Learning Provably Useful Representations, with Applications to Fairness

Daniel McNamara

A thesis submitted for the degree of
Doctor of Philosophy
College of Engineering and Computer Science
The Australian National University

August 2019

Except where otherwise indicated, this thesis is my own original work.

Daniel McNamara
9 August 2019

Acknowledgments

I would like to thank the funding agencies who supported my research. I was a recipient of an Australian Government Research Training Program Scholarship (formerly an Australian Postgraduate Award). I was also supported by a CSIRO Data61 (formerly NICTA) Top-Up Scholarship. I was privileged to receive a Fulbright Postgraduate Scholarship, which supported an eight-month research visit to the United States in 2016-17.

I would like to acknowledge the Australian National University and CSIRO Data61 for creating a supportive and stimulating work environment during the course of my PhD. I am lucky to have been part of lectures, seminars, reading groups and research retreats, and to have interacted with many staff and students in the ANU Research School of Computer Science and in the CSIRO Data61 Analytics Research Group. In particular, Aditya Menon offered invaluable discussions during the course of my research.

I would like to thank the Machine Learning Department at Carnegie Mellon University for having me for a very fruitful stay in Pittsburgh. In particular I would like to acknowledge A/Prof Nina Balcan and her group of graduate students, who generously hosted me and helped me to develop my research direction. It was satisfying that Nina and I were able to publish a paper based on our collaboration.

Prof Marnie Hughes-Warrington, Richelle Hilton and Leone Nurbasari helped me to arrange a productive collaboration with the Planning and Performance Measurement Unit at the Australian National University. I enjoyed working with the team there as part of my research. In particular, it was fantastic to co-present a paper with Leone at the Australasian Association for Institutional Research Forum in November 2018.

I am fortunate to have had the chance to collaborate with Tim Graham and Ellen Broad on a case study about recidivism prediction. They contributed substantially to our joint work on this topic, which forms the basis of Chapter 6 and has been released as a chapter in an edited book.

I would like to thank the three examiners who reviewed my thesis. Their feedback has been invaluable and has been incorporated into the final version of the thesis.

My supervisory panel provided ongoing support and guidance through my PhD. My adviser Dr Lizhen Qu provided useful perspectives and discussions. My co-supervisor Dr Cheng Soon Ong provided many insights and encouragements along the way, as well as teaching and internship opportunities. My lead supervisor Prof Bob Williamson has been a formidable intellectual sparring partner as well as a valued mentor. I am particularly grateful for his instrumental role in setting up my research visit to the US.

To my friends who helped me through the journey, thank you. Study retreats with Dan Musil and Ananth Gopal – nose surgeries and all – were a great morale booster amidst the thesis slog. My book club helped me to keep all the different parts of my brain active, and has always been an eagerly anticipated event in my calendar. In my years calling Canberra home I am grateful for the community of people around me, who have seen me through the highs and lows.

Thanks to my cousin Natasha for her support as I approached the finish line. Thanks to Mum, who gave me as much love now as she did during her own PhD, which she completed while raising me. I am grateful to Dad, who has always been my strongest supporter and has never stopped believing in me.

To Ayes – it's been quite a journey, but I am so grateful to have had you by my side. Your love means the world to me. I can't wait for the next era of our life together!

Abstract

Representation learning involves transforming data so that it is useful for solving a particular supervised learning problem. The aim is to learn a representation function which maps inputs to some representation space, and an hypothesis which maps the representation space to targets. It is possible to learn a representation function using unlabeled data or data from a probability distribution other than that of the main problem of interest, which is helpful if labeled data is scarce. This approach has been successfully applied in practice, for example through pre-trained neural networks in computer vision and word embeddings in natural language processing. This thesis explores when it is possible to learn representations that are provably useful.

We consider learning a representation function from unlabeled data, and propose an approach to identifying conditions where this technique will be useful for a subsequent supervised learning task. The approach requires shared structure in the labeled and unlabeled distributions, as well as a compatible representation function class and hypothesis class. We provide an example where representation learning can exploit cluster structure present in the data.

We also consider learning a representation function from a source task distribution and re-using it on a target task of interest, and again propose conditions where this approach will be successful. In this case the conditions depend on shared structure between source and target task distributions. We provide an example involving the transfer of weights in a two-layer feedforward neural network.

Representation learning can be applied to another topic of interest: fairness in machine learning. The issue of fairness arises when machine learning systems make or provide advice on decisions about people. A common approach to defining fairness is measuring differences in decisions made by an algorithm for one demographic group compared to another. One approach to preventing discrimination against particular groups is to learn a representation of the data from which it is not possible for an adversary to determine an individual's group membership, but which preserves other useful information. We quantify the costs and benefits of such an approach with respect to several possible fairness definitions. We also examine the relationships between different definitions of fairness and show cases where they cannot simultaneously be satisfied.

We explore the use of representation learning for fairness through two case studies: predicting domestic violence recidivism while avoiding discrimination on the basis of race, and predicting student outcomes at university while avoiding discrimination on the basis of gender. Our case studies reveal both the utility of fair representation learning and the trade-offs between accuracy and the definitions of fairness considered.

Contents

Acknowledgments v

Abstract vii

List of Figures xiii

List of Tables xv

1 Introduction 1
- 1.1 Representations . . . 1
- 1.2 Learning Representations . . . 2
- 1.3 Provably Useful Representations . . . 4
- 1.4 Applications to Fairness . . . 6
- 1.5 Contributions of this Thesis . . . 7
- 1.6 Structure of this Thesis . . . 7
- 1.7 Research Outputs Produced during PhD Candidature . . . 9

I Learning and Re-Using Representations 13

2 Unsupervised Representation Learning to Provably Improve Task Performance 15
- 2.1 Introduction . . . 15
- 2.2 Background . . . 16
- 2.3 When Unsupervised Representation Learning is Provably Useful . . . 18
- 2.4 Cluster Example . . . 23
- 2.5 Conclusion . . . 30

3 Risk Bounds for Transferring Representations With and Without Fine-Tuning 33
- 3.1 Introduction . . . 33
- 3.2 Background . . . 35
- 3.3 Representation Function Fixed by Source Task . . . 36
- 3.4 Representation Function Fine-Tuned on Target Task . . . 41
- 3.5 Applications . . . 48
- 3.6 Conclusion . . . 51
- 3.7 Appendix . . . 51

II The Use and Limits of Representation Learning for Fairness 57

4 Costs and Benefits of Fair Representation Learning 59
4.1 Introduction 59
4.2 Background 62
4.3 Fair Classification vs Fair Representation Learning 64
4.4 Costs of Fair Representation Learning 68
4.5 Benefits of Fair Representation Learning 72
4.6 Experiments 73
4.7 Conclusion 76
4.8 Appendix 77

5 Equalized Odds Implies Partially Equalized Outcomes Under Realistic Assumptions 87
5.1 Introduction 87
5.2 Problem Formalization 90
5.3 The Relationship Between Equalized Odds and Equalized Outcomes . . 92
5.4 Generalization of Calibration-Equalized Odds Impossibility Result . . . 97
5.5 Conclusion 99
5.6 Appendix 99

III Applying Representation Learning to Fairness: Two Case Studies 105

6 Trade-offs in Algorithmic Risk Assessment: an Australian Domestic Violence Case Study 107
6.1 Introduction 107
6.2 Algorithmic Risk Assessment in an Australian Domestic Violence Context 108
6.3 Designing Fair Algorithmic Risk Assessments 112
6.4 Predicting DV Recidivism with the BOCSAR Dataset 115
6.5 Conclusion 120

7 Using Cohort Analysis and Predictive Modeling to Inform Targeted Student Support 123
7.1 Introduction 123
7.2 Data Inputs and Student Outcomes 124
7.3 Analysis of Student Cohorts 126
7.4 Individualized Predictions of Student Outcomes 134
7.5 Incorporating Fairness into Predictive Model Design 138
7.6 Conclusion 141

8 Conclusion 143
8.1 When Representation Learning is Provably Useful 143
8.2 Methodological Insights 145
8.3 Issues in Applying Representation Learning to Fairness 146
8.4 Future Work 147

Bibliography 153

List of Figures

1.1 Using an intermediate representation in prediction 4
1.2 Summary of problem settings considered in this thesis 8

2.1 Measuring the effect of unsupervised representation learning 16
2.2 Relationship between semi-supervised learning and unsupervised representation learning 19
2.3 Cluster example of unsupervised representation learning 23

3.1 Approaches to learning a representation on a target task 34
3.2 Neural network learning from scratch and with weights transferred . . 39
3.3 Risk bounds with and without fine-tuning 49

4.1 Summary of fair classification and fair representation learning 60
4.2 Estimates of costs and benefits of fair representation learning on Adult and ProPublica datasets 76
4.3 Examples illustrating the cost of mistrust 84

5.1 Visualization of relationship between equalized odds and equalized outcomes 95
5.2 Motivation: equalized odds appears related to equalized outcomes . . . 104

6.1 Learning fair representations with an adversary 115
6.2 Results of applying pre-processing to the BOCSAR dataset 117

7.1 Cohort analysis of first year attrition 128
7.2 Cohort analysis of failing at least one first year course 129
7.3 Relationship of gender to probability of failing at least one first year course 130
7.4 Cohort analysis of first year grade point average 132
7.5 Descriptive results for first year grade point average using high school subject features 133
7.6 Performance of several models and feature sets in predicting attrition and failing at least one course 136
7.7 Performance of several models and feature sets in predicting GPA . . . 137
7.8 Addressing potential for gender discrimination when predicting at least one year 1 course failure 140

8.1 Three situations where representation learning is provably useful 144

List of Tables

1.1 Accepted peer-reviewed papers based on work in this thesis 10
1.2 Other research outputs based on work in this thesis 11

2.1 Approach to identifying sufficient conditions for successful unsupervised representation learning 22

3.1 Evaluation of transferring representations 51

4.1 Fair representation learning: summary of problems 77
4.2 Fair representation learning: summary of costs 78
4.3 Fair representation learning: summary of benefits 78

5.1 Summary of fairness definitions and impossibility results 88
5.2 Test set results for a recidivism prediction model on the ProPublica dataset 103

6.1 Independent features in the BOCSAR dataset 116

7.1 Features used in the analysis of student outcomes at university 125
7.2 Feature sets used for individualized predictions 134

Chapter 1

Introduction

Representation learning is a technique at the core of many of the most successful machine learning systems. But what is it, and why does it work? And what is its connection to a topic of increasing importance: fairness in machine learning? We provide an introduction to these questions, including a high-level review of prior work, leaving more detailed literature reviews to the relevant chapters.

1.1 Representations

Data is a representation of something in the world. However, there is more than one way to represent a particular thing. We face a succession of choices about how to represent the thing, including selecting what to measure, how to measure it and to what level of precision, and how to store and organize those measurements. The tension between the seemingly objective existence of the thing, and the subjective nature of its representation, has long been discussed in fields such as philosophy [Floridi, 2010], linguistics [Saussure, 2011], cultural studies [Gitelman, 2013] and physics [Wheeler, 1990].

Consider the case that the *something* is the opinions of a group of people. These opinions might be measured via a survey, with the responses stored in a spreadsheet. The spreadsheet – a matrix whose cells contain the responses of each individual to each survey question – is data representing the respondents' opinions. We might use the data to produce a summary table aggregating average opinion across several questions. This summary table is another representation of the respondents' opinions.

Which representation is better, the spreadsheet or the summary table? A human might find the summary table easier to interpret and hence more useful for gaining a high-level understanding of respondents' opinions. The spreadsheet has some advantages, however: it contains all the information in the summary table, plus more. One could reconstruct the summary table from the spreadsheet, but going in the other direction would be impossible.

This example gives us an intuition that there is no general answer to the question 'what makes a good representation?'. When we change the way data is represented, we don't add any new information and we may well destroy some. This idea has been formalized in the *data processing inequality* (Theorem 2.8.1 of [Cover and Thomas, 2012]). We might be tempted to conclude that moving from one representation to another never helps. And yet, if we have in mind *a particular purpose*, changing the way that the data is represented can be very helpful. We must recognize this paradox if we want to understand representations.

1.2 Learning Representations

A representation is the result of applying a function f – which we refer to as a *representation function* – to an *input variable* X, yielding a new variable $Z := f(X)$. We refer to Z as a *representation variable*. It is useful to assume that the data is drawn from a probability distribution over X and *target variable* Y, which we refer to as μ_{XY}. Hence, the application of the representation function to this distribution results in an *induced distribution* μ_{ZY}. Typically we would like to learn an hypothesis mapping X to Y, which is known as *supervised learning*, and we hope that the introduction of the intermediate representation Z helps to achieve this.[1]

In machine learning, the representation function is often handcrafted, a process known as *feature extraction* or *feature engineering*. This approach allows a domain expert – or someone with statistical training – to incorporate their knowledge into the representation, prior to handing over the data to a learning algorithm. A disadvantage of this approach is that it provides no performance guarantees and requires a custom implementation for each new problem.

Representation learning – also known as feature learning – involves automatically learning such a representation function from data [Bengio et al., 2013]. As we shall see, representation learning is intimately connected to several machine learning techniques, including deep learning, unsupervised learning, transfer learning, manifold learning and kernel learning.

Deep learning – the dominant paradigm in machine learning in recent years – can be seen as a type of representation learning [Goodfellow et al., 2016]. An example is a feedforward neural network, which learns successive layers of representations. Each node in a layer is constructed using a weighted sum of nodes in the previous layer, which is passed through a fixed activation function. The target is predicted using a weighted sum of nodes in the penultimate layer. The weights are learned by minimizing some loss function over a training set. Some neural network architectures can be interpreted as probabilistic models, such as deep belief networks [Hinton et al., 2006], with the hidden units acting as latent variables. In general, from a probabilistic perspective, representation learning is equivalent to learning a relationship between observed and latent variables.

There has been considerable research on using deep learning techniques to learn representations from unlabeled data – an approach known as *unsupervised representa-*

[1] In this chapter we introduce common notation widely used in the thesis, where it helps to clarify our discussion. Each subsequent chapter also contains a complete description of relevant notation. While notation is mostly consistent across the thesis, where a symbol has a different meaning in a particular chapter this is explained at the beginning of the chapter.

tion learning – or data which has been labeled for another task – an approach known as *transfer representation learning* [Bengio, 2012]. These techniques are of interest since often these types of data are abundant, while labeled data for the task of interest is scarce. For example, the weights in a neural network may be trained to reconstruct unlabeled examples provided as inputs, then re-used for a supervised learning task, a technique which is known as *pre-training* and is used in computer vision [Hinton and Salakhutdinov, 2006]. A popular class of methods for learning re-usable representations from unlabeled images is known as 'self-supervision', which involves learning neural network weights from tasks which do not require human-created class labels, such as predicting the position of an image patch [Doersch et al., 2015] or the pixels of a missing image patch [Pathak et al., 2016]. Another example is learning vector representations of words by training a neural network to predict word co-occurence, then re-using the vector representations for a supervised learning task, a technique which is known as *word embeddings* and is used in natural language processing [Mikolov et al., 2013]. In some cases the learned representation function is made publicly available, allowing others to re-use it. Hence, representation learning can be seen as a mechanism to modularize supervised learning.

Manifold learning is another important approach to unsupervised representation learning, which involves transforming the input data to a lower-dimensional space. This is motivated by computational efficiency, and the objective of exploiting structure present in the unlabeled data. Feature selection – where certain dimensions are discarded – is a simple type of dimensionality reduction. Principal components analysis (PCA) is a linear manifold learning technique for learning a lower-dimensional subspace on which to project the input data. A number of non-linear manifold learning techniques have been developed which build upon PCA, by learning a lower-dimensional representation optimizing the preservation of pairwise distances between points in the sample [Bengio et al., 2004]. A variant on manifold learning involves learning sparse representations – which can be of higher dimensionality than the input data but may be more efficient to work with, since fewer dimensions are required to describe each data point – via techniques such as clustering and sparse coding [Olshausen and Field, 1996].

Methods for learning kernel functions can be seen as another approach to representation learning. Kernel methods involve using a kernel function $k(x, x')$ which has the property that $k(x, x') = f(x) \cdot f(x')$, where f is a representation function into some high (possibly infinite) dimensional space, x and x' are input samples, and $\cdot$ denotes the dot product. Hence, learning k is equivalent to learning f. While often the kernel is learned using labeled data [Ong et al., 2005], other techniques involve unsupervised kernel learning [Zhuang et al., 2011]. Kernel methods are motivated by the fact that using k may be more computationally efficient than explicitly computing f if the dimensionality of $f(x)$ is large.

Representation learning can be used to solve machine learning problems across a wide variety of data types. Representations of entities can be learned from data describing relationships between entities. For example, singular value decomposition can be used to extract representations of words and documents in natural language

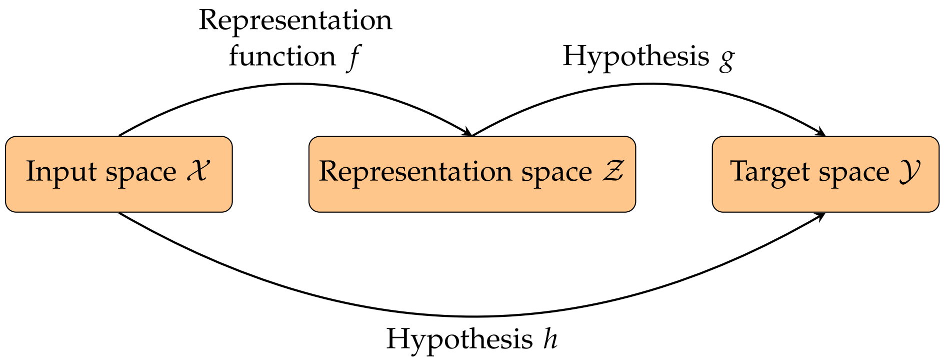

Representation function f
Hypothesis g
Input space $\mathcal{X}$
Representation space $\mathcal{Z}$
Target space $\mathcal{Y}$
Hypothesis h

Figure 1.1: Using an intermediate representation in prediction. We may obtain a mapping from input space $\mathcal{X}$ to target space $\mathcal{Y}$ can via an hypothesis $h : \mathcal{X} \to \mathcal{Y}$. An alternative is via an intermediate representation space $\mathcal{Z}$, a representation function $f : \mathcal{X} \to \mathcal{Z}$, an hypothesis $g : \mathcal{Z} \to \mathcal{Y}$ and the function composition $g \circ f$.

processing [Deerwester et al., 1990], or of users and products in recommender systems [Sarwar et al., 2000]. Representation learning is used to analyze time series data [Keogh and Pazzani, 1998], to enable the integration of images and text via shared embeddings [Socher et al., 2013], and in the sequence-to-sequence models used in machine translation systems [Sutskever et al., 2014]. With the rise of automated machine learning systems [Feurer et al., 2015] – where there is no human in the loop anywhere in the system's design – the value of representation learning continues to grow.

1.3 Provably Useful Representations

We would like to be able to *prove* that a representation is useful. To do so, we must define 'useful': what we will be using the representation for? Can we move beyond informal descriptions – such as the intuition that representing an image by edges and image segments may be more useful than raw pixels since they provide a higher level abstraction of its contents – to a formal description of the value of a representation?

To make progress, we focus on understanding when a representation is useful for prediction.[2] Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ be input, target and representation spaces corresponding to the variables X, Y and Z respectively. In supervised learning we learn an hypothesis $h : \mathcal{X} \to \mathcal{Y}$ which makes predictions of the target variable using the input variable. Alternatively, consider transforming the input via a representation function $f : \mathcal{X} \to \mathcal{Z}$, then learning an hypothesis $g : \mathcal{Z} \to \mathcal{Y}$ which makes predictions of the target variable using this representation. The final form of such a function may be written $g \circ f$, where $\circ$ denotes function composition. These two approaches to mapping $\mathcal{X}$ to $\mathcal{Y}$ are shown in Figure 1.1.

[2] A useful distinction can be made between the *predictive* and *explanatory* power of mathematical models [Shmueli, 2010]. While we focus on representation learning in the context of prediction, it may be possible to formalize its usefulness for explanation as well.

It is straightforward to see that for any given f, the class of hypotheses that may be written $g \circ f$ is a subset of the set all hypotheses mapping $\mathcal{X}$ to $\mathcal{Y}$. Hence, the best hypothesis using the representation will be no better, and possibly worse, than the best hypothesis using the input data. Should we conclude that all representations are useless?

Past theoretical results have examined cases when we are not too much worse off using a representation instead of the original input. For example, if it is possible to approximately reconstruct the input from the representation, and the target function is not too sensitive to small changes in its inputs, then accuracy predicting the target will not be too much worse using the representation rather than the input [Van Rooyen and Williamson, 2015]. The field of information theory has described the trade-off between the compactness and fidelity of data encoding [Cover and Thomas, 2012]. For example, the Johnson-Lindenstrauss lemma shows that it is possible to compress a finite set of high dimensional points to a low dimensional representation while bounding the distortion in pointwise Euclidean distances [Johnson and Lindenstrauss, 1984]. Reduced computational complexity might be one reason to use a different representation, even if there is a cost in terms of accuracy. However, we are interested in whether representation learning can also be useful for *improving* the accuracy of supervised learning.

One approach is to compare the original hypothesis class H from which h is drawn, to the hypothesis class G from which g is drawn. The representation function f may be useful if for some $g \in G$, the function $g \circ f$ more accurately predicts the target variable than any hypothesis in H. In other words, f is considered to be useful insofar as it compensates for the defects of H, and a representation function class F is useful if it contains such an f. For example, in a feedforward neural network, the uppermost layer is linear; the representations learned in the lower layers apply non-linear transformations to the data which cannot be learned at the uppermost layer. With sufficiently many hidden units, feedforward neural networks with a single hidden layer are universal function approximators [Hornik, 1991], while linear models are not. A representation is not useful in isolation; it is useful as a module in a pipeline whose other modules have limitations. This thesis attempts to formalize this intuition, with a focus on unsupervised and transfer representation learning.

