R(\delta)=\left(Y\cdot\beta^\mathsf{T}X\right)_+$, in addition to the structural risk minimization term $\lambda\|\beta\|_2^2$. A straightforward generalization of theorem 4 implies that the consistency properties of FO are not altered by the addition of a smooth, convex regularizer like this.

Alternatively, extending this to regression yields the form

$$\begin{aligned} \min\ & E_n\left(\left(Y-\beta^\mathsf{T}X\right)^2\right) \\ \text{s.t. }& \left|\mathbb{E}_n\left(Z^m\cdot(\beta^\mathsf{T}X)^q\right) - \mathbb{E}_n\left(Z^m\right)\cdot\mathbb{E}_n\left((\beta^\mathsf{T}X)^q\right)\right| \le \Delta_{m,q}, \\ & \qquad \text{for } (m,q)\in[\mathfrak{g}]\times[\mathfrak{h}]. \end{aligned} \tag{3.7}$$

This formulation is convenient because it removes concerns about multi-dimensionality of Z and outputs $\delta(X)$. Intuitions and interpretations provided in following sections will use the FO SVM formulation (3.6) as an example, but are largely generalizable to other FO formulations as well.

Finally, we also consider extensions of the FO hierarchy to deep-learning classifiers. In this case, our R the a simple logistic loss function commonly used in deep-learning classifiers. Since these lack the convenient convex formulation of the previous techniques, many of the interpretations and intuitions given do not easily extend to this setting. Indeed, this reflects the fact that there is currently little theory for the operation and learning process of deep-learning models, as well as why stochastic gradient descent seems to work well on these models. We are also required to include our fairness constraints as penalizations in the objective when extending this to the deep-learning setting, as handling constraints is not trivial. This is expanded on further in Section 3.8. While we are not able to provide the same theoretical intuitions for the deep-learning formulation as we are with the SVM or regression formulations, we provide experimental results to show that the same general ability to improve fairness hold in practice.

3.5 Interpretations

In this section, we provide a series of interpretations for our fairness constraints. Specifically, we view them as approximations to a bi-level programming problem, as implementations of a Maximum A Posteriori (MAP) formulation, and as strategic modifications made in a dual space. Finally, we briefly discuss comparisons between the FO hierarchy and the Lasserre hierarchy, which is another optimization hierarchy that can be thought of as optimizing over distributions.

Polynomial Approximations to Bi-level Programming

Note that the order-(1,1) FO for eq. (3.6) is

$$\begin{aligned} \min\ & \frac{\lambda}{2}\|\beta\|_2^2 + E_n\left(\left(Y\cdot\beta^\mathsf{T}X\right)_+\right) \\ \text{s.t.}\ & \left|\beta^\mathsf{T}\left(E_n(ZX) - E_n(Z)\cdot E_n(X)\right)\right| \leq \Delta_{1,1}. \end{aligned} \tag{3.8}$$

This is equivalent to bounding the correlation between Z and the score function (element-wise in Z), and reflects a popular approach in the in-training fairness literature [25, 191, 261, 270]. [191] also show that it has an interesting interpretation in the context of bi-level optimization. Specifically, if $\beta^\mathsf{T}X$ is assumed to be bounded over some range $[-M, M]$, then the order-(1,1) FO constraint is the convex relaxation that arises from upper- and lower-bounding the indicator functions implicit in eq. (3.2) with linear functions of the form $\text{ubound}(x) = \frac{1}{M}x + 1$ and $\text{lbound}(x) = \frac{1}{M}x$, respectively. Similarly, second-order constraints can be thought of as bounds of the form

$$\max\left\{\frac{1}{8}-\frac{1}{2M}\left(x-\frac{M}{2}\right)^2,\ -\frac{1}{8}+\frac{1}{2M}\left(x+\frac{M}{2}\right)^2\right\}\leq \mathbf{1}\{x\geq 0\}$$
$$\leq \min\left\{\frac{9}{8}-\frac{1}{2M}\left(x-\frac{M}{2}\right)^2,\frac{7}{8}+\frac{1}{2M}\left(x+\frac{M}{2}\right)^2\right\}$$

Higher-order constraints can similarly be translated to consecutively tighter higher-order polynomial bounds on the disparate-impact constraint 3.2. While these bounds, beyond the linear bounds, are no longer convex, they are more easily handled through existing optimization techniques than the original infinite-dimensional fairness constraint requiring independence.

Information Projections

In the case where a single predictions is made with binary protected attribute, we present a result that bounds the bias of a predictor. This result provides a guarantee of fairness in terms of the distance of the resulting conditional distributions of the activation function to the exponential family of distributions. Here, $\mathcal{KL}$ refers to the Kullback-Liebler divergence.

Theorem 6. *Let Z be a binary protected attribute, and let $\mathcal{P}_+$ and $\mathcal{P}_-$ be the distributions of $\delta(X)|Z=+1$ and $\delta(X)|Z=-1$ for some univariate function δ such that $E_{\mathcal{P}_+}(\delta(X)^q)=E_{\mathcal{P}_-}(\delta(X)^q)$ for $1\leq q\leq \mathfrak{h}$. Furthermore, let $\mathcal{Q}$ be the member of the exponential family of distributions with sufficient statistics $T(w)=[w^q]_{q=1}^{\mathfrak{h}}$ and such that the first $\mathfrak{h}$ moments of $\mathcal{Q}$ are the equal to those of $\mathcal{P}_+$ and $\mathcal{P}_-$. Then, the disparate impact of δ, $KS(\delta)$, is bounded by:*

$$KS(\delta)\leq\sqrt{\frac{1}{2}\left(\mathcal{KL}(\mathcal{P}_+||\mathcal{Q})+\mathcal{KL}(\mathcal{Q}||\mathcal{P}_-)\right)}. \tag{3.9}$$

Remark 12. This result implies that we can obtain better guarantees on fairness the closer our distributions $\mathcal{P}_+$ and $\mathcal{P}_-$ end up being to the exponential family of distributions. Furthermore, increasing the number of parameters used to define a subset of the exponential family of distributions leads to a larger subset; Since $\mathcal{Q}$ is also a minimizer of $\mathcal{KL}(\mathcal{P}_+||\mathcal{Q})$, this intuitively implies that Theorem 6 will become tighter as the number of fairness constraints is increased.

Proof. First, note that the left-hand-side of eq. (3.9) can be upper-bounded by:

$$\max_{t\in\mathbb{R}}|\mathcal{P}(\delta(X)>t|Z=+1)-\mathcal{P}(\delta(X)>t|Z=+1)|. \tag{3.10}$$

This follows from the fact that any classification decision is simply a thresholding of an activation function. Note, then, that eq. (3.10) merely describes the Total Variation distance

between $\mathcal{P}_+$ and $\mathcal{P}_-$. By Pinsker's inequality, this can be upper-bounded by $\sqrt{\frac{1}{2}\mathcal{KL}(\mathcal{P}_+||\mathcal{P}_-)}$. Finally, it remains to show that $\mathcal{KL}(\mathcal{P}_+||\mathcal{P}_-) \leq \mathcal{KL}(\mathcal{P}_+||\mathcal{Q}) + \mathcal{KL}(\mathcal{Q}||\mathcal{P}_-)$. This follows from the fact that $\mathcal{Q}$ is the information projection of $\mathcal{P}_-$ onto the linear space of distributions whose first $\mathfrak{h}$ moments match those of $\mathcal{P}_-$ (in fact, it holds with equality) [189, 255].

□

Maximum A Posteriori Estimation

Just as a many standard statistical learning methods can be interpreted through the lens of MAP estimation, FO can be understood as a Bayesian estimation method that takes advantage of exogenously-known relationships amongst variables. Consider a model $Y = \delta(X) + \varepsilon$, where (X, Y) has some joint distribution $f_{X,Y}$ and conditional distribution $f_{Y|X}(y|x)$, and ε is a noise term for which we know the distribution. Furthermore, suppose the existence of a protected attribute Z with distribution f_Z such that it is exogenously known that $f_{Y,Z|X}(y, z|x) = f_{Y|X}(y|x)f_Z(z) \forall x, y, z$. By the tower property and Bayes' Theorem,

$$P(\delta|Y, X) = E_Z\left[\frac{P(Y|\delta, X, Z)P(\delta|X, Z)}{P(Y|X, Z)}\right]. \tag{3.11}$$

The relationship between Z and X, Y would motivate the choice of $P(\delta|X, Z)$ such that $\delta(X) \perp Z$ to maximize the overall likelihood term $P(\delta|Y, X)$. This is precisely the fairness constraint 3.2. This means that $P(Y|\delta, X, Z) = P(Y|\delta, X)$, and so we re-obtain the standard MAP, $P(\delta|Y, X) \propto P(Y|\delta, X)P(\delta|X)$.

SVM Duality

MAP interpretations of SVM are not trivial [103, 236], so we also provide intuition for fairness constraints in terms of the dual formulation of SVM. Consider the order-(1,1) FO SVM, parametrized by hyperparameter λ controlling the ℓ_2 regularization term and with $\Delta_{1,1} = 0$. For simplicity, let $\phi = E_n(ZX) - E_n(Z) \cdot E_n(X)$. Then, the fairness constraint may be dualized with associated dual variable γ, and as such may be written as

$$\min \frac{\lambda}{2}\|\beta\|_2^2 + E_n\left(\left(Y \cdot \beta^\mathsf{T} X\right)_+\right) + \gamma \cdot \beta^\mathsf{T}\phi\phi^\mathsf{T}\beta.$$

Standard applications of Lagrangian duality for SVM show that, in the case of unconstrained SVM, the optimal separator β can be written as a weighted combination $\sum_{i=1}^n s_i Y_i X_i$ for some variables $0 \leq s_i \leq \frac{1}{\lambda n}$ [250]. As $\frac{\gamma}{\lambda} \to \infty$, the matrix $I - \frac{\gamma/\lambda}{1+\gamma/\lambda}\phi\phi^\mathsf{T}$ approaches the projection matrix $I - \phi\phi^\mathsf{T}$. Then, by Lagrangian duality, we have that the optimality condition $\left(\lambda I + \gamma\phi\phi^\mathsf{T}\right)\beta = \sum_{i=1}^n s_i Y_i X_i$ implies

$$\begin{aligned}\beta &= \left(\lambda I + \gamma\phi\phi^\mathsf{T}\right)^{-1}\sum_{i=1}^n s_i Y_i X_i \\ &= \frac{1}{\lambda}\left(I - \frac{\gamma/\lambda}{1+\gamma/\lambda}\phi\phi^\mathsf{T}\right)\sum_{i=1}^n s_i Y_i X_i.\end{aligned}$$

Effectively, as γ grows with respect to λ, the ϕ-component of the support vectors that comprise β are increasingly disregarded. Note that, by its design, ϕ represents the correlations between Z and each of the covariates in X. So, the component that would most increase the first-order interaction term, or correlation, is steadily minimized by increasing γ. Similar intuition can be extended to higher-order interaction terms.

Lasserre Hierarchy

Sum of Squares (SoS) optimization, also known as the Lasserre hierarchy [7, 146], is another hierarchy of convex optimization problems aimed at solving polynomial optimization problems. In effect, the hierarchy attempts to obtain reconstructions of potentially non-convex polynomials as sums of squares of other polynomials (referred to as *sum-of-squares proofs*), thereby proving non-negativity (this result is referred to as the *Positivstellensatz* [139]). Higher orders of the hierarchy include higher-order polynomials through which to construct these SoS proofs. Interestingly, the dual problem to this is to obtain distribution-like constructs called *pseudodistributions*, which act like distributions over the feasible region and effectively certify the non-inclusion of a polynomial in the set of sum-of-squares polynomials. Notably, this is done by explicitly treating the moments of the variables under such a pseudodistribution as optimization variables, and it can be shown that these pseudodistributions are real distributions when enough moments are considered in the optimization problem. While the Lasserre hierarchy describes a critically different paradigm from FO, it has ideological connections which warrant mention. While FO optimizes over distributions (in particular, that of $\delta(X, Z)$) to satisfy some moment bounds, SoS can be understood to optimize over moments to (ultimately) obtain a conforming distribution. As such, critical results in SoS may have analogs in, and can lend inspiration for, further work in FO. For example, much work has been done to prove the effectiveness of the Lasserre hierarchy in providing approximate solutions which can then be rounded to exact solutions [19, 113, 203], and the tools arising from this analysis can prove helpful in analyzing similar issues in the framework of FO.

3.6 Kernel Transformations

We note that FO generalizes easily to kernel methods. In this section, we consider the particular context of SVM, although equivalent extensions also exist for other margin-based learning methods. Furthermore, while the results of this section are also presented for scalar-valued Z to simplify notation, note that they easily generalize to $p > 1$. Recall that the *kernel trick* for SVM comprises of solving the dual optimization:

$$\begin{aligned} \min\ & \tfrac{\lambda}{2}(\vec{Y}\circ\alpha)^\mathsf{T} K(\vec{X},\vec{X})(\vec{Y}\circ\alpha) - \textstyle\sum_{i=1}^n \alpha_i \\ \text{s.t.}\ & \sum_{i=1}^n \alpha_i Y_i = 0 \\ & 0 \le \alpha_i \le \frac{1}{n}, \qquad \text{for } i \in [n], \end{aligned} \tag{3.12}$$

where $\vec{Y}\circ\alpha$ denotes the element-wise multiplication of Y and dual variables α. When $K(x,x') = x^\mathsf{T}x'$, this is equivalent to linear SVM. The advantage of the kernel trick is that complicated transformations of the data matrix X can be represented easily in terms of the kernel K without increasing computational complexity. In this case, the margin function $\delta(x) = \sum_{i=1}^n \alpha_i^* Y_i K(X_i, x) - b^*$, where α^* is the optimal solution to problem 3.12 and $b^* = \delta(X_i) - Y_i$ for any i such that $0 < \alpha_i^* < \frac{1}{n}$. With these transformations in mind, it is clear that the fairness constraints in problem 2.17 can be rewritten as:

$$\left| \left\langle \frac{1}{n}\sum_{i=1}^n (Z_i - E(Z))^m K(\vec{X}, X_i)^{\otimes q}; Y\circ\alpha \right\rangle \right| \le \Delta_{m,q}^{\text{kernel}}, \qquad \text{for } (m,q) \in [\mathfrak{g}]\times[\mathfrak{h}]. \tag{3.13}$$

These would then be added to the above problem. Note that the same intuitions hold regarding the convexity of the problem when $\mathfrak{g} = 1$.

3.7 Computational Properties

This section will elaborate on some results formalized in Section 2.4, with an emphasis on simple examples in the realm of the FO SVM formulation. In particular, it will consider the computational properties of lower levels of the FO hierarchy. The FO(1,1) problem shown in Equation (3.8) is clearly still a convex problem, and can be easily solved using off-the-shelf solvers. Furthermore, we note that constraints on higher-order interactions of the form $E\left[Z^m \cdot \beta^\mathsf{T}X\right]$ are also linear, and so easy to incorporate (as formalized in proposition 1). The difficulty arises in higher orders of $\mathfrak{h}$, which introduce further nonlinearity into the problem of selecting β. Here, we will first explore some efficient approaches for solving FO(2,2), and then proceed to a brief exposition of possible methods for solving higher-order FO problems.

Consider the FO(2,2) problem with selectively-dualized constraints:

$$\begin{aligned} \min\ & \frac{\lambda}{2}\|\beta\|_2^2 + E_n\left(\left(Y\cdot\beta^\mathsf{T}X\right)_+\right) + \mu_1 t_1 + \mu_2 t_2 \\ \text{s.t.}\ & \left|\beta^\mathsf{T}\left(E_n(ZX) - E_n(Z)\cdot E_n(X)\right)\right| \le \Delta_{1,1} \\ & \left|\beta^\mathsf{T}\left(E_n\left(Z\cdot XX^\mathsf{T}\right) - E_n(Z)\cdot E_n\left(XX^\mathsf{T}\right)\right)\beta\right| \le t_1 \\ & \left|\beta^\mathsf{T}\left(E_n\left(Z^2X\right) - E_n\left(Z^2\right)\cdot E_n(X)\right)\right| \le \Delta_{2,1} \\ & \left|\beta^\mathsf{T}\left(E_n\left(Z^2\cdot XX^\mathsf{T}\right) - E_n\left(Z^2\right)\cdot E - n\left(XX^\mathsf{T}\right)\right)\beta\right| \le t_2. \end{aligned} \tag{3.14}$$

As mentioned above, penalizing the interactions terms in the objective is functionally equivalent to explicit constraints. Especially with higher-order interaction terms, it may be beneficial in a practical setting to utilize the penalized representation, both in terms of interpretability and to avoid being too harshly restricting the feasible region. We also employ this rendition for purposes of later notational convenience. In Proposition 8, we describe a condition under which problem 3.14 can be solved through convex techniques.

Proposition 8. *Consider the FO(2,2) with ℓ_2 regularization. If $\frac{\lambda}{\mu_1+\mu_2} \geq 2\left\|E_n\left(Z \cdot XX^\intercal\right)\right\|_2$, then there exists a polynomial-time algorithm to solve this to global optimality.*

Proof. We first consider the case where $Z \in \{\pm 1\}$, which allows us to simplify the problem to

$$\begin{aligned} \min\ & \frac{\lambda}{2}\|\beta\|_2^2 + E_n\left(\left(Y \cdot \beta^\intercal X\right)_+\right) + \mu t \\ \text{s.t.}\ & \left|\beta^\intercal \phi\right| \leq \Delta_{1,1} \\ & \beta^\intercal E_n\left(Z \cdot XX^\intercal\right)\beta \leq t, \end{aligned} \tag{3.15}$$

where $\phi = E_n(ZX) - E_n(Z) \cdot E_n(X)$. Note that the final inequality is two-sided in the original formulation; however, at most one of these two inequalities will ever be tight, so the problem (3.15) can simply be run twice, once with each constraint, to achieve the same result. Thus, we may proceed with the problem (3.15) without loss of generality.

Now, consider the eigenvalue decomposition $E_n\left(Z \cdot XX^\intercal\right) = VDV^\intercal$, where V is unitary and D is diagonal. We may make the transformation $w = V\beta$ and, since $\|V\beta\|_2 = \|\beta\|_2$, we may instead solve the following

$$\begin{aligned} \min\ & \frac{\lambda}{2}\|w\|_2^2 + E_n\left(\left(Y \cdot w^\intercal VX\right)_+\right) + \mu t \\ \text{s.t.}\ & \left|\beta^\intercal V\phi\right| \leq \Delta_{1,1} \\ & \sum_{i=1}^{d} D_{ii} w_i^2 \leq t \end{aligned} \tag{3.16}$$

A standard lifting argument yields a convex relaxation of this.

$$\begin{aligned} \min\ & \frac{\lambda}{2}\mathbf{e}^\intercal u + E_n\left(\left(Y \cdot w^\intercal VX\right)_+\right) + \mu t \\ \text{s.t.}\ & \left|\beta^\intercal V\phi\right| \leq \Delta_{1,1} \\ & \sum_{i=1}^{d} D_{ii} u_i \leq t \\ & w_i^2 \leq u_i, i = 1, \ldots, d. \end{aligned} \tag{3.17}$$

When $\frac{\lambda}{\mu} \geq 2\left\|E_n\left(Z \cdot XX^\intercal\right)\right\|_2$, it is clear that, for any value of w, the objective value can always be improved by decreasing any element of u such that $w_i^2 < u_i$. Thus, we will

have $u_i = w_i^2$, and the result follows for the case of binary Z. For non-binary Z, we can simply restate the same argument to account for the second quadratic constraint that arises. □

The intuition behind Proposition 8 is that the convexity of the regularization term can counteract the non-convexity of the interaction term, if it is weighted highly enough. In our empirical studies, we find that smaller values of μ tend to yield the best trade-off between fairness and accuracy, which suggests that this condition may be of use in a practical setting. Similar relaxations based on Semidefinite Programming (SDP) can be applied even when the assumption of Proposition 8 is not satisfied, but these necessarily become weak relaxations in these cases, and so often do not produce good solutions.

Alternative techniques exist for solving FO(2,2) problem (3.14) in these cases, albeit only to local optimality. The non-convexity of FO(1,2) or FO(2,2) arises wholly from the constraints that involve order-2 moments of the feature vectors, X, and each of these can be decomposed into a difference of convex functions. For example, consider the following decomposition of the higher-order interaction term in the FO(1,2) problem:

$$E_n\left(Z \cdot XX^\mathsf{T}\right) - E_n(Z) \cdot E_n\left(XX^\mathsf{T}\right) = VDV^\mathsf{T}. \tag{3.18}$$

Let $U_+ = V \max\{D, 0\}V^\mathsf{T}$ and $U_- = V \max\{-D, 0\}V^\mathsf{T}$. Then, it is clear that

$$\beta^\mathsf{T}\left(E_n\left(Z \cdot XX^\mathsf{T}\right) - E_n(Z) \cdot E_n\left(XX^\mathsf{T}\right)\right)\beta = \beta^\mathsf{T}U_+\beta - \beta^\mathsf{T}U_-\beta, \tag{3.19}$$

where both $\beta^\mathsf{T}U_+\beta$ and $\beta^\mathsf{T}U_-\beta$ are convex in β. Then, the iterative Convex-Concave Procedure [235, 249, 268] may be applied to FO(1,2) (or equivalently, to FO(2,2)) in order to derive locally optimal solutions. This proceeds by turning the non-convex constraint into two convex constraints by linearizing each term in one of the resulting constraints. For example, consider the following rendition of FO(1,2)

$$\begin{aligned}
\min\ & \tfrac{\lambda}{2}\|\beta\|_2^2 + E_n\left(\left(Y \cdot \beta^\mathsf{T}X\right)_+\right) + \mu \cdot t \\
\text{s.t.}\ & -\Delta_{1,1} \le \beta^\mathsf{T}\left(E_n(ZX) - E_n(Z) \cdot E_n(X)\right) \le \Delta_{1,1} \\
& \beta^\mathsf{T}U_+\beta - \beta_k^\mathsf{T}U_-\beta_k - 2\beta_k^\mathsf{T}U_-^\mathsf{T}(\beta - \beta_k) \le t \\
& \beta^\mathsf{T}U_-\beta - \beta_k^\mathsf{T}U_+\beta_k - 2\beta_k^\mathsf{T}U_+^\mathsf{T}(\beta - \beta_k) \le t
\end{aligned} \tag{3.20}$$

Here, it is clear that $-\beta_k^\mathsf{T}U_-\beta_k - 2\beta_k^\mathsf{T}U_-^\mathsf{T}(\beta - \beta_k)$ and $-\beta_k^\mathsf{T}U_+\beta_k - 2\beta_k^\mathsf{T}U_+^\mathsf{T}(\beta - \beta_k)$ are the linearizations of $\beta^\mathsf{T}U_-\beta$ and $\beta^\mathsf{T}U_+\beta$, respectively, at some β_k, thus making the above a convex problem. This subproblem is then solved efficiently in iteration k, to obtain an optimal solution β_{k+1}^*, which is used as β_{k+1} in iteration $k+1$. Note that this procedure may be extended to any number of convex-concave constraints, and we have only included one for notational simplicity. The following result formalizes this method's convergence to locally optimal solutions.

Table 3.2: List of Datasets Used in Numerical Experiments

Dataset	p	n	Z Type	Task	Source
Adult Income	58	32561	Binary	Classification	[154]
Biodeg	40	1055	Categorical	Classification	[167]
Communities	96	1994	Continuous	Regression	[72–74, 210]
EEG	12	4000	Binary	Regression	[88]
Energy	8	768	Categorical	Classification	[246]
German Credit	49	1000	Continuous	Classification	[154]
Letter	15	20000	Continuous	Classification	[93]
Music	68	1034	Continuous	Regression	[275]
Parkinson's	18	5875	Binary	Classification	[154]
Pima	7	768	Continuous	Classification	[234]
Recidivism	6	5278	Binary	Classification	[11]
SkillCraft	17	3338	Continuous	Both	[243]
Statlog	35	3486	Binary	Classification	[154]
Steel	25	1941	Categorical	Classification	[154]
Taiwan Credit	22	29623	Binary	Classification	[266]
Wine Quality	11	6497	Binary	Both	[65]

Theorem 7 ([235]). *The CCP problem (3.20) gives iterates β_k that converge to a local minimum.*

Such a method does not easily extend to higher-order constraints, however. The difficulty arises due to vagueness and complexity of calculating the analog of eigenvalue decompositions for tensors of order greater than 2. While it has been shown that any polynomial can be decomposed into the sum of convex and concave components [8,256], even determining whether two given polynomials g and $f - g$ are a convex-concave decomposition of a polynomial f is strongly NP-Hard [8]. For higher-order problems, we are restricted to existing relaxation methods for optimization over higher-order polynomials. Of these, the two most popular methods are Sum of Squares optimization (introduced earlier and also referred to as the Lasserre hierarchy) [7, 146], as well as the Relaxation-Linearization Technique (RLT) [231]. Both operate by lifting arguments; SoS relaxes the problem of polynomial optimization to one of optimization over structures approximating probability distribution and formulates this as an SDP, while RLT relaxes this further by removing some semi-definiteness constraints. Full exposition of these methods is beyond the scope of this thesis, but we refer the reader to the sources cited for tutorials of both methods.

3.8 Numerical Experiments

In this section, we implement various levels of the FO problem (2.17) for: classification, regression, and decision-making (recall Proposition 3 says that for binary Z, we only need to try the level-$(1, \mathfrak{h})$ FO since level-$(\mathfrak{g}, \mathfrak{h})$ is equivalent to level-$(1, \mathfrak{h})$ in this case). Classification

tasks are defined both for convex-margin classifiers (in particular SVM, unless mentioned otherwise) as well as deep-learning classifiers. Unless mentioned otherwise, classifier accuracy is measured by area-under-the-curve (AUC), which represents the area under the receiver operator characteristic (ROC) curve for a given binary classifier. Regression accuracy is measured by mean-squared-error (MSE), unless indicated otherwise. In all cases, fairness is measured using disparate impact (3.2). All results are averaged over 50 iterations, where 70% of the data is used for training and the 30% is used for measuring the aforementioned metrics (reshuffled every iteration). Unless mentioned explicitly, all hyperparameters are chosen using 10-fold cross-validation. Unless otherwise noted, all experiments were carried out using the Mosek 8 optimization package [184]. We first present convex classification and regression implementations of FO on a series of datasets from the UC Irvine Machine Learning Repository [154], the full list of which is in Table 3.2. Next, we investigate the sensitivity of FO to various choices of hyperparameters. We then present similar results on a few datasets when extending the fairness constraints in FO to a deep-learning context. Finally, we present a case study on the use of FO to perform fair morphine and heparin dosing.

Comparison Methods

In the following subsections, we compare FO to three other methods. The methods of [24] and [125] are designed for fair classification and fair regression, respectively, and are similar to our method in that they enforce fairness at training time. We also compare FO to the method of [48], although this takes a pre-processing approach.

Berk et al. [24] The method of [24] is one of the few comparable methods for fair regression. They also take an in-training approach, defining two regularization terms that enforce fairness. Let $P_z = \{i \in [n] : Z_i = z\}$, and note $\#P_z$ refers to the cardinality of these sets. Given a binary protected attribute Z, they define a regularizer for group fairness

$$\left((\#P_{-1} \cdot \#P_{+1})^{-1} \sum_{i \in P_{-1}} \sum_{j \in P_{+1}} d(Y_i, Y_j) \cdot (X_i^\mathsf{T}\beta - X_j^\mathsf{T}\beta)\right)^2, \tag{3.21}$$

for some distance measure $d(\cdot, \cdot)$. Note that this is similar to the term constrained in FO for $(m, q) = (1, 1)$. They also define the following regularizer for individual fairness:

$$(\#P_{-1} \cdot \#P_1)^{-1} \sum_{i \in P_{-1}} \sum_{j \in P_{+1}} d(Y_i, Y_j) \cdot (X_i^\mathsf{T}\beta - X_j^\mathsf{T}\beta)^2. \tag{3.22}$$

This term is similar to a term in FO for $(m, q) = (1, 2)$, although not equivalent. It has the benefit of being convex, although the double-summation term can be computationally prohibitive for large datasets. In our implementation, we estimate this term from a sub-sample (10%) of the data when this issue arises. Since the first term is similar to a term arising in FO, we implement it as a constraint instead and require that it is equal to zero, as we do with the analogous term in FO. Finally, we note that this method can only accommodate binary-valued protected attributes, so we cannot provide comparisons to many of the datasets we use for fair regression.

Calmon et al. [48] This work is comparable to that of [271]. Both of these works formulate nonparametric optimization problems whose solution yields a conditional distribution $f_{\widehat{X},\widehat{Y}|X,Y,Z}$ that then probabilistically transforms the data. We only compare our method to the approach introduced in [48], since their formulation directly builds on that of [271].

Given a predefined notion of deviation amongst distributions, this method minimizes the overall deviation of $f_{\widehat{X},\widehat{Y}}$ from $f_{X,Y}$. In the original work, the authors chose to minimize $\frac{1}{2}\sum_{x,y}|f_{\widehat{X},\widehat{Y}}(x,y) - f_{X,Y}(x,y)|$. They also include constraints on pointwise distortion $E_{\hat{X},\hat{Y}|X,Y}[\theta((X,Y),(\hat{X},\hat{Y})]$ for some user-defined function $\theta : \{\mathbb{R}^p \times \{\pm 1\}\}^2 \to \mathbb{R}_{\geq 0}$. There are also bounds on the dependency of the new main label $\widehat{Y}$ on the original protected label $J(f_{\widehat{Y}|Z}[y|z], f_Y(y))$, where $J(a,b) = |\frac{a}{b} - 1|$ is defined to be the probability ratio measure. Thus, the final formulation is

$$\begin{aligned} \min\ & \tfrac{1}{2}\textstyle\sum_{x,y}|f_{\widehat{X},\widehat{Y}}(x,y) - f_{X,Y}(x,y)| \\ \text{s.t. } & \mathbb{E}_{\widehat{X},\widehat{Y}|X,Y}[\theta((X,Y),(\widehat{X},\widehat{Y})|x,y] \leq c, \quad \text{for all } x,y \\ & |f_Y(y)^{-1} f_{\widehat{Y}|Z}[y|z] - 1| \leq d, \qquad\quad \text{for all } y,z \\ & f_{\widehat{X},\widehat{Y}|X,Y,Z} \text{ are all distributions.} \end{aligned} \tag{3.23}$$

Following the procedure used by the authors, we approximate $f_{X,Y,Z}$ with the empirical distribution of the original data, separated into a pre-selected number of bins. Note that the resulting optimization problem will have $8(\#\text{bins})^{2p}$ parameters, which can quickly become computationally infeasible when the dataset is high-dimensional. To account for this, we again follow the original work and choose the 3 features most correlated with the main label Y. Each dimension is split into 8 bins. We choose $\theta((x',y'),(x,y))$ to be 0 if $y = y'$ and $x = x'$, 0.5 if $y = y'$ and x, x vary by at most one in any dimension, and 1 otherwise: This is similar to the θ chosen in the original paper itself. Finally, c and d were set at 0.1 and 0.3, respectively.