Statistical learning theory is a tool that can be used to prove when representation learning is useful. This approach describes conditions under which learning a task from a particular representation requires fewer labeled data points than learning the task from the original inputs. Past works have shown specific cases of this for representations learned from unlabeled data [Arora and Risteski, 2017] or several other tasks [Balcan et al., 2015]. However, strong assumptions on the data-generating distribution are often required: for example, in [Arora and Risteski, 2017] it is assumed that the task labels are generated by a linear separator over a finite representation space, and in [Balcan et al., 2015] that the task labels are generated by a linear separator over a low dimensional subspace of the input space. This thesis further develops the statistical learning theory approach to understanding the value of representations learned from unlabeled data and from other tasks.

1.4 Applications to Fairness

We know from the data processing inequality that a representation function can destroy but not create information. Are there cases where this is a feature, not a bug? In the context of fairness in machine learning, the answer is yes. An increasingly important research topic is the design of systems which make or provide advice on decisions about people – such as whether to grant someone a loan, or whether to release someone on bail – which accurately predict some target variable, but do not discriminate against particular groups. Group membership is described by the *sensitive variable* S, which along with X and Y can be used to construct the distribution μ_{XYS}. The task of learning a mapping from X to Y, without discriminating too much on the basis of S, is known as *fair supervised learning* [Madras et al., 2018].

One approach to ensuring that decisions do not discriminate against particular groups is changing the way that the data is represented to the system, removing information about S [Zemel et al., 2013]. This approach, known as *fair representation learning*, is not as simple as removing the column specifying group membership, since it may be possible to infer an individual's group membership based on the other columns – a phenomenon known as *redundant encoding* [Dwork et al., 2012]. Methods have been developed to solve this problem, such as using a neural network to learn a representation from which a separate adversary neural network cannot predict individuals' group membership [Edwards and Storkey, 2016]. We investigate applying these methods in two settings of interest: predicting domestic violence recidivism and predicting student outcomes at university.

We arrive at another paradox: if we want our decisions to be both accurate and fair, why not simply trade off these objectives in a joint optimization [Menon and Williamson, 2018] instead of using representation learning? Once again, representation learning narrows the hypothesis space and hence cannot offer us better performance than using the original data. However, we retain the benefit of modularization – with fairness taken care of at the representation learning stage, we are free to use whatever learner we like on the representation, knowing that the resulting decisions will not discriminate on the basis of group membership. This is particularly valuable from a governance and regulatory perspective. Supposing that the decision-maker is not trusted to be fair, the fairness of their decisions is nevertheless guaranteed by the representation that they access. Representation learning is a means of ensuring fairness by keeping the group membership of individuals private. This thesis provides a novel formalization of the costs and benefits of using fair representation learning relative to alternative approaches to fair supervised learning.

To apply representation learning to fairness, we need to define what it means to be fair. While this is a deep question, researchers have proposed several quantitative definitions which are useful in the context of fairness in machine learning [Mitchell and Shadlen, 2018]. We focus on those that consider differences in decisions made for different groups, known as *parity* metrics. Not only are there multiple definitions of fairness, but it is possible to show that some of them are in conflict [Kleinberg et al., 2017b]. We investigate this issue and contribute new results about such conflicts, focusing on the relationship between two common parity metrics: *equalized odds* and *equalized outcomes*.

1.5 Contributions of this Thesis

The major contributions of the thesis are to:

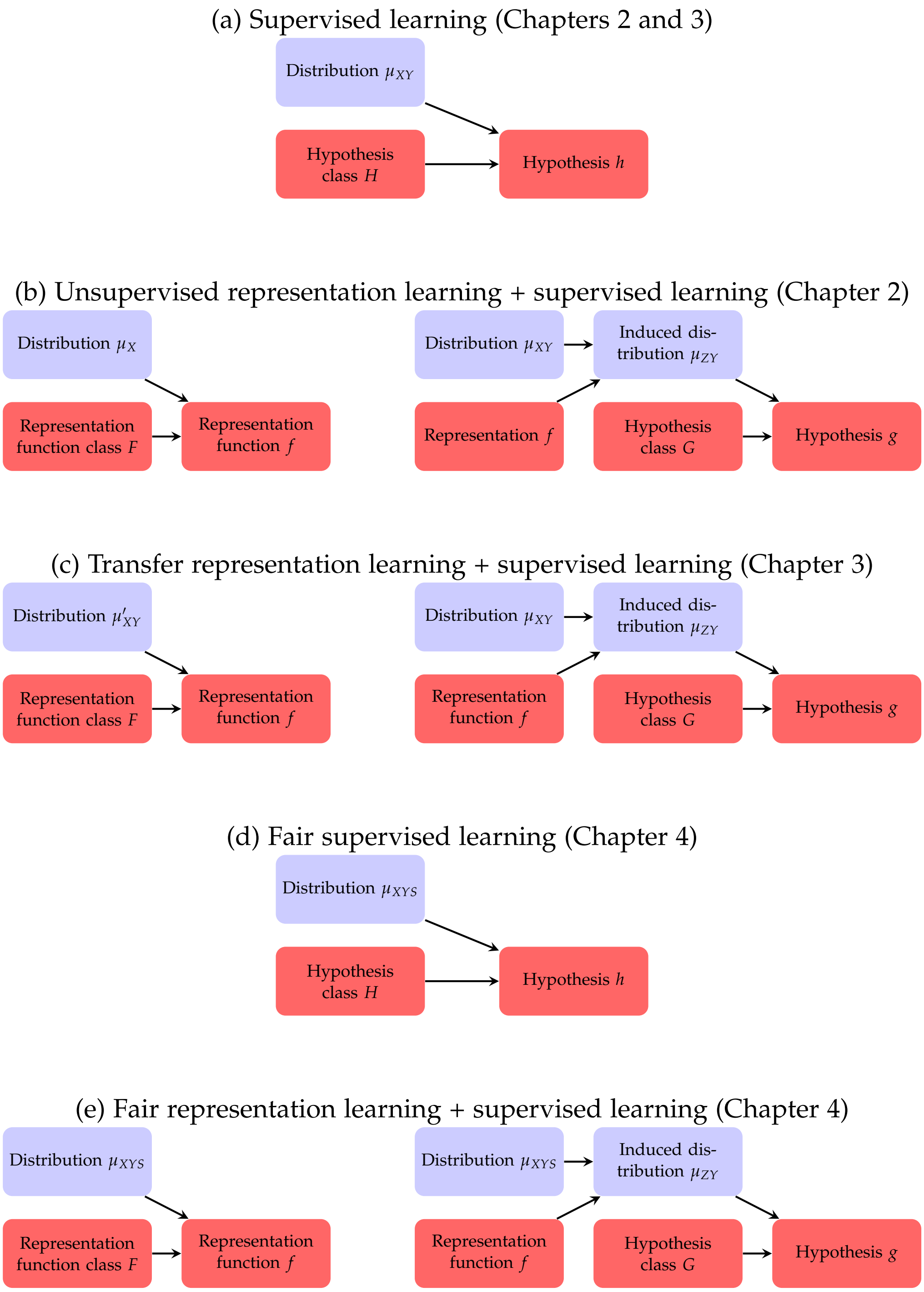
- identify conditions under which unsupervised representation learning is provably useful (Part I, Chapter 2)
- identify conditions under which transfer representation learning is provably useful (Part I, Chapter 3)
- formalize the problem of fair representation learning and quantify the costs and benefits of using a given representation (Part II, Chapter 4)
- quantify the relationship between two common notions of fairness, equalized odds and equalized outcomes (Part II, Chapter 5)
- demonstrate the use and performance of fair representation learning in the contexts of predicting recidivism (Part III, Chapter 6) and student outcomes at university (Part III, Chapter 7).

Another way to understand the contributions of the thesis is through the several problem settings which we consider, as summarized in Figure 1.2. We compare supervised learning (a) to unsupervised representation learning (b) and transfer representation learning (c) (see Part I). We compare fair supervised learning (d) – which adds the sensitive variable S encoding group membership to the standard supervised learning problem – to fair representation learning (e) (see Part II). We also explore case studies of fair representation learning (e) (see Part III).

1.6 Structure of this Thesis

This thesis is structured in three parts. Part I presents a theoretical approach to understanding when learning and re-using representations is useful. Part II considers the application of representation learning to fairness and the challenges of conflicting definitions of fairness. Part III introduces two case studies – one about predicting recidivism in a criminal justice context and the other about predicting student outcomes at university – which incorporate fair representation learning, and are of interest in their own right.

In Part I, we first provide a theoretical analysis of when unsupervised representation learning is useful in Chapter 2. We describe an approach to identifying sufficient conditions for unsupervised representation learning to provide a benefit, and give an example of these conditions where cluster structure is present in the



(a) Supervised learning (Chapters 2 and 3)

(b) Unsupervised representation learning + supervised learning (Chapter 2)

(c) Transfer representation learning + supervised learning (Chapter 3)

Distribution μ'_{XY}
Representation function class F
Representation function f
Distribution μ_{XY}
Induced distribution μ_{ZY}
Representation function f
Hypothesis class G
Hypothesis g

(d) Fair supervised learning (Chapter 4)

(e) Fair representation learning + supervised learning (Chapter 4)

Distribution μ_{XYS}
Representation function class F
Representation function f
Distribution μ_{XYS}
Induced distribution μ_{ZY}
Representation function f
Hypothesis class G
Hypothesis g

Figure 1.2: Summary of problem settings considered in this thesis and the chapters where they are introduced. The diagram includes functions and function classes (red), and probability distributions (blue). The notation is described in the text.

data. We subsequently provide a theoretical analysis of when transferring representations from a source to a target task is useful in Chapter 3. Given a fixed number of target task samples, under certain conditions the target task risk can be more tightly upper bounded using a transferred representation function compared to learning the target task from scratch. We consider the case where the transferred representation function is fixed, and the case where it is fine-tuned on the target task. We show examples of our risk bounds using feedforward neural networks. The theorems on unsupervised and transfer representation learning we present in Chapters 2 and 3 are all novel results.

In Part II, we first provide an analysis of the costs and benefits of fair representation learning in Chapter 4. We show that fair representation learning incurs a cost compared to optimally trading off fairness and accuracy. We quantify the benefit of fair representation learning by showing that any subsequent use of a particular representation will not be too unfair. We also show that a novel regulatory model with desirable characteristics is made possible by this approach. In Chapter 5, we describe the limits imposed on fair representation learning – or indeed on any algorithm – by competing definitions of fairness. We examine the relationship between equalized odds – where the true positive rates and false positive rates of an algorithm are the same across groups – and equalized outcomes – where the difference in predicted outcomes between groups is less than the difference observed in the training data. We show that under realistic assumptions, equalized odds implies partially equalized outcomes.

In Part III we present two case studies. In Chapter 6 we analyze recidivism prediction in the criminal justice system, and implications for fairness across racial groups. In Chapter 7 we examine predicting student outcomes at university, and consequences for the fair provision of academic support across genders. While we analyze specific aspects of the two problem settings in some detail, we also explore the common need to incorporate fairness into algorithm design and examine the suitability of representation learning for this purpose. In both cases we find there is a trade-off between maximizing absolute utility and equalizing the relative utility of different groups.

1.7 Research Outputs Produced during PhD Candidature

During the course of the PhD, a range of outputs have been produced, as summarized in Table 1.1. Collaborators are shown for these outputs where applicable, and the corresponding chapters include some work contributed by these collaborators. However, Daniel McNamara was the lead researcher and first author for all work included in this thesis. He has also delivered number of research talks based on the work in this thesis, as summarized in Table 1.2.

Table 1.1: Accepted peer-reviewed papers based on work in this thesis.

Output	Related Chapter
Papers in Conference Proceedings	
"Risk Bounds for Transferring Representations With and Without Fine-Tuning". Daniel McNamara and Maria-Florina Balcan. *International Conference on Machine Learning*, 2017.	3
"Costs and Benefits of Fair Representation Learning". Daniel McNamara, Cheng Soon Ong and Robert C. Williamson. *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*, 2019.	4
"Equalized Odds Implies Partially Equalized Outcomes Under Realistic Assumptions". Daniel McNamara. *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*, 2019.	5
Book Chapters	
"Trade-offs in Algorithmic Risk Assessment: an Australian Domestic Violence Case Study". Daniel McNamara, Timothy Graham, Ellen Broad and Cheng Soon Ong. *Theory on Demand #29: Good Data*, pp. 96-116, 2019.	6
Workshop Papers	
"Risk Bounds for Transferring Representations With and Without Fine-Tuning". Daniel McNamara and Maria-Florina Balcan. Principled Approaches to Deep Learning Workshop at *International Conference on Machine Learning*, 2017.	3
"Performance Guarantees for Transferring Representations". Daniel McNamara and Maria-Florina Balcan. Workshop Track at *International Conference on Learning Representations*, 2017.	3

Table 1.2: Other research outputs based on work in this thesis.

Output	**Related Chapter**
Technical Papers	
“A Modular Theory of Feature Learning”. Daniel McNamara, Cheng Soon Ong and Robert C. Williamson. *arXiv:1611.03125*, 2016.	2
“Using Cohort Analysis and Predictive Modelling to Inform Targeted Student Support”. Daniel McNamara, Robert C. Williamson and Leone Nurbasari. *Australasian Association for Institutional Research Forum*, 2018.	7
Posters at Student Symposiums	
“Learning Features to Provably Improve Task Performance”. *Australian Joint Conference on Artificial Intelligence Student Symposium*, 2015.	2
“Algorithmic Stereotypes: Implications for Fairness of Generalizing from Past Data”. *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society Student Track*, 2019.	5
“Learning Features to Improve the Performance of Machine Learning Algorithms”. *Fulbright Scholar Presentation*, 2016.	2
Talks	
“Risk Bounds for Transferring Representations With and Without Fine-Tuning”. Representation Learning Workshop, Simons Institute for the Theory of Computing, University of California Berkeley, 2017.	3
“Carnegie Mellon University Research Visit Summary". Carnegie Mellon University, 2017.	3
“Performance Guarantees for Transferring Representations”. Microsoft, 2017.	3
“Risk Bounds for Transferring Representations With and Without Fine-Tuning”. Canon Information Systems Research Australia, 2017.	3
“Risk Bounds for Transferring Representations With and Without Fine-Tuning”. CSIRO Data61, 2017.	3
“Teaching Machines to Play Fair”. Australian National University Machine Learning Retreat, 2018.	4
“Teaching Machines to Play Fair”. Australian Fulbright Alumni Association Conference, 2017.	4
“Teaching Machines to Play Fair”. School of Regulation and Global Governance, Australian National University, 2017.	4
“Algorithmic Stereotypes”. Ethical Algorithms Symposium, The University of Sydney, 2019 (forthcoming).	5
“Algorithmic Stereotypes”. AI, ML and Friends, Australian National University, 2018.	5
“Using Cohort Analysis and Predictive Modelling to Inform Targeted Student Support”. Academic Quality Assurance Committee, Australian National University, 2018.	7
“Using Cohort Analysis and Predictive Modelling to Inform Targeted Student Support”. Planning and Performance Measurement Division, Australian National University, 2018.	7

Part I

Learning and Re-Using Representations

Chapter 2

Unsupervised Representation Learning to Provably Improve Task Performance

2.1 Introduction

Representation learning techniques using unlabeled data are common in the field of machine learning [LeCun et al., 2015]. For example, they have been used to achieve empirical advances in areas such as computer vision [Hinton and Salakhutdinov, 2006] and natural language processing [Mikolov et al., 2013]. However, there are few theoretical results concerning when such techniques offer a benefit relative to standard supervised learning.

This chapter describes an approach to identifying sufficient conditions under which unsupervised representation learning provably improves task performance. This approach 'factorizes' the problem into finding separate conditions which apply to the unlabeled distribution, the labeled distribution, the representation function class, and the hypothesis class used for prediction. We provide an example where it is possible to determine whether these conditions are met, using an unlabeled data sample, analysis of the proposed representation function class and hypothesis class, and suitable assumptions about shared structure between the unlabeled and labeled distributions.

The novelty of this work is its generality beyond any single representation learning technique and its theoretical rather than empirical approach. Furthermore, we demonstrate the importance of considering the subsequent task for which the representation will be used, including the hypothesis class and loss function, in the definition of what makes a 'good' representation. We show that unsupervised representation learning is useful when it induces an hypothesis class containing an hypothesis with lower risk than any hypothesis in the original hypothesis class.

There are two other important features of the work worth calling out. First, we analyze a processing *pipeline*, not just a single step. The use of sequential pipelines is common in practice, but rarely addressed theoretically. Our approach is novel in this regard. Second, we analyze the problem via the *risk gap* between using unsupervised

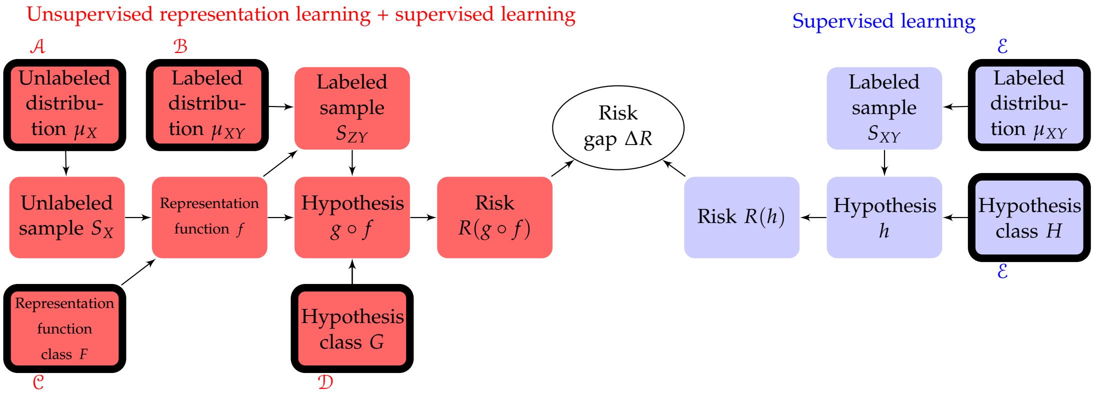


Figure 2.1: Measuring the effect of unsupervised representation learning (see Section 2.3 for details). The red path (left) shows unsupervised representation learning + supervised learning, the blue path (right) shows supervised learning, and the risk gap measures the difference between the risks of the two paths. The arrows indicate dependencies. Source nodes are shown with a black border and are annotated with corresponding conditions from Table 2.1.

representation learning and directly solving the problem with supervised learning. This is in contrast to the common approach in statistical learning theory of comparing the risk upper bounds or sample complexity of learners (e.g. [Balcan and Blum, 2010]). Our approach is illustrated in Figure 2.1.

The remainder of the chapter is structured as follows. In Section 2.2 we briefly review existing applied and theoretical approaches to unsupervised representation learning. In Section 2.3 we mathematically formalize the task of unsupervised representation learning, develop objectives describing what it means for unsupervised representation learning to be successful, and present an approach to verifying when these objectives are achievable. In Section 2.4, we instantiate this approach through an example involving exploiting cluster structure in unlabeled data to solve a supervised learning problem.

2.2 Background

Many representation learning techniques have been developed, including those using unlabeled data. Collectively they have achieved considerable empirical success [Bengio et al., 2013], but provide few theoretical guarantees concerning their effect on task performance [Sutskever et al., 2015]. While the details of these techniques are not important for the current work, we give a few examples to motivate our analysis, and describe some previous theoretical approaches to unsupervised representation learning.

Low dimensional manifold embeddings are a popular class of unsupervised representation learning techniques, motivated by the desire for computational efficiency.

Principal components analysis (PCA) – where the data is projected onto the lower dimensional space which minimizes reconstruction error – is a well-known technique of this kind. Variants which involve optimizing the preservation of pairwise distances between points include Isomap, Laplacian eigenmaps and local linear embedding [Mohri et al., 2012]. Clustering can also be seen as a manifold-based approach to unsupervised representation learning. Theoretical results concerning manifold learning have typically focused on what kinds of embeddings are possible while preserving information in the original data. For example, it is possible to compress a finite set of high dimensional points to a low dimensional representation while bounding the distortion in pointwise Euclidean distances [Johnson and Lindenstrauss, 1984]. However, in general there are no guarantees that using manifold embeddings will improve the performance of a subsequent learner.

Empirical results in the field of deep learning have shown the power of learning multiple levels of representations. Unsupervised pre-training techniques such as the autoencoder – where unlabeled data is used to learn weights in a neural network, which are then re-used for a supervised learning task – were important in initial deep learning advances [Hinton and Salakhutdinov, 2006]. The effect of unsupervised pre-training has been studied empirically [Erhan et al., 2010], with benefits observed in terms of both reduced training set error and improved generalization. While attempts have been made to theorize unsupervised representation learning in studies such as Saxe et al. [2014] — which concluded that a certain kind of random initialization could achieve the same condition as unsupervised pre-training — mostly experimental results have outpaced theory. Such techniques often learn representations in a higher dimension than the original inputs, which make the data more linearly separable.

Despite these advances, theoretical results about the value of representations have tended to be pessimistic. It is straightforward to show that a representation function can never decrease the risk of the optimal classifier. This is because for any given representation function, the set of hypotheses composed with the representation function is a subset of all possible hypotheses mapping inputs to targets. A result in a similar spirit is the data processing inequality (Theorem 2.8.1 of [Cover and Thomas, 2012]): given random variables X, Y and Z, if Z is conditionally independent of Y given X (i.e. $Y \to X \to Z$ is a Markov chain), then $I(Z, Y) \leq I(X, Y)$, where I is mutual information. In particular, if representation variable Z is constructed from input variable X, then this conditional independence property is satisfied with respect to target variable Y. This result formalizes the intuition that changing the way input data is represented cannot increase the amount of information about the target variable it contains.

A recent work, published after the submission of this thesis, proved that unsupervised representation learning can be useful in a particular problem setting [Arora et al., 2019]. This work considers ‘contrastive’ unsupervised representation learning, which involves drawing related pairs of unlabeled samples – such as co-occurring pairs of words in word2vec – and learning a representation function such that the representations of related samples are similar, while the representations of unrelated samples are dissimilar. This work showed that if the related pairs are from a common latent class among several possible latent classes, then the performance of a representation function on the unsupervised task is indicative of the performance of a linear classifier applied to the induced representation on a supervised classification task over the same set of latent classes. The result relies on a detailed set of assumptions about the relationship between the unsupervised task used for representation learning, and the supervised task on which the representation function will be re-used.

2.3 When Unsupervised Representation Learning is Provably Useful

Our goal is to determine under what conditions unsupervised representation learning enhances the performance of a subsequent supervised learner. This objective is pertinent to a range of common machine learning scenarios. Do the features learned by an autoencoder enhance the performance of a linear classifier compared to using the original inputs? Does a particular kernel function outperform a linear kernel when used with an hypothesis class of linear separators (recalling that kernel functions implicitly specify a representation space)? Do vector representations of words outperform one-hot unigram representations for natural language processing tasks?

As a step towards our goal, we first formalize the problem setting. We provide a comparison to a previous formalization of semi-supervised learning. We then state the objectives of unsupervised representation learning, and a high-level approach to determining the conditions under which these objectives are met.

2.3.1 Problem Setting

Let $\mathcal{X}$ and $\mathcal{Y}$ be input and target spaces respectively. Let X and Y be input and target random variables respectively. Let μ_{XY} be a probability distribution over $\mathcal{X} \times \mathcal{Y}$ and μ_X be its marginal distribution over $\mathcal{X}$. Let $p(\cdot)$ refer to the probability that a point drawn from μ_{XY} satisfies some condition, and let $p(x) := p(X = x)$. Let S_{XY} be a sample drawn from μ_{XY} and let S_X be a sample drawn from μ_X. Let H be an hypothesis class whose elements are of type $h : \mathcal{X} \to \mathcal{Y}$.

Let $\mathcal{Z}$ be the representation space. Let F be a representation function class whose elements are of type $f : \mathcal{X} \to \mathcal{Z}$. In unsupervised representation learning, we use F and S_X and learn some $f \in F$. Let the representation variable $Z := f(X)$. Let G be an hypothesis class whose elements are of type $g : \mathcal{Z} \to \mathcal{Y}$.

Let $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ be a loss function. Let the risk of some $h \in H$ be

$$R(h) := \mathbb{E}_{X,Y \sim \mu_{XY}}[l(h(X), Y)].$$

Similarly, let the risk of hypothesis $g \in G$ using the representation function $f \in F$ be

$$R(g \circ f) := \mathbb{E}_{X,Y \sim \mu_{XY}}[l(g(f(X)), Y)].$$

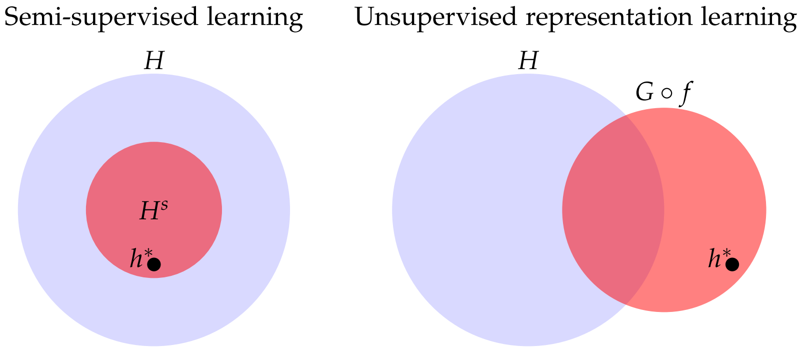


Figure 2.2: Relationship between semi-supervised learning proposed as formalized in [Balcan and Blum, 2010] (left) and unsupervised representation learning (right). In semi-supervised learning, unlabeled data is used to prune the hypothesis class H to the subset $H^s \subset H$. In unsupervised representation learning, unlabeled data is used to discover f and the hypothesis class changes to $G \circ f$. In both cases we hope that the target function h^* is within our new hypothesis class.

We are interested in computing and comparing these two quantities. Of particular interest are the cases where G and H are the same, or are related to each other through a straightforward change of type signature. For example, G and H are both the classes of linear separators for their respective input types.

2.3.2 Comparison to Semi-Supervised Learning

Semi-supervised learning involves using unlabeled data to help supplement scarce labeled data in a supervised learning problem. By this broad definition, unsupervised representation learning can be considered a kind of semi-supervised learning. However, previous formal analysis of semi-supervised learning from the perspective of statistical learning theory proposed by Balcan and Blum [2010] has used a somewhat narrower definition of semi-supervised learning. In this case, unsupervised representation learning as we describe it is conceptually different. The differences between the two approaches are shown in Figure 2.2.

In semi-supervised learning, we aim to prune H to some $H^s \subset H$ using only unlabeled data. Supposing that there is some target function $h^* \in H$ we are trying to learn, we also wish to ensure that $h^* \in H^s$. The smaller hypothesis class may make learning a subsequent supervised task more straightforward and enable a tighter generalization error bound [Balcan and Blum, 2010]. However, if the target function lies outside the original hypothesis class, semi-supervised learning will not help to discover it.

In unsupervised representation learning, the hypothesis class changes and hence it is possible to learn hypotheses not included in the original hypothesis class. We learn some $f \in F$ from unlabeled data, and then consider all hypotheses of the form $g \circ f$ for some $g \in G$, where $\circ$ denotes function composition. Our new hypothesis class is denoted by $G \circ f$. In this case, we hope that $h^* \in G \circ f$. This is particularly useful when $h^* \notin H$.

2.3.3 Objectives of Unsupervised Representation Learning

We introduce a set of objectives of unsupervised representation learning. While these objectives are intuitive, our formalization is novel. They help us to precisely answer the question: when is unsupervised representation learning guaranteed to be useful? Once we have described these objectives, we describe conditions under which they can be achieved.

We would like to show that using a representation function learned from an unlabeled sample guarantees that the risk of some hypothesis in our hypothesis class using this representation is not too large, as shown in Objective 2.1.

Objective 2.1 (Risk upper bound for unsupervised representation learning + supervised learning)**.** *Fix μ_{XY}, F, G and l. Draw an unlabeled sample S_X from μ_X. Find some $f \in F$ and $\epsilon^f_{\max}$ depending upon S_X, which with probability at least $1 - \delta$ over samples S_X satisfies*

$$\min_{g \in G} R(g \circ f) \leq \epsilon^f_{\max}. \tag{2.1}$$

We would also like to guarantee that we cannot achieve a small risk using the original hypothesis class, as shown in Objective 2.2. We may not want to bother with unsupervised representation learning if the task is solvable using H.[1]

Objective 2.2 (Risk lower bound for supervised learning)**.** *Fix μ_{XY}, H and l. Draw a labeled sample S_{XY} from μ_{XY}. Find some $\epsilon_{\min}$ depending upon S_{XY}, which with probability at least $1 - \delta$ over samples S_{XY} satisfies*

$$\min_{h \in H} R(h) \geq \epsilon_{\min}. \tag{2.2}$$

Finally, we would like to show that we can achieve smaller risk using unsupervised representation learning compared to the original hypothesis class. In formalizing this objective, it is useful to first define a *risk gap*.

Definition 2.3 (Risk gap)**.** *Fix μ_{XY}, F, G, H and l. Let the risk gap of some representation function f be*

$$\Delta R(f) := \min_{h \in H} R(h) - \min_{g \in G} R(g \circ f).$$

We now formalize what it means for unsupervised representation learning to be useful in Objective 2.4.

Objective 2.4 (Positive risk gap)**.** *Fix μ_{XY}, F, G, H and l. Draw an unlabeled sample S_X from μ_X and draw a labeled sample S_{XY} from μ_{XY}. Find some $f \in F$ and $\epsilon^f_{\max}$ depending upon S_X and some $\epsilon_{\min}$ depending upon S_{XY}, which with probability at least $1 - 2\delta$ over pairs of samples S_X and S_{XY} satisfies*

$$\Delta R(f) \geq \epsilon_{\min} - \epsilon^f_{\max} > 0. \tag{2.3}$$

[1] It is possible that working with H is disadvantageous for computational and/or sample complexity reasons. However, in this chapter we focus on comparing the minimum risk achievable using $G \circ f$ versus using H.

2.3.4 Approach to Verifying that the Objectives are Achieved

Our approach to understanding unsupervised representation learning is to search for combinations of F, G, H, μ_{XY} and l for which we can verify that Objectives 2.1, 2.2 and 2.4 are met.

Objective 2.1 can be verified empirically by drawing an unlabeled sample, using it to find some $f \in F$, then drawing a labeled sample, using it to find some $g \in G$, estimating $R(g \circ f)$ and hence upper bounding $\min_{g \in G} R(g \circ f)$. However, this does not give us any insight into *why* risk can be upper bounded for unsupervised representation learning.

We propose an alternative approach, which involves identifying and verifying four conditions that are sufficient for Objective 2.1 to hold. We write these conditions as $\mathcal{A}(\mu_X)$, $\mathcal{B}(\mu_{XY})$, $\mathcal{C}(F)$ and $\mathcal{D}(G)$, since they depend on the unlabeled distribution μ_X, the labeled distribution μ_{XY}, the representation function class F, and the hypothesis class G respectively (see Section 2.4 for an example). We would like to be able to verify that the conditions hold with high probability using a sample of μ_X, assumptions about the relationship between μ_X and μ_{XY}, and analysis of F and G. In this way we 'factorize' the problem of when the risk of unsupervised representation learning is upper bounded into several modular components. This provides more insight on the problem and helps us to determine whether to deploy unsupervised representation learning in problem settings where we can verify whether these conditions hold.

Objective 2.2 can be verified empirically by drawing a labeled sample, running empirical minimization over H on the sample, and using a generalization error bound to lower bound $\min_{h \in H} R(h)$ with high probability. We introduce this as a separate condition $\mathcal{E}(\mu_{XY}, H)$, noting that in this case it does not appear possible to 'factorize' the problem any further. We observe that if we satisfy Objectives 2.1 and 2.2 and $\epsilon^f_{\max} < \epsilon_{\min}$, by the union bound we also satisfy Objective 2.4.

Our approach to identifying conditions which verify Objectives 2.1 and 2.2 is summarized in Table 2.1. We provide an informal description of each condition, an example (developed further in Section 2.4) and our suggested approach to verifying whether the condition holds. We motivate our conditions by noting that each independent aspect of the prediction context (the source nodes in Figure 2.1) is associated with a condition. Thus, although we do not formally demonstrate that the conditions are *necessary* for unsupervised representation learning to be useful, it appears unlikely that the conditions can be reduced further. $\mathcal{A}(\mu_X)$, $\mathcal{B}(\mu_{XY})$, $\mathcal{C}(F)$ and $\mathcal{D}(G)$ and $\mathcal{E}(\mu_{XY}, H)$ can be viewed as "meta-conditions", since their precise definition depends on the particular problem setting. It appears unlikely that there exist more specific, problem-independent conditions for successful unsupervised representation learning given the problem-dependence of Objectives 2.1, 2.2 and 2.4.

Table 2.1: Approach to identifying sufficient conditions for the achievement of Objectives 2.1 and 2.2.

Condition	**Description**	**Example**	**Verification**
Objective 2.1: Risk upper bound for unsupervised representation learning + supervised learning			
$\mathcal{A}(\mu_X)$	Marginal distribution has some structure	Data lies in clusters (Definition 2.5)	Analysis of unlabeled sample
$\mathcal{B}(\mu_{XY})$	Joint distribution shares marginal distribution structure	Points within clusters share labels (Definition 2.6)	Assumption (in principle could be verified on a labeled sample)
$\mathcal{C}(F)$	Learned representation exploits marginal distribution structure	Cluster regions are mapped to distinct points by some representation function (Definition 2.7)	Analysis of F
$\mathcal{D}(G)$	Hypothesis class can exploit learned representation	All possible labelings of distinct points are separable by the hypothesis class (Definition 2.8)	Analysis of G
Objective 2.2: Risk lower bound for supervised learning			
$\mathcal{E}(\mu_{XY}, H)$	Hypothesis class performs poorly on original inputs	Inputs not separable by the hypothesis class (Definition 2.11)	Analysis of performance of H on a labeled sample

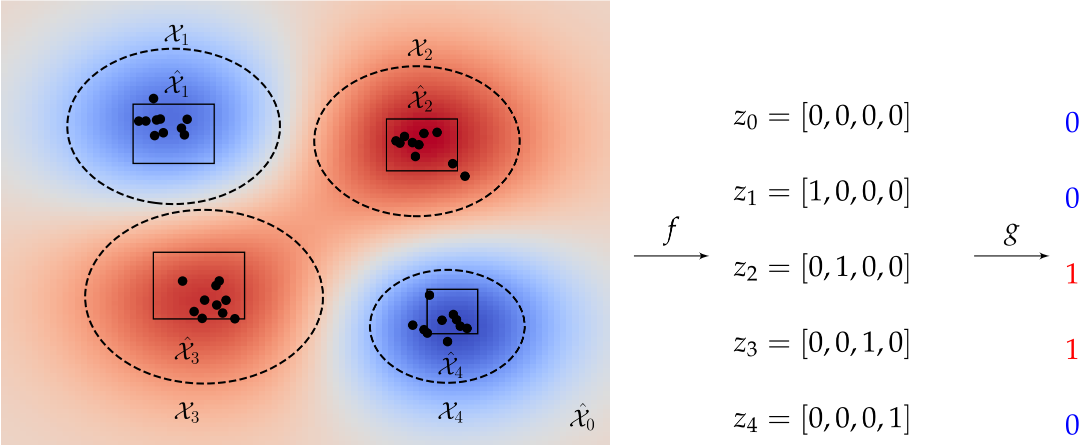


Figure 2.3: Cluster example considered in Section 2.4. For visualization purposes we set $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{Z} = \mathbb{R}^4$. The conditional probabilities $p(Y = 1|X = x)$ (left – higher values are red and lower values are blue) obey the cluster assumption, i.e. regions where the unlabeled density $p(X = x)$ is higher ($\mathcal{X}_i$, boundaries shown in dashed lines) tend to have common labels. Points from the unlabeled sample S_X are shown as black dots, which are used to construct approximate cluster regions ($\hat{\mathcal{X}}_i$, boundaries shown in solid lines) from the set of axis-aligned rectangles. The representation function class F maps each $\hat{\mathcal{X}}_i \subset \mathcal{X}$ to a fixed point $z_i \in \mathcal{Z}$. The hypothesis classes G and H are the sets of linear separators over $\mathcal{X}$ and $\mathcal{Z}$ respectively. We can find some $f \in F$ from the unlabeled sample, for which there exists some $g \in G$ such that $g \circ f$ approximately separates positive and negative labeled points. Conversely, all linear separators $h \in H$ perform poorly.

2.4 Cluster Example

We consider an example where we use cluster structure present in the data to solve a classification problem that is not linearly separable. The example is inspired by formalizations of the cluster assumption proposed in past works Rigollet [2007]; Singh et al. [2009], although our work is novel because it analyzes the relationship between the cluster assumption and unsupervised representation learning. We define conditions $\mathcal{A}(\mu_X)$, $\mathcal{B}(\mu_{XY})$, $\mathcal{C}(F)$ and $\mathcal{D}(G)$ and show that together they imply that Objective 2.1 is achieved. We also show that these conditions can be established with high probability using a sample of μ_X, assumptions about μ_{XY}, and analysis of F and G. Furthermore, we define a condition $\mathcal{E}(\mu_{XY}, H)$and verify it using labeled data to show that Objective 2.2 is achieved.

The example assumes the unlabeled distribution is concentrated in clusters, which share structure with the labeled distribution. We assume $\mathcal{X}$ is a Euclidean space, $\mathcal{Y} \in \{0, 1\}$ and l is the $0/1$ loss:

$$l(\hat{y}, y) := \begin{cases} 0 & \text{if } y = \hat{y} \\ 1 & \text{if } y \neq \hat{y}. \end{cases}$$

We give a visualization of the example in Figure 2.3.

2.4.1 Defining the conditions $\mathcal{A}(\mu_X)$, $\mathcal{B}(\mu_{XY})$, $\mathcal{C}(F)$ and $\mathcal{D}(G)$

We provide definitions of each of $\mathcal{A}(\mu_X)$, $\mathcal{B}(\mu_{XY})$, $\mathcal{C}(F)$ and $\mathcal{D}(G)$. This is a first step towards understanding their relationship to successful unsupervised representation learning.

Condition $\mathcal{A}(\mu_X)$ requires that almost all of the probability mass of μ_X lies in $\hat{k}$ non-overlapping high density regions, i.e. informally this means that the data lies in clusters. We will show in Section 2.4.3 how to verify this using an unlabeled sample S_X.

Definition 2.5 (Condition $\mathcal{A}(\mu_X)$: data lies in clusters). *Let $\hat{k}$ be a positive integer and let $\epsilon_{\mathcal{A}}$ and τ be non-negative constants. Suppose $\exists \hat{\mathcal{X}}^s \subseteq \mathcal{X}$ such that*

$$\hat{\mathcal{X}}^s = \hat{\mathcal{X}}_1 \cup \cdots \cup \hat{\mathcal{X}}_{\hat{k}} \tag{2.4}$$

where $\forall i \in \{1, \ldots, \hat{k}\}, \hat{\mathcal{X}}_i$ is a connected set, $\hat{\mathcal{X}}_i \cap \hat{\mathcal{X}}_j = \varnothing$ if $i \neq j$ and $\hat{\mathcal{X}}_0 := \mathcal{X} \setminus \hat{\mathcal{X}}^s$,

$$\forall x \in \hat{\mathcal{X}}^s, p(x) \geq \tau \tag{2.5}$$

and

$$\int\limits_{x \in \hat{\mathcal{X}}_0} p(x) dx \leq \epsilon_{\mathcal{A}}. \tag{2.6}$$

Condition $\mathcal{B}(\mu_{XY})$ requires that there is some low-risk hypothesis whose predictions are constant within regions of high density, or in other words that points within clusters tend to share labels. This is an example of shared structure between the marginal and joint distributions. While in principle we may verify this condition using labeled data, in practice it is likely to be treated as an assumption; since verification appears as difficult as solving the supervised learning problem, if this were possible unsupervised representation learning may not be useful.

Definition 2.6 (Condition $\mathcal{B}(\mu_{XY})$: points within clusters share labels). *Let k be a positive integer, let $\epsilon_{\mathcal{B}}$ a non-negative constant and let τ be the non-negative constant from Definition 2.5. Suppose $\exists \mathcal{X}^s \subseteq \mathcal{X}$, $h^* : \mathcal{X} \to \mathcal{Y}$ such that*

$$\mathcal{X}^s = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_k \tag{2.7}$$

where $\min\limits_{x_i \in \mathcal{X}_i, x_j \in \mathcal{X}_j} \|x_i - x_j\|_2 > 0$ if $i \neq j$ and $\mathcal{X}_0 := \mathcal{X} \setminus \mathcal{X}^s$,

$$\forall x \in \mathcal{X}_0, p(x) < \tau, \tag{2.8}$$

$$\forall i \in \{1,\ldots,k\}, \forall x, x' \in \mathcal{X}_i, h^*(x) = h^*(x') \tag{2.9}$$

and

$$R(h^*) \leq \epsilon_{\mathcal{B}}. \tag{2.10}$$

Condition $\mathcal{C}(F)$ requires that there is some representation function $f \in F$ which maps each region $\hat{\mathcal{X}}_i$ to a particular distinct point. We will show in Section 2.4.4 an example of where it is possible to verify that this condition holds by inspecting F.

Definition 2.7 (Condition $\mathcal{C}(F)$: cluster regions mapped to distinct points by some representation function)**.** *Let $\hat{k}$ be the positive integer and $\hat{\mathcal{X}}_i$ be the regions from Definition 2.5. Let $z_0, \ldots, z_{\hat{k}} \in \mathcal{Z}$, where $z_i \neq z_j$ if $i \neq j$. Suppose there is some $f \in F$ such that*

$$\forall i \in \{0,\ldots,\hat{k}\}, f(x) = z_i, \textit{ if } x \in \hat{\mathcal{X}}_i. \tag{2.11}$$

Condition $\mathcal{D}(G)$ requires that there is some hypothesis $g \in G$ which is capable of labeling a particular set of $\hat{k}+1$ points. We will show in Section 2.4.4 an example of where it is possible to verify that this condition holds by inspecting G.