Kamishima et al. [125] Another comparable method is that of [125], which also aims to enforce fairness at training time. As opposed to our approach of bounding interaction moments, they instead regularize with a mutual information term. Also, this method differs from our framework notably in that it imposes different treatments for different protected classes, violating the principle of individual fairness; as a result, it is also unable to handle continuous protected attributes. The authors implement their regularizer in the context of logistic regression. Let σ be a sigmoid function and $g_\beta[y|x,z] = y\sigma(x^\mathsf{T}\beta) + (1-y)(1-\sigma(x^\mathsf{T}\beta))$, and note that the notation β_z indicates that this approach has a different set of coefficients for each possible value of Z. the authors approximate the mutual information as

$$n^{-1}\sum_{i=1}^n \sum_{y\in\{\pm 1\}} g_{\beta_{Z_i}}[y|X_i,Z_i] \log \frac{\widehat{P}[y|Z_i]}{\widehat{P}(y)}, \tag{3.24}$$

with $\widehat{P}[y|z] = (\#P_z)^{-1}\sum_{i\in P_z} g_{\beta_z}[y|X_i,z]$ and $\widehat{P}(y) = \frac{1}{n}\sum_{i=1}^n g_{\beta_{z_i}}[y|X_i,Z_i]$. This is then weighted and added to the objective as a regularizer. We include this method as a comparison

to our fair SVM, while noting the core differences mentioned above. We also note that the logarithmic terms cause some numerical instability and make it difficult to implement this method in standard solvers [184]. All convex-margin classifier experiments for this method were done using the sequential least squares programming approach [136], with up to 500 iterations. A weight of 0.1 was used for the fairness regularization term.

Note that this does not generalize to a stochastic gradient descent setting: To account for this, we calculate $\widehat{P}[y|Z]$ and $\widehat{P}(z)$ per batch for deep learning experiments. In these cases, we consider several hyperparameter choices. During training, the logarithmic terms caused notable numerical instability and somewhat divergent results for some hyperparameter choices, so we ran each experiment five times and present the best result.

Penalization Formulation of FO Hierarchy

It can be advantageous from a numerical computation standpoint to solve the FO problem (2.17) where some of the constraints are included as a penalty function in the objective. Specifically, consider the level-$(\mathfrak{g}, \mathfrak{h})$ FO penalty formulation (presented in the more general notation introduced in Chapter 2)

$$\begin{aligned} \min_{B\in\mathcal{B};t_{m,q}\in\mathbb{R}_{\geq 0}} \; & R_n(B\cdot\omega(x,z)) + \textstyle\sum_{(m,q)\in I}\mu_{m,q}t_{m,q} \\ \text{s.t.} \; & \left\|\mathbb{E}_n\big(Z^{\otimes m}(B\Omega)^{\otimes q}\big) - \mathbb{E}_n\big(Z^{\otimes m}\big)\otimes\mathbb{E}_n\big((B\Omega)^{\otimes q}\big)\right\| \leq t_{m,q}, \\ & \text{for } (m,q)\in[\mathfrak{g}]\times[\mathfrak{h}]. \end{aligned} \tag{3.25}$$

where $I \subseteq [\mathfrak{g}]\times[\mathfrak{h}]$ is a subset of the indices. The numerical benefit of the penalty formulation is it makes finding an initial feasible point easier since there is no maximum bound on $\Delta_{m,q}$ for $(m,q)\in I$, whereas the original formulation of FO (2.17) involves constraints that must be satisfied for a fixed value of $\Delta_{m,q}$ for all $(m,q)\in[\mathfrak{g}]\times[\mathfrak{h}]$ in order to ensure feasibility. For the convex-margin methods that we consider in this section, the level-$(\mathfrak{g},1)$ formulations are convex optimization problems and are solved with $I=\emptyset$. For the level-$(1,2)$ formulations, we use $I=\{(1,2)\}$ and use the constrained convex-concave procedure [235,249,268] to solve the optimization problem. For deep-learning applications, all fairness constraints are included as penalty functions with a standard logistic loss, and the resulting problem is solved using standard stochastic gradient descent techniques.

Fair SVM

We first consider classification problems using a series of datasets, and formulate various hierarchies of fair SVM using the penalization formulation of FO (3.25) with $\Delta_{i,1}=0$ for $i=1,2,3$ and $\mu_{1,2}=1000$. The results are in Table 3.3. Since the mutual-information-based method of [125] cannot accommodate continuous protected classes, results are not reported for this method for the associated datasets. We note our method often improves fairness with less cost (in terms of accuracy) than the method of [48]. This is to be expected, as such pre-processing approaches do not take into account the downstream task that the transformed

Table 3.3: Classifier Comparison of Various Levels of FO to [48] and [125]

	SVM		FO(1,1)		FO(1,2)		FO(2,1)	
	AUC	KS	AUC	KS	AUC	KS	AUC	KS
Adult Income	0.894	0.333	0.875	0.232	0.628	0.106	–	–
Biodeg	0.917	0.294	0.914	0.236	0.788	0.142	0.911	0.228
Energy	0.528	0.114	0.525	0.113	0.529	0.120	0.525	0.110
German Credit	0.767	0.148	0.761	0.127	0.743	0.119	0.760	0.125
Letter	0.739	0.154	0.738	0.149	0.728	0.141	0.738	0.152
Parkinson's	0.643	0.157	0.642	0.157	0.617	0.079	–	–
Pima	0.821	0.161	0.807	0.136	0.729	0.131	0.807	0.136
Recidivism	0.727	0.286	0.559	0.060	0.577	0.049	–	–
SkillCraft	0.878	0.100	0.826	0.060	0.776	0.060	0.827	0.060
Statlog	0.998	0.331	0.992	0.331	0.937	0.216	–	–
Steel	0.764	0.127	0.763	0.124	0.757	0.107	0.763	0.118
Taiwan Credit	0.727	0.061	0.729	0.056	0.728	0.056	–	–
Wine Quality	0.798	0.273	0.788	0.103	0.778	0.069	–	–

	FO(3,1)		Calmon		Kamishima	
	AUC	KS	AUC	KS	AUC	KS
Adult Income	–	–	0.515	0.231	0.862	0.212
Biodeg	0.912	0.238	0.604	0.144	0.884	0.190
Energy	0.525	0.117	0.518	0.114	0.525	0.123
German Credit	0.760	0.125	0.630	0.113	–	–
Letter	0.738	0.151	0.648	0.185	–	–
Parkinson's	–	–	0.530	0.104	0.660	0.217
Pima	0.807	0.136	0.544	0.149	–	–
Recidivism	–	–	0.554	0.080	0.716	0.107
SkillCraft	0.826	0.060	0.625	0.069	0.871	0.083
Statlog	–	–	0.675	0.163	0.980	0.539
Steel	0.763	0.124	0.553	0.118	0.632	0.154
Taiwan Credit	–	–	0.745	0.068	0.741	0.072
Wine Quality	–	–	0.665	0.093	0.794	0.071

data is to be used for. Our method is also able to match or improve the fairness results of the mutual information approach. Recall that this method maintains explicitly different treatments for different protected classes, while ours adheres to the principle of individual fairness. Given this, it is unsurprising that the method of [125] can often achieve fairness at a lower cost to accuracy, although our method even outperforms on this metric for a number of datasets. Further, this feature of disparate treatments can yield fairness values notably worse than even a standard SVM. Finally, we note that our level-three interaction constraints provide only a marginal benefit.

Fair Regression

We next consider regression problems using another series of datasets, and implement various levels of the FO hierarchy for regression using the penalization formulation of FO (3.25) with

Table 3.4: Regression Comparison of Various Levels of FO to [24]

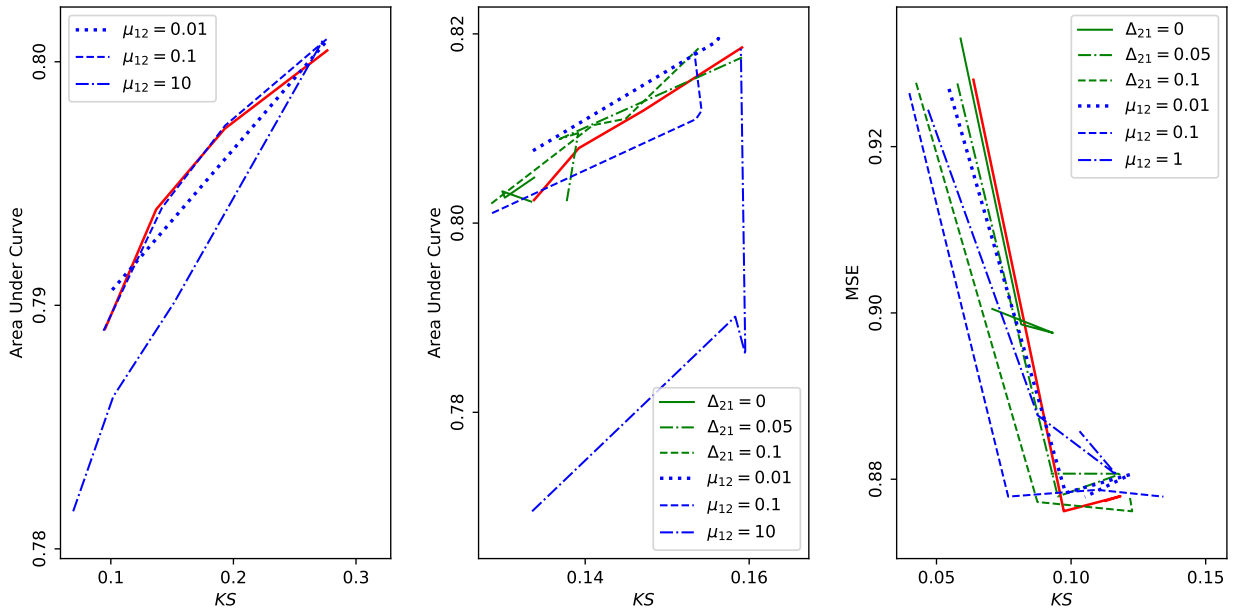
	OLS		FO(1,1)		FO(1,2)		FO(2,1)		FO(3,1)	
	MSE	KS	MSE	KS	MSE	KS	MSE	KS	MSE	KS
Commun.	0.355	0.414	0.836	0.128	0.897	0.049	0.843	0.107	0.851	0.116
EEG	0.062	0.085	0.069	0.086	0.201	0.080	–	–	–	–
Music	0.876	0.122	0.925	0.062	0.912	0.058	0.932	0.047	0.934	0.061
SkillCraft	0.930	0.025	0.960	0.008	0.960	0.007	0.964	0.007	0.967	0.007
Wine	0.712	0.285	0.734	0.080	0.737	0.065	–	–	–	–

*Note [24] yielded MSE of 0.069 and 0.734, and KS of 0.087 and 0.080 for the EEG and Wine datasets, respectively. Its formulation does not allow it to be used on the other datasets.

$\Delta_{i,1} = 0$ for $i = 1, 2, 3$ and $\mu_{1,2} = 0.1$. The results are displayed in Table 3.4. For the method of [24], the group fairness term is implemented as a constraint and required to be equal to 0, while the individual fairness term is left as a penalty in the objective, also with a weight of 0.1. As the method of [24] is unable to accommodate non-binary protected attributes, we only provide results for the appropriate datasets. Again, we note that our method is able to reduce the bias of a typical regression, often without considerable loss in accuracy. It again seems that lower levels of the hierarchy are largely sufficient.

Hyperparameter Sensitivity

We next explore the sensitivity of FO to the hyperparameters. Fig. 3.2 shows results for a selection of the datasets. Red lines are results of the level-(1,1) FO, green lines with the level-(2,1) FO, and blue lines with the level-(1,2) FO. Each line plots the average accuracy and fairness over 50 runs of the associated FO for $\Delta_{1,1} \in \{0.0, 0.05, 0.1, 0.2\}$, and $\Delta_{2,1}$ and $\mu_{1,2}$ indicated, as appropriate, in the legend. There is generally a negative tradeoff between fairness and accuracy as a function of $\Delta_{1,1}$, but higher levels of the FO hierarchy can flout this tradeoff: In all of the examples shown, it is possible to uniformly outperform the level-(1,1) FO in terms of both accuracy and fairness with higher levels of the hierarchy. This suggests fairness is not purely a detriment to predictive strength, and implies that loosely-enforced higher-order moment constraints can act as regularizers that protect against overfitting. The Pima Diabetes dataset in Fig. 3.2b is relatively small, which may explain the large benefit of fairness constraints as regularizers for that dataset; it is easier to overfit with a small dataset, and the enforcement of exogenously-available information (like independence of Y from Z) through regularization can provide a big benefit.



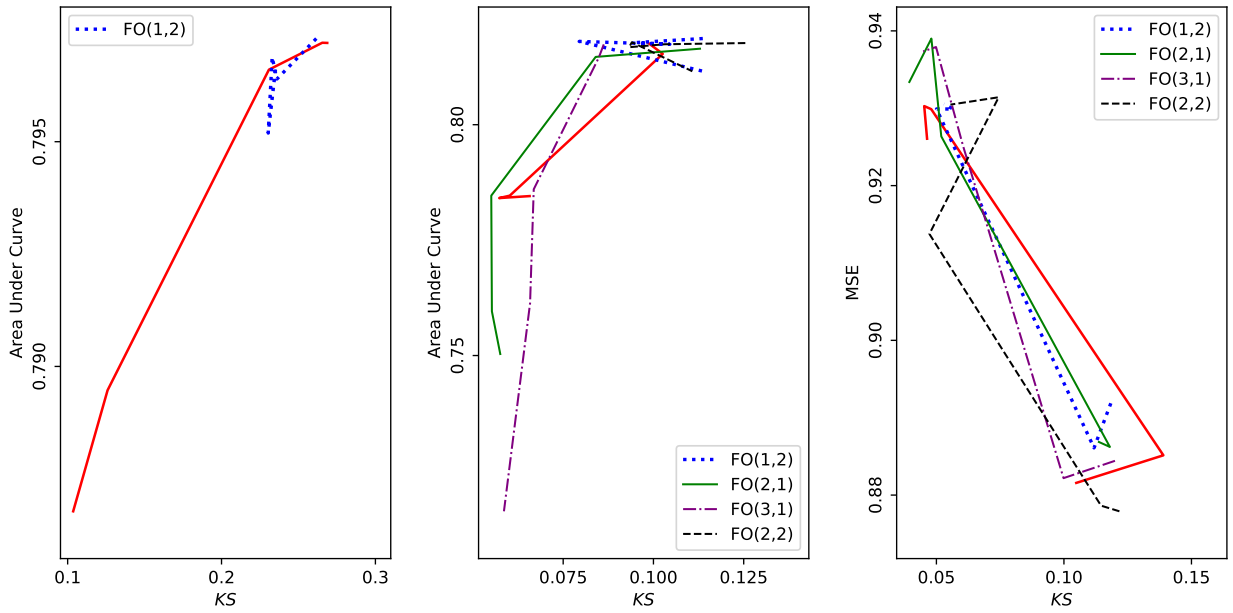
(a) Wine quality dataset (b) Pima Diabetes dataset (c) Music dataset

Figure 3.2: The sensitivity of FO to changes in its hyperparameters. Red lines indicate the level-(1,1) FO, green lines the level-(2,1) FO, and blue lines the level-(1,2) FO. Each line represents the evolution of accuracy and fairness for $\Delta_{1,1} \in \{0, 0.05, 0.1, 0.2\}$. Note that Figure 3.2c reflects an instance of fair regression, and so accuracy is measure via mean-squared error.

Single Parameter Tuning

One of the obvious questions with using the FO hierarchy is how to choose the tuning parameters. The theory associated with Theorems 3, 4, and 5 from Chapter 2 implies that in practice an FO problem with a small level-$(\mathfrak{g}, \mathfrak{h})$ should be used since $\mathfrak{g}$ and $\mathfrak{h}$ are required to grow very slowly at a double-logarithmic rate. Even if we derived faster bounds for these theorems that are based on α, these terms could grow no faster than a sub-logarithmic rate. Thus from a practical standpoint, the more relevant question is how to tune the $\Delta_{m,q}$, and potentially $\mu_{m,q}$ for $(m, q) \in I$, hyperparameters. Towards this end, Theorem 5 suggests one possible approach. For a level-$(\mathfrak{g}, \mathfrak{h})$ FO, consider the penalty formulation variant given by

$$\begin{aligned} &\min_{B \in \mathcal{B}} \; R_n(B \cdot \omega(x, z)) + \mu \cdot \epsilon \\ &\text{s.t.} \; \left\| \mathbb{E}_n\left(Z^{\otimes m}(B\Omega)^{\otimes q}\right) - \mathbb{E}_n\left(Z^{\otimes m}\right) \otimes \mathbb{E}_n\left((B\Omega)^{\otimes q}\right) \right\| \leq \epsilon \cdot (m+q)!, \\ &\qquad \text{for } (m, q) \in [\mathfrak{g}] \times [\mathfrak{h}]. \end{aligned} \tag{3.26}$$



(a) Wine quality dataset (b) Pima Diabetes dataset (c) Music dataset

Figure 3.3: The sensitivity of FO with single parameter tuning. Red lines indicate the level-(1,1) FO, green lines the level-(2,1) FO, and blue lines the level-(1,2) FO. Each line represents the evolution of accuracy and fairness as the value of μ changes. Note that Figure 3.3c reflects an instance of fair regression, and so accuracy is measure via mean-squared error.

This formulation has just one tuning parameter μ. A sensitivity analysis with respect to this parameter is shown in Fig. 3.3.

Classification with a single-layer perceptron

We also consider the penalized FO formulation applied to a fully-connected one-layer network with 20 nodes. We use adaptive gradient descent as the learning algorithm. Learning rates are selected to optimize performance of the unmodified learner, and then kept uniform for each dataset. The datasets were randomly split into training and testing sets, with 80% of the data used in training. All reported results are calculated on the test set. Training sets were randomly sub-sampled to improve predictions of the unmodified learner, which is a common technique for unbalanced data. All experiments for deep-learning applications are run using Tensorflow 1.10 [1] on an Intel Core i7 processor.

We first consider the Wine Quality and Adult Income datasets, on which we measure the ability of our method to temper disparate impact. For this section, we use a single-layer network with 20 nodes. Tables 3.5a and 3.5b provide full results and comparisons for both

datasets. Here, the parameters for models that are not parameterized (i.e. the unmodified learner and the pre-processing approach) are inputed as 0. For the in-training comparison, there is only one parameter, as there is only one regularization term. For our FO approach, the parameters shown represent μ_1 and μ_2. As the appropriate penalty terms for our fairness constraints may be positive or negative, we run our method with the positive and negative of each value in a pre-specified set of considered hyper-parameters, and only report the direction (positive or negative) that serves to improve fairness. In practical settings, we note that the positivity or negativity of hyper-parameters can be forecast via empirical estimations of the higher-order moment terms from the training data.

We note that our method is able to notably reduce disparate impact with respect to the unmodified model in both datasets. Furthermore, for the proper choice of hyper-parameters, this modification does not come at a significant cost in terms of accuracy and AUC. In fact, we observe that the addition of fairness constraints can even improve accuracy, an interesting fact that may reflect on over-fitting in the unmodified model. Each of the moment constraints can be observed to have different levels of utility for different datasets, reflecting underlying structural relationships between the variables themselves. For example, the addition of second-order constraints notably impacts accuracy for the Wine Quality dataset, but not for the Adult Income dataset. Finally, we note that our method can yield better fairness, at less of a cost, than pre-processing. Even though the in-training benchmark maintains high accuracy, we remind the reader that this is because these methods allow differing treatments across protected classes; on the other hand, our method provides similar results while also satisfying *individual fairness* by maintaining one treatment for all.

On the other hand, the Recidivism dataset was originally brought to the attention of the fairness community not due to a preponderance of disparate impact; in fact, a main topic of the original article of [11] regarded the *error rates* among different protected classes. Thus, we report results of specific experiments in Table 3.6 that investigate the Equal Opportunity (EO) level of the results, with respect to convicts that were determined to be likely to recommit a crime. We note similar patterns to the previous results, our method decreasing discrepancy in error rates with more controlled loss in accuracy. It is particularly interesting that the increasing of the magnitude of penalty on the second-order constraint seems to be the main driver of fairness in this case, while modifications of the penalty on the first-order constraint seem relatively inconsequential.

Case Study: Dosing

In this section, we investigate the practicality of our FO hierarchy in the setting of automated medication dosage. Automated dosage has steadily gained attention as a key area where machine learning can improve efficiency in healthcare [105, 116, 265]. Any application of automation techniques which require statistical learning are susceptible to over-learning biases and undesirable correlations in the data, and so can perpetuate unfair decisions. To that end, we show that our mechanism for fair learning can address this critical issue. Since there are medical justifications for the consideration of gender and ethnicity in dosing, we

Table 3.5: Results for the first two datasets with a single-layer perceptron.

(a) Adult Income

version	metric params	Accuracy	AUC	KS
Unmod.	0, 0	0.7883	0.6238	0.0489
Calmon	0, 0	0.7149	0.6186	0.3173
Kamishima	0.01	0.7939	0.5817	0.0380
	0.001	0.7872	0.5917	0.0414
	0.0001	0.7952	0.6252	0.0999
FO(1,1)	-100, 0	0.7470	0.6226	0.0498
	-1000, 0	0.7799	0.6189	0.0459
FO(1,2)	-100, 1	0.7894	0.5893	0.0337
	-100, 10	0.7824	0.5818	0.0283
	-1000, 1	0.7942	0.5925	0.0336
	-1000, 10	0.7819	0.5814	0.0292

(b) Wine Quality

version	metric params	Accuracy	AUC	KS
Unmod.	0, 0	0.7246	0.7268	0.2009
Calmon	0, 0	0.6333	0.6231	0.1925
Kamishima	0.01	0.5059	0.5756	0.0286
	0.001	0.7261	0.7213	0.1483
	0.0001	0.7323	0.7319	0.1215
FO(1,1)	100, 0	0.7372	0.7362	0.1925
	1000, 0	0.7259	0.7272	0.1979
FO(1,2)	100, -1	0.7307	0.7313	0.1870
	100, -10	0.7108	0.7192	0.1154
	1000, -1	0.7296	0.7298	0.1617
	1000, -10	0.6880	0.6780	0.0173

Table 3.6: Results for the Recidivism dataset on a single-layer network architecture

version	metric params	Accuracy	AUC	EO
Unmod.	0, 0	0.6761	0.6758	0.3073
Calmon	0, 0	0.5476	0.5486	0.0471
Kamishima	0.01	0.5980	0.5996	0.1342
	0.001	0.6725	0.6737	0.2792
	0.0001	0.6731	0.6743	0.2938
FO(1,1)	100, 0	0.6705	0.6706	0.2994
	1000, 0	0.6761	0.6756	0.3056
FO(1,2)	100, -1	0.6676	0.6678	0.2798
	100, -10	0.6098	0.6126	0.1747
	1000, -1	0.6723	0.6722	0.3001
	1000, -10	0.6089	0.6122	0.1071

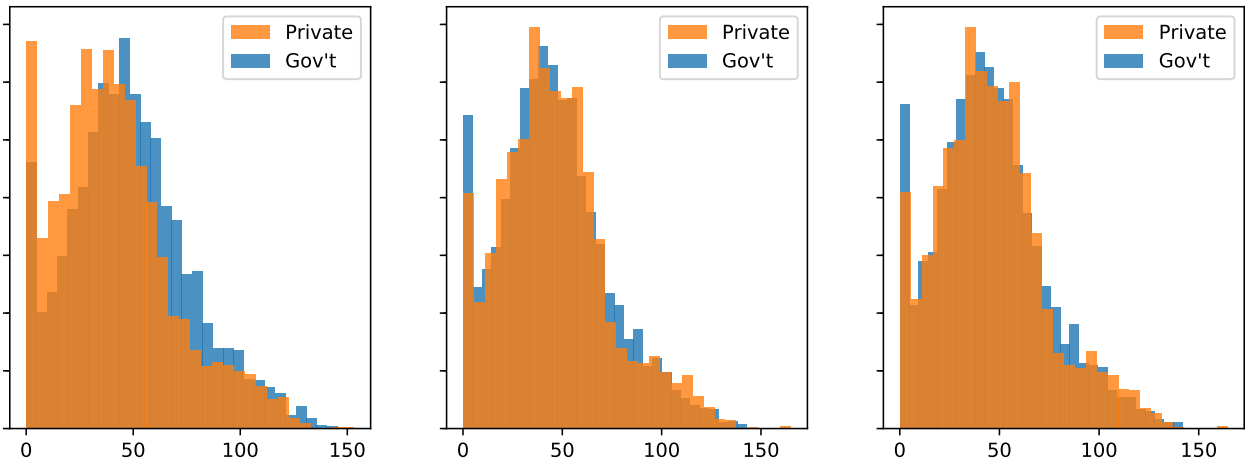
decide to instead consider insurance type (government or private) as our protected variable in this analysis.

Loss function Standard loss functions heretofore considered, such as the squared loss function of regression, logistic loss function of logistic regression and hinge-loss of SVM, are not necessarily applicable in the case of medication dosage. This is because they may not accurately represent the risks of over- and under-dosing. One possible risk function for dosing builds on the classic newsvendor problem from the operations research community, where supply must be chosen beforehand to meet random demand and undersupply/oversupply are penalized differently. Recent work has formulated and justified a data-driven newsvendor model for dosing, where demand is predicted via a quantile regression problem [217]. Similarly, we treat dosage as a matter of supply, with demand being the amount of medication that a specific patient needs.

MIMIC III Date for patients in the following studies were drawn from the publicly-available Multiparameter Intelligent Monitoring in Intensive Care (MIMIC III) database [218]. This is a large, deidentified database of health data drawn from over 40,000 patients that stayed in the critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. It comprises highly-granular including dosage levels, vital sign measurements taken bedside approximately hourly, caregiver notes and mortality, among other things. For our purposes, we were able to find patients in the Intensive Care Units (ICU) that were given morphine or heparin, and to track the vital signs of these patients (as well as their mortality status) at the time of prescription (and throughout their stay, for the heparin case).

One-time morphine dosage via quantile regression Opioid overdoses, including from illicit heroine and synthetic fentanyl, have become the leading cause of death in Americans under 50 [219]. Today, Americans comprise 4.6% of the global population, but 51.2% of global morphine usage. This has largely arisen due to misguided views in the 1990's on the danger of opioids [145, 174]. So there has been much recent interest in regulated and disciplined methods for dosing [166]. At the same time, recent reports have indicated that women and low-income patients are more likely to be under-diagnosed for pain or made to wait longer for a diagnosis [33, 76]. Thus, we seek to employ FO in order to train an individualized dosing policy that adapts to each patient's measurements and status, but can be made certifiably fair with regards to protected labels.

For this case, we extracted data for 7156 morphine prescriptions made to 4612 unique patients extracted from the MIMIC III database. For each patient, we collected age (at the time of prescription), heart rate, breath rate, blood pressure (both systolic and diastolic), weight and temperature. In all cases, measurements are the latest possible within 48 hours of prescription. We also collect, as categorical variables, admission type (ER, urgent care or other), service type (surgery or medical), ethnicity (black, white or other), gender (male or female) and insurance type (private or governmental). We also note the presence of embolism



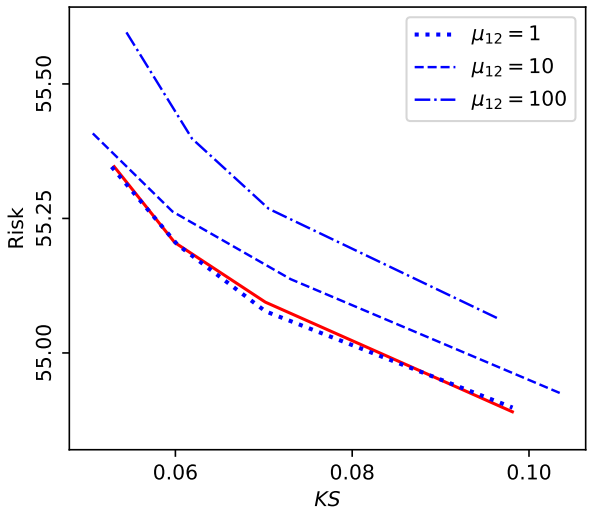
(a) Quantile regression (b) FO(1,1) (c) FO(1,2)

Figure 3.4: The distributions of morphine dosage, conditional on insurance type, for varying levels of FO. Figure 3.4a reflects a standard quantile regression with no fairness constraints. Figure 3.4b reflects the level-(1,1) FO with $\Delta_{1,1} = 0$, and Figure 3.4c with $\Delta_{1,1} = 0$ and $\mu_{1,2} = 10$. Histograms are generated by running each method five times over random training-testing splits of the data, and compiling results. We can see that increasing orders of FO can yield more similar distributions. Note that all negative dosage recommendations from the respective models are replaced with zero.

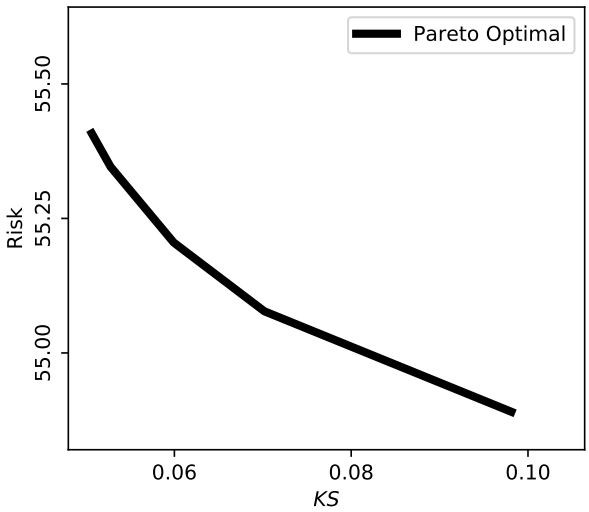
or obesity amongst the diagnoses of the patients at admission. We exclude all patients who are not prescribed Morphine Sulfate to be taken intravenously, and all patients for whom the appropriate measurements were not available. To begin, we conduct a standard linear regression to determine if insurance type does currently play a role in, or is at least highly correlated with, morphine dosage, conditional on all other variables considered. The results found that insurance type had a large magnitude coefficient with $p < 0.001$, which provides some statistical evidence that insurance type is correlated to dosing even after adjusting for the other predictor variables.

As mentioned above, we model the problem of dosing as one of quantile regression, motivated by data-driven formulations of the newsvendor problem. In our case, we impose a linearly increasing cost to both under-prescription and over-prescription, with the cost to over-prescription increasing half as quickly as that of under-prescription. This reflects the short-term nature of the risks of under-prescription, and the long-term nature of the risks of over-prescription. Given the features described above (excluding insurance payer), we then formulate varying levels of our FO to solve the quantile regression problem that specifies dosing.

The results of our analysis are displayed in Fig. 3.5 and Fig 3.4. In Fig. 3.5, the



(a) AUC vs. Disparate impact for morphine dosage.

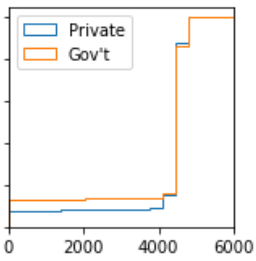


(b) Pareto optimal curve for morphine dosage.

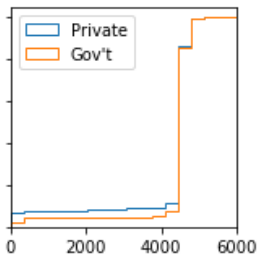
Figure 3.5: The accuracy vs. disparate impact of the learned dosage rule under varying orders of FO and for different hyper-parameters. In Figure 3.5a, the red curve reflects the level-(1,1) FO, while the blue curves reflect the level-(1,2) FO for the $\mu_{1,2}$ parameters indicated. All curves represent $\Delta_{1,1} \in \{0, 0.05, 0.1, 0.2\}$. The curve in Figure 3.5b reflects the Pareto optimal curve, showing the lowest loss achievable for all levels of disparate impact.

Figure 3.6: Heparin case study results.

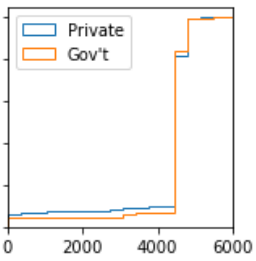
version	metric params	loss	KS
Unmod.	0, 0	243	0.0585
First	-0.1, 0	259	0.0429
	-1.0, 0	711	0.0587
Second	-0.1, 1E-6	234	0.0555
	-1.0, 1E-6	390	0.0617
	-0.1, 1E-5	244	0.0331
	-1.0, 1E-5	412	0.0036



(a) $(0, 0)$



(b) $(-0.1, 0)$



(c) $(-0.1, 1E-5)$

tradeoff between risk and fairness is displayed, as well as the range of best possible dosage rules. In Fig. 3.5a, the solid red curve represents results from the level-(1,1) FO with $\Delta_{1,1} \in \{0, 0.05, 0.1, 0.2\}$, and each broken blue curve represents results from the level-(1,2) FO with $\Delta_{1,1} \in \{0, 0.05, 0.1, 0.2\}$ and $\mu_{1,2}$ as indicated in the legend. As also visible from

Fig. 3.5b, we note that decreased disparate impact comes at the cost of increasing risk, and that this marginal cost increases as more fairness is demanded. We also note that the level-(1,2) FO can achieve better fairness results than the level-(1,1) FO, for proper choice of hyperparameters. Furthermore, the level-(1,2) FO can even achieve better *accuracy* results than the level-(1,1) FO, albeit only slightly. We attribute this to a possible regularization affect that the fairness constraints can have, preventing overfitting to uninformative elements in the data that have a dependence on the protected attribute. Visual evidence of the reduction in disparate impact is shown in Fig. 3.4, which presents the difference in the distribution of dosage levels across insurance types for standard Quantile Regression (QR), the level-(1,1) FO and the level-(1,2) FO. There is a clear disparity between the distributions in Fig. 3.4a, but this difference is significantly reduced in Fig. 3.4b and even more so in Fig. 3.4c.

Sequential heparin dosing via LSTM Heparin is one of the most common drugs administered in Intensive Care Units [69]. However, setting the correct dose of Heparin for each patient is challenging since the way in which the drug is metabolized varies greatly between patients and its not possible to directly measure the concentration of Heparin in a patient's bloodstream [58]. As such, several machine learning based approaches have been proposed to assist in setting Heparin doses by predicting what dose of Heparin individuals would react to best [99, 188]. In this section we use a similar model to showcase how our approach can be applied in this setting.

We extracted daily measurements for patients that were prescribed heparin from from the publicly-available Multiparameter Intelligent Monitoring in Intensive Care (MIMIC III) database [218]. For each patient and for each day that the patient spent in the hospital, we collected age, heart rate, breath rate, blood pressure (both systolic and diastolic), weight and temperature. We also collect measurements of albumin, arterial CO2, arterial pH, bilirubin, Blood Urea Nitrogen (BUN), creatinine, Glasgow Coma Score (GCS), hematocrit, hemoglobin, platelet count, aPTT, troponin-t and white blood cell count. Measurements are the latest possible within 48 hours with the exception of GCS, which is the latest within the last month. We collect, as categorical variables, admission type (ER, urgent care or other), ethnicity (black, white or other), gender (male or female), insurance type (private or governmental) and service type (surgery or medical). We also note whether the patient has been diagnosed with embolism or obesity. We exclude all patients only administered heparin flush, and all patients for whom the appropriate measurements were not available. In all, we have measurements for 549 patients over a combined 12196 days, with a maximum of 395 days of measurements for any one patient. We use insurance type as the protected variable in this analysis, as there may be medical justification for taking gender and ethnicity into account for dosage. Our goal is to develop an automated procedure for heparin dosage, so this problem is continuous.

We trained an LSTM to predict the level of Heparin to dose a patient for every day that the patient spends in the hospital. Specifically, we use a one-layer LSTM followed by

a 64-node dense layer. Inputs to the model include the latest medical measurements taken from the patient on a daily basis. Procedures for the training-set sizes and hyper-parameter selection are the same as in Section 3.8. As noted above, the problem of dosage is similar to the newsvendor problem, as a level of supply must be acquired to meet random demand and undersupply and oversupply can incur different costs. In this case, supply is the level of Heparin dosed and demand is the level of Heparin that the patient requires. We assume linearly increasing costs to over- and under-dosage, with under-dosage 1.5 times as costly as over-dosage. These design assumptions are flexible and can be altered; our goal here is to provide a proof-of-concept, not a definitive and fully-autonomous end-product. Results are presented in fig. 3.6. Note that, with the proper hyper-parameter choice, our method is able to generate dosage rules that decrease bias without precipitously increasing loss. To further visualize this, the empirical cdf's of the levels of Heparin dosed (conditional on the insurance type) are presented in figs. 3.6a to 3.6c. In each case, the captions are of the form (μ_1, μ_2), indicating the hyper-parameters used to learn those dosage rules. We note that, by adding fairness constraints, it is possible to decrease the distance between the cdf's, which implies greater fairness in the sense of eq. (3.2).

3.9 Conclusion

In this chapter, we discussed various notions of fairness in supervised learning and consider a series of hierarchical relaxations of these notions oriented around matching conditional moments of data. We generate regularizers to emulate these relaxations and show how they may be applied to deep learning techniques to enforce fairness at training time. We also discuss the benefits of enforcing fairness at the time of training, as opposed to in pre- or post-processing stages. Finally we present experimental results showing the benefits and flexibility of our method with respect to different network architectures and fairness notions. In this work, we only consider the first two moment constraints, but would be interested in future work evaluating the benefits of including additional higher-order moment constraints.

Chapter 4

Fairness in Unsupervised Learning

4.1 Introduction

Despite the success of machine learning in informing policies and automating decision-making, there is growing concern about the fairness (with respect to protected classes like race or gender) of the resulting policies and decisions [11, 176, 185, 214]. Hence, several groups have studied how to define fairness for supervised learning [45, 77, 109, 276] and developed supervised learners that maintain high prediction accuracy while reducing unfairness [25, 57, 109, 191, 270].

However, fairness in the context of unsupervised learning has only recently received more attention in some early works [56, 108, 192, 221]. One reason is that fairness is easier to define in the supervised setting, where positive predictions can often be mapped to positive decisions (e.g., an individual who is predicted to not default on a loan maps to the individual being offered a loan). Such notions of fairness cannot be used for unsupervised learning, which does not involve making predictions. A second reason is that it is not obvious why fairness is an issue of relevance to unsupervised learning, since predictions are not made.

Relevance of fairness to unsupervised learning

Fairness is important to unsupervised learning: First, unsupervised learning is often used to generate qualitative insights from data. Examples include visualizing high-dimensional data through dimensionality-reduction and clustering data to identify common trends or behaviors. If such qualitative insights are used to generate policies, then there is an opportunity to introduce unfairness in the resulting policies if the results of the unsupervised learning are unequal for different protected classes (e.g., race or gender). We present such an example in Section 4.6 using individual health data.

Second, unsupervised learning is often used as a preprocessing step for other learning methods. For instance, dimensionality reduction is sometimes performed prior to clustering, and hence fair dimensionality reduction could indirectly provide methods for fair clustering. Similarly, there are no fairness-enhancing versions of most supervised learners. Consequently,

techniques for fair unsupervised learning could be combined with state-of-the-art supervised learners to develop new fair supervised learners. In fact, the past work most related to this chapter concerns techniques that have been developed to generate fair data transformations that maintaining high prediction accuracy for classifiers that make predictions using the transformed data [77, 86, 271]; however, these past works are most accurately classified as supervised learning because the data transformations are computed with respect to a label used for predictions.

We briefly review this work. [77] propose a linear program that maps individuals to probability distributions over possible classifications such that similar individuals are classified similarly. [271] [48] generate an intermediate representation for fair clustering using a non-convex formulation that is difficult to solve. [86] propose an algorithm that scales data points such that the distributions of features, conditioned on the protected attribute, are matched; however, this approach makes the restrictive assumption that predictions are monotonic with respect to each dimension. [56] directly perform fair clustering by approximating an NP-hard preprocessing step; however, this approach only applies to specific clustering techniques whereas the approach we develop can be used with arbitrary clustering techniques. Finally, a series of work has emerged using auto-encoders in the the context of deep classification. This area is promising, but suffers from a lack of theoretical guarantees and is further oriented almost entirely around an explicit classification task [27, 272]. In contrast, our method has applications in both supervised and unsupervised learning tasks, and well-defined convergence and optimality guarantees.

Outline and novel contributions

This chapter studies fairness for principal component analysis (PCA), and we make three main contributions: First, in Section 4.3 we propose and motivate a novel quantitative definition of fairness for dimensionality reduction. Second, in Section 4.5 we develop convex optimization formulations for fair PCA and fair kernel PCA. Third, in Section 4.6 we demonstrate the efficacy of our semidefinite programming (SDP) formulations using several datasets, including using fair PCA as preprocessing to perform fair (with respect to age) clustering of health data that can impact health insurance rates.

4.2 Preliminaries

Let $[n] = \{1, \ldots, n\}$, $\mathbf{1}(u)$ be the Heaviside function, and let $\mathbf{e}$ be the vector whose entries are all 1. A positive semidefinite matrix U with dimensions $q \times q$ is denoted $U \in \mathbb{S}_+^q$ (or $U \succeq 0$ when dimensions are clear). We use the notation $\langle \cdot, \cdot \rangle$ to denote the inner product and $\mathbb{I}$ the identity matrix.

Our data consists of 2-tuples (X, Y), the realizations of which are denoted as X_i and Y_i for $i = 1, \ldots, n$. Here the $X_i \in \mathbb{R}^p$ are a set of features, and the $Z_i \in \{-1, 1\}$ label a protected class. For a matrix W, the i-th row of W is denoted W_i. Let $\vec{X} \in \mathbb{R}^{n \times p}$ and

$\vec{Z} \in \mathbb{R}^n$ be the matrices so that $\vec{X}_i = (X_i - E - n(x))^\mathsf{T}$ and $Z_i = z_i$. Also, we use the notation $\Pi : \mathbb{R}^p \to \mathbb{R}^d$ to refer to a function that performs dimensionality reduction on the covariates X, where d is the dimension of the dimensionality-reduced data.

Let $P = \{i : Z_i = +1\}$ be the set of indices where the protected class is positive, and similarly let $N = \{i : Z_i = -1\}$ be the set of indices where the protected class is negative. We use $\#P$ and $\#N$ for the cardinality of these sets. Furthermore, we define $\vec{X}_+$ to be the matrix whose rows are X_i^T for $i \in P$, and we similarly define $\vec{X}_-$ to be the matrix whose rows are $\vec{X}_i^\mathsf{T}$ for $i \in N$. Next, let $\widehat{\Sigma}_+$ and $\widehat{\Sigma}_-$ be the sample covariances matrices of $\vec{X}_+$ and $\vec{X}_-$, respectively.

For a kernel function $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+$, let $K(\vec{X}, \vec{X}') = [k(X_i, X'_j)]_{ij}$ be the transformed Gram matrix. Since the *kernel trick* involves replacing $X_i^\mathsf{T} X_j$ with $K(X_i, X_j)$, the benefit of the above notation is it allows us to replace $\vec{X}(\vec{X}')^\mathsf{T}$ with $K(\vec{X}, \vec{X}')$ as part of applying the kernel trick.

4.3 Fairness for dimensionality reduction

Definitions of fairness for supervised learning [25, 45, 57, 77, 86, 109, 276] specify that predictions conditioned on the protected class are roughly equivalent. However, these fairness notions cannot be used for dimensionality reduction because predictions are not made in unsupervised learning. This section discusses fairness for dimensionality reduction. We first provide and motivate a general quantitative definition of fairness, and then present several important cases of this definition.

General definition

Consider a fixed classifier $h(u, t) : \mathbb{R}^d \times \mathbb{R} \to \{-1, +1\}$ that inputs features $u \in \mathbb{R}^d$ and a threshold t, and predicts the protected class $Z \in \{-1, +1\}$. We say that a dimensionality reduction $\Pi : \mathbb{R}^p \to \mathbb{R}^d$ is $\Delta(h)$-fair if

$$\Big|\mathbb{P}\left[h(\Pi(X), t) = +1 \middle| Z = +1\right] - \mathbb{P}\left[h(\Pi(X), t) = +1 \middle| Z = -1\right]\Big| \leq \Delta(h), \ \forall t \in \mathbb{R}. \tag{4.1}$$

Moreover, let $\mathcal{F}$ be a family of classifiers. Then we say that a dimensionality reduction $\Pi : \mathbb{R}^p \to \mathbb{R}^d$ is $\Delta(\mathcal{F})$-fair if it is $\Delta(h)$-fair for all classifiers $h \in \mathcal{F}$.

Our fairness definition can be interpreted via classification: Observe that the first term in the left-hand-side of (4.1) is the true positive rate of the classifier h in predicting the protected class using the dimensionality-reduced variable $\Pi(x)$ at threshold t, and the second term is the corresponding false positive rate. Thus, $\Delta(h)$ in our definition (4.1) can be interpreted as bounding the accuracy of the classifier h in predicting the protected class using the dimensionality-reduced variable $\Pi(x)$.

Note that eq. (4.1) is analogous to *disparate impact* for classifiers [45,86], where we require that treatment not vary at all between protected classes. This has often been criticized as too

strict of a notion in classification, and so alternate notions of fairness have been developed, such as *equalized odds* and *equalized opportunity* [109]. Instead of equalizing all treatment across protected classes, these notions instead focus on equalizing error rates; for example, in the case of lending, equalized odds would require nondiscrimination *among all applicants of similar FICO scores*, whereas disparate impact would require nondiscrimination among all applicants. This may be preferred in cases where the target variable that is ultimately to be predicted is strongly correlated to the protected attribute, Z. In any case, it can easily be incorporated into our model by simply further conditioning the two terms on the left-hand-side of eq. (4.1) on a target label, Y. These notions are explored in more detail in Chapter 3.

Motivation

The above is a meaningful definition of fairness for dimensionality reduction because it implies that a supervised learner using fair dimensionality-reduced data will itself be fair. This is formalized below:

Proposition 9. *Suppose we have a family of classifiers $\mathcal{F}$ and a dimensionality reduction Π that is $\Delta(\mathcal{F})$-fair. Then any classifier that is selected from $\mathcal{F}$ to predict a label $Y \in \{-1, +1\}$ using $\Pi(X)$ as features will have disparate impact less than $\Delta(\mathcal{F})$.*

Proposition 9 follows directly from our definition of fairness. We anticipate that in most situations the goal of the dimensionality reduction would not be to explicitly predict the protected class. Thus, our approach of bounding intentional discrimination on Z represents a conservative bound on any discrimination that may incidentally arise when performing classification using the family $\mathcal{F}$ or when deriving qualitative insights form the results of unsupervised learning.

Special cases

An important special case of our definition occurs for the family $\mathcal{F}_c = \{h(u,t) = \mathbf{1}(u \leq w + t) : w \in \mathbb{R}^d\}$, where the inequality in this expression should be interpreted element-wise. In this case, our definition can be rewritten as $\sup_u \left|F_{\Pi(X)|Z=+1}(u) - F_{\Pi(X)|Z=-1}(u)\right| \leq \Delta(\mathcal{F}_c)$, where F is the cumulative distribution function (c.d.f.) of the random variable in the subscript. Restated, for this family our definition is equivalent to saying $\Delta(\mathcal{F})$ is a bound on the Kolmogorov distance between $\Pi(X)$ conditioned on $Z = \pm 1$ (i.e., the left-hand side of the above equation).

Other important cases are the family of linear support vector machines (SVM's) $\mathcal{F}_v = \{h(u,t) = \mathbf{1}(w^\mathsf{T} u - t \leq 0) : w \in \mathbb{R}^d\}$ and the family of kernel SVM's $\mathcal{F}_k$ for a fixed kernel k. These important cases are used in Section 4.5 to propose formulations for fair PCA and fair kernel PCA. These cases are important because they are used in Section 4.5 to propose formulations for fair PCA and fair kernel PCA.

Next, we briefly discuss empirical estimation of $\Delta(\mathcal{F})$. An empirical estimate of $\Delta(h)$ is given by $\widehat{\Delta}(h) = \sup_t |\frac{1}{\#P}\sum_{i\in P}\mathbf{1}(h(\Pi(X_i),t) = +1) - \frac{1}{\#N}\sum_{i\in N}\mathbf{1}(h(\Pi(X_i),t) = +1)|$. Similarly, we define $\widehat{\Delta}(\mathcal{F}) = \sup\{\widehat{\Delta}(h) \mid h \in \mathcal{F}\}$. Last, note that we can provide high probability bounds of the actual fairness level in terms of these empirical estimates:

Proposition 10. *Consider a fixed family of classifiers $\mathcal{F}$. If the samples (X_i, Z_i) are i.i.d., then for any $\delta > 0$ we have with probability at least $1 - \exp(-n\delta^2/2)$ that $\Delta(\mathcal{F}) \leq \widehat{\Delta}(\mathcal{F}) + 8\sqrt{\mathcal{V}(\mathcal{F})/n} + \delta$, where $\mathcal{V}(\mathcal{F})$ is the VC dimension of the family $\mathcal{F}$.*

This result follows from the triangle inequality, bounding $\Delta(\mathcal{F})$ with $\hat{\Delta}(\mathcal{F})$ plus a generalization error, for which there are standard bounds via Dudley's entropy integral [254].

Remark 13. Recall that $\mathcal{V}(\mathcal{F}_c) = d + 1$ [233], and that $\mathcal{V}(\mathcal{F}_v) = d + 1$ [254]. This means $\widehat{\Delta}(\mathcal{F}_c)$ and $\widehat{\Delta}(\mathcal{F}_v)$ will be accurate when n is large relative to d.

4.4 Projection defined by PCA

Our approach to designing an algorithm for fair PCA will begin by first studying the convex relaxation of a non-convex optimization problem whose solution provides the projection defined by PCA. First, note that computation of the first d PCA components v_i for $i = 1, \ldots, d$ can be written as the following non-convex optimization problem: $\max\{\sum_{i=1}^d v_i^\mathsf{T}\vec{X}^\mathsf{T}\vec{X}v_i \mid \|v_i\|_2 \leq 1, \quad v_i^\mathsf{T}v_j = 0, \text{ for } i \neq j\}$. Now suppose we define the matrix $P = \sum_{i=1}^d v_i v_i^\mathsf{T}$, and note $\sum_{i=1}^d v_i^\mathsf{T}\vec{X}^\mathsf{T}\vec{X}v_i = \sum_{i=1}^d \langle \vec{X}^\mathsf{T}\vec{X}, v_i v_i^\mathsf{T}\rangle = \langle \vec{X}^\mathsf{T}\vec{X}, P\rangle$. Thus, we can rewrite the above optimization problem as

$$\max\left\{\langle \vec{X}^\mathsf{T}\vec{X}, P\rangle \;\middle|\; \text{rank}(P) \leq d,\ \mathbb{I} \succeq P \succeq 0\right\}. \tag{4.2}$$

In the above problem, we should interpret the optimal P^* to be the projection matrix that projects $X \in \mathbb{R}^p$ onto the d PCA components (still in the original p-dimensional space). Next, we consider a convex relaxation of (4.2). Since $\mathbb{I} - P \succeq 0$, the usual nuclear norm relaxation is equivalent to the trace [209]. So our convex relaxation is

$$\max\left\{\langle \vec{X}^\mathsf{T}\vec{X}, P\rangle \;\middle|\; \text{trace}(P) \leq d,\ \mathbb{I} \succeq P \succeq 0\right\}. \tag{4.3}$$

Note that this base model is the same as that used by [13]. The following result shows that we can recover the first d PCA components from any P^* that solves (4.3).

Theorem 8. *Let P^* be an optimal solution of (4.3), and consider its diagonalization: $P^* = \sum_{i=1}^p \lambda_i^* v_i v_i^\mathsf{T}$, where v_i is an orthonormal basis, and (without loss of generality) the λ_i^* are in non-increasing order. Then the positive semidefinite $P^{**} \triangleq \sum_{i=1}^d v_i v_i^\mathsf{T}$ is an optimal solution to (4.2).*

Proof. We consider two cases. First, if $\text{rank}(P^*) \leq d$ then $\lambda_i^* \in \{0, 1\}$ or $v_i^\mathsf{T} \vec{X}^\mathsf{T} \vec{X} v_i = 0$ for all i, since otherwise we could increase λ_i^* if $v_i^\mathsf{T} \vec{X}^\mathsf{T} \vec{X} v_i > 0$ (or vice versa) to improve the objective while maintaining feasibility. It follows that $\langle \vec{X}^\mathsf{T} \vec{X}, P^* \rangle = \langle \vec{X}^\mathsf{T} \vec{X}, P^{**} \rangle$. This means that P^{**} is optimal for (4.3); since it is also feasible for (4.2), we are done. Second, if $\text{rank}(P^*) > d$ then $0 < \lambda_d^* < 1$ since the λ_i^* are ordered. Consider $\tilde{P} \triangleq (P^* - cP^{**})/(1-c), c = \min\{\lambda_d^*, 1 - \lambda_d^*\}$. Note that $\tilde{P}$ is feasible for (4.3), and that P^* is a strict convex combination of P^{**} and $\tilde{P}$. All points between $\tilde{P}$ and P^{**} are feasible by convexity, and so the optimality of P^* implies that P^{**} and $\tilde{P}$ must also be optimal for (4.3) by linearity of the objective (i.e., at least one must have objective value no less than that of P^*, but if one had a strictly better objective value than the other, then no strict convex combination of the two could be optimal). The result then follows from the optimality of P^{**} for (4.3) and feasibility for (4.2). □

We conclude this section with two useful results on the spectral norm $\|\cdot\|_2$ of a symmetric matrix.

Theorem 9. *Let Q be a symmetric matrix, and suppose $\varphi \geq \|Q\|_2$. Then $\|Q\|_2 = \max\{\|Q + \varphi\mathbb{I}\|_2, \|-Q + \varphi\mathbb{I}\|_2\} - \varphi$.*

Proof. First diagonalize $Q = \sum_{i=1}^p \lambda_i v_i v_i^\mathsf{T}$, with orthonormal basis v_i and (without loss of generality) λ_i in non-increasing order. Then $+Q + \varphi\mathbb{I} = \sum_{i=1}^p (+\lambda_i + \varphi) v_i v_i^\mathsf{T}$, $-Q + \varphi\mathbb{I} = \sum_{i=1}^p (-\lambda_i + \varphi) v_i v_i^\mathsf{T}$. But by construction $\lambda_i + \varphi \geq 0$ and $-\lambda_i + \varphi \geq 0$ for all $i = 1, \ldots, p$. Thus $\|Q + \varphi\mathbb{I}\|_2 = \lambda_1 + \varphi$ and $\|-Q + \varphi\mathbb{I}\|_2 = -\lambda_p + \varphi$. The result follows since $\|Q\|_2 = \max\{\lambda_1, -\lambda_p\}$. □

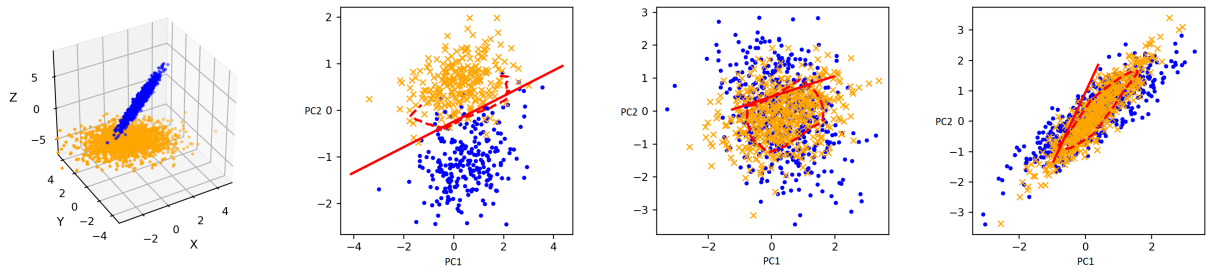
Corollary 1. *Let Q be a symmetric matrix, and suppose $\varphi \geq \|Q\|_2$. If V is such that $V^\mathsf{T} V = \mathbb{I}$, then $\|V^\mathsf{T} Q V\|_2 = \max\{\|V^\mathsf{T}(Q + \varphi\mathbb{I})V\|_2, \|V^\mathsf{T}(-Q + \varphi\mathbb{I})V\|_2\} - \varphi$.*

Proof. First note that $V^\mathsf{T}(Q + \varphi\mathbb{I})V = V^\mathsf{T} Q V + \varphi\mathbb{I}$ and that $V^\mathsf{T}(-Q + \varphi\mathbb{I})V = -V^\mathsf{T} Q V + \varphi\mathbb{I}$. Since the spectral norm is submultiplicative, this means $\|V^\mathsf{T} Q V\|_2 \leq \|V^\mathsf{T}\|_2 \|Q\|_2 \|V\|_2 \leq \|Q\|_2$. So $\varphi \geq \|V^\mathsf{T} Q V\|_2$, and the result follows by applying Theorem 9 to $V^\mathsf{T} Q V$. □

Recall that using the Schur complement allows representation of $\|VRV^\mathsf{T}\|_2$ as a positive semidefinite matrix constraint when R is positive semidefinite [42]. So the above corollary is useful because it means we can represent $\|VQV^\mathsf{T}\|_2$ using positive semidefinite matrix constraints since $(Q + \varphi\mathbb{I})$ and $(-Q + \varphi\mathbb{I})$ are positive semidefinite by construction.

4.5 Designing formulations for fair PCA

Consider the linear dimensionality reduction $\Pi(X) = V^\mathsf{T} X$ for $V \in \mathbb{R}^{p \times d}$ such that $V^\mathsf{T} V = \mathbb{I}$. Then for linear classifier $h(u, t) = \mathbf{1}(w^\mathsf{T} u - t \leq 0)$, definition (4.1) simplifies to $\Delta(h) = \sup_t \left| \mathbb{P}[w^\mathsf{T} V^\mathsf{T} X \leq t | Z = +1] - \mathbb{P}[w^\mathsf{T} V^\mathsf{T} X \leq t | Z = -1] \right|$. But the right-hand side is the Kolmogorov distance between $w^\mathsf{T} V^\mathsf{T} X$ conditioned on $Z = \pm 1$, which is upper bounded (as can



(a) Original data (b) PCA (c) FPCA - Mean con. (d) FPCA - Both con.

Figure 4.1: Comparison of PCA and FPCA on synthetic data. In each plot, the thick red line is the optimal linear SVM separating by color, and the dotted line is the optimal Gaussian kernel SVM.

be seen trivially from its definition) by the total variation distance. Consequently, applying Pinsker's inequality [171] gives $\Delta(h) \leq \sqrt{\frac{1}{2}\mathcal{KL}\left(w^\mathsf{T}V^\mathsf{T}X_- \,\middle\|\, w^\mathsf{T}V^\mathsf{T}X_+\right)}$, where $\mathcal{KL}(\cdot\|\cdot)$ is the Kullback-Leibler divergence, X_+ is the random variable $X|Z=+1$, and X_- is the random variable $X|Z=-1$. For the special case $X_+ \sim \mathcal{N}(\mu_+, \Sigma_+)$ and $X_- \sim \mathcal{N}(\mu_-, \Sigma_-)$, we have [141]:

$$\Delta(h) \leq \sqrt{\frac{1}{4}\left(\frac{s_-}{s_+} + \frac{(m_+ - m_-)^2}{s_+} + \log\frac{s_+}{s_-} - 1\right)}. \tag{4.4}$$

where $s_+ = w^\mathsf{T}V^\mathsf{T}\Sigma_+Vw$, $s_- = w^\mathsf{T}V^\mathsf{T}\Sigma_-Vw$, $m_+ = w^\mathsf{T}V^\mathsf{T}\mu_+$, and $m_- = w^\mathsf{T}V^\mathsf{T}\mu_-$. The key observation here is that (4.4) is minimized when $s_+ = s_-$ and $m_+ = m_-$, and we will use this insight to propose constraints for FPCA.