Definition 2.8 (Condition $\mathcal{D}(G)$: all possible labelings of distinct points are separable by the hypothesis class)**.** *Let $\hat{k}$ be the positive integer from Definition 2.5. Let $z_0, \ldots, z_{\hat{k}}$ be the set of points from Definition 2.7. Let $h : z_0 \cup \cdots \cup z_{\hat{k}} \to \{0,1\}$ be a labeling function. Suppose that for any labeling function h, $\exists g \in G$ such that*

$$\forall i \in 0,\ldots,\hat{k}, g(z_i) = h(z_i). \tag{2.12}$$

2.4.2 Using Conditions $\mathcal{A}(\mu_X)$, $\mathcal{B}(\mu_{XY})$, $\mathcal{C}(F)$ and $\mathcal{D}(G)$ to Verify that Objective 2.1 is Satisfied

We wish to verify that Objective 2.1 is satisfied. In Theorem 2.9, we show that this is possible in the case where $\mathcal{A}(\mu_X)$, $\mathcal{B}(\mu_{XY})$, $\mathcal{C}(F)$ and $\mathcal{D}(G)$ hold. The result shows that we can find conditions under which we may upper bound the risk of using unsupervised representation learning.

Theorem 2.9 (Objective 2.1 is satisfied assuming conditions hold)**.** *Suppose conditions $\mathcal{A}(\mu_X)$, $\mathcal{B}(\mu_{XY})$, $\mathcal{C}(F)$ and $\mathcal{D}(G)$ hold. Let $\epsilon_{\mathcal{A}}$ be the constant from Definition 2.5, $\epsilon_{\mathcal{B}}$ be the constant from Definition 2.6 and f be the representation function from Definition 2.7. Then*

$$\min_{g' \in G} R(g' \circ f) \leq \epsilon_{\mathcal{A}} + \epsilon_{\mathcal{B}}.$$

Proof. First we show that

$$\forall i \in \{1,\ldots,\hat{k}\}, \exists j \in \{1,\ldots,k\} \text{ such that } \hat{\mathcal{X}}_i \subseteq \mathcal{X}_j. \tag{2.13}$$

Recall that $\forall i \in \{1,\ldots,\hat{k}\}$, $\forall x \in \hat{\mathcal{X}}^i, p(x) \geq \tau$ by (2.5) and that $\forall x \in \mathcal{X}_0, p(x) < \tau$ by (2.8). Therefore $\forall i \in \{1,\ldots,\hat{k}\}$, $\hat{\mathcal{X}}_i$ cannot intersect $\mathcal{X}_0$.

Also recall that $\forall i \in \{1, \ldots, \hat{k}\}$, $\hat{\mathcal{X}}_i$ is a connected set by (2.4). Furthermore, recall that $\min_{x_i \in \mathcal{X}_i, x_j \in \mathcal{X}_j} \|x_i - x_j\|_2 > 0$ if $i \neq j$ by (2.7), which implies that any connected set intersecting both $\mathcal{X}_i$ and $\mathcal{X}_j$ must contain a point in $\mathcal{X}_0$.

Since $\forall i \in \{1, \ldots, \hat{k}\}$, $\hat{\mathcal{X}}_i$ cannot intersect $\mathcal{X}_0$ nor intersect both $\mathcal{X}_i$ and $\mathcal{X}_j$, we conclude that $\exists j \in \{1, \ldots, k\}$ such that $\hat{\mathcal{X}}_i \subseteq \mathcal{X}_j$.

Combining (2.9) and (2.13), we have $\forall i \in \{1, \ldots, \hat{k}\}$, h^* is constant on $\hat{\mathcal{X}}_i$. Let

$$h(x) := \begin{cases} h^*(x) & \text{if } x \in \hat{\mathcal{X}}^s \\ 0 & \text{if } x \in \hat{\mathcal{X}}_0. \end{cases} \tag{2.14}$$

Therefore $\forall i \in \{0, \ldots, \hat{k}\}$, h is constant on $\hat{\mathcal{X}}_i$. From (2.11), $\forall i \in \{0, \ldots, \hat{k}\}$, $\hat{\mathcal{X}}_i$ is mapped by f to z_i. By (2.12), $\exists g \in G$ matching any labeling of $z_0, \ldots, z_{\hat{k}}$. Combining these facts about h, f and G, we have $\exists g \in G$ such that $g(f(x)) = h(x)$.

We now complete the proof.

$$\begin{aligned}
&\min_{g' \in G} R(g' \circ f) \\
\leq\, &R(h) && \text{since we showed that it is possible to select } f \in F \text{ and } g \in G \text{ such that } h = g \circ f \\
\leq\, &\int_{x : h(x) \neq h^*(x)} p(x)dx + R(h^*) && \text{by basic probability} \\
\leq\, &\int_{x \in \hat{\mathcal{X}}_0} p(x)dx + R(h^*) && \text{by (2.14)} \\
\leq\, &\epsilon_{\mathcal{A}} + R(h^*) && \text{by (2.6)} \\
\leq\, &\epsilon_{\mathcal{A}} + \epsilon_{\mathcal{B}}. && \text{by (2.10)}
\end{aligned}$$

□

2.4.3 Verifying Condition $\mathcal{A}(\mu_X)$

We consider the task of establishing that Condition $\mathcal{A}(\mu_X)$ holds with high probability from an unlabeled sample S_X of m points. This includes verifying that (2.5) holds. With a finite sample, this is not possible without further assumptions since we cannot expect to sample every point $x \in \hat{\mathcal{X}}^s$. However, verification is possible if we introduce a smoothness assumption on μ_X.

Furthermore, we suppose that from S_X it is possible to find a high probability lower bound on the probability mass contained in each of $\hat{\mathcal{X}}_1, \ldots, \hat{\mathcal{X}}_{\hat{k}}$. We subsequently discuss approaches to achieving this.

We observe that verifying Condition $\mathcal{A}(\mu_X)$ requires fixing $\hat{\mathcal{X}}_1, \ldots, \hat{\mathcal{X}}_{\hat{k}} \subseteq \mathcal{X}$. Each $\hat{\mathcal{X}}_i$ may be learned using an unlabeled sample and a set of candidate regions, as is discussed further below. Once $\hat{\mathcal{X}}_1, \ldots, \hat{\mathcal{X}}_{\hat{k}}$ and the representation function class F are fixed, then some representation function f satisfying (2.11) in Definition 2.7 can be identified.

Theorem 2.10 (Verifying condition $\mathcal{A}(\mu_X)$)**.** *Let $\hat{k}$ be a positive integer.*

Let $\hat{\mathcal{X}}_1, \ldots, \hat{\mathcal{X}}_{\hat{k}} \subseteq \mathcal{X}$, where $\forall i \in \{1, \ldots, \hat{k}\}, \hat{\mathcal{X}}_i$ is a connected set and $\hat{\mathcal{X}}_i \cap \hat{\mathcal{X}}_j = \varnothing$ if $i \neq j$. Let $\hat{\mathcal{X}}^s := \hat{\mathcal{X}}_1 \cup \cdots \cup \hat{\mathcal{X}}_{\hat{k}}$ and $\hat{\mathcal{X}}_0 := \mathcal{X} \setminus \hat{\mathcal{X}}^s$. For each $i \in \{1, \ldots, \hat{k}\}$, let

$$\begin{aligned} \hat{V}_i &:= \int\limits_{x \in \hat{\mathcal{X}}_i} dx && \textit{(volume) and} \\ d_i^{\max} &:= \max_{x, x' \in \hat{\mathcal{X}}_i} \|x - x'\|_2 && \textit{(maximum distance between two points).} \end{aligned}$$

Suppose for each region there is some constant $\hat{P}_i^{\min}$ such that with probability at least $1 - \frac{\hat{k}}{\delta}$,

$$\int\limits_{x \in \hat{\mathcal{X}}_i} p(x) dx \geq \hat{P}_i^{\min}. \tag{2.15}$$

Let

$$\tau := \min_{i \in \{1, \ldots, \hat{k}\}} \frac{\hat{P}_i^{\min}}{\hat{V}_i} - \lambda d_i^{\max} \tag{2.16}$$

and

$$\epsilon_{\mathcal{A}} := 1 - \sum_{i \in \{1, \ldots, \hat{k}\}} \hat{P}_i^{\min}. \tag{2.17}$$

Suppose that there is some non-negative constant λ such that

$$\forall x, x' \in \mathcal{X}, |p(x) - p(x')| \leq \lambda \|x - x'\|_2. \tag{2.18}$$

Then with probability at least $1 - \delta$, Condition $\mathcal{A}(\mu_X)$ holds.

Proof. To satisfy Condition $\mathcal{A}(\mu_X)$, we must satisfy (2.4), (2.5) and (2.6). Observe that we satisfied (2.4) in our definition of $\hat{\mathcal{X}}_1, \ldots, \hat{\mathcal{X}}_{\hat{k}}$.

By the union bound, (2.15) holds $\forall i \in \{1, \ldots, \hat{k}\}$ with probability at least $1 - \delta$.

In that case, $\forall i \in \{1, \dots, \hat{k}\}$,

$$\begin{aligned}
&\min_{x \in \hat{\mathcal{X}}_i} p(x) \\
&\geq \max_{x \in \hat{\mathcal{X}}_i} p(x) - \lambda d_i^{\max} && \text{by (2.18)} \\
&\geq \frac{1}{\hat{V}_i} \int\limits_{x \in \hat{\mathcal{X}}_i} p(x) dx - \lambda d_i^{\max} && \text{the maximum of } p(x) \text{ is at least the average of } p(x) \\
&\geq \frac{\hat{P}_i^{\min}}{\hat{V}_i} - \lambda d_i^{\max} && \text{by (2.15)} \\
&\geq \tau && \text{by (2.16).}
\end{aligned}$$

We have shown that (2.5) holds.
Furthermore,

$$\begin{aligned}
&\int\limits_{x \in \hat{\mathcal{X}}_0} p(x) dx \\
&= 1 - \sum_{i \in \{1, \dots, \hat{k}\}} \int\limits_{x \in \hat{\mathcal{X}}_i} p(x) dx && \text{by the definition of } \hat{\mathcal{X}}_0 \\
&\leq 1 - \sum_{i \in \{1, \dots, \hat{k}\}} \hat{P}_i^{\min} && \text{by (2.15)} \\
&= \epsilon_{\mathcal{A}} && \text{by (2.17).}
\end{aligned}$$

We have shown that (2.6) holds. □

We would like to find some $\hat{P}_i^{\min}$ such that with probability at least $1 - \frac{\delta}{\hat{k}}$ over m unlabeled points drawn from μ_{XY}, (2.15) holds. We achieve this using the following definition, which is motivated by a straightforward application of Hoeffding's inequality [Hoeffding, 1963]:

$$\hat{P}_i^{\min} := \frac{1}{m} \sum_{x \in S_X} \mathbf{1}(x \in \hat{\mathcal{X}}_i) - \sqrt{\frac{\log \frac{\hat{k}}{\delta}}{2m}}. \tag{2.19}$$

We may apply (2.19) to the case where $\hat{\mathcal{X}}_i$ is fixed – for example, if we had found $\hat{\mathcal{X}}_i$ using some unlabeled sample separate to S_X. However, we are also interested in the case where we estimate $\hat{\mathcal{X}}_i$ from S_X. In this case, from several candidate choices of $\hat{\mathcal{X}}_i$ construct the hypothesis class $\hat{H}_i := \{\hat{h}_i : \hat{h}_i(x) = \mathbf{1}(x \in \hat{\mathcal{X}}_i)\}$. Let $VC(\hat{H}_i)$ be the VC-dimension of $\hat{H}_i$. For example, if the candidates $\hat{\mathcal{X}}_i$ are the set of axis-aligned hyper-rectangles in $\mathbb{R}^{\hat{k}}$, then $VC(\hat{H}_i) = 2\hat{k}$ [Ben-David and Borbely, 2008]. Using a standard VC-dimension based bound on generalization error [Mohri et al., 2012], we have for all $\hat{h}_i \in \hat{H}_i$ simultaneously, with probability at least $1 - \frac{\delta}{\hat{k}}$, (2.15) holds for

$$\hat{P}_i^{\min} := \frac{1}{m} \sum_{x \in S_X} \mathbf{1}(x \in \hat{\mathcal{X}}_i) - 2\sqrt{\frac{2VC(\hat{H}_i)\log(2em/VC(\hat{H}_i)) + 2\log(4\hat{k}/\delta)}{m}}, \tag{2.20}$$

where e is Euler's number.

2.4.4 Verifying that $\mathcal{C}(F)$ and $\mathcal{D}(G)$ are Satisfied

We provide an example where it is possible to establish that both $\mathcal{C}(F)$ and $\mathcal{D}(G)$ hold by inspecting F and G.

Let $\mathcal{Z} := \mathbb{R}^{\hat{k}}$. $\forall i \in \{0, \ldots, \hat{k}\}$ let $z_i := [z_{i1}, \ldots, z_{i\hat{k}}] \in \mathcal{Z}$ and

$$z_{ij} := \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{2.21}$$

Suppose we choose each of $\hat{\mathcal{X}}_1, \ldots, \hat{\mathcal{X}}_{\hat{k}}$ from a fixed class of regions, for example the class of axis-aligned hyper-rectangles. This also determines the remaining area outside the regions, $\hat{\mathcal{X}}_0$. Suppose that $\hat{\mathcal{X}}_i \cap \hat{\mathcal{X}}_j = \varnothing$ if $i \neq j$. Let F be the set of functions $f : \mathcal{X} \to \mathcal{Z}$ which satisfy

$$f(x) = z_i, \text{ if } x \in \hat{\mathcal{X}}_i$$

for some choice of $\hat{\mathcal{X}}_0, \ldots, \hat{\mathcal{X}}_{\hat{k}}$. Hence (2.11) holds and $\mathcal{C}(F)$ is satisfied.

Furthermore, let G be the set of functions $g : \mathcal{Z} \to \{0, 1\}$ which satisfy

$$g(z) = \mathbf{1}(w \cdot z \geq 0) \tag{2.22}$$

for some choice of $w \in \mathbb{R}^{\hat{k}}$. Then (2.12) holds and hence $\mathcal{D}(G)$ is satisfied. This is due to the property of linear separators that for any labeling of the points $z_0, \ldots, z_{\hat{k}}$ from (2.21), there is some linear separator matching this labeling [Abu-Mostafa, 2012].

2.4.5 Defining and Verifying $\mathcal{E}(\mu_{XY}, H)$ to show that Objective 2.2 is Satisfied

We define the condition $\mathcal{E}(\mu_{XY}, H)$ for our clustering example. It is straightforward to observe that if this condition is satisfied, then Objective 2.2 is achieved.

Definition 2.11 (Condition $\mathcal{E}(\mu_{XY}, H)$: hypothesis class performs poorly on inputs)**.** *Fix μ_{XY}, H and l. Let $\epsilon_{\min}$ be a non-negative constant. With probability at least $1 - \delta$ over samples S_{XY} of μ_{XY}, suppose*

$$\min_{h \in H} R(h) \geq \epsilon_{\min}.$$

It would appear that to verify this condition we require access to a labeled sample S_{XY} of μ_{XY}. The verification is possible when we may conduct empirical risk minimization over H, for example when this minimization is convex.

Theorem 2.12 (Verifying Condition $\mathcal{E}(\mu_{XY}, H)$)**.** *Fix μ_{XY} and H. Let S_{XY} be a sample of m points from μ_{XY}. For some $h \in H$, let its empirical risk be*

$$\hat{R}(h) := \frac{1}{m} \sum_{\{x,y\} \in S_{XY}} \mathbf{1}(h(x) \neq y).$$

Let $VC(H)$ be the VC-dimension of H. Then with probability at least $1 - \delta$ over samples S_{XY}, Condition $\mathcal{E}(\mu_{XY}, H)$ holds for

$$\epsilon_{\min} := \min_{h \in H} \hat{R}(h) - 2\sqrt{\frac{2VC(H)\log(2em/VC(H)) + 2\log(4/\delta)}{m}}.$$

Proof. With probability at least $1 - \delta$, for all hypotheses $h \in H$ simultaneously,

$$\begin{aligned}
&R(h) \\
&\geq \hat{R}(h) - 2\sqrt{\frac{2VC(H)\log(2em/VC(H)) + 2\log(4/\delta)}{m}} && \text{[Mohri et al., 2012], p. 48} \\
&\geq \min_{h \in H} \hat{R}(h) - 2\sqrt{\frac{2VC(H)\log(2em/VC(H)) + 2\log(4/\delta)}{m}} \\
&= \epsilon_{\min}. && \text{by the definition of } \epsilon_{\min}
\end{aligned}$$

□

2.5 Conclusion

We have developed an approach to determining when unsupervised representation learning provably improves task performance. We have abstracted away from particular unsupervised representation learning techniques, instead considering the problem more generally. First, we formally defined the objectives of unsupervised representation learning. We then proposed an approach to guaranteeing that these objectives are achieved via a set of sufficient conditions, which depend separately on structure in the unlabeled distribution, shared structure in the labeled distribution, and properties of the representation function class and the hypothesis class which consumes the representation.

Our results require all of these elements to be present in order to show that unsupervised representation learning will be successful. If the structure in the unlabeled distribution is not shared in the labeled distribution, it may not be useful. If the representation function class cannot adequately exploit the structure in the unlabeled distribution, once again that structure may not be useful. And if the hypothesis class cannot exploit the structure captured by the representation function, it may not be able to learn effectively (e.g. a linearly separable representation may not be useful if the subsequent hypothesis class is not the class of linear separators). Finally, if the problem is easy for the original hypothesis class and input data, unsupervised representation learning may not be necessary. Hence, unsupervised representation learning relies on an intricate set of relationships between the different components of the problem setting in order to be successful.

We demonstrated the feasibility of our approach with an example based on the cluster assumption. In this example, we proposed methods to verify that unsupervised representation learning will be successful, both in cases where the unlabeled distribution is known (Theorem 2.9) and where it must be approximated from an unlabeled sample (Theorem 2.10). We also proposed a method to verify that supervised learning with the original data would be unsuccessful (Theorem 2.12). Combined, these may allow us to establish that there is a *positive risk gap* induced by unsupervised representation learning compared to solving the problem from scratch with supervised learning. Our methods gave us more insight into *why* unsupervised representation learning is useful in some cases, rather than simply relying on trial-and-error tests of particular techniques on particular problem settings.

We have seen that in order to prove that unsupervised representation learning helps, it is necessary to theoretically show several different claims. While feasible for the example we considered, extension to more complex situations and model classes such as multi-layer neural networks will be challenging. It will involve formally identifying instantiations of each of the proposed conditions, which are at present unknown. Nevertheless, we have developed the first complete theoretical analysis of when unsupervised representation learning is useful. We cannot rule out a simpler theory that is easier to apply to wider class of models, but such a theory remains elusive.

One approach to using unlabeled data is to construct an 'artificial' task from the data, such as reconstructing an image or predicting which words will frequently co-occur. A neural network is trained to make predictions on this task, and the weights in the network are then transferred for use on a subsequent task [Hinton and Salakhutdinov, 2006; Mikolov et al., 2013]. This kind of unsupervised representation learning can be seen as a special case of transfer learning, where a representation function is learned on a source task and re-used on a target task. We explore this situation further in Chapter 3.

Chapter 3

Risk Bounds for Transferring Representations With and Without Fine-Tuning

3.1 Introduction

A widely used machine learning technique is the transfer of a representation function learned from a source task, for which labeled data is abundant, to a target task, for which labeled data is scarce [Pan and Yang, 2010; Bengio, 2012]. This may be effective if both tasks can be learned using a common representation function mapping inputs to a representation space, followed by a task-specific hypothesis acting on the representation space drawn from a common hypothesis class. For example:

- features learned from an image of a human face to predict age may also be useful for predicting gender
- word embeddings learned to predict word contexts may also be useful for part of speech tagging
- features learned from financial data to predict loan default may also be useful for predicting insurance fraud.

Source and target task learning are often conducted by two separate organizations. The organization that conducts representation learning on the source task may have greater access to data, computational and human resources relative to the organization that wishes to learn the target task. Examples are the Google word2vec package [Mikolov et al., 2013], and downloadable pre-trained neural networks,[1] which were created by leading research groups and have been re-used by a range of organizations. Under this 'representation-as-a-service' model, a user may expect to access the representation function itself, as well as information about its performance on the source task data on which it was trained. We aim to convert this into a guarantee

[1] See http://code.google.com/archive/p/word2vec, http://caffe.berkeleyvision.org/model_zoo and http://vlfeat.org/matconvnet/pretrained for examples.

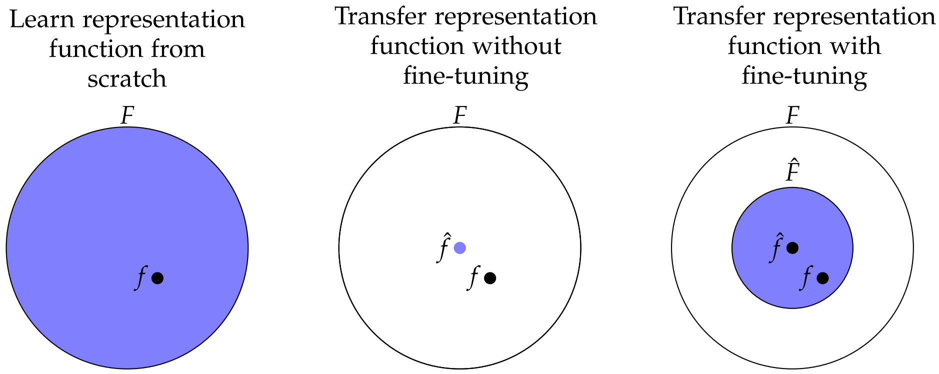

Learn representation function from scratch
F
f
Transfer representation function without fine-tuning
F
$\hat{f}$
f
Transfer representation function with fine-tuning
F
$\hat{F}$
$\hat{f}$
f

Figure 3.1: A comparison of approaches to learning a representation function f on a target task, where the search space in each case is the shaded area. Learning from scratch, we search a representation function class F. Without fine-tuning, we fix a representation function $\hat{f}$ learned from the source task. With fine-tuning, we narrow the search to $\hat{F} \subseteq F$ near $\hat{f}$, which still contains f.

of the usefulness of the representation function on other tasks, a guarantee which is known *in advance* without the effort or cost of testing the representation function on the target task(s). Our analysis also covers the case where the source task is constructed from unlabeled data, as in neural network unsupervised pre-training.