We first design constraints for the non-convex formulation (4.2) so that $\hat{m}_+ - \hat{m}_- = w^\mathsf{T}V^\mathsf{T}\psi$ has small magnitude, where $\psi = \hat{\mu}_+ - \hat{\mu}_- = \frac{1}{\#P}\sum_{i\in P} X_i - \frac{1}{\#N}\sum_{i\in N} X_i$. Note we make the identification $P = VV^\mathsf{T}$ because of the properties of P in (4.2) and since $V^\mathsf{T}V = \mathbb{I}$. Observe that $w^\mathsf{T}V^\mathsf{T}\psi$ is small if $V^\mathsf{T}\psi$ is small, which can be formulated as

$$\|V^\mathsf{T}\psi\|^2 = \langle VV^\mathsf{T}, \psi\psi^\mathsf{T}\rangle = \langle P, \psi\psi^\mathsf{T}\rangle \leq \delta^2, \tag{4.5}$$

where $\|\cdot\|$ is the ℓ_2-norm, and δ is a bound on the norm. This (4.5) is a linear constraint on P.

We next design constraints for the non-convex formulation (4.2) so that $\hat{s}_+ - \hat{s}_- = w^\mathsf{T}V^\mathsf{T}(\widehat{\Sigma}_+ - \widehat{\Sigma}_-)Vw$ has small magnitude. Recall the identification $P = VV^\mathsf{T}$ because of the properties of P in (4.2) and since $V^\mathsf{T}V = \mathbb{I}$. Next observe that $w^\mathsf{T}V^\mathsf{T}(\widehat{\Sigma}_+ - \widehat{\Sigma}_-)Vw$ is

small if $V^\mathsf{T}(\widehat{\Sigma}_+ - \widehat{\Sigma}_-)V$ is small. Let $Q = \widehat{\Sigma}_+ - \widehat{\Sigma}_-$, then using Corollary 1 gives

$$\begin{aligned}\mu + \varphi \geq \|V^\mathsf{T} Q V\|_2 + \varphi &= \max\{\|V^\mathsf{T}(Q + \varphi\mathbb{I})V\|_2, \|V^\mathsf{T}(-Q + \varphi\mathbb{I})V\|_2\} \\ &= \max\{\|VV^\mathsf{T}(Q + \varphi\mathbb{I})VV^\mathsf{T}\|_2, \|VV^\mathsf{T}(-Q + \varphi\mathbb{I})VV^\mathsf{T}\|_2\} \\ &= \max\{\|P(Q + \varphi\mathbb{I})P\|_2, \|P(-Q + \varphi\mathbb{I})P\|_2\}, \quad (4.6)\end{aligned}$$

where $\varphi \geq \|\widehat{\Sigma}_+ - \widehat{\Sigma}_-\|_2$, and μ is a bound on the norm. Note (4.6) can be rewritten as SDP constraints using a standard reformulation for the spectral norm [42].

We design an SDP formulation for FPCA by combining the above elements. Though (4.2) with constraint (4.5) and (4.6) is a non-convex problem for FPCA, we showed in Theorem 8 that (4.3) was an exact relaxation of (4.2) after extracting the d largest eigenvectors of the solution of (4.3). Thus, we propose the following SDP formulation for FPCA:

$$\begin{aligned}\max\ & \langle \vec{X}^\mathsf{T}\vec{X}, P\rangle - \mu t && (4.7a)\\ \text{s.t. }& \operatorname{trace}(P) \leq d,\ \mathbb{I} \succeq P \succeq 0 && (4.7b)\\ & \langle P, \psi\psi^\mathsf{T}\rangle \leq \delta^2 && (4.7c)\\ & \begin{bmatrix} t\mathbb{I} & PM_+ \\ M_+^\mathsf{T}P & \mathbb{I}\end{bmatrix} \succeq 0, && (4.7d)\\ & \begin{bmatrix} t\mathbb{I} & PM_- \\ M_-^\mathsf{T}P & \mathbb{I}\end{bmatrix} \succeq 0 && (4.7e)\end{aligned}$$

where $M_i M_i^\mathsf{T}$ is the Cholesky decomposition of $iQ + \varphi\mathbb{I}$ ($i \in \{-, +\}$), $\varphi \geq \|\widehat{\Sigma}_+ - \widehat{\Sigma}_-\|_2$, (4.7c) is called the *mean constraint* and denotes the use (4.5), and (4.7d) and (4.7e) are called the *covariance constraints* and are the SDP reformulation of (4.6). Note that these are analogous to the fairness constraints in the FO(1,2) formulation introduced in Chapter 2, and that they could easily be extended to a FO($\mathfrak{g}$,2) formulation for any $\mathfrak{g}$. Our convex formulation for FPCA consists of solving (4.7) and then extracting the d largest eigenvectors from the optimal P^*.

Furthermore, this method may be extended to multiple protected attributes by replicating constraints (4.7c), (4.7d) & (4.7e) appropriately. That is, for secondary protected attribute Z', we may define the appropriate ψ, M_+ and M_- values and add the analogous constraints. Note that this will only abet "pairwise fairness", or fairness with respect to each of the protected attributes individually. To attain "joint fairness", or fairness with respect to both terms simultaneously, we would need to recreate constraints (4.7c), (4.7d) & (4.7e) for Z' as well as the interaction between Z and Z'. This notion is covered in more detail in Chapter 3, so we forego further discussion here.

SDP Formulation for Fair Kernel PCA (F-KPCA)

We can apply the kernel trick to (4.7) to develop an SDP for F-KPCA. This is useful for extracting nonlinear patterns [227]. Here, we only present the resulting SDP:

$$\begin{aligned}
\max\ & \langle K(\vec{X},\vec{X}), P\rangle + \mu t && \text{(4.8a)}\\
\text{s.t.}\ & \operatorname{trace}(P) \le d && \text{(4.8b)}\\
& \mathbb{I} \succeq P \succeq 0 && \text{(4.8c)}\\
& \langle P, \phi_k \phi_k^\top \rangle \le \delta^2 && \text{(4.8d)}\\
& \begin{bmatrix} t\mathbb{I} & N_+^\top P \\ PN_+ & \mathbb{I} \end{bmatrix} \succeq 0 && \text{(4.8e)}\\
& \begin{bmatrix} t\mathbb{I} & N_-^\top P \\ PN_- & \mathbb{I} \end{bmatrix} \succeq 0 && \text{(4.8f)}
\end{aligned}$$

where $N_i N_i^\top$ is the Cholesky decomposition of $iQ_k+\varphi\mathbb{I}$, $\phi_k = \frac{1}{\#P}K(\vec{X},X_+)\mathbf{e}-\frac{1}{\#N}K(\vec{X},X_-)\mathbf{e}$, $Q_k = K(\vec{X},X_+)K(X_+,\vec{X}) - K(\vec{X},X_-)K(X_-,\vec{X})$, and $\varphi \ge \|Q_k\|_2$. Our convex formulation for K-FPCA consists of solving the convex SDP (4.8) and then extracting the d largest eigenvectors from the optimal P^*.

4.6 Experimental results

We use synthetic and real datasets from the UC Irvine Machine Learning Repository [154] to demonstrate the efficacy of our SDP formulations. We also show how FPCA can be used to minimize discrimination in health insurance rates (with respect to age). For any SVM run, tuning parameters were chosen using 5-fold cross-validation, and data was normalized to have unit variance in each field. All results presented in this section are after averaged over 5 rounds of 70-30 training-testing splits, where an approach was trained on a random 70% of the data and evaluated based on the specified metrics using the remaining 30% of the data. In each case, the data was dimensionality-reduced using the top 5 principal components, fair or otherwise. All results follow after normalizing data columns, a practice that is common for datasets in which different features are of incomparable units. All results here use $\delta = 0, \mu = 0.01$.

Benchmarks

To the best of our knowledge, there are very few methods that are directly comparable to ours. Most existing work is married to an explicit classification task, while ours is a general pre-processing step that makes it amenable to any type of analysis. Among the few comparable approaches are those of [271] and [48]. Both design non-parametric optimization problems that yield a conditional distribution, $f_{\hat{X},\hat{Y}|X,Y,Z}$, which can then be used to

transform data in a probabilistic way. We compare our method to that of [48], as their formulation is an extension of that of [271].

This method minimizes some pre-defined notion of overall deviation of $f_{\hat{X},\hat{Y}}$ from $f_{X,Y}$. In the original work, the authors choose to minimize $\frac{1}{2}\sum_{x,y}\left|f_{\hat{X},\hat{Y}}(x,y) - f_{X,Y}(x,y)\right|$. They subjects this to constraints on point-wise distortion $E_{\hat{X},\hat{Y}|X,Y}[\delta((X,Y),(\hat{X},\hat{Y})]$ for some function $\delta : \{\mathbb{R}^p \times \{\pm 1\}\}^2 \to \mathbb{R}_+$. It also bounds the dependency of the new main label $\hat{Y}$ on the original protected label, $J\left(f_{\hat{Y}|Z}(y|z), f_Y(y)\right)$, where they define J to be the probability ratio measure:

$$J(a,b) = \left|\frac{a}{b} - 1\right|.$$

Thus, the final formulation is as follows:

$$\begin{aligned} \min \quad & \frac{1}{2}\sum_{x,y}\left|f_{\hat{X},\hat{Y}}(x,y) - f_{X,Y}(x,y)\right| \\ \text{s.t.} \quad & E_{\hat{X},\hat{Y}|X,Y}[\delta((X,Y),(\hat{X},\hat{Y})|x,y] \leq c, \forall x,y \\ & \left|\frac{1}{f_Y(y)} f_{\hat{Y}|Z}(y|z) - 1\right| \leq d, \forall y,z \\ & f_{\hat{X},\hat{Y}|X,Y,Z} \text{ are all distributions.} \end{aligned}$$

Following the authors, we approximate $f_{X,Y,Z}$ with the empirical distribution of the original data, separated into a pre-selected number of bins. Note that the resulting optimization problem will have $8(\#\text{bins})^{2p}$ parameters, and so can become computationally infeasible for high-dimensional datasets. To account for this, we follow the example of the original work and choose the 3 features most correlated with the main label, y. Each dimension is split into 8 bins. We choose $\delta((x',y'),(x,y))$ to be 0 if $y = y'$ and $x = x'$, 0.5 if $y = y'$ and x, x vary by at most one in any dimension, and 1 otherwise, which is similar to the δ chosen by the authors themselves. Finally, c and d were set at 0.1 and 0.3, respectively.

Synthetic Data

We sampled 1000 points each from $\vec{X}_+$ and $\vec{X}_-$ distributed as different 3-dimensional multivariate Gaussians, and these points are shown in Figure 4.1a. Figure 4.1b displays the results of dimensionality reduction using the top two unconstrained principal components of X: the resulting separators for linear and Gaussian kernel SVM's are also shown. It is clear that the two sub-populations are readily distinguishable in the lower-dimensional space. Figure 4.1c displays the analogous information after FPCA with only the mean constraint, and Figure 4.1d after FPCA with both constraints. Figures 4.1c and 4.1d clearly display better mixing of the data, and the SVM's conducted afterwards are unable to separate the

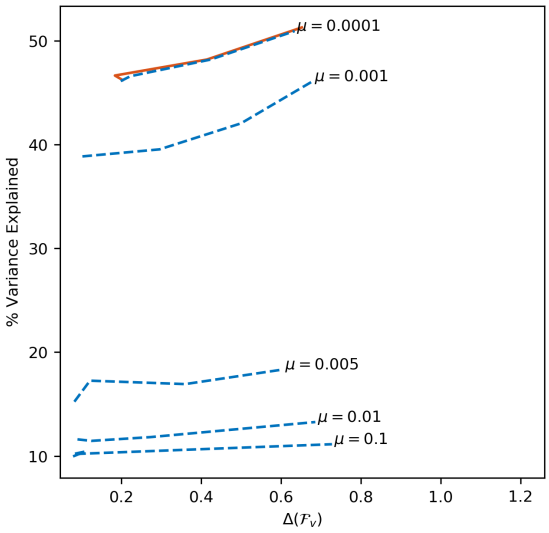


Figure 4.2: The sensitivity of FPCA to the δ and μ for the wine quality dataset. The full red line represents FPCA with only the mean constraint, and the dotted blue lines denote FPCA with both constraints. For each curve, $\delta \in \{0, 0.1, 0.3, 0.5\}$ was considered.

sub-groups as cleanly as they can in Figure 4.1b; furthermore, the addition of the covariance constraints (4.7d) incentivizes the choosing of a dimensionality reduction that better matches the skew of the entire data set.

Real data

We next consider a selection of datasets from UC Irvine's online Machine Learning Repository [154]. For each of the datasets, one attribute was selected as a protected class, and the remaining attributes were considered part of the feature space. After splitting each dataset into separate training (70%) and testing (30%) sets, the top five principal components were then found for the training sets of each of these datasets three times: once unconstrained, once with (4.7) with only the mean constraints (and excluding the covariance constraints) with $\delta = 0$, and once with (4.7) with both the mean and covariance constraints with $\delta = 0$ and $\mu = 0.01$; the test data was then projected onto these vectors. All data was normalized to have unit variance in each feature, which is common practice for datasets with features of incomparable units. For each instance, we estimated $\Delta(\mathcal{F})$ using the test set and for the families of linear SVM's $\mathcal{F}_v$ and Gaussian kernel SVM's $\mathcal{F}_k$. Finally, for each set of principal components V, the proportion of variance explained by the components was calculated as

Table 4.1: Δ-fairness for both linear and Gaussian kernel SVM for PCA and FPCA. Best results for each fairness metric are bolded.

	Unconstrained			FPCA - Mean Con.			FPCA - Both Con.		
Data Set	%var	Lin.	Gaus.	%var	Lin.	Gaus.	%var	Lin.	Gaus.
Adult Income	11.41	0.54	0.54	9.27	0.14	0.35	5.33	**0.07**	**0.15**
Biodeg [167]	31.16	0.2	0.35	30.46	0.14	0.29	21.45	**0.10**	**0.28**
E. Coli [115]	65.01	0.65	0.80	54.31	0.46	0.59	53.75	**0.24**	**0.54**
Energy [246]	84.08	0.10	0.20	66.48	**0.07**	0.20	62.11	**0.07**	**0.16**
German Credit	11.19	0.21	0.31	10.91	0.14	0.33	8.84	**0.11**	**0.29**
Image	62.68	0.18	0.32	52.78	**0.14**	0.33	48.55	0.15	**0.28**
Letter	42.33	0.58	0.58	29.29	**0.07**	0.22	23.76	**0.07**	**0.19**
Magic [38]	61.91	0.32	0.33	29.57	**0.11**	**0.21**	25.36	0.12	0.30
Pima [234]	49.00	0.30	0.37	43.98	**0.17**	0.26	43.26	0.18	**0.25**
Recidivism [11]	56.28	0.24	0.26	46.58	**0.08**	**0.16**	39.34	**0.08**	0.21
Skillcraft [243]	40.62	0.15	0.19	29.95	**0.07**	**0.14**	25.48	**0.07**	0.17
Statlog	87.80	0.79	0.79	21.77	0.23	0.69	7.76	**0.13**	**0.44**
Steel	46.05	0.64	0.71	40.79	0.19	0.51	11.86	**0.09**	**0.22**
Taiw. Credit [266]	45.52	0.11	0.17	30.07	0.08	0.16	20.08	**0.06**	**0.14**
Wine Quality [65]	50.21	0.97	0.96	37.34	0.21	0.51	10.12	**0.06**	**0.13**

$\operatorname{trace}(V\widehat{\Sigma}V^{\mathsf{T}}))/\operatorname{trace}(\widehat{\Sigma})$, where $\widehat{\Sigma}$ is the centered sample covariance matrix of training set X. Table 4.1 displays all of these results averaged over 5 different training-testing splits.

We may observe that our additional constraints are largely helpful in ensuring fairness by all definitions. Furthermore, in many cases, this increase in fairness comes at minimal loss in the explanatory power of the principal components. There are a few datasets for which (4.7d) appear superfluous. In general, gains in fairness are stronger with respect to $\mathcal{F}_v$; this is to be expected, as $\mathcal{F}_k$ is a highly sophisticated set, and thus more robust to linear projections. Kernel FPCA may be a better approach to tackling this issue, but we leave this for future work.

In table 4.2, we present additional fairness results using the family $\mathcal{F}_c$ of multivariate CDF's described in Section 4.3 (analogous to Kolmogorov-Smirnov distance) as a metric. We run this for unconstrained PCA, FPCA with only the mean constraint, FPCA with both constraints, and the method of [48]. We observe that our methods greatly improve fairness by this metric as well.

FPCA as a Preprocessing Step

Recall that a major use-case of PCA is as a pre-processing step in problems that suffer from the curse-of-dimensionality, or as a flexible way of ensuring fairness regardless of the type of task to be carried out (or algorithm used) on data. Thus, we present data showing that our method is competitive in this realm as well. In table 4.3, we present statistics for clustering done transformed data. Again, the methods used to transform the data are

Table 4.2: Δ-fairness levels for the multivariate KS distance, for PCA, FPCA. and the method of [48]. Best results for each fairness metric are bolded.

Data Set	Unconstrained	FPCA - Mean	FPCA - Both	Calmon et al.
Adult Income	0.25	0.16	**0.07**	0.25
Biodeg	0.16	**0.15**	0.17	**0.15**
Ecoli	0.64	0.29	0.32	**0.25**
Energy	0.16	0.12	**0.1**	0.18
German Credit	0.17	0.16	0.16	**0.13**
Image Seg	0.19	**0.16**	0.17	0.21
Letter Rec	0.57	**0.09**	**0.09**	0.24
Magic	0.14	**0.09**	0.12	0.16
Pima Diabetes	0.33	0.19	**0.18**	**0.18**
Recidivism	0.20	0.09	**0.07**	0.08
SkillCraft	0.12	**0.08**	**0.08**	**0.08**
Statlog	0.45	0.17	**0.12**	0.18
Steel	0.48	**0.10**	**0.10**	0.58
Taiwanese Credit	0.12	**0.07**	0.08	0.13
Wine Quality	0.58	0.20	**0.07**	0.44

PCA, FPCA with only the mean constraint, FPCA with both constraints, and the method of [48]. Reducing dimensionality prior to clustering is a common technique [4,142], so this is a relevant metric of comparison. For each case, we display the average squared distance from the closest cluster as a measure of accuracy, and the standard deviation of the proportion of each cluster that is of a certain protected class (the same metric reported in Section of 6.4 of the main document). That is, we consider the proportion of each cluster that is of protected class $Z = +1$ (in percentage points), and return the standard deviation of these figures (so the units would also be percentage points for these columns). In a given clustering, it is intuitive that the most fair outcome would be for every cluster to have the same composition in terms of protected classes (thus standard deviation of zero as mentioned above), so we maintain that this is a reasonable proxy for fairness. We observe that our method greatly reduces the unfairness within clusters, while not significantly decreasing the value of the clustering compared to a typical clustering. In some cases, we note that our method does even better in terms of accuracy; this may arise due to the fact that we are evaluating based on testing error as opposed to training error (i.e. we find cluster centers on training data and then find the closest cluster center for each point in the testing set). This suggests that our method may even act to aid in reducing generalization error.

Finally, we present an analysis of our method as a preprocessing step for classification in table 4.4. Here, we define a classification task on the datasets, and show the performance of linear SVM after dimensionality reduction via PCA, FPCA with the mean constraint and FPCA with both constraints. We compare these all with the method of [48], as before, but we also compare to the FO SVM (labeled "FSVM" for "Fair SVM") method presented in Chapter 3 (run with hyperparameters $\Delta_{1,1} = 0, \mu_{1,2} = 0.1$ on non-dimensionality-reduced

Table 4.3: Average squared distance from cluster center, as well as standard deviation of the proportion of each cluster that is of a certain protected class, for PCA, FPCA and the method of [48]. Best fairness results for each dataset are bolded.

	Unconstrained		FPCA - Mean		FPCA - Both		Calmon et al.	
Data Set	Score	Std. Dev	Score	Std. Dev	Score	Std. Dev	Score	Std. Dev
Adult Income	0.19	12.43	0.23	7.57	0.29	**2.28**	0.05	11.32
Biodeg	0.27	6.87	0.27	6.16	0.27	**5.34**	0.16	5.49
Ecoli	0.08	19.66	0.05	12.2	0.09	**10.69**	0.18	11.78
Energy	0.08	3.99	0.13	3.75	0.13	**3.57**	0.10	5.02
German Credit	0.25	6.4	0.25	4.82	0.28	**3.88**	0.03	4.16
Image Seg	0.10	8.46	0.09	**4.82**	0.11	5.95	0.12	10.85
Letter Rec	0.27	16.33	0.25	3.38	0.23	**3.28**	0.37	8.65
Magic	0.20	9.26	0.31	**5.15**	0.35	5.42	0.18	8.77
Pima Diabetes	0.24	9.09	0.27	6.36	0.26	5.96	0.28	**5.72**
Recidivism	0.26	7.6	0.17	**3.7**	0.19	3.8	0.05	4.69
SkillCraft	0.21	4.57	0.21	**2.27**	0.24	2.88	0.38	3.21
Statlog	0.09	21.99	0.23	16.06	0.31	**10.18**	0.13	11.12
Steel	0.16	18.49	0.19	9.85	0.24	**4.22**	0.22	17.97
Taiwanese Credit	0.17	3.85	0.24	2.99	0.29	**2.67**	0.03	3.64
Wine Quality	0.22	22.41	0.29	11.77	0.35	**2.11**	0.34	11.70

data), which was specifically designed for such a task. We compared the datasets based on fairness, as well as Area Under the Curve (AUC), which is measured as the area under the ROC curve of a classifier that takes a threshold as an input. We note that our method often produces more fair results. In some cases, our method matches or even beats the accuracy of FSVM. It is of importance that our method is a flexible method, while FSVM is specifically tailored to margin classifiers. Thus, it is to be expected that our method would not be strictly better in terms of accuracy. However, the comparison with regards to fairness is often quite favorable for our method.

Hyperparameter sensitivity

Next, we consider the sensitivity of our results to hyperparameters δ, μ, for the Wine Quality dataset. The data was split into training (70%) and testing (30%) sets, and the top three fair principle components were found using (4.7) with only the mean constraint for each candidate δ and using (4.7) with both constraints for all combinations of candidate δ and μ. All data was normalized to have unit variance in each independent feature. We calculated the percentage of the variance explained by the resulting principle components, and we estimated the fairness level $\Delta(\mathcal{F}_v)$ for the family of linear SVM's. This process was run 10 times for random data splits, and the averaged results are plotted in Figure 4.2. Here, the solid red line represents (4.7) with only the mean constraint. On the other hand, the dotted blue lines represent the (4.7) with both constraints, for the indicated μ.

Table 4.4: Comparison of accuracy and fairness on classification task using linear SVM. Results shown for linear SVM after dimensionality-reduction via PCA, FPCA with just the mean constraint and FPCA with both constraints, and are compared to the FSVM method of [191] (run with $\delta = 0, \mu = 0.1$ on non-dimensionality-reduced data) and the non-parametric method of [48]. Best fairness results are bolded.

	FSVM (no PCA)		Unconstrained		FPCA - Mean		FPCA - Both		Calmon et al.	
Data Set	AUC	Δ	AUC	Δ	AUC	Δ	AUC	Δ	AUC	Δ
Adult Income	0.86	0.13	0.66	0.17	0.69	**0.07**	0.57	0.08	0.51	0.23
Biodeg	0.85	0.12	0.82	0.20	0.81	0.13	0.79	**0.11**	0.60	0.14
Ecoli	0.74	**0.17**	0.84	0.50	0.69	0.23	0.72	0.29	0.63	0.30
Energy	0.55	0.09	0.51	0.09	0.56	0.08	0.55	**0.07**	0.54	0.13
German Credit	0.76	0.11	0.62	0.11	0.57	**0.10**	0.58	0.14	0.63	0.11
Image Seg	0.99	0.19	0.99	0.16	0.99	0.19	0.98	**0.15**	0.79	0.20
Letter Rec	0.72	**0.07**	0.58	0.60	0.50	0.09	0.49	0.10	0.65	0.19
Magic	0.83	0.13	0.74	0.14	0.82	0.13	0.72	**0.12**	0.65	0.13
Pima Diabetes	0.80	0.14	0.75	0.21	0.73	**0.11**	0.76	0.15	0.54	0.15
Recidivism	0.54	0.08	0.69	0.24	0.54	**0.06**	0.52	0.07	0.55	0.08
SkillCraft	0.82	0.06	0.85	0.10	0.82	**0.05**	0.80	**0.05**	0.62	0.07
Statlog	0.99	0.31	1.00	0.33	0.99	0.33	0.85	0.18	0.67	**0.16**
Steel	0.73	0.15	0.53	0.37	0.62	0.19	0.61	**0.12**	0.55	0.15
Taiwanese Credit	0.73	**0.07**	0.60	0.11	0.60	0.09	0.64	**0.07**	0.75	**0.07**
Wine Quality	0.78	0.10	0.69	0.75	0.69	0.19	0.67	**0.05**	0.66	0.09

Adding the covariance constraints and further tightening μ generally improves fairness and decreases the proportion of variance explained. However, observe that the relative sensitivity of fairness to δ is higher than that of the variance explained, at least for this dataset. Similarly, increasing μ decreases the portion of variance explained while resulting in a less discriminatory dataset after the dimensionality reduction. We note that increasing μ past a certain point does not provide much benefit, and so smaller values of μ are to be preferred. We found that increasing μ past 0.1 did not substantively change results further, so the largest μ that we consider is 0.1. In general, hyperparameters may be set with cross-validation, although (4.4) may serve as guidance.

Fair clustering of health data

Health insurance companies are considering the use of patterns of physical activity as measured by activity trackers in order to adjust health insurance rates of specific individuals [196, 220]. In fact, a recent clustering analysis found that different patterns of physical activity are correlated with different health outcomes [96]. The objective of a health insurer in clustering activity data would be to find qualitative trends in an individual's physical activity that help categorize the risks that that customer portends. That is, individuals within these activity clusters are likely to incur similar levels of medical costs, and so it would be

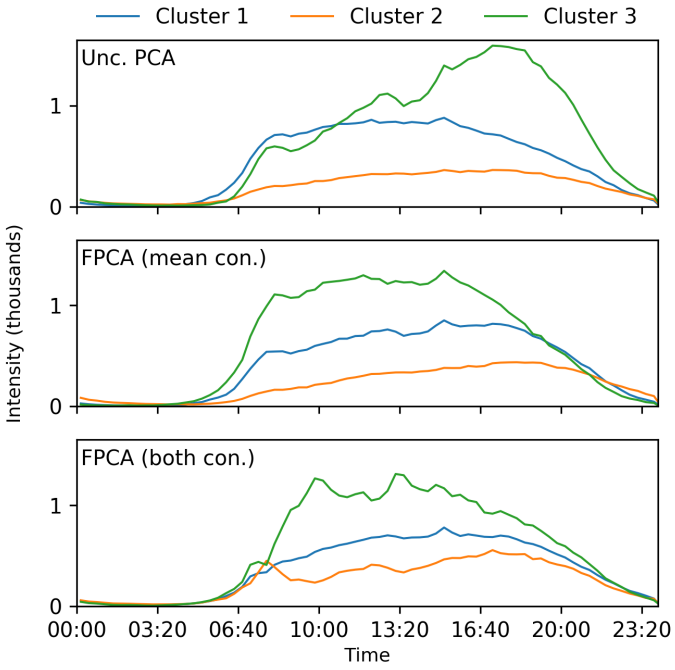


Figure 4.3: The mean physical activity intensities, plotted throughout a day, of the clusters generated after dimensionality reduction through PCA, FPCA with the mean constraint, and FPCA with both constraints. In each plot, each line represents the average activity level of the members of one cluster.

beneficial to engineer easy-to-spot features that can help insurers bucket customers. However, health insurance rates must satisfy a number of legal fairness considerations with respect to gender, race, and age. This means that an insurance company may be found legally liable if the patterns used to adjust rates result in an unreasonably-negative impact on individuals of a specific gender, race, or age. Thus, an insurer may be interested in a feature engineering method to bucket customers while minimizing discrimination on protected attributes. Motivated by this, we use FPCA to perform a fair clustering of physical activity. *Our goal is to find discernible qualitative trends in activity which are indicative of an individual's activity patterns, and thus health risks, but fair with respect to age.*

We use minute-level data from the the National Health and Nutrition Examination Survey (NHANES) from 2005–2006 [52], on the intensity levels of the physical activity of about 6000 women, measured over a week via an accelerometer. In this example, we consider age to be our protected variable, specifically whether an individual is above or below 40 years of

Table 4.5: The proportion of each cluster that are over 40 years of age. 36.05% of all respondents are over 40. The final row displays the standard deviation of the numbers in the first three. The most fair solution would be the same age composition in all clusters, so this is a reasonable fairness metric.

	Unc.	Mean	Both
Cluster 1	43.18%	33.54%	35.61%
Cluster 2	32.94%	38.64%	36.11%
Cluster 3	8.71%	33.32%	37.28%
Std. Dev	14.87%	2.46%	1.79%

age. We exclude weekends from our analysis, and average, over weekdays, the activity data by individual into 20-minute buckets. Thus, for each participant, we have data describing her average activity throughout an average day. We exclude individuals under 12 years of age, and those who display more than 16 hours of zero activity after averaging. The top 1% most active participants, and corrupted data, were also excluded. Finally, data points corrupted or inexact due to accelerometer malfunctioning were excluded. This preprocessing mirrors that of [96] and reflects practical concerns of insurers as well as the patchiness of accelerometer data.

PCA is sometimes used as a preprocessing step prior to clustering in order to expedite runtime. In this spirit, we find the top five principal components through PCA, FPCA with mean constraint, and FPCA with both constraints, with $\delta = 0$ and $\mu = 0.1$ throughout. Then we conduct k-means clustering (with $k = 3$) on the dimensionality-reduced data for each case. Figure 4.3 displays the averaged physical activity patterns for the each of the clusters in each of the cases. Furthermore, Table 4.5 documents the proportion of each cluster comprised of examinees over 40. We note that the clusters found under an unconstrained PCA are most distinguishable after 3:00 PM, so an insurer interested in profiling an individual's risk would largely consider their activity in the evenings. However, we may observe in Table 4.5 that this approach results in notable age discrimination between buckets, opening the insurer to the risk of illegal price discrimination. On the hand, the second and third plots in Figure 4.3 and columns in Table 4.5 suggest that clustering customers based on their activity during the workday, between 8:00 AM and 5:00 PM, would be less prone to discrimination.

4.7 Conclusion

This chapter has proposed a quantitative definition of fairness for dimensionality reduction, developed convex SDP formulations for fair PCA, and then demonstrated its effectiveness using several datasets. Many avenues remain for future research on fair unsupervised learn-

ing, including developing additional approaches for fair PCA. For instance, one algorithm for PCA involves a sequential calculation [70]. This insight was used to develop a deflation approach for sparse PCA [68]. Another technique for sparse PCA is based on a reformulation of PCA as a non-convex regression problem that can be solved by alternating minimization [277]. We believe that our formulations in this chapter may have suitable modifications that can be used to develop analogous deflation and regression approaches for fair PCA.

Chapter 5

Covariance-Robust Dynamic Watermarking

5.1 Introduction

Whereas previous chapters have concerned fairly well-developed areas of statistical learning, this chapter will extend notions of fairness to the realm of hypothesis testing. While this was touched upon in example 2 in Chapter 2, the notion of fair hypothesis testing is expanded here in a study of new, distributionally-robust dynamic watermarking schemes. Dynamic Watermarking is a hypothesis-testing framework for detecting attacks on dynamical systems, but it relies heavily on knowledge of the underlying system, and thus complete knowledge of the null hypothesis set. One major implication of the results of this chapter are that we expose a fundamental connection between fairness and robustness: By making a hypothesis test "fair", we can make it robust to errors in the null hypothesis. The material in this chapter borrows from material by the author in the following paper [193].

5G Applications as Motivation

The development of 5G, the "fifth generation" of wireless technology, brings with it increased bandwidth, massive-scale device-to-device (D2D) connections, lower latency and high reliability, all of which require more robust cyber-security measures for the relevant control systems [89]. A key feature of the cyber-security challenges posed by 5G is large numbers of evolving, parallel systems. For example, as the "point to multi-point" model of communications between base stations and devices is broken down in favor of decentralized, software-defined digital routing, the number of communication channels active at any time will increase exponentially [104, 120, 252]. Similarly, "beamforming" and radio transmission at millimeter wave frequencies, two more technologies core to 5G, will also necessitate constant multi-point transmission and control account for channel impairments and to compensate for severe path fading, respectively [148, 172, 207]. Furthermore, 5G's latency reductions open the door to growth in cyber-physical systems (CPS), which involve the intercommuni-

cation and real-time management of large numbers of physical sensors and actuators, often in shifting environments [89]. Opportunities for system vulnerabilities abound in all of these technologies [3, 49, 144].

When system dynamics are fixed and known, existing work in secure CPS and watermarking provides effective watermarking techniques for detecting attacks on LTI systems [178, 222, 257]. However, the dynamics of systems like those described above are often subject to certain types of "benign" changes. Existing tests for malicious activity that rely on fixed, or at least known, system dynamics can thus lose their power when actually in practice. Consider the case of controlling the transmission power of many mobile devices (referred to as uplink power control) to meet Signal-to-Noise requirements, while minimizing co-channel interference among devices on the same frequency band and reducing battery usage [6, 273]. Such control systems are the essence of mission-critical, yet constantly undergo shifts in dynamics and the distribution of noise due to something as simple as a mobile user walking past a building.

This chapter presents a novel, distributionally-robust dynamic watermarking scheme to test for adversarial attacks on LTI systems. Recent work has established watermarking as a key active method for detecting sensor attacks [178–181, 222, 223, 257]. Within these, a key segment has developed dynamic watermarking techniques, which ensure that only zero-average-power attacks can remain undetected [222, 223]. In this work, we consider a case where the observer noise in an LTI system does not have a fixed and known covariance, but is only known to be within a set of "reasonable" distributions. We design robust watermarking procedures for two sub-cases: First, the case where the covariance is fixed but unknown, and second, the case where the covariance can vary throughout time. The first sub-case reflects a scenario with many small systems, where it is not possible to estimate the state of, and design a new watermarking test for, each system individually; this could be relevant to millimeter-wave-frequency technology, as its high rate of attenuation will require denser networks with data streams split at multiple points before reaching transmission nodes, i.e. radio base stations. The second scenario is more relevant in larger systems that evolve over time; key examples of this include the uplink power control problem above, where observer noise comes from interference between devices and can clearly change as users move, or a self-driving car, whose sensors may experience hugely different noise in the case of a change of weather, road conditions, or sensor outages. Attack detection is critical in all of these cases, so we need statistical tests that retain their power in the face of predictable system changes.

Fairness

Robust data-driven decision-making has gained attention in the literature on algorithmic fairness. Motivated by machine learning tasks with societal applications, the fairness literature has sought to design learning methods that refrain from considering certain variables. To that extent, this body of work defines rigorous, mathematical notions of fairness for

supervised learning [24, 45, 57, 77, 109, 191, 270, 276], which have recently been extended to unsupervised learning by [56, 192].

The work in [16] outlines a general framework: Consider (X, Y, Z) with a joint distribution $\mathbb{P}$, where X are exogenous inputs, Y are endogenous "targets", and Z is a "protected attribute". The goal is to choose a *decision rule* $\delta(x)$ that makes a decision d using inputs X, in order to minimize some *risk function* $\mathcal{R}_{\mathbb{P}}(\delta, Y)$. In dynamic watermarking: X are measurements, Y is a binary variable that denotes if the system is under attack, and Z is the true system characterization; our decision rule δ for if the system is under attack is made without Y and Z, which are not observed. We then define a decision rule to be without *disparate impact* if

$$\delta^* \in \arg\min_{\delta} \left\{ R_{\mathbb{P}}(\delta, Y) \;\middle|\; \delta(X) \perp\!\!\!\perp Z \right\}, \tag{5.1}$$

where $\delta(X) \perp\!\!\!\perp Z$ means $\delta(X)$ is independent of Z. This increases fairness because it removes any impact of Z on the decision by imposing independence as a constraint. However, some [77, 109] have argued that this above definition of fairness can be too restrictive in some cases and that *equalized odds* is a better definition of fairness.

$$\delta^* \in \arg\min_{\delta} \left\{ R_{\mathbb{P}}(\delta, Y) \;\middle|\; (\delta(X) \perp\!\!\!\perp Z) \,\middle|\, Y \right\}, \tag{5.2}$$

That is, equalized odds ask for independence of $\delta(X)$ and Z when conditioned on Y. We can interpret equalized odds as requiring error rates to be similar across protected groups. Finally, a notion associated equalized odds is that of *equal opportunity*, which amounts to enforcing $(\delta(X) \perp\!\!\!\perp Z)|Y = y$, for some value y. This is relevant when one particular type of error is of more interest than another.

Relevance of Fairness to Watermarking

The topic of fair statistical learning has received increasing attention over the last few years [25, 56, 57, 109, 191, 192, 270]. Here, we argue that the notions presented in this literature are key to the problem of designing robust tests. Work in this field has sought to design learning methods that actively guarantee some notion of independence between the output of the model learned, and an exogenously chosen "protected variable" [16, 77, 109]. Importantly, the protected variable may be inherently tied to the desired output; in such situations, different notions of independence are derived which allow for more nuance and specificity [16].

Fairness is relevant to the design of robust tests for two reasons. First, it provides a well-established technical language with which to discuss our requirement of robustness. Past dynamic watermarking techniques require exact system knowledge, and as such the corresponding watermarking tests will have error rates that are biased over inevitable system perturbations or uncertainties. Fairness notions such as *equalized odds* and *equal opportunity* allow for more specific framing of the problem and thus give a framework to design more robust methods for dynamic watermarking.

Second, robust cyber-security methods will have improved social impacts, which is the most general way of interpreting "fairness". For example, complex CPS like smart homes comprise possibly hundreds of sensors, the integrity of which are critical to the well-being of the individuals who occupy the home. Changes in the distribution of sensor noise can correlate with factors such as climate, which correlates with geography and thus attributes like race, ethnicity or class. A systemic bias in the ability to detect threats thus yields, and possibly perpetuates, systemic bias in outcomes among these groups. The field of fair machine learning has largely existed to address such concerns, raised through notable, though anecdotal, recognition of the impact of biased machine learning [11, 20, 79, 81, 84, 214].

Outline

In Sect. 5.2, we outline key terminology and results in dynamic watermarking. In Sect. 5.3, we present our covariance-robust dynamic watermarking (CRDW) scheme for the case of fixed, but unknown, measurement noise covariance. This is then extended in Sect. 5.3 to the case where measurement noise covariance is allowed to slowly vary. Sect. 5.4 presents empirical results that demonstrate efficacy of our approach.

5.2 Preliminaries

We describe the LTI system and attack models, and then review existing results about dynamic watermarking.

LTI System Model

Consider a partially-observed MIMO LTI system

$$\begin{aligned} x_{n+1} &= Ax_n + Bu_n + w_n \\ y_n &= Cx_n + z_n + v_n \end{aligned} \tag{5.3}$$

for $x_n, w_n \in \mathbb{R}^p, u_n \in \mathbb{R}^q$ and $y_n, z_n, v_n \in \mathbb{R}^m$. Here w_n is mean-zero i.i.d. multivariate Gaussian process noise with covariance matrix Σ_W, and this is independent of z_n that is i.i.d. Gaussian measurement noise with mean-zero; but we assume that the covariance matrix for z_n is a linear function $\Sigma_Z(\theta)$ of a set of parameters $\theta \in \mathcal{P} \subset \mathbb{R}^d$ taking values in polyhedron $\mathcal{P}$. For now, θ is assumed constant but unknown for any fixed system. The v_n is an additive signal chosen by an attacker who seeks to corrupt sensor measurements.

Stabilizability of (A, B) and detectability of (A, C) imply the existence of a controller K and observer L such that $A+BK$ and $A+LC$ are Schur stable. The closed-loop system can be stabilized using the control input $u_n = K\hat{x}_n$, where $\hat{x}_n$ is the observer-estimated state. Define $\tilde{x}_n = \begin{bmatrix} x_n^\mathsf{T} & \hat{x}_n^\mathsf{T} \end{bmatrix}^\mathsf{T}$, $\underline{D} = \begin{bmatrix} I & 0 \end{bmatrix}^\mathsf{T}$, $\underline{L} = \begin{bmatrix} 0 & -L^\mathsf{T} \end{bmatrix}^\mathsf{T}$, and

$$\underline{A} = \begin{bmatrix} A & BK \\ -LC & A + BK + LC \end{bmatrix}. \tag{5.4}$$

We can write the closed-loop evolution of the state and estimated state when $v_n \equiv 0$ as $\tilde{x}_{n+1} = \underline{A}\tilde{x}_n + \underline{D}w_n + \underline{L}z_n$. Alternatively, we may define the observation error $\delta_n = \hat{x}_n - x_n$. Let $\breve{x}_n = \begin{bmatrix} x_n^\mathsf{T} & \delta_n^\mathsf{T} \end{bmatrix}^\mathsf{T}$, $\underline{\underline{D}} = \begin{bmatrix} I & -I \end{bmatrix}^\mathsf{T}$, $\underline{\underline{L}} = \underline{L}$, and

$$\underline{\underline{A}} = \begin{bmatrix} A + BK & BK \\ 0 & A + LC \end{bmatrix}. \tag{5.5}$$

The closed-loop system for this change of variables is $\breve{x}_{n+1} = \underline{\underline{A}}\breve{x}_n + \underline{\underline{D}}w_n + \underline{\underline{L}}z_n$. Note that $\underline{\underline{A}}$ is Schur stable since both $A + BK$ and $A + LC$ are Schur stable.

Attack Model

Following [111], we consider attacks where $v_n = \alpha(Cx_n + z_n) + C\eta_n + \zeta_n$ for a fixed $\alpha \in \mathbb{R}$ and i.i.d. Gaussian ζ_n with mean-zero and covariance matrix Σ_S. Here, the η_n are chosen to follow the process $\eta_{n+1} = (A+BK)\eta_n+\omega_n$, where ω_n are similarly i.i.d. Gaussian with mean-zero and covariance matrix Σ_O. The implication is that the attacker minimizes or mutes the true output Cx_n+z_n, and instead replaces it with a simulated output that follows the system dynamics and is thus not easily distinguishable as false. Furthermore, the attacker has access to process w_n and measurement noise z_n. With this attack, the closed-loop systems above become $\tilde{x}_{n+1} = \underline{A}\tilde{x}_n + \underline{D}w_n + \underline{L}(z_n + v_n)$ and $\breve{x}_{n+1} = \underline{\underline{A}}\breve{x}_n + \underline{\underline{D}}w_n + \underline{\underline{L}}(z_n + v_n)$.

(Nonrobust) Dynamic Watermarking

The steady-state distribution of δ_n in an unattacked system will be Gaussian with mean-zero and a covariance matrix of

$$\Sigma_\Delta = (A + LC)\Sigma_\Delta(A + LC)^\mathsf{T} + \Sigma_W + L\Sigma_Z(\theta)L^\mathsf{T}. \tag{5.6}$$

Dynamic watermarking adds a small amount of Gaussian noise e_n, the values unknown to the attacker, into the control input $u_n = K\hat{x}_n + e_n$. This private excitation has mean-zero and covariance matrix Σ_E. Defining $\underline{B} = \begin{bmatrix} B^\mathsf{T} & B^\mathsf{T} \end{bmatrix}^\mathsf{T}$ and $\underline{\underline{B}} = \begin{bmatrix} B^\mathsf{T} & 0 \end{bmatrix}^\mathsf{T}$, the closed-loop systems with watermarking are given by $\tilde{x}_{n+1} = \underline{A}\tilde{x}_n + \underline{B}e_n + \underline{D}w_n + \underline{L}(z_n + v_n)$ and $\breve{x}_{n+1} = \underline{\underline{A}}\breve{x}_t + \underline{\underline{B}}e_n + \underline{\underline{D}}w_n + \underline{\underline{L}}(z_n + v_n)$, respectively.

The watermarking noise e_n leaves a detectable signal in the measurements y_n, which can detect the presence of an attack v_n by comparing the observer error $C\hat{x}_n - y_n$ to previous values of the watermark e_{n-k} for some integer $k > 0$. Specifically, the work in [111] proposes the tests

$$\begin{aligned} &\text{as-lim}_{N\to\infty} \tfrac{1}{N} \textstyle\sum_{n=0}^{N-1} (C\hat{x}_n - y_n)(C\hat{x}_n - y_n)^\mathsf{T} = \\ &\qquad\qquad C\Sigma_\Delta C^\mathsf{T} + \Sigma_Z \end{aligned} \tag{5.7}$$

$$\text{as-lim}_{N\to\infty} \tfrac{1}{N} \textstyle\sum_{n=0}^{N-1} (C\hat{x}_n - y_n)e_{n-k'-1}^\mathsf{T} = 0, \tag{5.8}$$

where $k' = \min_{k\geq 1}\{C(A+BK)^k B^\mathsf{T} \neq 0\}$. Any modeled attack passing these tests can be shown to asymptotically have zero power as-$\lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} v_n^\mathsf{T} v_n = 0$ [111].

Finally, [111] also provides a test statistic for implementing the above test. Define $\psi_n = \begin{bmatrix}(C\hat{x}_n - y_n)^\mathsf{T} & e_{n-k'-1}^\mathsf{T}\end{bmatrix}^\mathsf{T}$ and $S_n = \sum_{i=n+1}^{n+\ell} \psi_n \psi_n^\mathsf{T}$. Then the negative log-likelihood of a Wishart distribution is

$$\begin{aligned}\mathcal{L} =& (m+q+1-\ell)\log\det S_n \\ &+ \operatorname{trace}\left\{\begin{bmatrix}\left(C\Sigma_\Delta C^\mathsf{T} + \Sigma_Z\right)^{-1} & 0 \\ 0 & \Sigma_E^{-1}\end{bmatrix} \times S_n\right\}.\end{aligned} \tag{DW}$$

This can be used to perform a statistical hypothesis test to detect attacks when using dynamic watermarking.

5.3 Covariance-Robust Dynamic Watermarking

We develop covariance-robust dynamic watermarking methods for two different cases. The first is where θ is fixed but unknown, and the second is where θ is slowly varying.

Fixed But Unknown Noise Covariance

We begin by stating our assumptions for this case. First, we assume that we have knowledge of a set of positive semidefinite matrices $\Sigma_{z,1}, \ldots, \Sigma_{z,d}$ such that these matrices are affinely independent and $\Sigma_Z(\theta) \in \operatorname{int}(\Omega^Z)$ for the set

$$\Omega^Z = \{\theta_1\Sigma_{z,1} + \cdots + \theta_d\Sigma_{z,d} : \mathbf{1}^T\theta = 1, \theta \geq \mathbf{0}\}. \tag{5.9}$$

Note that Ω^Z is a polyhedron, and that this set is defined to be the convex combination of $\Sigma_{z,1}, \ldots, \Sigma_{z,d}$. Our first result characterizes Ω^Δ, which is the set of possible $\Sigma_\Delta(\theta)$.

Lemma 1. *Let $\bar{\Sigma}_{\delta,k}$ satisfy $\bar{\Sigma}_{\delta,k} = (A+LC)\bar{\Sigma}_{\delta,k}(A+LC)^T + \Sigma_W + L\Sigma_{z,k}L^T$. For $\Sigma_Z(\theta) = \theta_1\Sigma_{z,1} + \cdots + \theta_d\Sigma_{z,d}$, the solution to (5.6) is $\Sigma_\Delta(\theta) = \theta_1\bar{\Sigma}_{\delta,1} + \cdots + \theta_d\bar{\Sigma}_{\delta,d}$.*

Proof. This immediately follows by noting that both sides of (5.6) are linear in the matrices Σ_Δ and $\Sigma_Z(\theta)$. □

Since $E[\psi_n\psi_n^\mathsf{T}] = \operatorname{blkdiag}\{C\Sigma_\Delta C^\mathsf{T} + \Sigma_Z, \Sigma_E\}$, we need to characterize the set Ω of feasible matrices in terms of θ.

Lemma 2. *Let $\bar{\Sigma}_k = \operatorname{blkdiag}\{C\bar{\Sigma}_{\delta,k}C^T + \Sigma_{z,k}, \Sigma_E\}$. Then $\Omega = \{\theta_1\bar{\Sigma}_k + \cdots + \theta_d\bar{\Sigma}_d : \mathbf{1}^T\theta = 1, \theta \geq \mathbf{0}\}$.*

Proof. This follows by the linearity in Σ_Δ and Σ_Z. □

The set Ω represents covariance matrices of ψ_n that are "acceptable", according to the original set Ω^Z of observation noise covariances that we should not mistake for attacks.

Lemma 3. *The set Ω is of dimension $d-1$.*

Proof. This follows from Lemma 1, the fact that L is of full column-rank, and the observability of $(A+LC, C)$, which in turn follows from the observability of (A, C). □

Finally, consider a modification of (DW) given by

$$\mathcal{L}(S_n, V) = (m+q+1-\ell)\log\det S_n + \operatorname{trace}\{VS_n\} - \ell\log\det V. \quad (5.10)$$

Note (5.10) is the negative log-likelihood of an $(m+q)\times(m+q)$ Wishart distribution with scale matrix V^{-1} and ℓ degrees of freedom. Now, we may present our test statistic. Let $\Omega^{-1} = \{V : V^{-1} \in \Omega\}$ and define the test statistic

$$T(S_n) = \min_{V\in\Omega^{-1}} \mathcal{L}(S_n, V) \quad (5.11)$$

for the composite null hypothesis $H_0 : E[\psi_n\psi_n^\mathsf{T}] \in \operatorname{int}(\Omega)$. For some $0 \le \nu$, consider the test

$$\begin{cases} \text{reject } H_0 & \text{if } T(S_n) > \nu \\ \text{accept } H_0 & \text{if } T(S_n) \le \nu. \end{cases} \quad (5.12)$$

Since $\arg\min_{V\in\Omega^{-1}} \mathcal{L}(S_n, V) = S_n^{-1}$, this proposed test is equivalent to the generalized likelihood ratio test.

Theorem 10. *For large enough ℓ, the decision rule (5.12) using test statistic $T(S_n)$ satisfies equal opportunity with respect to the null hypothesis and where the protected attribute is the true measurement noise covariance $\Sigma_Z(\theta) \in \operatorname{int}(\Omega^Z)$.*

Proof. Due to Lemma 3 and our assumption that $\Sigma_Z(\theta) \in \operatorname{int}(\Omega^Z)$, $T(S_n)$ satisfies the Le Cam regularity conditions required for the application of Wilk's Theorem [260]. This means $-2T(S_n)$ will be asymptotically distributed as a $\chi^2(m+q-p)$ random variable plus a fixed constant *regardless of the true value of* Σ_Δ, and thus implies that the event of a Type I error is independent of Σ_Δ. □

This is a useful result because it implies that, in the proper regime, our test can come arbitrarily close to satisfying the initial goal of remaining robust to some uncertainty in the distribution of the measurement noise. However, Ω^{-1} is a non-convex set, and so the computation of $T(S_n)$ is difficult. To this end, we propose the approximate test statistic

$$\begin{aligned} \bar{T}(S_n) = \min \quad & \mathcal{L}(S_n, V) \\ \text{s.t.} \quad & \textstyle\sum_{k=1}^p \theta_k \bar{\Sigma}_k^{-1} \succeq V, \\ & \begin{bmatrix} V & I \\ I & \sum_{k=1}^p \theta_k \bar{\Sigma}_k \end{bmatrix} \succeq 0, \\ & \mathbf{1}^\mathsf{T}\theta = 1, \\ & \theta \ge \mathbf{0}. \end{aligned} \quad (\text{CRDW})$$

Lemma 4. *For any $V \in \Omega^{-1}$, there exists a $\theta \in \mathbb{R}^p$ such that (V, θ) is a feasible solution to the optimization problem defining test (CRDW).*

Proof. First observe that any $V \in \Omega^{-1}$ can be written as $V = (\sum_{k=1}^p \theta_k \bar{\Sigma}_k)^{-1}$ for some nonzero θ such that $\mathbf{1}^\mathsf{T}\theta = 1$. Thus, it holds trivially that

$$(\textstyle\sum_{k=1}^p \theta_k \bar{\Sigma}_k)^{-1} \succeq V \succeq (\sum_{k=1}^p \theta_k \bar{\Sigma}_k)^{-1} \tag{5.13}$$

The right-most constraint in (5.13) can be restated using the Schur complement, and this reformulation is exact. Since $\sum_{k=1}^p \theta_k \bar{\Sigma}_k \succeq 0$, the Schur complement implies the second constraint in (CRDW) is equivalent to $V - (\sum_{k=1}^p \theta_k \bar{\Sigma}_k)^{-1} \succeq 0$.

The first constraint in (CRDW) follows from the convexity of the matrix inverse for positive semidefinite matrices: Letting $X(\tau) = (1-\tau)X_1 + \tau X_2$ for positive definite $n \times n$ matrices X_1, X_2 and $0 \leq \tau \leq 1$, we have $\frac{\nabla^2}{\nabla \tau^2} X(\tau)^{-1} = 2X^{-1}(\tau)X'(\tau)X^{-1}(\tau)X'(\tau)X^{-1}(\tau)$. For any $a \in \mathbb{R}^n$, the function $\phi_a(\tau) = a^\mathsf{T} X^{-1}(\tau) a$ will have second derivative $\phi_a''(\tau) = 2a^\mathsf{T} X^{-1}(\tau)X'(\tau)X^{-1}(\tau)X'(\tau)X^{-1}(\tau)a \geq 0$ due to the positive-semidefiniteness of $X(\tau)^{-1}$, so $(1-\tau)\phi_a(0) + \tau\phi_a(1) \geq \phi_a(\tau)$. Since this holds for any a, we have that

$$\textstyle\sum_{k=1}^p \theta_k \bar{\Sigma}_k^{-1} \succeq (\sum_{k=1}^p \theta_k \bar{\Sigma}_k)^{-1}. \tag{5.14}$$

The first constraint in (CRDW) follows from (5.13) and (5.14). □

Remark 14. It was shown in [111] that test (5.8) ensures $\alpha = 0$ in any attack such that it holds true. In that case, we have

$$\begin{aligned}\text{as-lim}_{N\to\infty} \tfrac{1}{N} \textstyle\sum_{n=0}^{N-1} (C\hat{x}_n - y_n)(C\hat{x}_n - y_n)^\mathsf{T} \\ = C\Sigma_\Delta(\theta)C^\mathsf{T} + \Sigma_Z(\theta) + \\ \Sigma_S + \text{as-lim}_{N\to\infty} \tfrac{1}{N} \textstyle\sum_{n=0}^{N-1} C\eta_n\eta_n^\mathsf{T}C^\mathsf{T},\end{aligned} \tag{5.15}$$

since the Schur stability of $A + BK$ implies that any effect of x_0 and η_0 are reduced to zero asymptotically. Since Σ_S and $\text{as-lim}_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} C\eta_n\eta_n^\mathsf{T}C^\mathsf{T}$ are both positive semidefinite, meaning that

$$\begin{aligned}\text{as-lim } \tfrac{1}{N} \textstyle\sum_{n=0}^{N-1} (C\hat{x}_n - y_n)(C\hat{x}_n - y_n)^\mathsf{T} \succeq \\ C\Sigma_\Delta(\theta)C^\mathsf{T} + \Sigma_Z(\theta).\end{aligned} \tag{5.16}$$

Inverting both sides of this implies that, in the case that $\Sigma_S + \text{as-lim}_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} C\eta_n\eta_n^\mathsf{T}C^\mathsf{T} \neq 0$, we can generally expect that $S_n^{-1} \preceq (C\Sigma_\Delta(\theta)C^\mathsf{T} + \Sigma_Z(\theta))^{-1} \in \Omega^{-1}$. The takeaway is that the looseness of the upper bound (5.14) should not greatly decrease the power of the modified test in the presence of test (5.8), as the tight lower bound is more germane to situations where the system is actually being attacked.

Remark 15. If the dimension $m+q$ is large, then the optimization (CRDW) may be expensive to solve from scratch each time. Furthermore, S_n will likely not change drastically between runs when ℓ is large. So, lighter-weight first-order methods such as ADMM can be used instead [258]. These generally take longer to converge to high levels of accuracy, but have the advantage of being able to be readily warm-started.

Slowly Varying Unknown Noise Covariance

A key difference between this setting and that of the static distribution is that a shift in the observer noise covariance in one period can have impacts on Σ_Δ over the next few periods that do not easily fit into our previous representation of the Ω. This is because it will take many steps before the covariance of δ_n approaches its asymptotic limit in Ω. Thus, to accommodate a dynamically changing distribution of z_n, we must use an expansion of the set Ω.

We modify our setup for this subsection. The true covariance of δ_n and z_n are Σ_{Δ_n} and Σ_{Z_n}, respectively. Let $\Psi_n = \Sigma_{Z_n} - \Sigma_{Z_{n-1}}$ and $\Phi_n^j = (A+LC)^j L\Psi_n L^\mathsf{T}(A+LC)^{j\mathsf{T}}$. Note that all Σ_{Z_n} are still assumed to be in Ω^Z. Finally, we make some additional assumptions. Since the spectral radius of $A+LC$ is less than one, there exists some induced norm (denote this $\|\cdot\|$) such that $\|A+LC\| < 1$ [114]. We assume θ changes every step but $\Sigma_{Z_0} \in \Omega$ and all Ψ_n satisfy $\|\Psi_n\| \leq \xi$ for some known value of $\xi > 0$. We also assume the system starts at steady state in the sense $\Sigma_{\Delta_0} = (A+LC)\Sigma_{\Delta_0}(A+LC)^\mathsf{T} + \Sigma_W + L\Sigma_{Z_0}L^\mathsf{T}$. Under these assumptions we have:5

Lemma 5. *Let $\varepsilon \in \mathbb{R}$ be defined as*

$$\varepsilon = \frac{\xi\|C\|^2\|L\|^2\|A+LC\|^2\sqrt{m}}{\left(1-\|A+LC\|^2\right)^2} \tag{5.17}$$

Then $C\Sigma_{\Delta_n}C^\mathsf{T} + \Sigma_{Z_n} \in \Omega \oplus \{E : -\varepsilon I \preceq E \preceq \varepsilon I\}$, where $\oplus$ is the Minkowski sum for all n.

Proof. Let $\Omega_{m\times m}$ be the set of $m \times m$ upper-left submatrices of elements of Ω, associated with $C\Sigma_\Delta(\theta)C^\mathsf{T} + \Sigma_Z(\theta)$ terms. We start by noting that

$$\begin{aligned}\Sigma_{\Delta_1} =& (A+LC)\Sigma_{\Delta_0}(A+LC)^\mathsf{T} + \Sigma_W + L\Sigma_{Z_1}L^\mathsf{T} \\ =& \Sigma_{\Delta_0} + \Phi_1^0.\end{aligned} \tag{5.18}$$

Similarly, we can see that $\Sigma_{\Delta_2} = \Sigma_{\Delta_0} + L\left(\Psi_0+\Psi_1\right)L^\mathsf{T} + \Phi_0^1 = \Sigma_{\Delta_0} + \Phi_2^0 + \Phi_1^0 + \Phi_1^1$. Continuing this recursion relation leads to the fact that

$$\Sigma_{\Delta_n} = \Sigma_{\Delta_0} + \textstyle\sum_{i=0}^{n-1}\sum_{j=0}^{i}\Phi_{n-i}^j. \tag{5.19}$$

Due to the Schur stability of $A+LC$, the following limit exists, and can be represented as in Lemma 1.

$$\Sigma_{\Delta_\infty^{k'}} = \lim_{k\to\infty}\left(\Sigma_{\Delta_0} + \textstyle\sum_{i=k-k'}^{k-1}\sum_{j=0}^{i}\Phi_{k-i}^j\right) \tag{5.20}$$

Note that $\Sigma_{\Delta_\infty^{k'}}$ is the steady state that Σ_{Δ_n} would ultimately reach if θ (and therefore Σ_{Z_n} does not shift after step k'; thus, it solves (5.6) for $\Sigma_{Z_{k'}}$ and exists in $\Omega_{m\times m}$. Denote $\Upsilon_i = \Sigma_{\Delta_\infty^i} - \Sigma_{\Delta_\infty^{i-1}}$. Then,

$$\begin{aligned}\Sigma_{\Delta_n} &= \lim_{k\to\infty}\left(\Sigma_{\Delta_0} + \textstyle\sum_{i=k-n}^{k-1}\sum_{j=0}^{i-k+n}\Phi_{k-i}^j\right) \\ &= \Sigma_{\Delta_\infty^n} - \lim_{k\to\infty}\left(\textstyle\sum_{i=k-n}^{k-1}\left(\sum_{j=i-k+n+1}^{i}\Phi_{k-i}^j\right)\right) \\ &= \Sigma_{\Delta_\infty^n} - \textstyle\sum_{i=1}^{n}(A+LC)^{n-i+1}\Upsilon_i(A+LC)^{n-i+1\mathsf{T}}\end{aligned} \tag{5.21}$$

Note that the term in the limit in the first equality is a constant in k due to a simple re-indexing of (5.19). This is convenient because we can now break Σ_{Δ_n} into an element known to be in $\Omega_{m\times m}$ and an error term. Our goal is now to choose ε large enough to bound

$$\min_{P\in\Omega_{m\times m}} \left\|C\Sigma_{\Delta_n}C^\mathsf{T}+\Sigma_{Z_n}-P\right\|_2, \tag{5.22}$$

over all paths that Σ_{Z_n} can take. An easy bound on the minimization is to simply set $P=C\Sigma_{\Delta_\infty^n}C^\mathsf{T}+\Sigma_{Z_n}$. Then, ε only needs to exceed

$$\left\|\textstyle\sum_{i=1}^n C(A+LC)^{n-i+1}\Upsilon_i(A+LC)^{n-i+1^\mathsf{T}}C^\mathsf{T}\right\|_2 \tag{5.23}$$

By sub-multiplicativity of induced norms,

$$\begin{aligned}\|\Upsilon_i\| &= \left\|\textstyle\sum_{j=0}^\infty \Phi_i^j\right\| \le \textstyle\sum_{j=0}^\infty \|(A+LC)\|^{2j}\|L\|\|\Psi_{n'+k_i}\| \\ &= \xi\|L\|^2\left(1-\|A+LC\|^2\right)^{-1}\end{aligned} \tag{5.24}$$

Finally, using the fact that $\|\cdot\|_2\le\sqrt{m}\|\cdot\|$ [87] and applying (5.24) to the error term from (5.22) yields the desired result. □

Remark 16. Due to the topological equivalence of induced norms, the dependence of our choice of norm $\|\cdot\|$ on $A+LC$ can only affect the value of ξ required by a constant $\sqrt{m}$.