We consider two approaches to transferring a representation function learned from a source task to a target task, as shown in Figure 3.1. We may either treat the representation function as fixed, or we may narrow the class of representation functions considered on the target task, which we refer to as *fine-tuning*. The fixed option may be attractive when very little labeled target task data is available and hence overfitting is a strong concern, while the advantage of fine-tuning is relatively greater hypothesis class expressiveness.

Let $\mathcal{X}, \mathcal{Y}$ and known as the input and target spaces respectively, and let X and Y be corresponding random variables. We focus on the binary classification setting where $\mathcal{Y} = \{-1, 1\}$. Let $\mathcal{Z}$ be a set known as the representation space. Let F be a representation function class, where $f : \mathcal{X} \to \mathcal{Z}$ for $f \in F$. Given a representation function f, let $Z := f(X)$ be the corresponding representation random variable. Let G be an hypothesis class acting on the representation space, where $g : \mathcal{Z} \to \mathcal{Y}$ for $g \in G$. Let the hypothesis class

$$H := \{h : \exists f \in F, g \in G \text{ such that } h = g \circ f\}.$$

For target task T, let μ_{XY} be its joint distribution over $\mathcal{X}$ and $\mathcal{Y}$ and μ_X be its marginal distribution over $\mathcal{X}$. Given a representation function f, let μ_{ZY} be the induced joint distribution over $\mathcal{Z}$ and $\mathcal{Y}$ and let μ_Z be the induced marginal distribution over $\mathcal{Z}$. Similarly, for source task S, let μ'_{XY} be its joint distribution and μ'_X be its marginal distribution, and let μ'_{ZY} and μ'_Z be the joint and marginal distributions induced by f respectively.

For some distribution μ let $p(\cdot)$ be the probability that a point sampled from the distribution satisfies some condition. Let $p_S(\cdot)$ be the probability of an event under distribution μ'_{XY} and $p_T(\cdot)$ be the probability of an event under distribution μ_{XY}.

Let l be a loss function $\mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$. For an hypothesis $h : \mathcal{X} \to \mathcal{Y}$, let its risks on S and T be

$$R_S(h) := \mathbb{E}_{X,Y \sim \mu'_{XY}}[l(h(X), Y)]$$

and

$$R_T(h) := \mathbb{E}_{X,Y \sim \mu_{XY}}[l(h(X), Y)]$$

respectively. Let $\hat{R}_S(h)$ and $\hat{R}_T(h)$ be the corresponding empirical (i.e. training set distribution) risks. We focus on the case where l is the $0/1$ loss:

$$l(\hat{y}, y) := \begin{cases} 0 & \text{if } y = \hat{y} \\ 1 & \text{if } y \neq \hat{y}. \end{cases} \tag{3.1}$$

Let m_S be the number of samples available for task S and m_T be the number of samples available for task T. Let $VC(\cdot)$ be the VC-dimension of an hypothesis class.

The remainder of the chapter is structured as follows. In Section 3.2 we summarize related work. In Sections 3.3 and 3.4 we analyze the cases where the transferred representation function is fixed and fine-tuned respectively. In Section 3.5 we apply the results and use them to motivate and test a practical approach to weight transfer in neural networks. We conclude in Section 3.6, and present lemmas used in the proofs of our theorems in Section 3.7.

3.2 Background

Empirical studies have shown the success of transferring representation functions between tasks [Donahue et al., 2014; Hoffman et al., 2014; Girshick et al., 2014; Socher et al., 2013; Bansal et al., 2014]. Word embeddings learned on a source task have been shown to perform better than unigram features on target tasks such as part of speech tagging, and comparably or better than embeddings fine-tuned on the target task [Qu et al., 2015]. Yosinski et al. [2014] learned neural network weights using half of the ImageNet classes, and then learned the other classes with a neural network initialized with these weights, finding a benefit compared to random initialization only with target task fine-tuning. The transfer of representation functions, both with and without fine-tuning, is widely and successfully used.

Previous work on domain adaptation [Ben-David et al., 2010; Mansour et al., 2009; Germain et al., 2013] has considered learning an hypothesis h on S and re-using it on T, bounding $R_T(h)$ using $R_S(h)$ (measured with labeled source data) and some notion of similarity between μ_X and μ'_X (measured with additional unlabeled target data). Such results motivate a joint optimization using labeled source and unlabeled target data [Ganin et al., 2016; Long et al., 2015] to learn separate mappings $f_S, f_T : \mathcal{X} \to \mathcal{Z}$, as well as an hypothesis $g : \mathcal{Z} \to \mathcal{Y}$ learned from the source labels which can be re-used on T. This approach assumes that if f_S applied to μ'_X and f_T applied to μ_X induce similar marginal distributions over $\mathcal{Z}$, then some g can be found such that

$g \circ f_S$ and $g \circ f_T$ are accurate hypothesis for tasks S and T respectively. We consider an alternative situation where there is some common representation function $f : \mathcal{X} \to \mathcal{Z}$ and two separate task-specific hypotheses $g_S, g_T : \mathcal{Z} \to \mathcal{Y}$ such that $g_S \circ f$ and $g_T \circ f$ are accurate hypotheses for tasks S and T respectively. We consider estimating f using a labelled sample from S and then estimating g_T from a small amount of labeled target data. Given the widespread use of 'downloadable' representations, where f and g_T are learned separately and there is no joint optimization over source and target data, this is a realistic setting.

Work on lifelong learning relates the past performance of a representation function over many tasks to its expected future performance. For a representation function $f \in F$ we construct $G \circ f := \{g \circ f : g \in G\}$. Suppose there is a distribution over tasks, known as an environment. Assume several tasks from this environment have been sampled, and that for each task some hypothesis in $G \circ f$ has been selected and its empirical risk evaluated. Previous work has provided bounds on the difference between the average empirical risk and the expected risk of the best hypothesis in $G \circ f$ for a new task drawn from the environment. Such bounds have been given by measuring the complexity of F and G using covering numbers [Baxter, 2000], a variant of the growth function [Galanti et al., 2016], and a distribution-dependent measure known as Gaussian complexity [Maurer et al., 2016]. All of these bounds rely on known past performance on a large number of tasks.[2] In practice, however, representation functions such as neural network weights or word embeddings are often learned using only a single source task, which is the setting we consider.

3.3 Representation Function Fixed by Source Task

Suppose samples from source task S are abundant, samples from target task T are scarce, and there exist some f, g_S, g_T such that $g_S \circ f$ and $g_T \circ f$ are accurate hypotheses for tasks S and T respectively. A natural approach to leveraging the source data is to learn $\hat{g}_S \circ \hat{f} \in H$ using data from task S, from which we assume we may recover $\hat{f} \in F$, then perform empirical risk minimization over $G \circ \hat{f} := \{g \circ \hat{f} : g \in G\}$ on T yielding $\hat{g}_T \circ \hat{f}$. While in general we cannot recover $\hat{f}$ with knowledge of $\hat{g}_S \circ \hat{f}$ alone, in the case of feedforward neural networks which we focus on, knowing the weights learned on S is sufficient for recovering $\hat{f}$.

Theorem 3.1 upper-bounds $R_T(\hat{g}_T \circ \hat{f})$ using four terms:

1. a function ω measuring a transferrability property obtained analytically from the problem setting;

[2] Pentina and Lampert [2014] extend this analysis to stochastic hypotheses (i.e. distributions over deterministic hypotheses), where for each task we learn a posterior given a prior and training data. The quality of the prior affects the learner's performance. The study proposes using source tasks to learn a 'hyperposterior', a distribution over priors which is sampled to give a prior for each task. Such a hyperposterior may focus the learner on a representation function shared across tasks. The study gives a PAC-Bayes bound on the expected risk of using a hyperposterior to learn a new task drawn from the environment, in terms of the average empirical risk obtained using the hyperposterior to learn the source tasks.

2. the source task empirical risk $\hat{R}_S(\hat{g}_S \circ \hat{f})$;

3. the generalization error of an hypothesis in H learned from m_S samples; and

4. the generalization error of an hypothesis in G learned from m_T samples.

Note that we do not settle for bounding $R_T(\hat{g}_T \circ \hat{f})$ in terms of $\hat{R}_T(\hat{g}_T \circ \hat{f})$, which may be large.

Theorem 3.1. *Let $\omega : \mathbb{R} \to \mathbb{R}$ be a non-decreasing function. Suppose μ_{XY}, μ'_{XY}, $\hat{f}$ and G have the property that*

$$\forall \hat{g}_S \in G, \min_{g \in G} R_T(g \circ \hat{f}) \leq \omega(R_S(\hat{g}_S \circ \hat{f})). \tag{3.2}$$

Let $\hat{g}_T := \arg\min_{g \in G} \hat{R}_T(g \circ \hat{f})$. Then with probability at least $1-\delta$ over pairs of training sets for tasks S and T,

$$R_T(\hat{g}_T \circ \hat{f}) \leq \omega(\hat{R}_S(\hat{g}_S \circ \hat{f}) + 2\sqrt{\frac{2VC(H)\log(2em_S/VC(H)) + 2\log(8/\delta)}{m_S}}) \\ + 4\sqrt{\frac{2VC(G)\log(2em_T/VC(G)) + 2\log(8/\delta)}{m_T}}.$$

Proof. Let $g_T^* := \arg\min_{g \in G} R_T(g \circ \hat{f})$. With probability at least $1-\delta$,

$$\begin{aligned}
& R_T(\hat{g}_T \circ \hat{f}) \\
& \leq \hat{R}_T(\hat{g}_T \circ \hat{f}) + 2\sqrt{\frac{2VC(G)\log(2em_T/VC(G)) + 2\log(8/\delta)}{m_T}} && (3.3) \\
& \leq \hat{R}_T(g_T^* \circ \hat{f}) + 2\sqrt{\frac{2VC(G)\log(2em_T/VC(G)) + 2\log(8/\delta)}{m_T}} && (3.4) \\
& \leq R_T(g_T^* \circ \hat{f}) + 4\sqrt{\frac{2VC(G)\log(2em_T/VC(G)) + 2\log(8/\delta)}{m_T}} && (3.5) \\
& \leq \omega(R_S(\hat{g}_S \circ \hat{f})) + 4\sqrt{\frac{2VC(G)\log(2em_T/VC(G)) + 2\log(8/\delta)}{m_T}} && (3.6) \\
& \leq \omega(\hat{R}_S(\hat{g}_S \circ \hat{f}) + 2\sqrt{\frac{2VC(H)\log(2em_S/VC(H)) + 2\log(8/\delta)}{m_S}}) \\
& \qquad + \ 4\sqrt{\frac{2VC(G)\log(2em_T/VC(G)) + 2\log(8/\delta)}{m_T}}. && (3.7)
\end{aligned}$$

Using m training points and an hypothesis class of VC-dimension $VC(\cdot)$, with probability at least $1-\delta$, for all hypotheses $h \in H$ simultaneously, the risk $R(h)$ and empirical risk $\hat{R}(h)$ satisfy

$$|R(h)-\hat{R}(h)| \leq 2\sqrt{\frac{2VC(\cdot)\log(2em/VC(\cdot))+2\log(4/\delta)}{m}} \tag{3.8}$$

[Mohri et al., 2012]. Applying (3.8) to G yields (3.3) and (3.5) with probability at least $1-\frac{\delta}{2}$. Applying (3.8) to H, and using the fact that ω is non-decreasing, yields (3.7) with probability at least $1-\frac{\delta}{2}$. (3.4) holds by the definition of $\hat{g}_T$ and (3.6) follows from the assumption (3.2). Applying the union bound achieves the result. □

While we refer to ω in a general form, we give an example in Section 3.3.1 and expect that others exist. We define ω by relating $R_S(\hat{g}_S \circ \hat{f})$ to $\min_{g\in G} R_T(g \circ \hat{f})$, since we expect this may be feasible analytically as in our example in Section 3.3.1. However, because we only observe $\hat{R}_S(\hat{g}_S \circ \hat{f})$, in Theorem 3.1 we use this to bound $R_S(\hat{g}_S \circ \hat{f})$ and then apply ω.

It is instructive to compare Theorem 3.1 to a standard VC-dimension based bound on the target task risk of an hypothesis h drawn from H learned using m_T training points [Mohri et al., 2012]: with probability at least $1-\delta$, for all hypotheses h simultaneously,

$$R_T(h) \leq \hat{R}_T(h) + 2\sqrt{\frac{2VC(H)\log(2em_T/VC(H))+2\log(4/\delta)}{m_T}}. \tag{3.9}$$

We conclude that if $\omega(R)=O(R)$; $\hat{R}_S(\hat{g}_S \circ \hat{f})$ is a small constant; $m_S \gg m_T$, i.e. labeled source task data is abundant while labeled target task data is scarce; and $VC(H) \gg VC(G)$, i.e. transferring the representation function $\hat{f}$ simplifies target task learning by virtue of the smaller hypothesis space it induces compared to searching F; then consequently, the VC-dimension-based upper bound on target task risk is smaller by transferring $\hat{f}$ from S compared to learning T from scratch using H.

We observe that a smaller upper bound on risk does not imply smaller risk; indeed, since $G \circ \hat{f} \subseteq H$, it follows that

$$\min_{h\in H} R_T(h) \leq \min_{g\in G} R_T(\hat{g} \circ \hat{f})$$

and hence we may 'get lucky' and find a low risk hypothesis h learning just with samples from task T. In this case we may not be able to verify that the hypothesis is low risk, however, given the scarcity of samples from T and the expressiveness of H. Conversely, transferring $\hat{f}$ from S and applying Theorem 3.1, we may more tightly bound target task risk with high probability. We observe that Theorem 3.1 can be used to select source task S given several options by picking the task corresponding to the lowest risk upper bound.

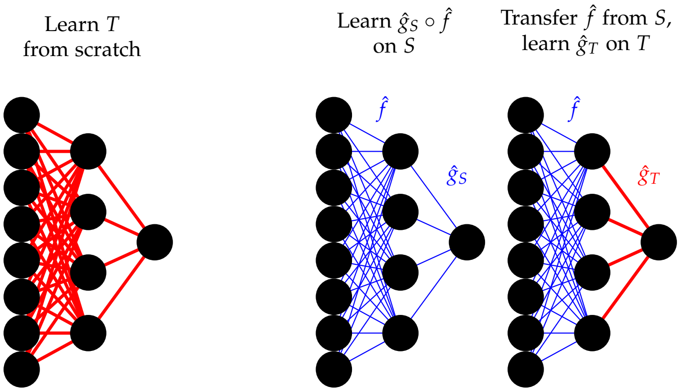


Figure 3.2: Neural network example learning T from scratch (left) and with weights transferred from S (right). Thin blue and thick red lines show weights trained on S and T respectively. Under certain assumptions, weight transfer yields low risk on T.

3.3.1 Neural Network Example with Fixed Representation

In Theorem 3.5, we give an example of the property required by Theorem 3.1, which is specific to a particular problem setting. We consider a feedforward neural network with a single hidden layer (see Figure 3.2). We propose transferring the lower-level weights (corresponding to $\hat{f}$) learned on S, so that only the upper-level weights (corresponding to G) have to be learned on T. We want to show $\hat{f}$ is also useful for T, i.e. that for some $g \in G$ we have small $R_T(g \circ \hat{f})$.

We introduce several assumptions required to show Theorem 3.5. Assumption 3.4 requires that some lower-level weights perform well on both tasks, which is clearly a necessary condition for the *specific* $\hat{f}$ we are transferring to perform well on both tasks. Our other two assumptions together guarantee that a point $x \in \mathcal{X}$ for which $\hat{f}(x)$ contributes to the risk on T cannot be 'hidden' from the risk of using $\hat{f}$ on S, either through low magnitude upper-level weights (prevented by Assumption 3.2) or low $\mu'_X(x)$ (prevented by Assumption 3.3). Hence $R_S(\hat{g}_S \circ \hat{f})$ reliably indicates the usefulness of $\hat{f}$ on T.

Assumption 3.2 (Restricted class of feedforward neural networks)**.** *Let* $\mathcal{X} = \mathbb{R}^n$, $\mathcal{Z} = \mathbb{R}^k$ *and* $a : \mathbb{R} \to \mathbb{R}$ *be a fixed activation function satisfying*

$$a(-x) = -a(x), \tag{3.10}$$

i.e. a is an odd function (examples include tanh, sign and identity). Let

$$F := \{f : \mathcal{X} \to \mathcal{Z} : f(x) = [a(w_1 \cdot x), \ldots, a(w_k \cdot x)], w_i \in \mathbb{R}^n \text{ for } 1 \leq i \leq k\} \tag{3.11}$$

and

$$G := \{g : \mathcal{Z} \to \mathcal{Y} : g(z) = sign(v \cdot z), v \in \{-1, 1\}^k\}, \tag{3.12}$$

where the symbol $\cdot$ denotes the dot product.

Assumption 3.3 (Relative rotation invariance between source and target unlabeled distributions). *Let $\hat{w}_i \in \mathbb{R}^n$ for $1 \leq i \leq k$ and let $\hat{f} \in F$ be defined as*

$$\hat{f}(x) := [a(\hat{w}_1 \cdot x), \ldots, a(\hat{w}_k \cdot x)]. \tag{3.13}$$

Suppose there exist finite nonzero constants c, $\alpha_1, \cdots, \alpha_k$ and $\beta_1, \cdots, \beta_k$ such that

$$\|w_i\| = \|\alpha_i \hat{w}_i - \beta_i w_i\|, \tag{3.14}$$

$$w_i \cdot (\alpha_i \hat{w}_i - \beta_i w_i) = 0, \tag{3.15}$$

the $2k \times n$ matrix

$$M := \begin{bmatrix} w_1 \\ \alpha_1 \hat{w}_1 - \beta_1 w_1 \\ \vdots \\ w_k \\ \alpha_k \hat{w}_k - \beta_k w_k \end{bmatrix} \tag{3.16}$$

is full rank,[3] *and*

$$\forall x_1, x_2 \in \mathcal{X} \text{ such that } \|Mx_1\| = \|Mx_2\|, \mu_X(x_1) \leq c\mu'_X(x_2), \tag{3.17}$$

which we call relative rotation invariance and implies μ'_X and μ_X have the same support.[4]

Assumption 3.4 (Shared representation exists). *Suppose there exist some*

$$\begin{aligned}
&f \in F : f(x) := [a(w_1 \cdot x), \ldots, a(w_k \cdot x)], \\
&g_S \in G : g_S(z) := sign(v_S \cdot z), \\
&g_T \in G : g_T(z) := sign(v_T \cdot z), \\
&\epsilon \geq 0
\end{aligned}$$

such that

$$\max[R_S(g_S \circ f), R_T(g_T \circ f)] \leq \epsilon. \tag{3.18}$$

We now state the target task risk bound for transferring representations in our neural network example.

[3] To see that this condition is necessary, consider the following example where M is not full rank. Let $n = 4, k = 2$, $y = sign(x_1)$ under μ'_{XY} and $y = sign(x_2)$ under μ_{XY}. For $f(x) = [x_1 + x_2, x_1 - x_2]$, $g_S(z) = sign(z_1 + z_2)$ and $g_T(z) = sign(z_1 - z_2)$, we have $R_S(g_S \circ f) = R_T(g_T \circ f) = 0$. On S we learn $\hat{f}(x) = [x_1 + x_3, x_1 - x_3]$ and $\hat{g}_S(z) = sign(z_1 + z_2)$, so that $R_S(\hat{g}_S \circ \hat{f}) = 0$, but in general $\min_{g \in G} R_T(g \circ \hat{f}) > 0$ since $\hat{f}$ ignores x_2.

[4] If M is an orthogonal matrix then $\forall x_1, x_2 \in \mathcal{X}$ such that $\|x_1\| = \|x_2\|, \mu_X(x_1) \leq c\mu'_X(x_2)$. For example, this equation is satisfied if μ_X and μ'_X are spherical Gaussians. Note that a zero-mean multivariate Gaussian distribution can be converted to a spherical Gaussian by the whitening transformation $x \to \Lambda^{-1/2} U^T x$, where the columns of U and entries of the diagonal matrix Λ are the eigenvectors and eigenvalues of the distribution's covariance matrix respectively.