Corollary 2. *If $\|A+LC\|_2<1$, then the statement in Lemma 5 holds for $\|\cdot\|=\|\cdot\|_2$ and*

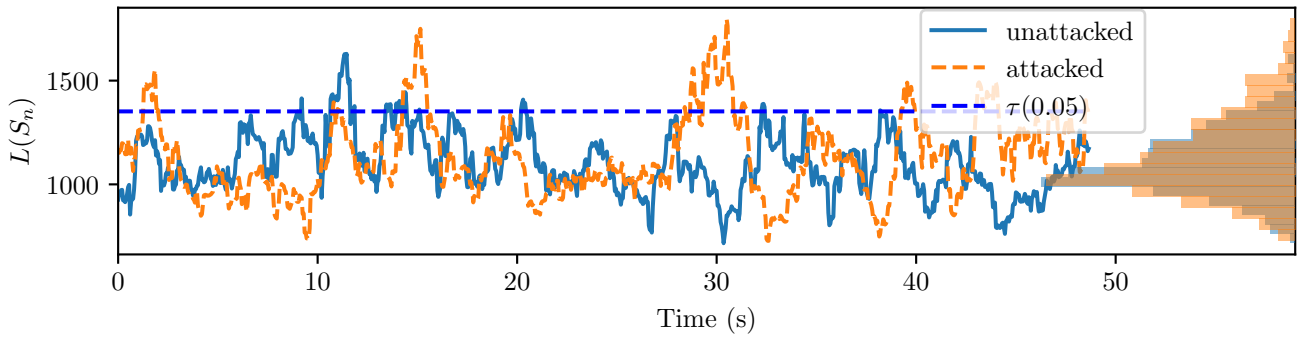
$$\varepsilon=\frac{\xi\|C\|_2^2\|L\|_2^2\|A+LC\|_2^2}{\left(1-\|A+LC\|_2^2\right)^2} \tag{5.25}$$

Proof. The proof of this result is almost identical to the proof of the previous lemma with the only changes that $\|\cdot\|=\|\cdot\|_2$ and that we stop after applying (5.24) to (5.22). □

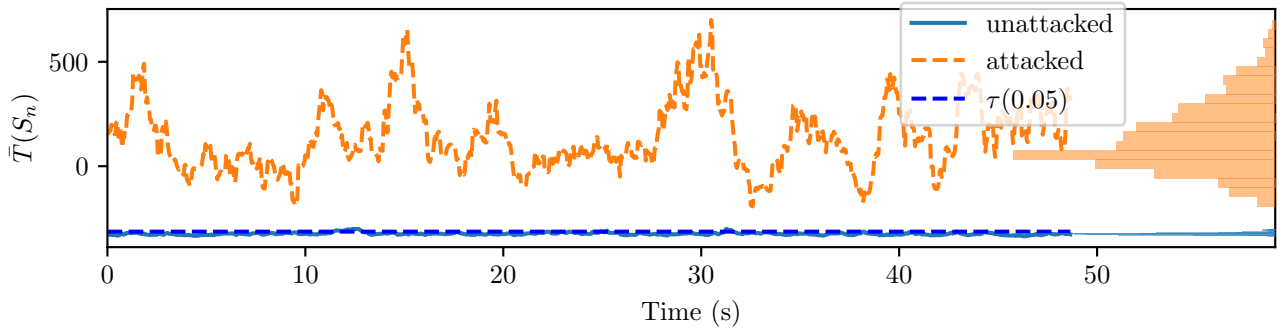
With this ε, it is straightforward to extend the previous test statistic (CRDW) to this new expansion of Ω as long as $\bar{\Sigma}_k-\varepsilon I$ remains positive definite for all k. In this case, we may define our new test statistic as

$$\begin{aligned}\underline{T}(S_n)=\min\quad & \mathcal{L}(S_n,V) \\ \text{s.t.}\quad & \textstyle\sum_{k=1}^p\theta_k\left(\bar{\Sigma}_k-\varepsilon I\right)^{-1} \succeq V, \\ & \begin{bmatrix}V & I\\ I & \varepsilon I+\sum_{k=1}^p\theta_k\bar{\Sigma}_k\end{bmatrix}\succeq 0, \\ & \mathbf{1}^\mathsf{T}\theta = 1, \\ & \theta \ge \mathbf{0}.\end{aligned} \tag{CRDW*}$$

Remark 17. If there is some k so $\bar{\Sigma}_k-\varepsilon I$ is not positive definite, then the first constraint above is not well-defined. Recalling that V is a surrogate for $\left(C\Sigma_{\Delta_n}C^\mathsf{T}+\Sigma_{Z_n}\right)^{-1}$, we note V trivially satisfies $\Sigma_{Z_n}^{-1}\succeq V$. Thus in this problematic case, we may replace the $\left(\bar{\Sigma}_k-\varepsilon I\right)$ in the first constraint with $\Sigma_{z,k}$, for all k. This issue is unlikely to be of practical concern for the same reasons discussed in Remark 14 regarding the relaxation of the set Ω. Specifically, the structure of the attacks makes it unlikely that the first constraint in (CRDW*) would be binding in any case.



(a) Test statistic (DW)



(b) Test statistic (CRDW)

Figure 5.1: The evolution and histogram of test statistics (DW) and (CRDW) on the attacked and unattacked systems where Σ_Z is fixed, but unknown to the tester. In this case, the nonrobust test statistic (DW) is unable to clearly distinguish the attacked from the unattacked system, whereas the new test statistic (CRDW) can.

5.4 Empirical Results

In this section, we present simulation results that showcase the strength of our method when compared with the original test statistic (DW). We present results for both the case where the noise distribution is fixed but unknown, and for the case where the noise covariance is unknown and slowly-varying.

We use the standard model for simulation of an autonomous vehicle in [248], where the error kinematics of lane keeping and speed control is given by $x^\mathsf{T} = \begin{bmatrix}\psi & y & s & \gamma & v\end{bmatrix}$ and $u^\mathsf{T} = \begin{bmatrix}r & a\end{bmatrix}$. Here, ψ is heading error, y is lateral error, s is trajectory distance, γ is vehicle angle, v is vehicle velocity, r is steering, and a is acceleration. We linearize and initialize with a straight trajectory and constant velocity $v_0 = 10$. We then performed exact discretization

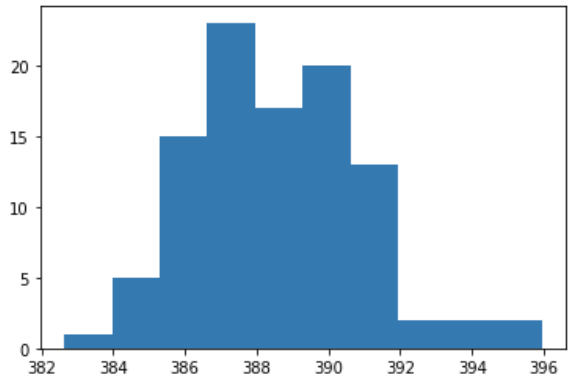


Figure 5.2: The same test as displayed in Fig. 5.1, run over 100 random instantiations of the true Σ_Z. The average difference between the test statistics (DW) and (CRDW) was recorded for each instantiation, and used to generate this histogram. Larger numbers here indicate that (CRDW) outperforms (DW).

with sampling period $t_s = 0.05$. This yields the system dynamics

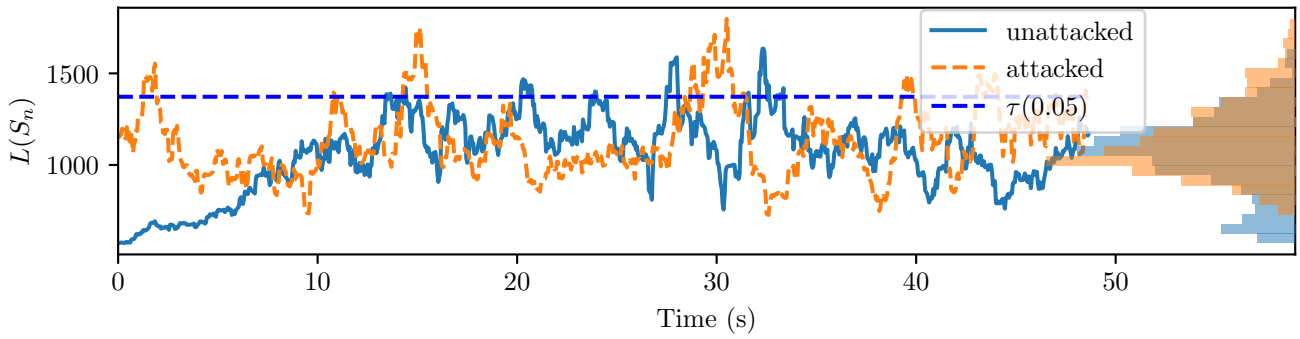
$$A = \begin{bmatrix} 1 & 0 & 0 & \frac{1}{10} & 0 \\ \frac{1}{2} & 1 & 0 & \frac{1}{40} & 0 \\ 0 & 0 & 1 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} \frac{1}{400} & 0 \\ \frac{1}{2400} & 0 \\ 0 & \frac{1}{800} \\ \frac{1}{20} & 0 \\ 0 & \frac{1}{20} \end{bmatrix} \tag{5.26}$$

with $C = \begin{bmatrix} I & 0 \end{bmatrix} \in \mathbb{R}^{3\times 5}$. We use process noise covariance $\Sigma_W = 10^{-8} \times I$.

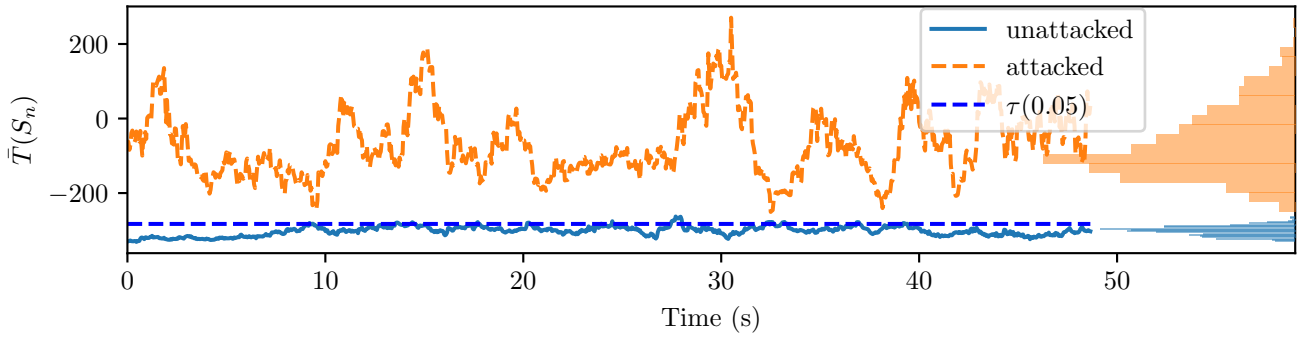
All tests use dynamic watermarking with variance $\Sigma_E = \frac{1}{2}I$, and K and L were chosen to stabilize the system without an attack. We conduct four simulations: attacked and non-attacked systems where the measurement noise covariance is fixed, and attacked and non-attacked systems where the measurement noise covariance is allowed to vary. We ran all four simulations for 1000 iterations, or 50 seconds. In all cases, we compare the test metrics using the hypothesis test described in (5.12), where the measurement noise covariance is assumed to be $10^{-5} \times I$. When simulating the attacked system, we choose an attacker with $\alpha = -1$, $\eta_0 = 0$, $\Sigma_O = 10^{-8} \times I$, and $\Sigma_S = 10^{-8} \times I$.

Fixed Covariance

We first show our test outperforms in the case where the true measurement noise covariance matrix is fixed but unknown to the tester. In our simulations, the true noise covariance is $\Sigma_Z = 10^{-5} \times \text{diag}\{0.18, 30, 0.18\}$. In all tests, Ω^Z is described by the $p = 4$ extreme points: $\Sigma_{Z,1} = 10^{-6} \times \text{diag}\{300, 1.8, 1.8\}$, $\Sigma_{Z,2} = 10^{-6} \times \text{diag}\{1.8, 300, 1.8\}$, $\Sigma_{Z,3} = 10^{-6} \times$



(a) Test statistic (DW)



(b) Test statistic (CRDW*)

Figure 5.3: The evolution and histogram of test statistics (DW) and (CRDW*) on the attacked and unattacked systems where Σ_Z varies as described, again unknown to the tester. Note the robust statistic (CRDW*) takes distinctly higher for the attacked values over almost the entire 1000 iterations than in the unattacked system, while the nonrobust statistic (DW) is again unable to clearly distinguish the two.

$\text{diag}\{9, 9, 12\}$, $\Sigma_{Z,4} = 10^{-6} \times \text{diag}\{9, 9, 9\}$. Both the true measurement noise covariance and that incorrectly assumed in test statistic (DW) are in the resulting set. The simulation is run for 1000 steps.

Fig. 5.1 shows the efficacy of our method under this new uncertainty. If test detection is consistent, the negative log likelihood values should be lower under regular conditions, and higher when the model is attacked. In particular, the nonrobust test statistic (DW) is shown in Fig. 5.1a to be wholly unable to distinguish an attacked system from an unattacked system when its assumption on the measurement noise covariation is violated, while Fig. 5.1b shows the robust test statistic (CRDW) to be able to do so. In fact, the distributions of the respective situations almost never overlap when using our proposed test.

Fig. 5.2 shows a histogram representing that same setup with a fixed but unknown Σ_Z, only this time run over 100 different instantiations of Σ_Z. For each instantiation, the same procedure is used as is depicted in Fig. 5.1, and the average difference between the statistics (DW) and (CRDW) are recorded. The histogram in Fig. 5.2 then depicts the distribution of these values over all 100 instantiations. Here, more positive numbers mean that our statistic (CRDW) outperforms the old statistic (DW). Furthermore, the characterization of Ω remains the same, and the true Σ_Z is chosen by taking a convex combination of the extreme points of Ω^Z, where the coefficients are uniform random variables that have been normalized to have a sum of 1. We can see that our method regularly outperforms, even for randomly chosen Σ_Z.

Varying Covariance

Unattacked and attacked simulations were also conducted with a measurement noise distribution that was allowed to vary. We set $\xi = 0.00002$, implying $\varepsilon = 7.205 \times 10^{-6}$. The true measurement noise is initialized at $\Sigma_{Z_0} = 10^{-5} \times \text{diag}\{0.9, 0.9, 1.2\}$. This shifts linearly over the course of 250 iterations to a new value of $\Sigma_{Z_{250}} = 10^{-5} \times \text{diag}\{15, 15, 0.18\}$, at which point it changes direction to shift linearly over 250 iterations to a value of $\Sigma_{Z_{500}} = 10^{-5} \times \text{diag}\{30, 0.18, 0.18\}$. The measurement noise covariance stays at this value for 150 iterations. It then shifts linearly over 200 iterations to a terminal value of $\Sigma_{Z_{850}} = 10^{-5} \times \text{diag}\{0.18, 30, 0.18\}$, which it takes for another 150 iterations before the simulation is terminated. The results for both the nonrobust and robust tests are shown in Fig. 5.3. As in the fixed covariance case, our test is able to distinguish between the attacked and unattacked systems better and more consistently than the nonrobust test that requires unsatisfied assumptions.

5.5 Conclusion

We developed covariance-robust dynamic watermarking tests for detecting sensor attacks on LTI systems in the presence of uncertainty about the measurement noise covariance. We considered cases where the covariance of measurement noise is unknown and either fixed or slowly-varying, and we required our test to be "fair" with respect to all possible values of the covariance in that it not be more or less powerful for some covariances over others. These reflect real-world needs that will increase as 5G is deployed, because there will be an increase in the deployment of smart CPS systems. In such systems, an "unfair" test can translate to disparate impact across different users in different environments, which is a problem of algorithmic bias.

Future research in this vein includes investigations of how standard watermarking techniques can be adapted to further uncertainties in system dynamics. This work represents a first step in the design of robust and fair watermarking techniques and, in a larger scope,

fair hypothesis tests. Further research is also necessary in the application of different modes of fairness to hypothesis testing.

Chapter 6

Average Margin Regularization for Classifiers

6.1 Introduction

The previous chapter made the link between fairness and robustness in the context of hypothesis testing. In this chapter, similar ideas of data-dependent regularizers (which are one interpretation of the constraints employed in the FO hierarchy) are applied to the problem of robust classification. In a way, the problem of robustness can be seen as an extension of the problem of fairness: While fairness requires that any decision functions learned from the data have minimal dependence (or conditional dependence) on some small subset of features, robustness requires that decision functions have minimal dependence *on all but* a small subset of informative features. In effect, if we could enforce independence of our decision function with respect to all non-informative variables, we could effectively ensure a decision function that does not overfit to noise and is thus more robust. Fortunately, this chapter will show that this can also be achieved with data-dependent regularization terms like those that appear in the FO hierarchy, yet in a much simpler way. The material in this chapter borrows from material by the author in the following paper [190].

There is renewed interest in robust learning due to the observed fragility of deep classifiers to imperceptible adversarial corruptions [32, 67, 85, 240]. Such classifiers are often key elements in a variety of cyber-physical systems (CPS) like self-driving cars, smart homes, or smart grids. Adversarial perturbations on sensors within CPS can have disastrous consequences, and this has been exhibited by several major attacks on mission-critical CPS [3, 49, 89, 144]. In response, numerous adversarial training approaches have been proposed, both in the context of linear margin classifiers [26, 143, 245] and deep classifiers [85, 102]. These approaches train classifiers so as to minimize loss with respect to adversarially-perturbed data.

Interestingly, such adversarial methods have been shown to be equivalent to a particular type of regularization in the context of linear margin classifiers and regression [26, 160, 262,

263]. While regularization protects against (though does not always eliminate) overfitting [64], such outcomes critically depend upon having regularization that is congruous to the underlying data distributions [5, 28, 85, 143, 228, 259, 263].

Here, we argue that adversarial training ignores notable attributes of the data. For instance, image data often has manifold structure [98, 202, 205]. Similar claims may be made about the dynamics of learned systems in system identification or reinforcment learning contexts [211, 232]. Yet adversarial training regularizes with respect to full-dimensional perturbations and not with respect to any underlying manifold structure. This is significant because the imperceptibility of the most successful adversarial perturbations suggests that they lie orthogonal to these manifolds [240]. Thus, any robust methodology that does not exploit this kind of structure will likely remain susceptible to adversarial attacks.

Recently, [247] have claimed that there is a strict trade-off between the accuracy of a classifier and its robustness to adversarial perturbations. They augment their argument with demonstrations of this inverse relationship on a specific dataset, and claim that adversarial training best minimizes the cost of robustness. This chapter shows that this is trade-off is not general and that, in fact, robustness and accuracy can grow concurrently for broad classes of datasets.

We make three contributions here: First, we develop a novel generalization bound that shows classifier accuracy depends on a tradeoff between minimum and average margin. Second, we propose a new regularization that we call average margin (AM) regularization. This regularization consists of a linear term added to the objective, and is hence amenable to efficient numerical computation. We prove that for certain distributions, AM regularization can improve both accuracy and adversarial robustness of a classifier. Third, we use synthetic and real data to empirically show that AM regularization can generate support vector machine (SVM) classifiers that strictly dominate (in terms of accuracy and robustness) classifiers computed with or without adversarial training. Taken together, these results suggest that the phenomenon of adversarial fragility is an issue of overfitting rather than a fundamental issue unique to adversarial attacks.

Robust Linear SVM

Linear SVM relies upon on maximizing training margin by minimizing the *hinge-loss*, and several methods have been proposed to improve its robustness. Some approaches truncate the hinge-loss function [22, 60, 138, 159, 239, 267], but a major issue with this approach is that it forfeits the convexity of the original problem. Other methods [170, 237] penalize outliers uniformly while maintaining the convexity of the hinge-loss, and [5] generates sparse projections to minimize the visibility of outliers. Instead of attempting to devalue outliers, [264] formulate a mixed-integer problem with the hinge-loss that removes them, and [259] redesigns the loss function entirely using bounds on the leave-one-out cross-validation error. Adversarial training has also been considered: [26, 143, 245] take a minimax approach, solving bilevel programs to design classifiers robust to worst-case perturbations of either a given

distribution or a given magnitude. Robustness of classifiers to noise in the label, as opposed to only the features, has also been considered [26, 31, 187].

Robust Deep Classifiers

Adversarial fragility is pronounced in deep classifiers. [240] identified the sensitivity of deep learners to adversarial noise, and [85, 102] followed up with empirical and theoretical examinations of this instability. It has been shown that relatively minute and visually unrecognizable perturbations (in the case of images) can significantly impact the accuracy of these learners, and some work has been done on characterizing these minimal deviations [32, 67]. Notably, [50] showed that many existing methods for adversarially-robust deep classification are not wholly effective. However, there has been a spate of promising recent results in this direction, some of which come with theoretical guarantees [126, 164, 206].

Outline

Section 6.2 describes our notation and presents the problem setup that will be considered. Next, Section 6.3 introduces a novel generalization bound, proposes the average margin (AM) regularization, and then theoretically studies properties of this regularization for a specific distribution. Section 6.3 presents empirical results comparing linear classifiers that have been computed using different regularization and adversarial training approaches, using multiple datasets.

6.2 Preliminaries

We use $\mathcal{N}(\mu, \Sigma)$ to refer to a multivariate normal distribution with mean μ and covariance matrix Σ. Also, let $\mathbf{0}_d$ and $\mathbf{e}_d$ refer to vectors of length d with all entries zero and one, respectively, and I_d to the $d \times d$ identity matrix. These subscripts will be dropped when the size is obvious due to context. In contrast, the function notation $\mathbf{1}(\cdot)$ refers to the indicator function. We use U_i to denote the i-th row of a matrix U. For some kernel function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, let the matrix $K(U, U')$ be such that $K(U, U')_{ij} = k(U_i, U'_j)$.

Consider data observations $(X, Y) \sim \mathcal{D}$ where $X \in \mathbb{R}^d$ and $Y \in \{+1, -1\}$. In binary classification, the goal is compute a classifier $h : \mathbb{R}^d \to \{+1, -1\}$ to predict a label Y from from a feature vector X. Actual observations of these random variables are denoted with the lowercase, (x_i, y_i) for $i = 1, \ldots, n$ (this is used in optimization formulations where the data are assumed fixed observations of random variables X and Y). For a margin classifier from a family $\mathcal{H}$, this is achieved by minimizing $\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i h(x_i))$, which is the sample average of some given loss function $\ell(\cdot)$. The expected classification error rate of a classifier h is defined as $\mathcal{L}(h) = \mathbb{E}(\mathbf{1}(Y \neq \text{sign}(h(X))))$.

Adversarial robustness for a classifier h refers to its ability to maintain accuracy in predicting a label y when given a corrupted corresponding feature vector $x+\delta$, where corruption

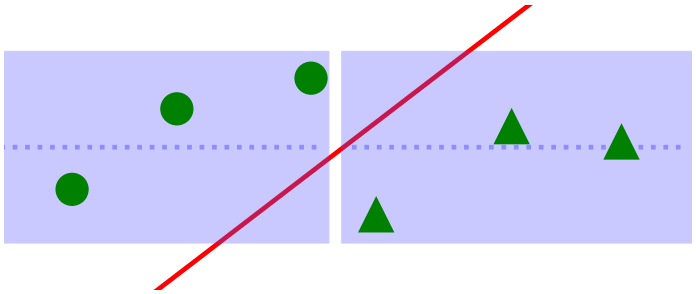


Figure 6.1: This example shows that maximizing the margin on training data can reduce classification accuracy since it does not use data far from the margin boundary. The marks are sampled data, the two supports of the data distributions for the two labels $Y \in \{-1, +1\}$ are the two shaded rectangles, and the dashed line is the maximum margin linear classifier.

δ is chosen by an adversary and has magnitude bounded by a quantity e. The expected adversarial classification error rate of a classifier h when the adversary can perturb data by e magnitude is $\mathcal{L}(h, e) = \mathbb{E}(\max_{\delta:\|\delta\|\leq e} \mathbf{1}(y \neq h(x+\delta)))$. The inner maximization is interpreted as an adversary choosing an attack. Also note that $\mathcal{L}(h, 0) = \mathcal{L}(h)$.

6.3 Regularization Method

Margin classifiers primarily use only data near the boundaries of different classes for the purpose of estimating the parameters of the classifier. However, in the low (relative to dimensionality) data regime this can be problematic. Figure 6.1 shows an example where the usual margin classifier has issues. Maximizing the minimum margin leads to a classifier with high expected classification error because only a small amount of the data lies at the boundaries of the two classes. The manifold-like structure of the two classes leads to a situation where much of the data lies away from the boundary. A natural question to ask is how margin classifiers may be modified in order to better use data away from the boundary to improve predictions in situations similar to the above shown example.

Given the manifold-like example above, one possiblity is to use manifold regularization techniques. In fact, manifold regularization can be useful for regression [15, 23, 29, 54]. However, a disadvantage of manifold regularization is it requires the indirect step of first estimating the manifold, and then using the estimated manifold for regularization. This indirect step can increase estimation error in a way that often reverses its regularizing effect. Our goal then is to design a regularizer that provides benefits in the manifold-like setting but does not require estimation of any manifolds.

In this section, we develop and study a new regularization for margin classifiers. We begin by proving a new generalization bound that demonstrates how maximizing the minimum margin does not always lead to minimal expected classification error. This generalization bound is used to motivate our new regularization for any margin classifier, which we call the

average margin (AM) regularization. Next, we provide a probabilistic interpretation of this regularization in the context of deep learning. We conclude the section by discussing AM regularization in the special context of SVM. It is shown how this regularization can be used for kernel SVM, and then a result is given showing how AM regularization can simultaneously improve expected classification error and robustness to adversarial perturbations; this is significant because it is in direct contrast to results on adversarial training [247] that find a strict trade-off between classifier accuracy and robustness to adversarial perturbations.

Average Margin Generalization Bound

Classical results on the generalization error of classifiers [22] provide justification for maximizing the minimum margin of classifiers. Below, we present a new generalization bound in terms of the average margin of a classifier.

Theorem 11. *Let $\mathcal{L}(h) = \mathbb{E}(\mathbf{1}(Y \neq \text{sign}(h(X))))$ be the expected classification error rate of h, $K_\gamma(h) = \#\{i : Y_i h(X_i) \leq \gamma\}/n$ be the fraction of data with γ-margin mistakes, $J(h) = E_n(Yh(X))$ be the average classification margin, and suppose that $\sup_x |h(x)| \leq c$ for all $h \in \mathcal{H}$. Then for any $\zeta \in [0,1]$ we have with probability at least $1 - 2\delta$ that*

$$\mathcal{L}(h) \leq \zeta \cdot (1 - J(h)/c) + (1-\zeta) \cdot K_\gamma(h) + 4\frac{\mathcal{R}_n(\mathcal{H})}{\gamma} + \sqrt{\frac{\log(\log_2 \frac{4c}{\gamma})}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}. \quad (6.1)$$

for all $\gamma \in (0, c]$ and all $h \in \mathcal{H}$.

Proof. This proof uses a similar argument to the proof of Theorem 2 by [123], with suitable modifications made to apply to our setting. Define the functions

$$l_\gamma(u) = \begin{cases} 1, & u \leq 0 \\ 1 - u/\gamma, & 0 < u < \gamma \\ 0, & u \geq \gamma \end{cases} \quad (6.2)$$

and $\ell_\gamma(u) = \zeta \cdot (1 - u/c) + (1-\zeta) \cdot l_\gamma(u)$. Let $\mathcal{L}_\gamma(h) = \mathbb{E}(\ell_\gamma(Yh(X)))$ and $\hat{\mathcal{L}}_\gamma(h) = E_n(\ell_\gamma(Yh(X)))$. We will consider the values $\gamma_k = c/2^k$ and $\delta_k = \delta/(k+1)^2$ for $k \in \{0, 1, \ldots\}$. Since $\ell_{\gamma_k}(u)$ is Lipschitz with constant $\zeta/c + (1-\zeta)/\gamma_k \leq 1/\gamma_k$, applying Theorem 7 from [22] gives that

$$\mathcal{L}_{\gamma_k}(h) \leq \hat{\mathcal{L}}_{\gamma_k}(h) + \frac{2}{\gamma_k}\mathcal{R}_n(\mathcal{H}) + \sqrt{\frac{\log(1/\delta_k)}{2n}} \quad (6.3)$$

holds with probability at least $1 - \delta_k$ for all $h \in \mathcal{H}$. Next observe that for any $\zeta \in [0,1]$, $\gamma > 0$, and any $h \in \mathcal{H}$; we have $\mathcal{L}(h) \leq \mathcal{L}_\gamma(h)$ and $\hat{\mathcal{L}}_\gamma \leq \zeta \cdot (1 - J(h)/c) + (1-\zeta) \cdot K_\gamma(h)$. Thus with probability at least $1 - \delta_k$ we have

$$\mathcal{L}(h) \leq \zeta \cdot (1 - J(h)/c) + (1-\zeta) \cdot K_{\gamma_k}(h) + \frac{2}{\gamma_k}\mathcal{R}_n(\mathcal{H}) + \sqrt{\frac{\log(1/\delta_k)}{2n}} \quad (6.4)$$

for all $h \in \mathcal{H}$. Applying the union bound over all $k \in \{0, 1, \ldots\}$ gives that (6.4) holds with probability at least $1 - \pi^2\delta/6 \geq 1 - 2\delta$ for all $k \in \{0, 1, \ldots\}$ and $h \in \mathcal{H}$. Now we will assume that this union event occurs. This implies (6.1) holds for $\gamma = c$. Next consider $\gamma \in (0, c)$, and choose the k such that $\gamma_k \leq \gamma < \gamma_{k-1}$. Note $k \leq \log_2(c/\gamma) + 1$, and so

$$\mathcal{L}(h) \leq \zeta \cdot (1 - J(h)/c) + (1 - \zeta) \cdot K_\gamma(h) + \frac{4}{\gamma}\mathcal{R}_n(\mathcal{H}) + \sqrt{\frac{\log(1/\delta) + 2\log(\log_2(4c/\gamma))}{2n}} \quad (6.5)$$

since $K_{\gamma_k}(h) \leq K_\gamma(h)$, $1/\gamma_k \leq 2/\gamma$, and $\log(1/\delta_k) \leq \log(1/\delta) + 2\log(\log_2(4c/\gamma))$. □

Average Margin Regularization

Given the above intuition, we propose that the average margin $\frac{1}{n}\sum_{i=1}^n y_i h(x_i)$ can be used as a regularization term for any margin classifier. Specifically, an AM-regularized classifier can be computed by solving

$$\min_{h \in \mathcal{H}} \tfrac{1}{n}\textstyle\sum_{i=1}^n \ell(y_i h(x_i)) - \mu \cdot \tfrac{1}{n}\sum_{i=1}^n y_i h(x), \quad (6.6)$$

where $\mu \geq 0$ is a tuning parameter. Note that we subtract the average margin term because we are minimizing.

Next, we consider deep learning with activation function $\exp(h(x))/(1 + \exp(h(x))$. The logistic loss is often used to construct classifiers, and the corresponding AM-regularized network is computed by solving

$$\min_{h \in \mathcal{H}} \tfrac{1}{n}\textstyle\sum_{i=1}^n \log(1 + \exp(-y_i h(x_i))) - \mu \cdot \tfrac{1}{n}\sum_{i=1}^n y_i h(x_i), \quad (6.7)$$

where $\mu \geq 0$ is again a tuning parameter. Now suppose we use the labels $t = (1 + y)/2 \in \{0, 1\}$, and we make the identifications that $\hat{p}_0(x) = 1/(1 + \exp(h(x)))$ and $\hat{p}_1(x) = \exp(h(x))/(1+\exp(h(x)))$. Then training the AM-regularized network is equivalent to solving

$$\min_{h \in \mathcal{H}} -\tfrac{1}{n}\textstyle\sum_{i=1}^n \left[t_i \log(\hat{p}_1(x)) + (1 - t_i)\log(\hat{p}_0(x))\right] - \mu \cdot \tfrac{1}{n}\sum_{i=1}^n t_i \log\frac{\hat{p}_1(x)}{\hat{p}_0(x)} - \mu \cdot \tfrac{1}{n}\textstyle\sum_{i=1}^n (1 - t_i)\log\frac{\hat{p}_0(x)}{\hat{p}_1(x)}. \quad (6.8)$$

The first term is cross-entropy, while the last two terms are negative log-likelihood ratios. Thus, for deep learning our AM regularization encourages larger log-likelihood ratios. Recalling hypothesis testing, a larger log-likelihood ratio makes it easier to distinguish between classes. Note that AM-regularization generalizes in a natural way to multi-class classification with deep neural networks.

Special Case of SVM

Here, we consider AM regularization in the special case of SVM. First, consider linear SVM with $h(x) = x^\mathsf{T}\beta + b$. Then the AM-regularized linear SVM is given by

$$\begin{aligned} &\min\ \lambda\|\beta\|^2 + \tfrac{1}{n}\textstyle\sum_{i=1}^n s_i - \mu \cdot \tfrac{1}{n}\sum_{i=1}^n y_i(x_i^\mathsf{T}\beta + b) \\ &\text{s.t. } y_i(x_i^\mathsf{T}\beta + b) \geq 1 - s_i, \quad \text{for } i = 1, \ldots, n \\ &\quad s_i \geq 0, \qquad\qquad\quad \text{for } i = 1, \ldots, n \end{aligned} \tag{6.9}$$

As seen above, one advantage is that AM regularization is simply a linear term in the objective function. Thus, AM regularization does not significantly affect the computational complexity of solving the linear SVM optimization problem. Another benefit of AM regularization is it can be dualized, which allows for its use in kernel SVM. A standard argument using the KKT conditions shows that the kernel SVM with AM regularization is computed by solving

$$\begin{aligned} &\min\ z^\mathsf{T}\big(yy^\mathsf{T} \circ K(X,X)\big)z - \mathbf{e}_n z^\mathsf{T} \\ &\text{s.t. } yz^\mathsf{T} = 0 \\ &\quad \tfrac{\mu}{1+\mu} \cdot \tfrac{1}{\lambda n} \leq z \leq \tfrac{1}{\lambda n} \end{aligned} \tag{6.10}$$

This kernel SVM formulation provides further intuition about AM regularization: It shows that increasing μ increases the impact of points away from the margin, thereby mixing the original support vectors with an average over all data points.

Last, we show that AM regularization can generate classifiers that both improve classification accuracy and robustness. Let $\mathcal{L}(h, e) = \mathbb{E}(\max_{\delta:\|\delta\|\leq e} \mathbf{1}(Y \neq h(Y + \delta)))$ be the generalization error of classifier h when the adversary can perturb data by e magnitude. We specifically prove a result that formalizes the intuition shown in Figure 6.1.

Proposition 11. *Let $\hat{h}_{AM}$ and $\hat{h}_{L2}$ be the linear classifiers computed by SVM with and without AM regularization, respectively, using $n = 4k \geq 4$ points sampled from a data distribution. Recalling (6.9), we assume λ is chosen so that all sampled points lie beyond the margin, that μ can be chosen based on the sampled data, and that without loss of generality the linear classifier has no intercept term (i.e. $h(x) = x^\mathsf{T}\beta$). There exists a data distribution such that*

$$\begin{aligned} &\mathcal{L}(\hat{h}_{AM}, e) < \mathcal{L}(\hat{h}_{L2}, e), \text{ for } e \in [0, 7\sqrt{5}/25) \\ &\mathcal{L}(\hat{h}_{AM}, e) = \mathcal{L}(\hat{h}_{L2}, e), \text{ for } e \in [7\sqrt{5}/25, 15\sqrt{5}/25) \\ &\mathcal{L}(\hat{h}_{AM}, e) > \mathcal{L}(\hat{h}_{L2}, e), \text{ for } e \in [15\sqrt{5}/25, \sqrt{5}) \\ &\mathcal{L}(\hat{h}_{AM}, e) = \mathcal{L}(\hat{h}_{L2}, e), \text{ for } e \in [\sqrt{5}, 2\sqrt{5}) \\ &\mathcal{L}(\hat{h}_{AM}, e) < \mathcal{L}(\hat{h}_{L2}, e), \text{ for } e \in [2\sqrt{5}, 110\sqrt{5}/25) \\ &\mathcal{L}(\hat{h}_{AM}, e) = \mathcal{L}(\hat{h}_{L2}, e), \text{ for } e \in [110\sqrt{5}/25, \infty) \end{aligned}$$

Proof. We consider a balanced data distribution with $y_{2i-1} = +1$ and $y_{2i} = -1$ for $i = 1, \dots, 2k$. Suppose $x_{2i-1} = (10, 0)$ and $x_{2i} = (-10, 0)$ for $i = k+1, \dots, 2k$. For $i = 1, \dots, k$: let $a_i = +1$ be the event that $x_{2i-1} = (1, 2)$ and $x_{2i} = (-1, -2)$, and let $a_i = -1$ be the event that $x_{2i-1} = (1, -2)$ and $x_{2i} = (-1, 2)$. We assume the a_i are independent Rademacher random variables.

Let $I = \{1, \dots, k\}$, and note the margin assumption on λ implies the classifier satisfies $Y_i X_i^\mathsf{T} \beta \geq 1$ for all $i = 1, \dots, n$. Define the conditional expectation $\mathcal{L}(h, e, \mathcal{A}) = \mathbb{E}[\max_{\delta:\|\delta\| \leq e} \mathbf{1}(Y \neq h(X + \delta))|\mathcal{A}]$ for some event $\mathcal{A}$. Clearly, choosing $\mu = 0$ removes the effect of AM regularization and ensures that $\mathcal{L}(\hat{h}_{AM}, e, \mathcal{A}) \leq \mathcal{L}(\hat{h}_{L2}, e, \mathcal{A})$.

Now consider the event $\mathcal{B}$ where $a_i = +1$ for all $i \in I$. Here, the margin assumption means the classifier must satisfy $10\beta_1 \geq 1$, $\beta_1 + 2\beta_2 \geq 1$. A straightforward calculation gives that $\hat{h}_{L2}$ has $\hat{\beta}_{L2} = (\frac{1}{5}, \frac{2}{5})$ and $\mathcal{L}(\hat{h}_{L2}, e, \mathcal{B}) = \frac{1}{4} + \frac{1}{4}\mathbf{1}(e \geq \sqrt{5}) + \frac{1}{2}\mathbf{1}(e \geq 2\sqrt{5})$. The AM-regularized SVM is

$$\begin{aligned} \min\ & \lambda\|\beta\|^2 - 2\mu k \cdot (\beta_1 + 2\beta_2) - 2\mu k \cdot (10\beta_1) \\ \text{s.t.}\ & 10\beta_1 \geq 1 \\ & \beta_1 + 2\beta_1 \geq 1 \end{aligned} \tag{6.11}$$

Now suppose we choose the largest μ such that the optimal solution satisfies $10\beta_1 > 1$ and $\beta_1 + 2\beta_1 = 1$. Then it can be easily verified that this largest value is $\mu = \lambda/(15k)$, and so a straightforward calculation gives that $\hat{h}_{AM}$ has $\hat{\beta}_{AM} = (\frac{11}{15}, \frac{2}{15})$ and $\mathcal{L}(\hat{h}_{AM}, e, \mathcal{B}) = \frac{1}{4}\mathbf{1}(e \geq 7\sqrt{5}/25) + \frac{1}{4}\mathbf{1}(e \geq 15\sqrt{5}/25) + \frac{1}{2}\mathbf{1}(e \geq 110\sqrt{5}/25)$. This means that $\mathcal{L}(\hat{h}_{AM}, e, \mathcal{B}) < \mathcal{L}(\hat{h}_{L2}, e, \mathcal{B})$ for $e \in [0, \sqrt{5})$ and $\mathcal{L}(\hat{h}_{AM}, e, \mathcal{B}) = \mathcal{L}(\hat{h}_{L2}, e, \mathcal{B})$ for $e \geq \sqrt{5}$. And as discussed earlier, choosing $\mu = 0$ for the event $\neg\mathcal{B}$ ensures that $\mathcal{L}(\hat{h}_{AM}, e, \neg\mathcal{B}) \leq \mathcal{L}(\hat{h}_{L2}, e, \neg\mathcal{B})$. Next note $\mathcal{B}, \neg\mathcal{B}$ partition the sample space, and that $\mathbb{P}(\mathcal{B}) > 0$. Hence the result follows from the law of total expectation. □

This is a more subtle result than that of [247], which considers $\mathcal{L}(\cdot, e)$ for adversarially trained classifiers at only two discrete values of e. Our analysis shows that AM regularization can both increase classifier accuracy (i.e., at $e = 0$) and robustness to adversarial perturbations for specific data distributions. This is seen because AM regularization generally has lower expected classification error over the whole range of e except for a very narrow range. The AM regularization is less robust when $e \in [15\sqrt{5}/25, \sqrt{5})$ because it has made a careful tradeoff between maximizing the margin and maximizing the average margin.

Another point to note is that for the data distribution in the above proposition, the adversarially trained SVM does not improve upon the standard SVM.

Proposition 12. *Let $\hat{h}_{D\gamma}$ be the linear classifier computed by adversarially trained SVM where the adversary can perturb data by $\gamma \in [0, \sqrt{5})$ and let $\hat{h}_{L2}$ be the linear classifier computed by SVM, using $n = 4k \geq 4$ points sampled from a data distribution. Recalling (6.9), we assume λ is chosen so all sampled (and perturbed) points lie beyond the margin, and that without loss of generality the linear classifier has no intercept term (i.e. $h(x) = x^\mathsf{T}\beta$).*

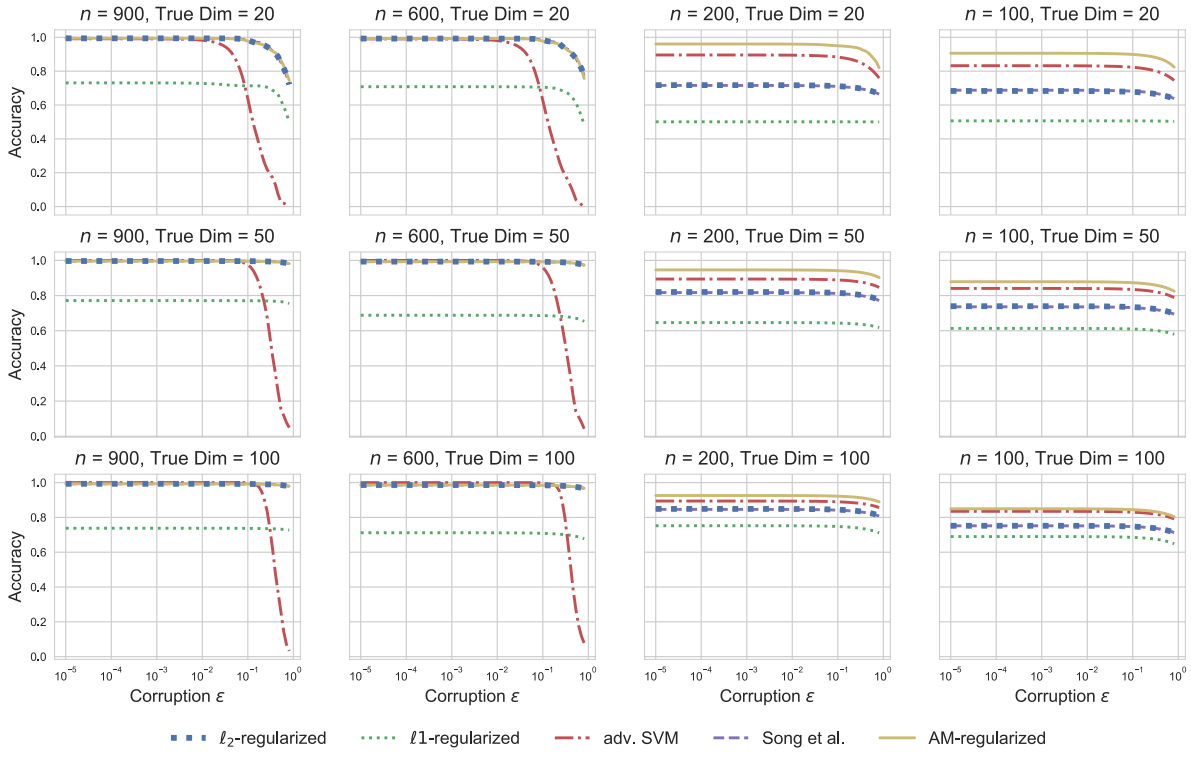


Figure 6.2: Robustness to adversarial corruptions of classifiers trained on synthetic data, with accuracy measured as the fraction correctly classified.

Then, for the distribution that is considered in the proof of Proposition 11, we have that $\mathcal{L}(\hat{h}_{D\gamma}, e) = \mathcal{L}(\hat{h}_{L2}, e)$ *for all* $e \geq 0$.

Proof. The proof of this proposition follows a similar argument to that for Proposition 11. □

6.4 Empirical results

Here, we use synthetic and real data to compare AM regularization to other regularizations, including adversarial training, for linear SVM. We first present the benchmark regularization methods that we compare AM regularization to. Next, we present empirical results for linear SVM using synthetic data, and we conclude this section by presenting numerical results for linear SVM applied to a real dataset.

Benchmarks

We compare linear SVM with AM regularization to: SVM with ℓ_2 regularization, SVM with ℓ_1 regularization, the robust SVM training method of [237], and the adversarial training method outlined by [26, 245, 263].

Song et al. This method relies on minimizing the impact of outliers on the design of the separator. Specifically, it avoids the over-reliance on such outliers by shifting the loss according to the distance of a point to its respective centroid:

$$\min \lambda\|\beta\|^2 + \tfrac{1}{n}\textstyle\sum_{i=1}^n \left(1 - y_i(\beta^\mathsf{T} x_i + b) - \gamma\|x_i - \mu_{y_i}\|_2^2\right)_+, \tag{6.12}$$

where μ_{y_i} is the centroid for class y_i, and $\lambda, \gamma \geq 0$ are tuning parameters.

Adversarial Training Adversarial training formulates SVM training as a bilevel program, where the loss is minimized not over the original data $\{X_i\}_{i=1}^n$, but rather over worst-case perturbations of the data $\{X_i + \delta_i\}_{i=1}^n$, where the perturbations δ_i are restricted to be in some space. If we restrict $\|\delta_i\|_q \leq \gamma$ for some norm $\|\cdot\|_q$, this is modeled as

$$\begin{aligned} \min\ & \lambda\|\beta\|^2 + \tfrac{1}{n}\textstyle\sum_{i=1}^n s_i \\ \text{s.t. }\ & y_i(\beta^\mathsf{T} x_i + b) - \gamma\|\beta\|_{q^*} \geq 1 - s_i, \quad \text{for } i = 1, \ldots, n \\ & s_i \geq 0, \qquad\qquad\qquad\qquad \text{for } i = 1, \ldots, n \end{aligned} \tag{6.13}$$

where $\|\cdot\|_{q^*}$ is the dual norm of $\|\cdot\|_q$. In this section, we consider the case where $q = 2$.

Synthetic Data

We next run randomized experiments to demonstrate situations where AM regularization is beneficial. We generate U_+, U_- by taking the Q from a QR decomposition done on a matrix of size $d \times m$ with standard normal entries, and $\mu_+, \mu_- \in \mathbb{R}^d$ to have uniformly-distributed entries between 0 and 2. Then, we set $X = \Pi_{U_+}\mu_+ + U_+ v_+ + \varepsilon_+$ when $Y = +1$, and $X = \Pi_{U_-}\mu_- + U_- v_- + \varepsilon_-$ when $Y = -1$, where $v_+ \sim \mathcal{N}(\mu_+, I_m)$, $\varepsilon_+ \sim \mathcal{N}(\mathbf{0}_d, \varepsilon I_d)$, $v_- \sim \mathcal{N}(\mu_-, I_m)$, and $\varepsilon_- \sim \mathcal{N}(\mathbf{0}_d, \varepsilon I_d)$. A training set of n samples was created. All models were then tested using 10,000 samples with adversarial corruption of various magnitude. This procedure was repeated 100 times with $\varepsilon = 0.01$ and $d = 200$, and the results for various training-set sizes and values of n are presented in Figure 6.2. Our method improves both robustness and accuracy in low-data and low-dimensionality regimes.

MNIST Experiments

Next, we use the classic MNIST dataset [149] to evaluate AM regularization. In Table 6.1, we compare AM-regularization to the benchmark methods on the fraction of test points correctly

Table 6.1: Fraction Correctly Classified when Classifying 0 vs. 1 in MNIST Dataset

Corruption	0.01	0.2	1.0
ℓ_2-regularized	0.940	0.938	0.921
ℓ_1-regularized	0.997	0.995	0.714
AM-regularized	0.998	0.997	0.983
Adversarial	0.996	0.995	0.979
Song et al.	0.994	0.992	0.974

classified. For simplicity, we focus on the binary classification problem of separating hand-drawn 0's from hand-drawn 1's, and report results from training on 10% of available training data. All hyperparameter values were set using 5-fold cross-validation on the training set. We repeat this process 50 times for each model, and report the accuracy results on a common testing set corrupted with various degrees of adversarial perturbations. We note that our method achieves better or almost equivalent accuracy with low corruption, and is best able to retain that level of accuracy at high levels of corruption.

6.5 Conclusion

Based on a novel generalization bound, we have proposed in this chapter a new form of regularization for margin classifiers that we call average margin (AM) regularization. Our theoretical and empirical results support its use by showing that AM regularization can increase both classifier accuracy and adversarial robustness. Taken together, our results suggest that adversarial fragility is an issue of overfitting, rather than a fundamental uniqueness of adversarial perturbations.

One future topic is to better understand AM regularization's theoretical properties. We believe AM regularization works best in the finite sample regime and that its asymptotic behavior when $\mu \not\to 0$ may be poor. Another topic is modifications of AM regularization. For instance, consider a hinged average margin (HAM) regularization that adds $-\mu \times E_n(\gamma - Yh(X))^+$, where $\mu, \gamma \geq 0$ are tuning parameters. HAM regularization may improve AM regularization by providing saturation, whereby points *very* far (specifically γ distance away) from the margin are not considered in the average margin calculation. Promisingly, this is convex in $Yh(X)$. Furthermore, we would like to investigate the efficacy of AM regularization for deep classifiers.

Chapter 7

Conclusions

This thesis has presented the FO hierarchy, an optimization framework for approximating independence and ensuring fairness. It has expanded this framework to a number of statistical learning settings.

7.1 Summary of Contributions

This thesis has made the following specific contributions to the literature:

- *Introduces and argues for the FO hierarchy as an approach to approximating independence in optimization and ensuring fairness in data-driven decision-making.* The intuition behind this framework is the concept of bounding moments. The framework is shown to be highly flexible, accommodating a range of statistical decision-making problems, including problems that involve making multiple decisions and which require ensuring fairness with respect to multiple protected attributes.

- *Provides results on consistency and non-asymptotic rates of convergence for FO.* Our framework will asymptotically guarantee the choice of a decision function that satisfies independence or fairness constraints as long as the order of the hierarchy grows as a double-log of the number of data points, n, and the bounds on each moment constraint shrink according to $n^{-1/4}$.

- *Examines empirical behavior and theoretical intuitions of FO in multiple supervised learning problems, including dynamical systems and case studies on automated and fair morphine and heparin dosage.* We show that the FO constraints can be interpreted as approximations to a bi-level programming problem, as information projections and as deflations in a dual space, when applied to certain supervised learning algorithms. We then show that its efficacy on a number of datasets and in two automated dosing case studies.

- *Defines a novel notion of fairness for unsupervised learning, and in particular dimensionality reduction problems, and extends the FO hierarchy to this setting.* This relies on an SDP relaxation of the PCA problem, as well as SDP relaxations of the relevant FO constraints. We show that this is effective both on standard metrics of dimensionality reduction, such as the percentage of data variance captured in a dimensionality reduction, and as a preprocessing step for further supervised and unsupervised learning tasks.

- *Provides the first analysis of fair hypothesis testing, and exploits these principles to design a distributionally-robust watermarking scheme for detecting attacks on dynamical systems.* Hypothesis testing is viewed as a fairness problem where the exact specifications of the null-hypothesis are viewed as the protected attribute; certain notions of fairness are then equivalent to showing that a certain test is robust to errors in the null hypothesis, in the sense that it retains its power. We incorporate this concept in the framework of dynamic watermarking.

- *Extends the notion of data-dependent regularization to propose an average-margin regularizer for robust classification in low-data regimes for data that lives in low-dimensional manifolds.* This comes from the intuition that simple, data-dependent regularizers can implicitly detect manifolds on which data may live, and thus be used to cancel out the chances of overfitting to errors outside of this manifold more efficiently than adversarial training techniques.

7.2 Future Work

Plenty of work remains to be done in the space of algorithmic fairness. From the perspective of the methodologies presented in this thesis, one major area of further work is the proper choice of hyperparameters. In particular, mechanisms that could determine the number of fairness constraints that are helpful given preliminary and efficient scans of data could prove very helpful in attaining the most value possible from the FO hierarchy. Furthermore, much work remains to be done in the nascent spaces of fair unsupervised learning and fair hypothesis testing. As shown in this thesis, these are problems that are amenable to fairness, and in fact are problems that can even benefit from the application of fairness concepts. The FPCA approach outlined in this thesis can run in to computational bottlenecks due to the SDP involved: A valuable line of work would involve exploiting the low-rank nature of the problem to make this more efficient. While FPCA works as a preprocessing technique, an alternative approach to fair unsupervised learning would be to directly redesign algorithms for unsupervised learning problems like clustering in order to take fairness into account at training time, much as is done for many fair supervised learning algorithms. There has been some initial work in this realm [40, 56], but much work remains to be done.

More work also remains to be done in task of defining fairness. This is easy to see in areas where fair statistical learning has only recently become of interest: unsupervised learning

and hypothesis testing. In these domains, it remains for the academic community to engage in the primary work of proposing definitions, considering the attributes of each proposal and coalescing around the few best notions. Yet even the notions of fairness used in supervised learning require further study. This is due to several reasons. First, many of these have been shown to be actively conflicting, and so further study is required to determine exactly when these conflicts arise and whether there are better or broader definitions that avoid these problems [131, 134]. Second, they can be extended to include more than just final decisions, but rather the uncertain societal impacts of these decisions [158]. Finally, these notions should be specified to become more reflective of issues in the real world [61]. The whole point of the field of fair machine learning is to redefine classical views of statistical learning to be more reflective of a complicated and unintuitive world: If the results of the fair machine learning literature are to ever be adopted widely to actually address the their motivating causes, they must reflect the needs, desires and restrictions of their users. To that end, they need to embrace the world in which they are to be deployed and the legal and philosophical foundations of fairness upon which citizens of that world will base their decision: Are automated decision-making algorithms worthy of our trust?

Bibliography

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

[2] M. Abbasi, S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. Fairness in representation: quantifying stereotyping as a representational harm. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 801–809. SIAM, 2019.

[3] M. Abrams and J. Weiss. Malicious control system cyber security attack case study–Maroochy water services, australia. *MITRE*, 2008.

[4] C. C. Aggarwal, A. Hinneburg, and D. A. Keim. On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer, 2001.

[5] C. C. Aggarwal and P. S. Yu. Outlier detection for high dimensional data. In *ACM Sigmod Record*, volume 30, pages 37–46. ACM, 2001.

[6] I. Ahmad, Z. Kaleem, R. Narmeen, L. D. Nguyen, and D.-B. Ha. Quality-of-service aware game theory-based uplink power control for 5g heterogeneous networks. *Mobile Networks and Applications*, 24(2):556–563, 2019.

[7] A. A. Ahmadi. Sum of squares (sos) techniques: An introduction. 2016.

[8] A. A. Ahmadi and G. Hall. Dc decomposition of nonconvex polynomials with algebraic techniques. *Mathematical Programming*, pages 1–26, 2017.

[9] E. S. Anderson. What is the point of equality? *Ethics*, 109(2):287–337, 1999.

[10] J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

[11] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica, May*, 23, 2016.

[12] R. J. Arneson. Equality and equal opportunity for welfare. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 56(1):77–93, 1989.

[13] R. Arora, A. Cotter, and N. Srebro. Stochastic optimization of pca with capped msg. In *Advances in Neural Information Processing Systems*, pages 1815–1823, 2013.

[14] A. Aswani. Statistics with set-valued functions: applications to inverse approximate optimization. *Mathematical Programming*, 174(1-2):225–251, 2019.

[15] A. Aswani, P. Bickel, and C. Tomlin. Regression on manifolds: Estimation of the exterior derivative. *The Annals of Statistics*, 39(1):48–81, 2011.

[16] A. Aswani and M. Olfat. Optimization hierarchy for fair statistical decision problems. *arXiv preprint arXiv:1910.08520*, 2019.

[17] I. Ayres. Outcome tests of racial disparities in police practices. *Justice research and Policy*, 4(1-2):131–142, 2002.

[18] S. Banach. Über homogene polynome in (l^{2}). *Studia Mathematica*, 7(1):36–44, 1938.

[19] B. Barak, J. A. Kelner, and D. Steurer. Rounding sum-of-squares relaxations. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 31–40. ACM, 2014.

[20] S. Barocas and A. D. Selbst. Big data's disparate impact. *California Law Review*, 104, 2016.

[21] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Large margin classifiers: convex loss, low noise, and convergence rates. In *Advances in Neural Information Processing Systems*, pages 1173–1180, 2004.

[22] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[23] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 7(Nov):2399–2434, 2006.

[24] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

[25] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: the state of the art. *arXiv preprint arXiv:1703.09207*, 2017.

[26] D. Bertsimas, J. Dunn, C. Pawlowski, and Y. D. Zhuo. Robust classification. *J. Mach. Learn. Res*, 2017.

[27] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

[28] J. Bi and T. Zhang. Support vector classification with input data uncertainty. In *NeurIPS*, pages 161–168, 2005.

[29] P. J. Bickel and B. Li. Local polynomial regression on unknown manifolds. In *Complex Datasets and Inverse Problems*, pages 177–186. Institute of Mathematical Statistics, 2007.

[30] D. Biddle. *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Gower Publishing, Ltd., 2006.

[31] B. Biggio, B. Nelson, and P. Laskov. Support vector machines under adversarial label noise. In *Asian Conference on Machine Learning*, pages 97–112, 2011.

[32] B. Biggio and F. Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

[33] J. Billock. Pain bias: The health inequality rarely discussed. *BBC*, May 2018.

[34] R. Binns. Fairness in machine learning: Lessons from political philosophy. *arXiv preprint arXiv:1712.03586*, 2017.