Theorem 3.5. *Suppose F and G satisfy Assumption 3.2; F, G, μ_{XY}, μ'_{XY} and ϵ satisfy Assumption 3.4; and $\hat{f}$, μ_X, μ'_X and c satisfy Assumption 3.3. Let $\omega : \mathbb{R} \to \mathbb{R}$ be defined as*

$$\omega(R) := cR + \epsilon(1 + c). \tag{3.19}$$

Then

$$\forall \hat{g}_S \in G, \min_{g \in G} R_T(g \circ \hat{f}) \leq \omega(R_S(\hat{g}_S \circ \hat{f})).$$

Proof. Let $g_S(z) := sign(v_S \cdot z)$, $g_T(z) := sign(v_T \cdot z)$, $\hat{g}_S(z) := sign(\hat{v}_S \cdot z)$ and $\hat{g}_T(z) := sign(d * \hat{v}_S \cdot z)$, where $d := v_S * v_T \in \{-1, 1\}^k$ and $*$ is the elementwise product. It is sufficient to show that

$$R_T(\hat{g}_T \circ \hat{f}) \leq cR_S(\hat{g}_S \circ \hat{f}) + \epsilon(1 + c).$$

$$\begin{aligned}
&R_T(\hat{g}_T \circ \hat{f}) \\
&= p_T(yd * \hat{v}_S \cdot \hat{f}(x) \leq 0) \\
&\leq p_T(yd * v_S \cdot f(x) d * v_S \cdot f(x) d * \hat{v}_S \cdot \hat{f}(x) \leq 0) \\
&\leq p_T(yd * v_S \cdot f(x) \leq 0) + p_T(d * v_S \cdot f(x) d * \hat{v}_S \cdot \hat{f}(x) \leq 0) \\
&\leq \epsilon + p_T(d * v_S \cdot f(x) d * \hat{v}_S \cdot \hat{f}(x) \leq 0) && (3.20) \\
&\leq \epsilon + cp_S(v_S \cdot f(x) \hat{v}_S \cdot \hat{f}(x) \leq 0) && (3.21) \\
&= \epsilon + cp_S(yv_S \cdot f(x) y\hat{v}_S \cdot \hat{f}(x) \leq 0) \\
&\leq \epsilon + cp_S(yv_S \cdot f(x) \leq 0) + p_S(y\hat{v}_S \cdot \hat{f}(x) \leq 0) \\
&= \epsilon + c[R_S(\hat{g}_S \circ \hat{f}) + R_S(g_S \circ f)] \\
&\leq cR_S(\hat{g}_S \circ \hat{f}) + \epsilon(1 + c). && (3.22)
\end{aligned}$$

(3.20) and (3.22) are due to the shared representation assumption (3.18). (3.21) holds by Lemma 3.11. The remaining lines apply simple rules of probability. □

3.4 Representation Function Fine-Tuned on Target Task

Consider learning $\hat{g}_S \circ \hat{f}$ on S, and then using $\hat{f}$ and $R_S(\hat{g}_S \circ \hat{f})$ to find $\hat{F} \subseteq F$, as in Figure 3.1. Let $\tilde{h}_{g \circ f}$ be a stochastic hypothesis (i.e. a distribution over H) associated with $g \circ f$ (e.g. $g \circ f$ is the mode of $\tilde{h}_{g \circ f}$). We propose learning T with the hypothesis class

$$\tilde{H}_{G \circ \hat{F}} := \{\tilde{h}_{g \circ f} : f \in \hat{F}, g \in G\}$$

and the prior $\tilde{h}_{\hat{g}_S \circ \hat{f}}$. Learning T from scratch we assume that we would instead use

$$\tilde{H}_{G \circ F} := \{\tilde{h}_{g \circ f} : f \in F, g \in G\}$$

and some fixed prior $\tilde{h}_0 \in \tilde{H}_{G \circ F}$. For some stochastic hypothesis $\tilde{h}$ let its risk on task T with respect to a loss function l be

$$R_T(\tilde{h}) := \mathbb{E}_{X,Y \sim \mu_{XY}, h \sim \tilde{h}}[l(h(X), Y)],$$

and let $\hat{R}_T(\tilde{h})$ be its risk on the training set distribution of T.

In Theorem 3.6 we show that if $\hat{F}$ is 'small enough' so that all $\tilde{h} \in \tilde{H}_{G \circ \hat{F}}$ have a small KL divergence from $\tilde{h}_{\hat{g}_S \circ \hat{f}}$, we may apply a PAC-Bayes bound to the generalization error of hypotheses in $\tilde{H}_{G \circ \hat{F}}$ involving four terms:

1. a function ω measuring a transferrability property
2. the empirical risk $\hat{R}_S(\hat{g}_S \circ \hat{f})$
3. the generalization error of an hypothesis in H learned from m_S points, and
4. a weak dependence on m_T.

Theorem 3.6. *Let $\omega : \mathbb{R} \to \mathbb{R}$ be non-decreasing. Suppose given $\hat{f} \in F$ and $R_S(\hat{g}_S \circ \hat{f})$ estimated from S, it is possible to construct $\hat{F} \subseteq F$ with the property*

$$\forall \tilde{h} \in \tilde{H}_{G \circ \hat{F}}, KL(\tilde{h} || \tilde{h}_{\hat{g}_S \circ \hat{f}}) \leq \omega(R_S(\hat{g}_S \circ \hat{f})). \tag{3.23}$$

Then with probability at least $1 - \delta$ over pairs of training sets for tasks S and T, $\forall \tilde{h} \in \tilde{H}_{G \circ \hat{F}}$,

$$R_T(\tilde{h}) \leq \hat{R}_T(\tilde{h}) + \sqrt{\frac{\omega(\hat{R}_S(\hat{g}_S \circ \hat{f}) + 2\sqrt{\frac{2VC(H)\log(2em_S/VC(H)) + 2\log(8/\delta)}{m_S}}) + \log 2m_T/\delta}{2(m_T - 1)}}.$$

Proof. With probability at least $1 - \delta$,

$$\begin{aligned} & R_T(\tilde{h}) \\ & \leq \hat{R}_T(\tilde{h}) + \sqrt{\frac{KL(\tilde{h} || \tilde{h}_{\hat{g}_S \circ \hat{f}}) + \log 2m_T/\delta}{2(m_T - 1)}} && (3.24) \\ & \leq \hat{R}_T(\tilde{h}) + \sqrt{\frac{\omega(R_S(\hat{g}_S \circ \hat{f})) + \log 2m_T/\delta}{2(m_T - 1)}}. && (3.25) \end{aligned}$$

(3.24) holds with probability at least $1 - \frac{\delta}{2}$ [Shalev-Shwartz and Ben-David, 2014]. (3.25) holds by the assumption (3.23). Furthermore,

$$R_S(\hat{g}_S \circ \hat{f}) \leq \hat{R}_S(\hat{g}_S \circ \hat{f}) + 2\sqrt{\frac{2VC(H)\log(2em_S/VC(H)) + 2\log(8/\delta)}{m_S}} \tag{3.26}$$

with probability at least $1 - \frac{\delta}{2}$ due to (3.8). The result follows due to (3.25), (3.26), the assumption that ω is non-decreasing, and the union bound. □

It is instructive to compare Theorem 3.6 to a standard PAC-Bayes based bound on the target task risk of a stochastic hypothesis $\tilde{h}$ learned using m_T training points and prior $\tilde{h}_0$ [Shalev-Shwartz and Ben-David, 2014]: with probability at least $1-\delta$,

$$R_T(\tilde{h}) \leq \hat{R}_T(\tilde{h}) + \sqrt{\frac{KL(\tilde{h}||\tilde{h}_0) + \log m_T/\delta}{2(m_T-1)}}. \tag{3.27}$$

We conclude that if $\omega(R) = O(R)$, $\hat{R}_S(\hat{g}_S \circ \hat{f})$ is a small constant, and $m_S \gg m_T$, we improve on the bound (3.27) applied to choices of $\tilde{h} \in \tilde{H}_{G\circ F}$ for which $KL(\tilde{h}||\tilde{h}_0)$ is large. Observe that using the restricted deterministic hypothesis class

$$G \circ \hat{F} := \{h : \exists f \in \hat{F}, g \in G \text{ such that } h = g \circ f\}$$

and a VC-dimension-based bound such as (3.8) may not improve on the bound for H, since possibly $VC(G \circ \hat{F}) = VC(H)$.

$\hat{F}$ is useful if it is also 'large enough' in the sense that for some small constant $\epsilon \geq 0$,

$$\exists \tilde{h}_{g_T \circ f} \in \tilde{H}_{G\circ\hat{F}} \text{ such that } R_T(\tilde{h}_{g_T\circ f}) \leq \epsilon.$$

The role of ω is to quantify how large the $\hat{F}$ we search on T must be in order to be 'large enough', in terms of $R_S(\hat{g}_S \circ \hat{f})$. While in general such an $\hat{F}$ and ω may not exist, we give an example where they do in Section 3.4.1.

As with Theorem 3.1, we observe that a smaller upper bound on risk does not imply smaller risk, and since $G \circ \hat{F} \subseteq G \circ F$, it follows that

$$\min_{\tilde{h}\in\tilde{H}_{G\circ F}} R_T(\tilde{h}) \leq \min_{\tilde{h}\in\tilde{H}_{G\circ\hat{F}}} R_T(\tilde{h}).$$

However, by transferring $\hat{F}$ from S, constructing the hypothesis class $\tilde{H}_{G\circ\hat{F}}$ and applying Theorem 3.6, we may more tightly bound target task risk compared to learning T from scratch with the hypothesis class $\tilde{H}_{G\circ F}$.

3.4.1 Neural Network Example with Fine-Tuning

We transfer and fine-tune weights in a feedforward neural network with one hidden layer to instantiate the property required by Theorem 3.6. We learn a deterministic hypothesis of this type on S and obtain k estimated lower-level weight vectors $\hat{w}_i$. Learning T we now consider only lower-level weights near $\hat{w}_i$, corresponding to $\hat{F}$. In Theorem 3.10, we show sufficient conditions under which it is possible to construct such an $\hat{F}$ to successfully learn T.

We introduce several assumptions required to show Theorem 3.10. Assumption 3.9 requires some lower-level weights w_i perform well on both S and T, which is clearly a necessary condition for the *specific* $\hat{F}$ we are transferring to contain lower-level weights that perform well on both tasks. We make $\hat{F}$ 'large enough' by using the risk observed using $\hat{w}_i$ on S to provide an upper bound on the angle between each pair w_i and $\hat{w}_i$, as formalized in (3.39), so that we know that searching $\hat{F}$ will include w_i. We make $\hat{F}$ 'small enough' by only including lower-level weights with small angles to $\hat{w}_i$, as formalized in (3.40).

We use a restricted class of feedforward neural networks to learn S, and a stochastic variant of this restricted class to learn T, as described in Assumption 3.7. In particular, on T we learn a stochastic hypothesis formed by taking a deterministic network, and adding independent sources of spherical Gaussian noise to the lower-level weights and sign-flipping noise to the upper-level weights. This choice of network architecture means that the KL divergence between two stochastic hypotheses is expressed using the angles between their lower-level weights[5] and a quantity computable from their upper-level weights.

Assumptions 3.7 and 3.8 together ensure that poor $\hat{w}_i$ cannot be 'hidden' from the risk on S, either through low magnitude higher-level weights (prevented by Assumption 3.7), or through low μ'_X density in the region where using $\hat{w}_i$ instead of w_i yields different predictions (prevented by Assumption 3.8). Hence the performance of $\hat{w}_i$ on S is a reliable indicator of the magnitude of the angle between w_i and $\hat{w}_i$.

Assumption 3.7 (Restricted class of feedforward neural networks). *Let $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Z} = \mathbb{R}^k$, where k is odd. Let*

$$F := \{f : \mathcal{X} \to \mathcal{Z} : f(x) = [sign(w_1 \cdot x), \ldots, sign(w_k \cdot x)], w_i \in \mathbb{R}^n \text{ for } 1 \leq i \leq k\} \tag{3.28}$$

Let

$$G := \{g : \mathcal{Z} \to \mathcal{Y} : g(z) = sign(v \cdot z), v \in \{-1, 1\}^k\}. \tag{3.29}$$

For some $f \in F, g \in G$, let

$$\tilde{h}_{g \circ f} := g' \circ f'$$

where

$$\begin{aligned}
&f(x) := [sign(w_1 \cdot x), \ldots, sign(w_k \cdot x)],\\
&f'(x) := [sign(w'_1 \cdot x), \ldots, sign(w'_k \cdot x)],\\
&w'_1, \ldots, w'_k \text{ are each drawn independently via } w'_i \sim \mathcal{N}(w_i, \sigma^2 I),\\
&\mathcal{N} \text{ is the multivariate normal distribution,}\\
&I \text{ is the } n \times n \text{ identity matrix and}\\
&\sigma \geq 0 \text{ is a constant;}
\end{aligned}$$

[5] Assuming that the lower-level weight vectors are of fixed magnitude, which is no loss of model expressiveness since we use the sign activation function at the hidden layer.

and

$$\begin{aligned}
&g(z) := sign(v \cdot z),\\
&g'(z) := sign(v' \cdot z),\\
&v'_1, \ldots, v'_k \textit{ are each drawn independently via } v'_i \sim (2\mathrm{Bern}(q) - 1)v_i,\\
&\mathrm{Bern} \textit{ is the Bernoulli distribution and}\\
&q \in [0.5, 1] \textit{ is a constant.}
\end{aligned}$$

Assumption 3.8 (Rotation invariance of unlabeled source distribution). *Let $\hat{w}_i \in \mathbb{R}^n$ for $1 \leq i \leq k$, let $\theta(w_i, \hat{w}_i)$ be the angle between w_i and $\hat{w}_i$, and suppose*

$$\forall i, \|\hat{w}_i\| = 1. \tag{3.30}$$

Let $\hat{f} \in F$ be defined as

$$\hat{f}(x) := [sign(\hat{w}_1 \cdot x), \ldots, sign(\hat{w}_k \cdot x)]. \tag{3.31}$$

Suppose there exist finite nonzero constants c, $\alpha_1, \cdots, \alpha_k$ and $\beta_1, \cdots, \beta_k$ such that

$$\|w_i\| = \|\alpha_i \hat{w}_i - \beta_i w_i\|, \tag{3.32}$$

$$w_i \cdot (\alpha_i \hat{w}_i - \beta_i w_i) = 0, \tag{3.33}$$

the $2k \times n$ matrix

$$M := \begin{bmatrix} w_1 \\ \alpha_1 \hat{w}_1 - \beta_1 w_1 \\ \vdots \\ w_k \\ \alpha_k \hat{w}_k - \beta_k w_k \end{bmatrix}, \tag{3.34}$$

and

$$\forall x_1, x_2 \in \mathcal{X} \textit{ such that } \|Mx_1\| = \|Mx_2\|, \mu'_X(x_1) \leq c\mu'_X(x_2), \tag{3.35}$$

which we call rotation invariance of μ'_X.[6]

Assumption 3.9 (Shared representation exists). *Suppose there exist some*

$$\begin{aligned}
&f \in F : f(x) := [sign(w_1 \cdot x), \ldots, sign(w_k \cdot x)],\\
&g_S \in G : g_S(z) := sign(v_S \cdot z),\\
&g_T \in G : g_T(z) := sign(v_T \cdot z),\\
&\epsilon \geq 0
\end{aligned}$$

[6] If M is an orthogonal matrix then $\forall x_1, x_2 \in \mathcal{X}$ such that $\|x_1\| = \|x_2\|, \mu'_X(x_1) \leq c\mu'_X(x_2)$. For example, μ'_X is a spherical Gaussian.

such that

$$\max[R_S(g_S \circ f), R_T(\tilde{h}_{g_T \circ f})] \leq \epsilon. \tag{3.36}$$

We now state the target task risk bound for transferring representations with fine-tuning in our neural network example.

Theorem 3.10. *Suppose F, G, k, q and σ satisfy Assumption 3.7; F, G, μ_{XY}, μ'_{XY} and ϵ satisfy Assumption 3.9; and $\hat{f}$, μ'_X and c satisfy Assumption 3.8. Let $\theta_{\max} : \mathbb{R} \to \mathbb{R}$ be defined as*

$$\theta_{\max}(R) := \min[\pi\sqrt{2(k-1)}c(R+\epsilon), \pi].$$

Given $\hat{f}$ and $R_S(\hat{g}_S \circ \hat{f})$ estimated from S, let

$$\hat{F} := \{f \in F : \forall i, \|w_i\| = 1 \wedge |\theta(w_i, \hat{w}_i)| \leq \theta_{\max}(R_S(\hat{g}_S \circ \hat{f}))\} \tag{3.37}$$

and let $\omega : \mathbb{R} \to \mathbb{R}$ be defined as

$$\omega(R) := \frac{k}{\sigma^2}[1 - \cos\theta_{\max}(R)] + k(2q-1)\log_2 \frac{q}{1-q}. \tag{3.38}$$

Then $\exists \tilde{h}_{g_T \circ f} \in \tilde{H}_{G \circ \hat{F}}$ such that

$$R_T(\tilde{h}_{g_T \circ f}) \leq \epsilon \tag{3.39}$$

and

$$\forall \tilde{h} \in \tilde{H}_{G \circ \hat{F}}, KL(\tilde{h}||\tilde{h}_{\hat{g}_S \circ \hat{f}}) \leq \omega(R_S(\hat{g}_S \circ \hat{f})). \tag{3.40}$$

Proof of (3.39)*: $\exists \tilde{h}_{g_T \circ f} \in \tilde{H}_{G \circ \hat{F}}$ such that $R_T(\tilde{h}_{g_T \circ f}) \leq \epsilon$.*
Recall that w_i are the weight vectors for f and $\hat{w}_i$ are those for $\hat{f}$. Observe that for any w_i such that $w_i \cdot \hat{w}_i < 0$, we have $-w_i \cdot \hat{w}_i > 0$ and $-v_i sign(-w_i \cdot x) = v_i sign(w_i \cdot x)$. Combining this with the assumption (3.36), we conclude $\exists f \in F, g_S, g_T \in G$ such that

$$\forall i, w_i \cdot \hat{w}_i \geq 0 \tag{3.41}$$

and

$$\max[R_S(g_S \circ f), R_T(\tilde{h}_{g_T \circ f})] \leq \epsilon.$$

Let $\hat{g}_S(z) := sign(\hat{v}_S \cdot z)$. Let μ be a distribution on $\mathcal{X}$ satisfying the rotation invariance property (3.35) for $c = 1$, and let $p(\cdot)$ denote the probability of an event under μ. To prove $\tilde{h}_{g_T \circ f} \in \tilde{H}_{G \circ \hat{F}}$, by the definition of $\hat{F}$ from (3.37) and observing that $|\theta(w_i, \hat{w}_i)| \leq \pi$, it is sufficient to show

$$\forall i, |\theta(w_i, \hat{w}_i)| \leq \pi\sqrt{2(k-1)}c(R_S(\hat{g}_S \circ \hat{f}) + \epsilon). \tag{3.42}$$

We show (3.42) holds as follows:

$$
\begin{aligned}
&\frac{\max_i |\theta(w_i, \hat{w}_i)|}{\pi\sqrt{2(k-1)}} \\
&\leq p(v_S \cdot f(x) v_S \cdot \hat{f}(x) \leq 0) && (3.43)\\
&\leq p(v_S \cdot f(x) \hat{v}_S \cdot \hat{f}(x) \leq 0) && (3.44)\\
&\leq c p_S(v_S \cdot f(x) \hat{v}_S \cdot \hat{f}(x) \leq 0) && (3.45)\\
&= c p_S(y v_S \cdot f(x) y \hat{v}_S \cdot \hat{f}(x) \leq 0) && (3.46)\\
&\leq c[p_S(y v_S \cdot f(x) \leq 0) + p_S(y \hat{v}_S \cdot \hat{f}(x) \leq 0)] && (3.47)\\
&= c[R_S(g_S \circ f) + R_S(\hat{g}_S \circ \hat{f})] && (3.48)\\
&\leq c[\epsilon + R_S(\hat{g}_S \circ \hat{f})]. && (3.49)
\end{aligned}
$$

(3.43) holds by Lemma 3.12. (3.44) holds by Lemma 3.13, using $\forall i, w_i \cdot \hat{w}_i \geq 0$ as shown in (3.41). (3.45) uses the rotation invariance of μ'_X assumed in (3.35). (3.46) and (3.47) use basic laws of probability. (3.48) follows from using 0/1 loss, defined in (3.1). (3.49) uses the assumption $R_S(g_S \circ f) \leq \epsilon$ as stated in (3.36). □

Proof of (3.40)*:* $\forall \tilde{h} \in \tilde{H}_{G \circ \hat{F}}, KL(\tilde{h} || \tilde{h}_{\hat{g}_S \circ \hat{f}}) \leq \omega(R_S(\hat{g}_S \circ \hat{f}))$.
For any $\tilde{h}_{g \circ f} \in \tilde{H}_{G \circ \hat{F}}$,

$$
\begin{aligned}
&KL(\tilde{h}_{g \circ f} || \tilde{h}_{\hat{g}_S \circ \hat{f}}) \\
&= \sum_{i=1}^{k} KL(\mathcal{N}(w_i, \sigma^2 I) || \mathcal{N}(\hat{w}_i, \sigma^2 I)) + \sum_{i=1}^{k} KL((2\text{Bern}(q) - 1) v_i || (2\text{Bern}(q) - 1)(\hat{v}_S)_i).
\end{aligned} \quad (3.50)
$$

This is due to the form of hypotheses $\tilde{h}_{g \circ f}$ given in (3.7), and the fact that the KL divergence of a product distribution is the sum of the KL divergences of its component distributions. We separately upper bound both terms on the right hand side of (3.50), and apply the definition of $\omega(R)$ from (3.38).

$$
\begin{aligned}
&\sum_{i=1}^{k} KL(\mathcal{N}(w_i, \sigma^2 I) || \mathcal{N}(\hat{w}_i, \sigma^2 I)) \\
&= \frac{1}{2\sigma^2} \sum_{i=1}^{k} \|w_i - \hat{w}_i\|^2 && (3.51)\\
&= \frac{1}{2\sigma^2} \sum_{i=1}^{k} (\|w_i\|^2 + \|\hat{w}_i\|^2 - 2\|w_i\| \|\hat{w}_i\| \cos |\theta(w_i, \hat{w}_i)|) && (3.52)\\
&= \frac{1}{\sigma^2} \sum_{i=1}^{k} (1 - \cos |\theta(w_i, \hat{w}_i)|) && (3.53)\\
&\leq \frac{k}{\sigma^2} (1 - \cos \theta_{\max}(R_S(\hat{g}_S \circ \hat{f})). && (3.54)
\end{aligned}
$$

(3.51) uses the KL divergence of Gaussian distributions. (3.52) uses the law of cosines. (3.53) is because $\forall i, \|w_i\| = \|\hat{w}_i\| = 1$ by the definition of $\hat{F}$ in (3.37). (3.54) follows by the definition of $\hat{F}$ from (3.37) and the fact that $1 - \cos\theta$ is non-decreasing for $\theta \in [0, \pi]$.

$$\begin{aligned}
&\sum_{i=1}^{k} KL((2\text{Bern}(q) - 1)v_i || (2\text{Bern}(q) - 1)(\hat{v}_S)_i) \\
&\leq k(q \log_2 \frac{q}{1-q} + (1-q)\log_2 \frac{1-q}{q}) && (3.55) \\
&= k(2q-1)\log_2 \frac{q}{1-q}. && (3.56)
\end{aligned}$$

(3.55) uses the KL divergence of Bernoulli distributions. (3.56) is a simplification.

□

3.5 Applications

We show the utility of the risk bounds, and present a novel technique and experiments motivated by our theorems.

3.5.1 Using the Risk Bounds

The results described yield tighter bounds on risk when transferring representations from S, compared to learning T from scratch. Examples are shown in Figure 3.3.[7]

We set $\delta = 0.05$. For the left part of Figure 3.3, we use the example from Section 3.3.1 and set $n = 10$, $k = 5$. Learning T from scratch with H, we use the bound (3.8). The VC-dimension of a network of $|E|$ edges using the sign activation is $O(|E| \log |E|)$ [Shalev-Shwartz and Ben-David, 2014], where in our case $|E| = nk + k$. We use $VC(H) = |E| \log |E|$ in the chart. Transferring a representation from S to T without fine-tuning, we consider the limit $\epsilon \to 0, \hat{R}_S(\hat{g}_S \circ \hat{f}) \to 0$, $m_S \to \infty$, and hence $\omega(\cdot) \to 0$ by Theorem 3.5. Furthermore, $VC(G) \leq k$ since G is finite and hence $VC(G) \leq \log_2 |G|$ [Shalev-Shwartz and Ben-David, 2014]. We use the bound from Theorem 3.1.

For the right part of Figure 3.3, we use the example from Section 3.4.1 and set $\sigma^2 = \frac{1}{10}$, $k = 499$, $q = \frac{2}{3}$. Learning T from scratch we use the stochastic hypothesis class $\{\tilde{h}_{g \circ f} : f \in F \text{ such that } \forall i, \|w_i\| = 1, g \in G\}$ and a prior $\tilde{h}_0$ where $\forall i$, $w_i = \mathbf{0}$ and $v \in \{-1, 1\}^k$ is arbitrary.[8] Using (3.50), we have $KL(\tilde{h}||\tilde{h}_0) \leq 10k + \frac{k}{3}$. We apply the PAC-Bayes bound (3.27). Transferring a representation from S and fine-tuning on T, we consider the limit $\epsilon \to 0$, $\hat{R}_S(\hat{g}_S \circ \hat{f}) \to 0$, $m_S \to \infty$. We have $KL(\tilde{h}||\tilde{h}_{\hat{g}_S \circ \hat{f}}) \leq \frac{k}{3}$ by Theorem 3.10. We use the bound from Theorem 3.6.

[7] Note that VC-dimension risk bounds are known for being rather loose, while PAC-Bayesian bounds are tighter and hence yield non-trivial results in higher dimensions with fewer samples.

[8] This class is as expressive as $\tilde{H}_{G \circ F}$ but by setting $\|w_i\| = 1$ the KL divergence of all hypotheses from any prior is bounded, allowing a fair comparison to $\tilde{H}_{G \circ \hat{F}}$. The choice of $\tilde{h}_0$ minimizes worst case KL divergence to an hypothesis in the class.

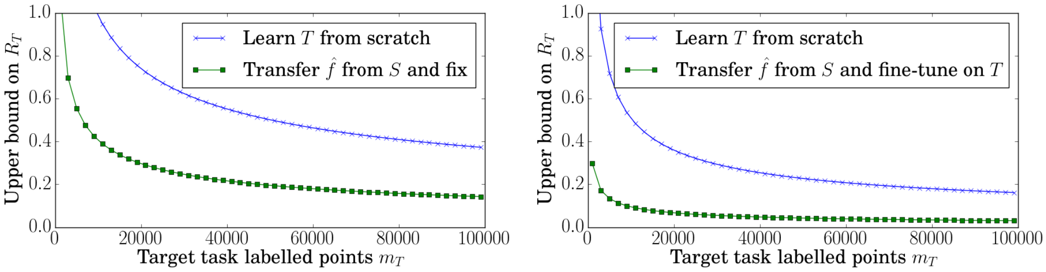


Figure 3.3: Risk bounds compared to learning T from scratch, without fine-tuning (left) and with fine-tuning (right). The two charts use different parameters (see Section 3.5.1).

3.5.2 Fine-Tuning through Regularization

We relax the hard constraint on $\hat{F}$ from Section 3.4.1 by using a modified loss function, which we find performs better in practice. Let y_i and $\hat{y}_i$ be the label and prediction respectively for the ith training point. In a fully-connected feedforward neural network with l layers of weights, let $W^{(j)}$ be the jth weight matrix, $\hat{W}^{(j)}$ be its estimate from S (excluding weights for bias units in both cases), and $\|\cdot\|_2$ be the entry-wise 2 norm. A typical loss function (3.57) used for training is composed of the sum of training set log loss and L2 regularization on the weights.

$$\sum_{i=1}^{m}[-y_i \log \hat{y}_i - (1-y_i)\log(1-\hat{y}_i)] + \frac{\lambda}{2}\sum_{j=1}^{l}\|W^{(j)}\|_2^2 \tag{3.57}$$

We replace the regularization penalty with (3.58).[9]

$$\sum_{j=1}^{l}[\frac{\lambda_1(j)}{2}\|W^{(j)} - \hat{W}^{(j)}\|_2^2 + \frac{\lambda_2(j)}{2}\|W^{(j)}\|_2^2] \tag{3.58}$$

This penalizes estimates of W far from the weights learned on S. Since we expect the tasks to share a low-level representation function (e.g. edge detectors for vision, word embeddings for text) but be distinct at higher levels (e.g. image components for vision, topics for text), we set $\lambda_1(\cdot)$ to be a decreasing function, while $\lambda_2(\cdot)$ controls standard L2 regularization. The technique is novel to our knowledge, although other approaches to transferring regularization between tasks exist [Evgeniou and Pontil, 2004; Raina et al., 2006; Argyriou et al., 2008; Ghifary et al., 2014].

[9] Basing our approach on (3.57), we follow the convention that weights connected to bias units are excluded from the regularization penalty. However, the inclusion of these weights in the $\|W^{(j)} - \hat{W}^{(j)}\|$ term of (3.58) is a plausible variant.

3.5.3 Experiments

We experiment on basic image and text classification tasks.[10] We show that learning algorithms motivated by our theoretical results can help to overcome a scarcity of labeled target task data. Note that we do not replicate the conditions specified in our theorems, nor do we attempt extensive tuning to achieve state-of-the-art performance.

We randomly partition label classes into sets S_+ and S_-, where $|S_+| = |S_-|$.[11] We construct T_+ by randomly picking from S_+ up to $\gamma := \frac{|S_+ \cap T_+|}{|S_+|}$, then randomly picking from S_- such that $|T_+| = |T_-|$. We let S be the task of distinguishing between S_+ and S_- and T be that of distinguishing T_+ and T_-. Constructing S_+ and T_+ as disjunctions of classes means that the class labels are a perfect representation shared between S and T.

We compare the accuracy on T of four options:

- learn T from scratch (BASE)
- transfer $\hat{f}$ from S, fine-tune f and train g on T using (3.58) (FINE-TUNE $\hat{f}$)
- transfer $\hat{f}$ from S and fix, train g on T (FIX $\hat{f}$)[12]
- transfer $\hat{g}_S \circ \hat{f}$ from S and fix (FIX $\hat{g}_S \circ \hat{f}$).[13]

We use $\lambda_1(1) = \lambda_2(2) = \lambda := 1$,[14] $\lambda_1(2) = \lambda_2(1) = 0$, $m_T = 500$ and the sigmoid activation function. For MNIST we use raw pixel intensities, a $784 \times 50 \times 1$ network and $m_S = 50000$. For NEWSGROUPS we use TF-IDF weighted counts of most frequent words, a $2000 \times 50 \times 1$ network and $m_S = 15000$. We use conjugate gradient optimization with 200 iterations.

The results are shown in Table 3.1.[15] When the tasks are non-identical, FINE-TUNE $\hat{f}$ is mostly the strongest, but performs better on MNIST than on NEWSGROUPS. FIX $\hat{f}$ outperforms BASE when $\gamma \geq 0.8$, i.e. when the tasks are similar. While FIX $\hat{f}$ outperforms FIX $\hat{g}_S \circ \hat{f}$ when the tasks are non-identical on MNIST, on NEWSGROUPS there is no evidence of benefit. When the tasks are identical, unsurprisingly FIX $\hat{g}_S \circ \hat{f}$ is the strongest.

It appears that learning an MNIST digit requires a dense weight vector and so $\hat{W}^{(1)}$ tends to encode single digits, which helps transferrability. However, it appears

[10] The MNIST and 20 Newsgroups datasets are available at http://yann.lecun.com/exdb/mnist and http://qwone.com/˜jason/20Newsgroups respectively.

[11] For MNIST there are 10 label classes and for 20 Newsgroups there are 20. In both cases the classes are approximately balanced. Note that we ignore the hierarchical structure of the 20 Newsgroups classes, which likely contributes to the lower accuracies reported for all methods for this dataset relative to MNIST.

[12] i.e. logistic regression with L2 regularization and $\hat{f}$ fixed.

[13] Used to isolate the benefit of transferring $\hat{f}$ rather than $\hat{g}_S \circ \hat{f}$.

[14] We explored tuning λ to lift the performance of BASE on MNIST, but found that the results did not materially improve. Potentially $\lambda_1(j)$ and $\lambda_2(j)$ in (3.58) could be tuned with cross validation on the target task.

[15] For $\gamma = 1$, the tasks are identical. We do not consider $\gamma < 0.5$, since that is equivalent to $1 - \gamma$ with the definitions of T_+ and T_- swapped.

Table 3.1: Evaluation of transferring representations. Entries are the test set accuracy of the technique (row) for the task (column) averaged over 10 trials, with the best result for each task bolded.

Technique	MNIST, $\gamma =$			Newsgroups, $\gamma =$		
	0.6	0.8	1	0.6	0.8	1
Base	88.4	87.9	87.9	**62.6**	63.2	66.1
Fine-tune $\hat{f}$	**91.9**	**93.9**	95.4	62.3	**72.3**	83.3
Fix $\hat{f}$	87.5	92.3	97.3	52.2	69.6	83.3
Fix $\hat{g}_S \circ \hat{f}$	67.4	85.6	**98.1**	55.5	70.7	**83.6**

that since we may learn a newsgroup with a sparse weight vector, $\hat{W}^{(1)}$ tends to encode disjunctions of newsgroups which somewhat reduces transferrability. When transferring representations does work, fine-tuning using the regularization penalty proposed in (3.58) improves performance.

3.6 Conclusion

We developed sufficient conditions for the successful transfer of representation functions both with and without fine-tuning. This is a step towards a principled explanation of the empirical success achieved by such techniques. A promising direction for future work is generalizing the neural network architectures considered (e.g. using multiple hidden layers) and relaxing the distributional assumptions required. Furthermore, in the fine-tuning case it may be possible to upper bound the target task generalization error of hypotheses in $G \circ \hat{F}$ using another measure such as the Rademacher complexity of $G \circ \hat{F}$, eliminating the need for stochastic hypotheses.

We proposed a novel form of regularization for neural network training motivated by our theoretical results, which penalizes divergence from source task weights and is stricter for lower-level weights. We validated this technique through applications to image and text classification. Future directions include experiments on more challenging tasks using deeper and more tailored network architectures (e.g. convolutional neural networks).

3.7 Appendix

We state and prove lemmas used in the proofs of Theorems 3.5 and 3.10.

Lemma 3.11. *Suppose F is a representation function class which satisfies Assumption 3.2. Let $f, \hat{f} \in F$, and let $v, \hat{v}, d \in \{-1, 1\}^k$. Suppose $\hat{f}$, μ_X, μ'_X and c satisfy Assumption 3.3. Then*

$$p_T(d * v \cdot f(x) d * \hat{v} \cdot \hat{f}(x) \leq 0) \leq c p_S(v \cdot f(x) \hat{v} \cdot \hat{f}(x) \leq 0).$$

Proof. Recall the definition of M in (3.16). Suppose there is a map $\phi : \mathbb{R}^n \to \mathbb{R}^n$, which is invertible, and satisfies

$$\forall x \in \mathcal{X}, \|Mx\| = \|M\phi(x)\| \tag{3.59}$$

and

$$d * v \cdot f(x) d * \hat{v} \cdot \hat{f}(x) = v \cdot f(\phi(x)) \hat{v} \cdot \hat{f}(\phi(x)). \tag{3.60}$$

Then the result follows since $\mu_X(x) \leq c\mu'_X(\phi(x))$ by (3.17).

Such a map is

$$\phi(x) := (M^T M)^{-1} M^T \tilde{d} * (Mx), \tag{3.61}$$

where $\tilde{d}$ is a vector of length $2k$ defined as

$$\tilde{d} := [d_1, d_1, \dots, d_k, d_k].$$

Rearranging (3.61) and using the fact that M is full rank we see that (3.59) is satisfied.

By the definition of M in (3.16) and $\phi(x)$ in (3.61), we have $\forall i$,

$$w_i \cdot \phi(x) = d_i w_i \cdot x \tag{3.62}$$

and

$$(\alpha_i \hat{w}_i - \beta_i w_i) \cdot \phi(x) = d_i(\alpha_i \hat{w}_i - \beta_i w_i) \cdot x,$$

and hence

$$\hat{w}_i \cdot \phi(x) = d_i \hat{w}_i \cdot x \tag{3.63}$$

for $\alpha_i, \beta_i \neq 0$.

Therefore we may show (3.60) as follows:

$$\begin{aligned} & d * v \cdot f(x) d * \hat{v} \cdot \hat{f}(x) \\ &= v \cdot d * f(x) \hat{v} \cdot d * \hat{f}(x) && (3.64) \\ &= v \cdot f(\phi(x)) \hat{v} \cdot d * \hat{f}(\phi(x)) && (3.65) \\ &= v \cdot f(\phi(x)) \hat{v} \cdot \hat{f}(\phi(x)). && (3.66) \end{aligned}$$

(3.64) is a property of the elementwise and dot products. For (3.65), apply (3.62) and the fact that a is an odd function to yield $a(w_i \cdot \phi(x)) = a(d_i w_i \cdot x) = d_i a(w_i \cdot x)$. Similarly for (3.66), apply (3.63) and the fact that a is an odd function to conclude $a(\hat{w}_i \cdot \phi(x)) = a(d_i \hat{w}_i \cdot x) = d_i a(\hat{w}_i \cdot x)$. □

Lemma 3.12. *Suppose F is a representation function class which satisfies Assumption 3.7. Let $f, \hat{f} \in F$ and let $v \in \{-1, 1\}^k$. Let μ be a distribution on $\mathcal{X}$, and let $p(\cdot)$ be the probability of an event under μ. Suppose that $\forall i, w_i \cdot \hat{w}_i \geq 0$, and that $\hat{f}$, μ and $c = 1$ satisfy Assumption 3.8. Then*

$$\frac{\max_i |\theta(w_i, \hat{w}_i)|}{\pi\sqrt{2(k-1)}} \leq p(v \cdot f(x) v \cdot \hat{f}(x) \leq 0).$$

Proof. For $i \in \{1, \cdots, k\}$, let

$$\begin{aligned}
&v_{-i} := [v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_k],\\
&f_i(x) := sign(w_i \cdot x),\\
&f_{-i}(x) := [f_1(x), \ldots, f_{i-1}(x), f_{i+1}(x), \ldots, f_k(x)],\\
&\hat{f}_i(x) := sign(\hat{w}_i \cdot x) \text{ and}\\
&\hat{f}_{-i}(x) := [\hat{f}_1(x), \ldots, \hat{f}_{i-1}(x), \hat{f}_{i+1}(x), \ldots, \hat{f}_k(x)].
\end{aligned}$$

$\forall i \in \{1, \cdots, k\}$ we have

$$\begin{aligned}
&p(v \cdot f(x) v \cdot \hat{f}(x) \leq 0)\\
&\geq p(v \cdot f(x) v \cdot \hat{f}(x) < 0)\\
&\geq p(v_{-i} \cdot f_{-i}(x) = 0) p(v \cdot f(x) v \cdot \hat{f}(x) < 0 | v_{-i} \cdot f_{-i}(x) = 0)\\
&= p(v_{-i} \cdot f_{-i}(x) = 0) p(v_i f_i(x) v_{-i} \cdot \hat{f}_{-i}(x) + f_i(x)\hat{f}_i(x) < 0 | v_{-i} \cdot f_{-i}(x) = 0)\\
&= p(v_{-i} \cdot f_{-i}(x) = 0)[p(v_i f_i(x) v_{-i} \cdot \hat{f}_{-i}(x) < -1, f_i(x)\hat{f}_i(x) = 1 | v_{-i} \cdot f_{-i}(x) = 0)\\
&\qquad + p(v_i f_i(x) v_{-i} \cdot \hat{f}_{-i}(x) < 1, f_i(x)\hat{f}_i(x) = -1 | v_{-i} \cdot f_{-i}(x) = 0)]\\
&\geq p(v_{-i} \cdot f_{-i}(x) = 0)[p(v_i f_i(x) v_{-i} \cdot \hat{f}_{-i}(x) < -1, f_i(x)\hat{f}_i(x) = -1 | v_{-i} \cdot f_{-i}(x) = 0)\\
&\qquad + p(v_i f_i(x) v_{-i} \cdot \hat{f}_{-i}(x) < 1, f_i(x)\hat{f}_i(x) = -1 | v_{-i} \cdot f_{-i}(x) = 0)] \qquad (3.67)\\
&= p(v_{-i} \cdot f_{-i}(x) = 0)[p(v_i f_i(x) v_{-i} \cdot \hat{f}_{-i}(x) < -1, f_i(x)\hat{f}_i(x) = -1 | v_{-i} \cdot f_{-i}(x) = 0)\\
&\qquad + p(v_i f_i(x) v_{-i} \cdot \hat{f}_{-i}(x) > -1, f_i(x)\hat{f}_i(x) = -1 | v_{-i} \cdot f_{-i}(x) = 0)] \qquad (3.68)\\
&= p(v_{-i} \cdot f_{-i}(x) = 0) p(f_i(x)\hat{f}_i(x) = -1 | v_{-i} \cdot f_{-i}(x) = 0)\\
&= p(v_{-i} \cdot f_{-i}(x) = 0) p(f_i(x)\hat{f}_i(x) = -1) \qquad (3.69)\\
&= \binom{k-1}{\frac{k-1}{2}} (\frac{1}{2})^{k-1} \frac{|\theta(w_i, \hat{w}_i)|}{\pi} \qquad (3.70)\\
&\geq \frac{2^{k-1}}{\sqrt{2(k-1)}} (\frac{1}{2})^{k-1} \frac{|\theta(w_i, \hat{w}_i)|}{\pi} \qquad (3.71)\\
&= \frac{|\theta(w_i, \hat{w}_i)|}{\pi \sqrt{2(k-1)}}.
\end{aligned}$$

(3.67) follows since μ satisfies the rotation invariance property (3.35) for $c = 1$ and $w_i \cdot \hat{w}_i \geq 0$. (3.68) and (3.69) use the fact that μ satisfies the rotation invariance property (3.35) for $c = 1$. (3.70) uses rotation invariance and the fact that k is odd. (3.71) is a standard lower bound for the central binomial coefficient. The other lines use basic simplifications and laws of probability. □

Lemma 3.13. *Suppose F is a representation function class which satisfies Assumption 3.7. Let $f, \hat{f} \in F$ and let $v, \hat{v} \in \{-1, 1\}^k$. Let μ be a distribution on $\mathcal{X}$, and let $p(\cdot)$ be the probability of an event under μ. Suppose that $\forall i, w_i \cdot \hat{w}_i \geq 0$, and that $\hat{f}$, μ and $c = 1$ satisfy*

Assumption 3.8. Then

$$p(v \cdot f(x) v \cdot \hat{f}(x) \leq 0) \leq p(v \cdot f(x) \hat{v} \cdot \hat{f}(x) \leq 0).$$

Proof. Let $\mathbb{E}[\cdot] := \mathbb{E}_{X \sim \mu}[\cdot]$. Let $f_i(x) := sign(w_i \cdot x)$ and $\hat{f}_i(x) := sign(\hat{w}_i \cdot x)$. Let $\tilde{f} \in \{-1,1\}^k$ and

$$p(\tilde{f}) := p([f_1(x)\hat{f}_1(x), \ldots, f_k(x)\hat{f}_k(x)] = \tilde{f}).$$

Let $d := \hat{v} * v \in \{-1,1\}^k$ and

$$\Delta(x) := \mathbf{1}(v \cdot f(x) \hat{v} \cdot \hat{f}(x) \leq 0) - \mathbf{1}(v \cdot f(x) v \cdot \hat{f}(x) \leq 0).$$

Assume $\hat{v} \neq v$ (if $\hat{v} = v$ then the lemma clearly holds). Let $a(\tilde{f}) := \sum_{i=1}^{k} \mathbf{1}(\tilde{f}_i = 1)$ and let $i^* := \min_{i: d_i = -1} i$. Let

$$\tilde{F} := \{\tilde{f} \in \{-1,1\}^k : a(\tilde{f}) > a(d * \tilde{f}) \vee (a(\tilde{f}) = a(d * \tilde{f}) \wedge \tilde{f}_{i^*} = 1)\}. \tag{3.72}$$

Let

$$\Phi(a) := \frac{1}{2^{k-1}} \sum_{b=0}^{\lfloor k/2 \rfloor} \sum_{j=\lceil a/2+b/2-k/4 \rceil}^{b} \binom{a}{j} \binom{k-a}{b-j}.$$

The term b counts coordinates where $v_i \hat{f}_i(x) = sign(v \cdot f(x))$, while j counts those where $v_i f_i(x) = sign(v \cdot f(x))$ and $f_i(x) = \hat{f}_i(x)$.

$$\begin{aligned}
& p(v \cdot f(x) \hat{v} \cdot \hat{f}(x) \leq 0) - p(v \cdot f(x) v \cdot \hat{f}(x) \leq 0) \\
&= \mathbb{E}[\mathbf{1}(v \cdot f(x) \hat{v} \cdot \hat{f}(x) \leq 0)] - \mathbb{E}[\mathbf{1}(v \cdot f(x) v \cdot \hat{f}(x) \leq 0)] \\
&= \mathbb{E}[\Delta(x)] && (3.73) \\
&= \sum_{\tilde{f} \in \tilde{F}} p(\tilde{f}) \mathbb{E}[\Delta(x) | \tilde{f}] + p(d * \tilde{f}) \mathbb{E}[\Delta(x) | d * \tilde{f}] && (3.74) \\
&= \sum_{\tilde{f} \in \tilde{F}} [p(\tilde{f}) - p(d * \tilde{f})] \mathbb{E}[\Delta(x) | \tilde{f}] && (3.75) \\
&= \sum_{\tilde{f} \in \tilde{F}} [p(\tilde{f}) - p(d * \tilde{f})][p(v \cdot f(x) v \cdot \hat{f}(x) \leq 0 | d * \tilde{f}) - p(v \cdot f(x) v \cdot \hat{f}(x) \leq 0 | \tilde{f})] \\
& && (3.76) \\
&= \sum_{\tilde{f} \in \tilde{F}} [p(\tilde{f}) - p(d * \tilde{f})][\Phi(a(d * \tilde{f})) - \Phi(a(\tilde{f}))] && (3.77) \\
&\geq 0. && (3.78)
\end{aligned}$$

(3.73) uses linearity of expectation. (3.74) uses the law of total expectation and the definition of $\tilde{F}$ from (3.72).