[35] T. M. Bisgaard and Z. Sasvári. When does e (xk· yl)= e (xk)· e (yl) imply independence? *Statistics & probability letters*, 76(11):1111–1116, 2006.

[36] A. Bloom and A. Kirsch. *The Republic of Plato*. Hachette UK, 2016.

[37] J. Bochnak and J. Siciak. Polynomials and multilinear mappings in topological vector-spaces. *Studia Mathematica*, 39(1):59–76, 1971.

[38] R. Bock, A. Chilingarian, M. Gaug, F. Hakl, T. Hengstebeck, M. Jiřina, J. Klaschka, E. Kotrč, P. Savickỳ, S. Towers, et al. Methods for multidimensional event classification: a case study using images from a cherenkov gamma-ray telescope. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 516(2):511–528, 2004.

[39] R. J. Bolton and D. J. Hand. Statistical fraud detection: A review. *Statistical science*, pages 235–249, 2002.

[40] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357, 2016.

[41] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

[42] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. *Linear matrix inequalities in system and control theory*, volume 15. SIAM, 1994.

[43] L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391):580–598, 1985.

[44] K. Burns, L. A. Hendricks, T. Darrell, and A. Rohrbach. Women also snowboard: Overcoming bias in captioning models. *arXiv preprint arXiv:1803.09797*, 2018.

[45] T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *Data mining workshops, 2009. ICDMW'09. IEEE international conference on*, pages 13–18. IEEE, 2009.

[46] T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang. Controlling attribute effect in linear regression. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 71–80. IEEE, 2013.

[47] T. Calders and S. Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

[48] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

[49] A. A. Cárdenas, S. Amin, and S. Sastry. Research challenges for the security of control systems. In *HotSec*, 2008.

[50] N. Carlini and D. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *ACM Workshop on Artificial Intelligence and Security*, pages 3–14. ACM, 2017.

[51] W. F. Cascio and H. Aguinis. The federal uniform guidelines on employee selection procedures (1978) an update on selected issues. *Review of Public Personnel Administration*, 21(3):200–218, 2001.

[52] Centers for Desease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National health and nutrition examination survey data, 2018.

[53] A. Chen and P. J. Bickel. Consistent independent component analysis and prewhitening. *IEEE Transactions on Signal Processing*, 53(10):3625–3632, 2005.

[54] M.-Y. Cheng and H.-T. Wu. Local linear regression on manifolds and its geometric interpretation. *Journal of the American Statistical Association*, 108(504):1421–1434, 2013.

[55] C.-F. Chien and L.-F. Chen. Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems with applications*, 34(1):280–290, 2008.

[56] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii. Fair clustering through fairlets. In *Advances in Neural Information Processing Systems*, pages 5036–5044, 2017.

[57] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv preprint arXiv:1703.00056*, 2017.

[58] R. J. Cipolle, R. D. Seifert, B. A. Neilan, D. E. Zaske, and E. Haus. Heparin kinetics: variables related to disposition and dosage. *Clinical Pharmacology & Therapeutics*, 29(3):387–393, 1981.

[59] G. A. Cohen. On the currency of egalitarian justice. *Ethics*, 99(4):906–944, 1989.

[60] R. Collobert, F. Sinz, J. Weston, and L. Bottou. Trading convexity for scalability. In *ICML*, pages 201–208, 2006.

[61] S. Corbett-Davies and S. Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[62] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806, 2017.

[63] F. M. Cornford. *Plato and Parmenides*. Routledge, 2014.

[64] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[65] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009.

[66] A. Cotter, M. Friedlander, G. Goh, and M. Gupta. Satisfying real-world goals with dataset constraints. *arXiv preprint arXiv:1606.07558*, 2016.

[67] N. Dalvi, P. Domingos, S. Sanghai, and D. Verma. Adversarial classification. In *KDD*, pages 99–108. ACM, 2004.

[68] A. d'Aspremont, L. El Ghaoui, M. Jordan, and G. Lanckriet. A direct formulation of sparse PCA using semidefinite programming. *SIAM Review*, 49(3), 2007.

[69] C. De Swart, B. Nijmeyer, J. Roelofs, and J. J. Sixma. Kinetics of intravenously administered heparin in normal humans. *Blood*, 60(6):1251–1258, 1982.

[70] J. W. Demmel. *Applied numerical linear algebra*, volume 56. Siam, 1997.

[71] S. Dempe. *Foundations of Bilevel Programming*. Springer, 2002.

[72] Department of Commerce, Bureau of the Census. Census of population and housing 1990 united states: Summary tape file 1a and 3a (computer files). 1992.

[73] Department of Justice, Bureau of Justice Statistics. Law enforcement and administrative statistics (computer file). 1992.

[74] Department of Justice, Federal Bureau of Investigation. Crime in the united states (computer file). *Source: http://www.fbi.gov/ucr/hc2004/openpage.htm*, 1995.

[75] L. Devroye and G. Wise. Detection of abnormal behavior via nonparametric estimation of the support. *SIAM Journal on Applied Mathematics*, 38(3):480–488, 1980.

[76] M. Dusenbury. "everbody was telling me there was nothing wrong". *BBC*, May 2018.

[77] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM, 2012.

[78] R. Dworkin. What is equality? part 1: Equality of welfare. *Philosophy & public affairs*, pages 185–246, 1981.

[79] Editorial. More accountability for big-data algorithms. *Nature*, 537, 2016.

[80] H. Edwards and A. Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

[81] L. Edwards. The gender gap in pain. *The New York Times*, 2013.

[82] B. Eidelson. *Discrimination and disrespect*. Oxford University Press, 2015.

[83] D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian. Runaway feedback loops in predictive policing. *arXiv preprint arXiv:1706.09847*, 2017.

[84] Executive Office of the President. *Big data: A report on algorithmic systems, opportunity, and civil rights*. 2016.

[85] A. Fawzi, O. Fawzi, and P. Frossard. Fundamental limits on adversarial robustness. In *ICML Workshop on Deep Learning*, 2015.

[86] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

[87] B. Q. Feng. Equivalence constants for certain matrix norms. *Linear algebra and its applications*, 374:247–253, 2003.

[88] S. Fernandez-Fraga, M. Aceves-Fernandez, J. Pedraza-Ortega, and S. Tovar-Arriaga. Feature extraction of eeg signal upon bci systems based on steady-state visual evoked potentials using the ant colony optimization algorithm. *Discrete Dynamics in Nature and Society*, 2018, 2018.

[89] M. A. Ferrag, L. Maglaras, A. Argyriou, D. Kosmanos, and H. Janicke. Security for 4g and 5g cellular networks: A survey of existing authentication and privacy-preserving schemes. *Journal of Network and Computer Applications*, 101:55–82, 2018.

[90] A. Feuerverger, R. A. Mureika, et al. The empirical characteristic function and its applications. *The annals of Statistics*, 5(1):88–97, 1977.

[91] R. A. Fisher. The goodness of fit of regression formulae, and the distribution of regression coefficients. *Journal of the Royal Statistical Society*, 85(4):597–612, 1922.

[92] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[93] P. W. Frey and D. J. Slate. Letter recognition using holland-style adaptive classifiers. *Machine learning*, 6(2):161–182, 1991.

[94] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

[95] R. G. Fryer Jr and G. C. Loury. Valuing diversity. *Journal of political Economy*, 121(4):747–774, 2013.

[96] Y. Fukuoka, M. Zhou, E. Vittinghoff, W. Haskell, K. Goldberg, and A. Aswani. Objectively measured baseline physical activity patterns in women in the mped trial: Cluster analysis. *JMIR Public Health and Surveillance*, 4(1):e10, 2018.

[97] F. Galton. Kinship and correlation. *The North American Review*, 150(401):419–431, 1890.

[98] S. Gerber, T. Tasdizen, S. Joshi, and R. Whitaker. On the manifold structure of the space of brain images. In *MICCAI*, pages 305–312, 2009.

[99] M. M. Ghassemi, S. E. Richter, I. M. Eche, T. W. Chen, J. Danziger, and L. A. Celi. A data-driven approach to optimized medication dosing: a focus on heparin. *Intensive care medicine*, 40(9):1332–1339, 2014.

[100] G. Goh, A. Cotter, M. Gupta, and M. P. Friedlander. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*, pages 2415–2423, 2016.

[101] Y. Goldberg and O. Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.

[102] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[103] Y. Grandvalet, J. Mariéthoz, and S. Bengio. A probabilistic interpretation of svms with an application to unbalanced classification. In *Advances in Neural Information Processing Systems*, pages 467–474, 2006.

[104] M. K. Griffith. 5g and security: There is more to worry about than huawei. Technical report, Wilson Center, November 2019.

[105] P. Gueth, D. Dauvergne, N. Freud, J. M. Létang, C. Ray, E. Testa, and D. Sarrut. Machine learning-based patient specific prompt-gamma dose monitoring in proton therapy. *Physics in Medicine & Biology*, 58(13):4563, 2013.

[106] A. Guntuboyina. Optimal rates of convergence for convex set estimation from support functions. *The Annals of Statistics*, 40(1):385–411, 2012.

[107] W. Hall. Representation of blacks, women, and the very elderly (aged¿ or= 80) in 28 major randomized clinical trials. *Ethnicity & disease*, 9(3):333–340, 1999.

[108] S. Har-Peled and S. Mahabadi. Near neighbor: Who is the fairest of them all? In *Advances in Neural Information Processing Systems*, pages 13176–13187, 2019.

[109] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

[110] D. Hellman. *When is discrimination wrong?* Harvard University Press, 2008.

[111] P. Hespanhol, M. Porter, R. Vasudevan, and A. Aswani. Dynamic watermarking for general lti systems. In *IEEE Conference on Decision and Control (CDC)*, pages 1834–1839, 2017.

[112] T. Hobbes. *Thomas Hobbes: Leviathan (Longman Library of Primary Sources in Philosophy)*. Routledge, 2016.

[113] S. B. Hopkins, P. K. Kothari, A. Potechin, P. Raghavendra, T. Schramm, and D. Steurer. The power of sum-of-squares for detecting hidden structures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 720–731. IEEE, 2017.

[114] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge university press, 2012.

[115] P. Horton and K. Nakai. A probabilistic classification system for predicting the cellular localization sites of proteins. In *Ismb*, volume 4, pages 109–115, 1996.

[116] Y.-H. Hu, C.-T. Tai, C.-F. Tsai, and M.-W. Huang. Improvement of adequate digoxin dosage: An application of machine learning approach. *Journal of healthcare engineering*, 2018, 2018.

[117] W. Huggins, P. Patil, B. Mitchell, K. B. Whaley, and E. M. Stoudenmire. Towards quantum machine learning with tensor networks. *Quantum Science and technology*, 4(2):024001, 2019.

[118] D. Hume. *Hume: Political Essays*. Cambridge University Press, 1994.

[119] R. Huseby. Can luck egalitarianism justify the fact that some are worse off than others? *Journal of Applied Philosophy*, 33(3):259–269, 2016.

[120] D.-T. Huynh, X. Wang, T. Q. Duong, N.-S. Vo, and M. Chen. Social-aware energy efficiency optimization for device-to-device communications in 5g networks. *Computer Communications*, 120:102–111, 2018.

[121] K. D. Johnson, D. P. Foster, and R. A. Stine. Impartial predictive modeling: Ensuring fairness in arbitrary models. *arXiv preprint arXiv:1608.00528*, 2016.

[122] M. Kac. Sur les fonctions indépendantes (i)(propriétés générales). *Studia Mathematica*, 6:46–58, 1936.

[123] S. M. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *NeurIPS*, pages 793–800, 2009.

[124] F. Kamiran and T. Calders. Classifying without discriminating. In *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on*, pages 1–6. IEEE, 2009.

[125] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Proceedings of the 2012th European Conference on Machine Learning and Knowledge Discovery in Databases-Volume Part II*, pages 35–50. Springer-Verlag, 2012.

[126] H. Kannan, A. Kurakin, and I. Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

[127] I. Kant. *Ethical philosophy: the complete texts of Grounding for the metaphysics of morals, and Metaphysical principles of virtue, part II of The metaphysics of morals, with On a supposed right to lie because of philanthropic concerns*. Hackett Publishing, 1994.

[128] O. Karan, C. Bayraktar, H. Gümüşkaya, and B. Karlık. Diagnosing diabetes using neural networks on small mobile devices. *Expert Systems with Applications*, 39(1):54–60, 2012.

[129] M. Kearns, S. Neel, A. Roth, and Z. S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*, 2017.

[130] G. S. Kirk et al. *Heraclitus: The cosmic fragments*. Cambridge University Press, 1954.

[131] J. Kleinberg. Inherent trade-offs in algorithmic fairness. In *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*, pages 40–40, 2018.

[132] J. Kleinberg, J. Ludwig, S. Mullainathan, and A. Rambachan. Algorithmic fairness. In *Aea papers and proceedings*, volume 108, pages 22–27, 2018.

[133] J. Kleinberg and S. Mullainathan. Simplicity creates inequity: implications for fairness, stereotypes, and interpretability. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 807–808, 2019.

[134] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[135] A. Korostelëv, L. Simar, and A. Tsybakov. Efficient estimation of monotone boundaries. *The Annals of Statistics*, pages 476–489, 1995.

[136] D. Kraft. A software package for sequential quadratic programming. *Forschungsbericht-Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt*, 1988.

[137] M. A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, 37(2):233–243, 1991.

[138] N. Krause and Y. Singer. Leveraging the margin more carefully. In *ICML*, page 63, 2004.

[139] J.-L. Krivine. Anneaux preordonnes. *Journal d'analyse mathematique*, 12(1):307–326, 1964.

[140] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[141] S. Kullback. *Information theory and statistics*. Courier Corporation, 1997.

[142] A. Kumar and R. Kannan. Clustering with spectral norm and the k-means algorithm. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 299–308. IEEE, 2010.

[143] G. R. Lanckriet, L. E. Ghaoui, C. Bhattacharyya, and M. I. Jordan. A robust minimax approach to classification. *Journal of Machine Learning Research*, 3(Dec):555–582, 2002.

[144] R. Langner. Stuxnet: Dissecting a cyberwarfare weapon. *IEEE Security & Privacy*, 9(3):49–51, 2011.

[145] L. Lasagna. Addicting drugs and medical practice: towards the elaboration of realistic goals and the eradication of myths, mirages and half-truths. *Narcotics. New York: McGraw*, 53:66, 1965.

[146] J. B. Lasserre. A sum of squares approximation of nonnegative polynomials. *SIAM review*, 49(4):651–669, 2007.

[147] J.-B. Lasserre. *Moments, positive polynomials and their applications*. World Scientific, 2010.

[148] K. Lau, T. Wigren, R. Delgado, and R. H. Middleton. Disturbance rejection properties for a 5g networked data flow delay controller. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 1681–1687. IEEE, 2017.

[149] Y. LeCun, C. Cortes, and C. Burges. MNIST handwritten digit database. *AT&T Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[150] A. M. Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot, 1805.

[151] E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[152] A. Lever. Racial profiling and the political philosophy of race. *The Oxford Handbook of Philosophy and Race*, page 425, 2016.

[153] S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

[154] M. Lichman. UCI machine learning repository, 2013.

[155] K. Lippert-Rasmussen. *Born free and equal?: A philosophical inquiry into the nature of discrimination*. Oxford University Press, 2014.

[156] Z. Lipton, J. McAuley, and A. Chouldechova. Does mitigating ml's impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems*, pages 8125–8135, 2018.

[157] P. J. Lisboa and A. F. Taktak. The use of artificial neural networks in decision support in cancer: a systematic review. *Neural networks*, 19(4):408–415, 2006.

[158] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt. Delayed impact of fair machine learning. *arXiv preprint arXiv:1803.04383*, 2018.

[159] Y. Liu and X. Shen. Multicategory ψ-learning. *Journal of the American Statistical Association*, 101(474):500–509, 2006.

[160] R. Livni, K. Crammer, and A. Globerson. A simple geometric interpretation of SVM using stochastic adversaries. In *AISTATS*, pages 722–730, 2012.

[161] J. Locke. *Second treatise of government: An essay concerning the true original, extent and end of civil government*. John Wiley & Sons, 2014.

[162] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

[163] D. Madras, E. Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018.

[164] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[165] A. Majumdar, R. Vasudevan, M. M. Tobenkin, and R. Tedrake. Convex optimization of nonlinear feedback controllers via occupation measures. *The International Journal of Robotics Research*, 33(9):1209–1230, 2014.

[166] L. Manchikanti, A. M. Kaye, N. N. Knezevic, H. McAnally, K. Slavin, A. M. Trescot, and J. Hirsch. Responsible, safe, and effective prescription of opioids for chronic non-cancer pain: American society of interventional pain physicians (asipp) guidelines. *Pain Physician*, 20(2S):S3–S92, 2017.

[167] K. Mansouri, T. Ringsted, D. Ballabio, R. Todeschini, and V. Consonni. Quantitative structure–activity relationship models for ready biodegradability of chemicals. *Journal of chemical information and modeling*, 53(4):867–878, 2013.

[168] J. Mao and A. K. Jain. Artificial neural networks for feature extraction and multivariate data projection. *IEEE transactions on neural networks*, 6(2):296–317, 1995.

[169] K. Marx and F. Engels. *The communist manifesto*. Penguin, 2002.

[170] H. Masnadi-Shirazi, V. Mahadevan, and N. Vasconcelos. On the design of robust classifiers for computer vision. In *CVPR*, pages 779–786, 2010.

[171] P. Massart. *Concentration inequalities and model selection*, volume 6. Springer, 2007.

[172] C. Masterson. Massive mimo and beamforming: The signal processing behind the 5g buzzwords. *10 Massive MIMO and Beamforming: The Signal Processing Behind the 5G Buzzwords*, page 10, 2017.

[173] G. Matheron. *Random sets and integral geometry*. John Wiley & Sons, 1975.

[174] H. McQuay. Opioids in pain management. *The Lancet*, 353(9171):2229–2232, 1999.

[175] J. S. Mill. *Utilitarianism*. Longmans, Green and Company, 1895.

[176] C. C. Miller. Can an algorithm hire better than a human. *The New York Times*, 25, 2015.

[177] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[178] Y. Mo, R. Chabukswar, and B. Sinopoli. Detecting integrity attacks on scada systems. *IEEE CST*, 22(4):1396–1407, 2014.

[179] Y. Mo, E. Garone, A. Casavola, and B. Sinopoli. False data injection attacks against state estimation in wireless sensor networks. In *Proc. of IEEE CDC*, pages 5967–5972, 2010.

[180] Y. Mo and B. Sinopoli. Secure control against replay attacks. In *Allerton Conference*, pages 911–918. IEEE, 2009.

[181] Y. Mo, S. Weerakkody, and B. Sinopoli. Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs. *IEEE Control Systems*, 35(1):93–109, 2015.

[182] R. Moddemeijer. On estimation of entropy and mutual information of continuous distributions. *Signal processing*, 16(3):233–248, 1989.

[183] I. Molchanov. *Theory of random sets*. Springer Science & Business Media, 2006.

[184] MOSEK, ApS. The mosek optimization tools version 3.2 (revision 8) user's manual and reference, 2002.

[185] C. Munoz, M. Smith, and D. Patil. Big data: A report on algorithmic systems, opportunity, and civil rights. *Executive Office of the President. The White House*, 2016.

[186] L. Nanni and A. Lumini. An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert systems with applications*, 36(2):3028–3033, 2009.

[187] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari. Learning with noisy labels. In *NeurIPS*, pages 1196–1204, 2013.

[188] S. Nemati, M. M. Ghassemi, and G. D. Clifford. Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2978–2981. IEEE, 2016.

[189] F. Nielsen. What is an information projection. *Notices of the AMS*, 65(3):321–324, 2018.

[190] M. Olfat and A. Aswani. Average margin regularization for classifiers. *arXiv preprint arXiv:1810.03773*, 2018.

[191] M. Olfat and A. Aswani. Spectral algorithms for computing fair support vector machines. In *International Conference on Artificial Intelligence and Statistics*, pages 1933–1942, 2018.

[192] M. Olfat and A. Aswani. Convex formulations for fair principal component analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 663–670, 2019.

[193] M. Olfat, S. Sloan, P. Hespanhol, M. Porter, R. Vasudevan, and A. Aswani. Covariance-robust dynamic watermarking. *arXiv preprint arXiv:2003.13908*, 2020.

[194] A. Ouattara and A. Aswani. Duality approach to bilevel programs with a convex lower level. In *2018 Annual American Control Conference (ACC)*, pages 1388–1395. IEEE, 2018.

[195] D. Pál, B. Póczos, and C. Szepesvári. Estimation of rényi entropy and mutual information based on generalized nearest-neighbor graphs. In *Advances in Neural Information Processing Systems*, pages 1849–1857, 2010.

[196] S. Paluch and S. Tuzovic. Leveraging pushed self-tracking in the health insurance industry: How do individuals perceive smart wearables offered by insurance organization? 2017.

[197] D. Parfit. Equality and priority. *Ratio*, 10(3):202–221, 1997.

[198] T. Patschkowski, A. Rohde, et al. Adaptation to lowest density regions with application to support recovery. *The Annals of Statistics*, 44(1):255–287, 2016.

[199] K. Pearson. The law of ancestral heredity. *Biometrika*, 2(2):211–228, 1903.

[200] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[201] C. Perlich, B. Dalessandro, T. Raeder, O. Stitelman, and F. Provost. Machine learning for targeted display advertising: Transfer learning in action. *Machine learning*, 95(1):103–127, 2014.

[202] G. Peyré. Manifold models for signals and images. *Computer Vision and Image Understanding*, 113(2):249–260, 2009.

[203] H. Peyrl and P. A. Parrilo. Computing sum of squares decompositions with rational coefficients. *Theoretical Computer Science*, 409(2):269–281, 2008.

[204] E. S. Phelps. The statistical theory of racism and sexism. *The american economic review*, 62(4):659–661, 1972.

[205] R. Pless and R. Souvenir. A survey of manifold learning for images. *IPSJ Transactions on Computer Vision and Applications*, 1:83–94, 2009.

[206] A. Raghunathan, J. Steinhardt, and P. Liang. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*, 2018.

[207] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez. Millimeter wave mobile communications for 5g cellular: It will work! *IEEE access*, 1:335–349, 2013.

[208] J. Rawls. *A theory of justice*. Harvard university press, 2009.

[209] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[210] M. Redmond and A. Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.

[211] A. Roberts. The utility of an invariant manifold description of the evolution of a dynamical system. *SIAM journal on mathematical analysis*, 20(6):1447–1458, 1989.

[212] R. J.-B. Rockafellar, R Tyrrell & Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[213] J. O. Royset and R. J.-B. Wets. Variational analysis of constrained m-estimators. *Annals of Statistics*, 2019. Accepted.

[214] C. Rudin. Predictive policing using machine learning to detect patterns of crime. *Wired Magazine, August*, 2013.

[215] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[216] B. Russell. *History of western philosophy: Collectors edition*. Routledge, 2013.

[217] A.-L. Sachs. The data-driven newsvendor with censored demand observations. In *Retail Analytics*, pages 35–56. Springer, 2015.

[218] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical care medicine*, 39(5):952, 2011.

[219] M. Salam. The opioid epidemic: a crisis years in the making. *The New York Times*, 26, 2017.

[220] J. Sallis, A. Bauman, and M. Pratt. Environmental and policy interventions to promote physical activity a. *American journal of preventive medicine*, 15(4):379–397, 1998.

[221] S. Samadi, U. Tantipongpipat, J. H. Morgenstern, M. Singh, and S. Vempala. The price of fair pca: One extra dimension. In *Advances in Neural Information Processing Systems*, pages 10976–10987, 2018.

[222] B. Satchidanandan and P. Kumar. Dynamic watermarking: Active defense of networked cyber-physical systems. *Proc. of IEEE*, 2016.

[223] B. Satchidanandan and P. Kumar. On minimal tests of sensor veracity for dynamic watermarking-based defense of cyber-physical systems. In *Communication Systems and Networks (COMSNETS), 2017 9th International Conference on*, pages 23–30. IEEE, 2017.

[224] T. M. Scanlon. *Moral dimensions*. Harvard University Press, 2009.

[225] F. F. Schauer. *Profiles, probabilities, and stereotypes*. Harvard University Press, 2003.

[226] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[227] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

[228] B. Scholkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[229] A. K. Sen. *Inequality reexamined*. Oxford University Press, 1992.

[230] R. S. Sexton, S. McMurtrey, J. O. Michalopoulos, and A. M. Smith. Employee turnover: a neural network solution. *Computers & Operations Research*, 32(10):2635–2651, 2005.

[231] H. D. Sherali and W. P. Adams. *A reformulation-linearization technique for solving discrete and continuous nonconvex problems*, volume 31. Springer Science & Business Media, 2013.

[232] L.-C. Shi and B.-L. Lu. Off-line and on-line vigilance estimation based on linear dynamical system and manifold learning. In *EMBC*, pages 6587–6590, 2010.

[233] G. R. Shorack and J. A. Wellner. *Empirical processes with applications to statistics*, volume 59. SIAM, 2009.

[234] J. W. Smith, J. Everhart, W. Dickson, W. Knowler, and R. Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 261. American Medical Informatics Association, 1988.

[235] A. J. Smola, S. Vishwanathan, and T. Hofmann. Kernel methods for missing variables. In *AISTATS*, 2005.

[236] P. Sollich. Probabilistic interpretations and bayesian methods for support vector machines. 1999.

[237] Q. Song, W. Hu, and W. Xie. Robust support vector machine with bullet hole image classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 32(4):440–448, 2002.

[238] C. J. Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, pages 1348–1360, 1980.

[239] S. Suzumura, K. Ogawa, M. Sugiyama, and I. Takeuchi. Outlier path: A homotopy algorithm for robust SVM. In *ICML*, pages 1098–1106, 2014.

[240] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

[241] G. J. Székely, M. L. Rizzo, et al. Brownian distance covariance. *The annals of applied statistics*, 3(4):1236–1265, 2009.

[242] E. R. Teoh and D. G. Kidd. Rage against the machine? google's self-driving cars versus human drivers. *Journal of safety research*, 63:57–60, 2017.

[243] J. J. Thompson, M. R. Blair, L. Chen, and A. J. Henrey. Video game telemetry as a critical tool in the study of complex skill learning. *PloS one*, 8(9):e75129, 2013.

[244] R. Tomioka and T. Suzuki. Spectral norm of random tensors. *arXiv preprint arXiv:1407.1870*, 2014.

[245] T. B. Trafalis and R. C. Gilbert. Robust support vector machines for classification and computational issues. *Optimisation Methods and Software*, 22(1):187–198, 2007.

[246] A. Tsanas and A. Xifara. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, 49:560–567, 2012.

[247] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. There is no free lunch in adversarial robustness (but there are unexpected benefits). *arXiv preprint arXiv:1805.12152*, 2018.

[248] V. Turri, A. Carvalho, H. Tseng, K. Johansson, and F. Borrelli. Linear model predictive control for lane keeping and obstacle avoidance on low curvature roads. In *Proc. of IEEE ITSC*, pages 378–383, 2013.

[249] H. Tuy. Dc optimization: theory, methods and algorithms. In *Handbook of global optimization*, pages 149–216. Springer, 1995.

[250] V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[251] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, 2015.

[252] J. S. Vardakas, I. T. Monroy, L. Wosinska, G. Agapiou, R. Brenot, N. Pleros, and C. Verikoukis. Towards high capacity and low latency backhauling in 5g: The 5g stepfwd vision. In *2017 19th International Conference on Transparent Optical Networks (ICTON)*, pages 1–4. IEEE, 2017.

[253] C. Wadsworth, F. Vera, and C. Piech. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199*, 2018.

[254] M. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. To appear., 2017.

[255] M. J. Wainwright, M. I. Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[256] S. Wang, A. Schwing, and R. Urtasun. Efficient inference of continuous markov random fields with polynomial potentials. In *Advances in neural information processing systems*, pages 936–944, 2014.

[257] S. Weerakkody, Y. Mo, and B. Sinopoli. Detecting integrity attacks on control systems using robust physical watermarking. In *Proc. of IEEE CDC*, pages 3757–3764, 2014.

[258] Z. Wen, D. Goldfarb, and W. Yin. Alternating direction augmented lagrangian methods for semidefinite programming. *Mathematical Programming Computation*, 2(3-4):203–230, 2010.

[259] J. Weston and R. Herbrich. Adaptive margin support vector machines. *NeurIPS*, pages 281–296, 1999.

[260] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics*, 9(1):60–62, 1938.

[261] B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*, 2017.

[262] H. Xu, C. Caramanis, and S. Mannor. Robust regression and lasso. In *NeurIPS*, pages 1801–1808, 2009.

[263] H. Xu, C. Caramanis, and S. Mannor. Robustness and regularization of support vector machines. *J. Mach. Learn. Res,*, 10(Jul):1485–1510, 2009.

[264] L. Xu, K. Crammer, and D. Schuurmans. Robust support vector machine training via convex outlier ablation. In *AAAI*, volume 6, pages 536–542, 2006.

[265] G. Yauney and P. Shah. Reinforcement learning with action-derived rewards for chemotherapy and clinical trial dosing regimen selection. In *Machine Learning for Healthcare Conference*, pages 161–226, 2018.

[266] I.-C. Yeh and C.-h. Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009.

[267] Y.-l. Yu, M. Yang, L. Xu, M. White, and D. Schuurmans. Relaxed clipping: A global training method for robust regression and classification. In *NeurIPS*, pages 2532–2540, 2010.

[268] A. L. Yuille and A. Rangarajan. The concave-convex procedure (cccp). In *Advances in neural information processing systems*, pages 1033–1040, 2002.

[269] G. U. Yule. On the theory of correlation. *Journal of the Royal Statistical Society*, 60(4):812–854, 1897.

[270] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

[271] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *Proceedings of the International Conference on Machine Learning*, pages 325–333, 2013.

[272] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. *arXiv preprint arXiv:1801.07593*, 2018.

[273] N. Zhang, J. Wang, G. Kang, and Y. Liu. Uplink nonorthogonal multiple access in 5g systems. *IEEE Communications Letters*, 20(3):458–461, 2016.

[274] P. Zhao, S. Mohan, and R. Vasudevan. Optimal control of polynomial hybrid systems via convex relaxations. *IEEE Transactions on Automatic Control*, 2019.

[275] F. Zhou, Q. Claire, and R. D. King. Predicting the geographical origin of music. In *2014 IEEE International Conference on Data Mining*, pages 1115–1120. IEEE, 2014.

[276] I. Zliobaite. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*, 2015.

[277] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.