(3.75) holds since

$$\begin{aligned}
&\mathbb{E}[\Delta(x)|d * \tilde{f}] \\
&= \sum_{f' \in \{-1,1\}^k} p(f(x) = f'|d * \tilde{f})\mathbb{E}[\Delta(x)|d * \tilde{f}, f(x) = f'] \\
&= -\sum_{f' \in \{-1,1\}^k} p(f(x) = f'|d * \tilde{f})\mathbb{E}[\Delta(x)|\tilde{f}, f(x) = f'] \\
&= -\sum_{f' \in \{-1,1\}^k} p(f(x) = f'|\tilde{f})\mathbb{E}[\Delta(x)|\tilde{f}, f(x) = f'] \\
&= -\mathbb{E}[\Delta(x)|\tilde{f}]
\end{aligned}$$

due to the fact that μ satisfies the rotation invariance property (3.35) for $c = 1$.

(3.76) holds by expanding $\Delta(x)$, linearity of expectation, and a similar argument to the one used to show (3.75), yielding

$$p(v \cdot f(x)\hat{v} \cdot \hat{f}(x) \leq 0|\tilde{f}) = p(v \cdot f(x)v \cdot \hat{f}(x) \leq 0|d * \tilde{f}).$$

(3.77) holds because μ satisfies the rotation invariance property (3.35) for $c = 1$, and k is odd.

For (3.78), $\Phi(a(d * \tilde{f})) - \Phi(a(\tilde{f}))$ is non-negative since $a(\tilde{f}) \geq a(d * \tilde{f})$ and Φ is non-increasing. $p(\tilde{f}) - p(d * \tilde{f})$ is also non-negative because μ satisfies the rotation invariance property (3.35) for $c = 1$, and $\forall i, w_i \cdot \hat{w}_i \geq 0$. □

Part II

The Use and Limits of Representation Learning for Fairness

Chapter 4

Costs and Benefits of Fair Representation Learning

4.1 Introduction

Machine learning algorithms are used to make or support decisions in a wide variety of contexts including financial and judicial risk assessments, applicant screening for employment, and online ad selection. Concerns about the fairness of these algorithms have arisen as a result [O'Neil, 2017; Barocas and Selbst, 2016; Angwin et al., 2016; Datta et al., 2015]. Decisions made by machine learning algorithms typically cannot be controlled or interpreted as straightforwardly as those made by rule-based systems. Furthermore, artefacts of previous discrimination in an algorithm's training data may affect its decisions. Researchers have responded by developing techniques to incorporate fairness into the design of machine learning algorithms [Barocas et al., 2018; Zliobaite, 2015; Romei and Ruggieri, 2014]. While these techniques often focus on achieving *group fairness* – i.e. not discriminating against particular groups – another important consideration is *individual fairness* – i.e. giving similar treatment to individuals who are similar [Dwork et al., 2012].

The problem of fair classification (see Figure 4.1(a)) involves making a *decision* (e.g. whether to grant a loan) based on an *input* (e.g. individual financial and demographic information) which accurately predicts a *target* of interest (e.g. loan default), while at the same time avoiding discrimination on the basis of an individual's group membership (e.g. race, gender) encoded in a *sensitive* variable. A single party, the *data user*, is trusted to access the sensitive variable in training and is responsible for making decisions that appropriately consider accuracy and fairness.

In contrast (see Figure 4.1(b)), the problem of fair representation learning involves producing a *cleaned* representation of the input which remains useful for predicting the target, but suppresses information which could be used to discriminate based on the sensitive variable. We now assume the data user is not trusted to access the sensitive variable in training, which may be appropriate if the data user could be either *adversarial*, i.e. interested in being unfair, or *indifferent*, i.e. interested only in target accuracy [Madras et al., 2018]. This problem setting involves three parties: a *data producer* who cleans the input data, a *data user* who makes decisions from the cleaned

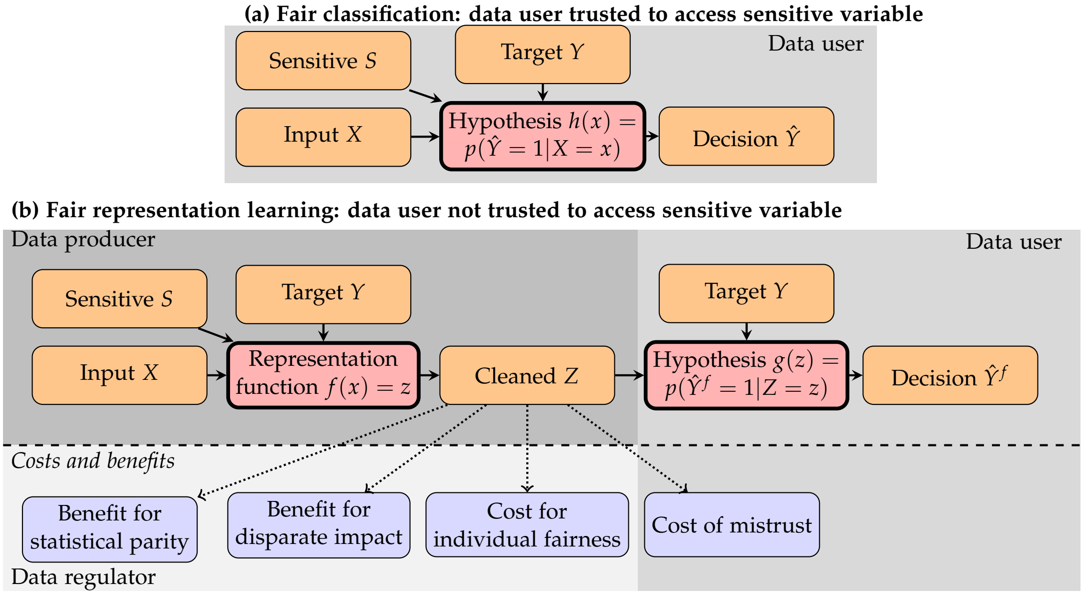

(a) Fair classification: data user trusted to access sensitive variable
Data user
Sensitive S
Target Y
Input X
Hypothesis $h(x) = p(\hat{Y} = 1|X = x)$
Decision $\hat{Y}$
(b) Fair representation learning: data user not trusted to access sensitive variable
Data producer
Data user
Sensitive S
Target Y
Target Y
Input X
Representation function $f(x) = z$
Cleaned Z
Hypothesis $g(z) = p(\hat{Y}^f = 1|Z = z)$
Decision $\hat{Y}^f$
Costs and benefits
Benefit for statistical parity
Benefit for disparate impact
Cost for individual fairness
Cost of mistrust
Data regulator

Figure 4.1: Summary of (a) fair classification and (b) fair representation learning, showing train time data processing for both, and costs and benefits of (b).

data, and a *data regulator* who oversees fair use of the data. For example, when deciding whether to give an individual a loan, the data producer might be a credit bureau, the data user a bank and the data regulator a government authority. Even within an organization, this separation of concerns has the advantage of providing checks and balances.

As we shall see, the fair representation learning approach offers both costs and benefits, and may be appropriate in some situations but not others. It is most useful when the data user is not trusted to achieve fairness. For example, if the data user has a financial incentive to prioritize decisions based on accurate predictions of the target variable regardless of fairness constraints, or if data is being publicly released and the objectives of data users may be diverse and hard to foresee. If there is good reason to trust a single data user to make decisions that consider both accuracy and fairness – for example, if the decisions are reported to regulators and violations of fairness constraints attract enforceable penalties – then alternative approaches to fair classification (e.g. [Menon and Williamson, 2018]) may be preferable.

4.1.1 Contributions of This Chapter

This chapter offers contributions that are have both scientific and policy significance, and are technically novel.

Scientific significance: A plethora of methods use fair representation learning [Zemel et al., 2013; Feldman et al., 2015; Edwards and Storkey, 2016; Louizos et al., 2016; Johndrow and Lum, 2017; Beutel et al., 2017; Madras et al., 2018] as a *technique* for fair classification. Recent work [Menon and Williamson, 2018] has solved in analytical form a canonical version of the fair classification problem. Is fair representation learning then to be relegated to a sub-optimal technique for a problem better solved through other means? Developing more fair representation learning techniques does not address this question. Instead, we show that fair representation learning in fact solves a different *problem* – i.e. how to guarantee that decisions made by an untrusted data user can be accurate but will not be unfair – and quantify the costs and benefits of such representations in terms of fairness and utility. This represents a progression in our scientific understanding, given that this problem had never previously been formally posed or analyzed.

Policy significance: Our approach makes possible a governance model involving a separation of concerns between a data producer, data user and data regulator (previous work assumes a single trusted data user). The model enables a regulator to guarantee fairness even if the data user is adversarial. This is an advance in the regulation of algorithmic fairness, given that no alternatives currently exist in the realistic setting where a data user is not trusted to be fair.

Novel technical results: We formalize the problem of fair representation learning as distinct from fair classification. By stating the data producer's optimization problem in (4.5) and showing that a proxy problem can be solved without access to the target variable (Theorem 4.6), we derive a principled way to select a fair representation learning objective function (this is heuristic in prior work).

We present a novel quantification of the costs of using a given representation (Section 4.4), a topic which had not previously been investigated. We identify costs both in terms of the accuracy-fairness trade-off (i.e. the *cost of mistrust* given in closed form in Theorem 4.8 and bounded without requiring access to the target variable in Theorem 4.9), and in terms of individual fairness (Theorem 4.13).

We present novel guarantees of the benefits of a given representation (Section 4.5). We do this for two common measures of fairness: *statistical parity* (Theorem 4.14) and *disparate impact* (Theorem 4.15), by computing the unfairness of an optimal adversary. Conditioning on the target variable, our analysis can be also be used to guarantee quantified versions of two other well-known fairness definitions: *equality of opportunity* and *equalized odds*.

4.1.2 Structure of This Chapter

The remainder of this chapter is structured as follows. We review related work in Section 4.2. We present our formalization of the problem of fair representation learning in Section 4.3. In Sections 4.4 and 4.5 we quantify the costs and benefits respectively of using a particular representation. In Section 4.6, we present experiments which demonstrate that our formalization of the fair representation learning problem can be used in practice, and which illustrate how the costs and benefits we identified can be estimated and interpreted. We conclude in Section 4.7. In the Appendix (Section 4.8) we provide proofs of our theoretical results (proof sketches are included in the main text), along with a summary of the problems, costs and benefits we consider, and examples of the cost of mistrust.

4.2 Background

We summarize prior works on quantitative definitions of fairness, and techniques for fair classification and fair representation learning, which are relevant to the theoretical analysis of fair representation learning presented in this chapter. Chapters 5 and 6 contain further details of existing research on fairness in machine learning related to the topics covered by those chapters respectively.

4.2.1 Quantitative Definitions of Fairness

Defining fairness is a deep and complex topic considered across several disciplines such as philosophy, law, politics and psychology (for example, see [Rawls, 1971] for one influential attempt). In order to assess the effects of algorithms which make or inform decisions about people's lives, we require *quantitative* definitions of fairness. While the accuracy of an algorithm in predicting the target variable is one potential aggregate measure of fairness, it is not sufficient for understanding the algorithm's impacts on particular groups in the population. Several proposed quantitative definitions of fairness formalize the idea of avoiding discrimination on the basis of a particular kind of group membership, such as race or gender. Three types of definition have emerged, which we state informally (see [Mitchell and Shadlen, 2018] for further details):

- **Parity**: Predictions should be similar for different groups
- **Independence**: Predictions should be independent of group membership
- **Causality**: Predictions should not be caused by group membership.

While each of these approaches has its advantages, our analysis focuses on definitions based on parity. A predictive model that achieves parity between groups is mathematically equivalent to one that is independent of group membership (see p. 43 of [Barocas et al., 2018]). However, (dis)parity may be measured on a continuous scale, unlike an all-or-nothing statement about independence. Unlike causality-based definitions [Kusner et al., 2017], parity measures can be computed using only an algorithm's outputs without the knowledge of its functional form, so that external auditing can be carried out without the co-operation of the algorithm's owner. Parity measures also do not require the selection of variables that are permitted to cause decisions (known as *resolving variables* [Kilbertus et al., 2017]), which potentially could include proxies for group membership (e.g. 'redlining' where neighborhood is used a proxy for race). Finally, parity-based measures are arguably the simplest to understand for a lay audience, which is significant given the risk of excluding participants from non-quantitative backgrounds in debates about fairness [Mitchell and Shadlen, 2018].

Several parity measures compare an algorithm's average decisions for different groups, i.e. the expected values of the decision variable conditioned on membership of particular groups. For example, we may take the difference between the average decisions for two groups – known as *statistical parity* [Calders and Verwer, 2010; Dwork et al., 2012] – or the ratio of the average decisions for two groups – known as *disparate impact* [United States Equal Opportunity Employment Commission, 1978; Feldman et al., 2015]. We may wish to compute a parity measure only on a population subset. If the population subset consists of individuals who are similar according to some metric, we have *individual fairness*, also known as avoiding *disparate treatment* [Dwork et al., 2012; Mitchell and Shadlen, 2018]. Constructing subsets by conditioning on particular values of the target variable yields variants [Hardt et al., 2016] such as *equality of opportunity* (conditioning only on the positive class) and *equalized odds* (conditioning separately on the positive and negative classes).[1]

There is no straightforward answer to the question of which subsets to measure parity on. Conditioning on the target variable is appealing since it allows us to measure whether an algorithm's tendency to make prediction errors differs between groups, an intuitive approach to measuring fairness. However, there are also some potential disadvantages to conditioning on the target variable. If the training data labels are collected in a way that is discriminatory towards one group [Barocas and Selbst, 2016], conditioning on the target variable may not be appropriate [Zafar et al., 2017a]. Another issue is that the target variable may be affected by a complex historical process which has disadvantaged one group. For example, in a criminal justice context the target variable may measure whether an individual reoffended, and certain groups may have higher reoffence rates in the training data as a result of long-term structural disadvantage. Decisions which achieve parity between groups conditioned on the target variable but not overall may reinforce this disadvantage, for example by justifying higher incarceration rates for some groups compared to others. This chapter focuses on parity measures which apply to the whole population (i.e. statistical parity and disparate impact), but we flag where our results can also be straightforwardly applied to parity measures conditioned on population subsets conditioned on the target variable (i.e. approximations of equality of opportunity and equalized odds).

Several mathematical results have shown that, for a particular set of fairness definitions, it is impossible for a predictive model to simultaneously satisfy all definitions in the set [Chouldechova, 2017; Kleinberg et al., 2017b; Lipton et al., 2018; Pleiss et al., 2017]. Within a particular context, different definitions are aligned to the interests of particular stakeholders [Nayaranan, 2018]. Furthermore, when predictions are also measured on their accuracy, the definitions of accuracy and fairness are in general not aligned [Corbett-Davies et al., 2017; Menon and Williamson, 2018; Corbett-Davies and Goel, 2018]. We explore the relationships between different measures of fairness, and between fairness and accuracy, in more detail in Chapter 5.

[1] Satisfying equalized odds has also previously been referred to as avoiding *disparate mistreatment* [Zafar et al., 2017a]. A formal definition of equalized odds is given in Chapter 5.

4.2.2 Techniques for Fair Classification and Fair Representation Learning

Recent work on quantitative fairness has, in addition to proposing fairness definitions, developed techniques for fair classification. These techniques can be divided into three categories [Mitchell and Shadlen, 2018]:

- **Pre-processing**: modify the data that the algorithm learns from, i.e. fair representation learning (e.g. [Zemel et al., 2013])
- **In-processing**: modify the algorithm's objective function to incorporate a fairness constraint or penalty (e.g. [Menon and Williamson, 2018])[2]
- **Post-processing**: modify the predictions produced by the algorithm (e.g. [Hardt et al., 2016]).

Several techniques have been proposed to achieve fair representation learning. One approach to fair representation learning is to design the cleaned variable Z such that the distributions of Z conditioned on different values of the sensitive variable S are similar [Feldman et al., 2015; Johndrow and Lum, 2017]. In addition to this requirement, the pre-processing procedure may optimize the independence of Z and S [Louizos et al., 2016]. Another approach to fair representation learning is to design Z such that it is maximally informative about the target variable Y, subject to a constraint that it is uninformative about S [Ghassami et al., 2018]. Adversarial approaches [Edwards and Storkey, 2016; Beutel et al., 2017; Madras et al., 2018] use a neural network to learn a representation function such that an adversary network cannot accurately predict the sensitive variable from the cleaned data. A problem variant, where the target is also modified and the input is discrete, has been formulated as a convex optimization problem [Calmon et al., 2017].

What existing approaches to fair representation learning typically do not offer (Theorem 4.1 from [Feldman et al., 2015] is an exception) is a guarantee that all uses of the cleaned data will be fair, or a quantification of the costs of the cleaning process. We seek to provide a stronger theoretical foundation for fair representation learning. This objective is similar in spirit to that of privacy aware learning, which is concerned with the mathematical trade-off between the privacy and utility of data [Wainwright et al., 2012]. We also show that fair representation learning in fact addresses a problem that is distinct from fair classification, which is of interest when the data user is not trusted to access the sensitive variable.

4.3 Fair Classification vs Fair Representation Learning

We introduce and compare the problems of fair classification and fair representation learning. This formal comparison is itself novel and is necessary for our subsequent analysis of the costs and benefits of fair representation learning.

[2] Several over works have proposed variants of this approach, including: [Donini et al., 2018; Zafar et al., 2017b,a; Dwork et al., 2018; Bechavod and Ligett, 2017].

4.3.1 Fair Classification

In fair classification (Figure 4.1(a)), the data user trains on samples of input variable X, target variable Y and sensitive variable S. The samples are drawn from a distribution over $\mathcal{X} \times \mathcal{Y} \times \mathcal{S}$, where $\mathcal{X}$ is the set of possible inputs, $\mathcal{Y}$ is the set of possible labels and $\mathcal{S}$ is the set of possible sensitive variable values. We focus on the setting where $\mathcal{Y} \in \{0,1\}$, corresponding to binary classification, and $\mathcal{S} \in \{0,1\}$, corresponding to some common sensitive variable examples such as gender or race. Let $\pi_Y := p(Y=1)$ and $\pi_S := p(S=1)$ be prior probabilities, and $\eta_Y(x) := p(Y=1|X=x)$ and $\eta_S(x) := p(S=1|X=x)$ be conditional probabilities, for the positive classes of Y and S respectively.

The data user learns a stochastic hypothesis $h : \mathcal{X} \to [0,1]$ which is used to construct decision variable $\hat{Y} \in \{0,1\}$, where $h(x) := p(\hat{Y}=1|X=x)$. Let $\mu_{XYS\hat{Y}}$ be the joint distribution of the input, target, sensitive and decision variables.

At test time, the data user makes a decision using a sample of X, which may contain information about S. The quality of an hypothesis h in predicting Y can be measured by a risk $R_Y : [0,1]^{\mathcal{X}} \to [0,1]$, where we prefer hypotheses with a small value of $R_Y(h)$. A common choice is the cost-sensitive risk.

Definition 4.1 (Cost-sensitive risk [Elkan, 2001; Zhao et al., 2013; Menon and Williamson, 2018])**.** *The cost-sensitive risk of hypothesis h with respect to Y is*

$$R_Y(h) := \pi_Y(1-c_Y)p(\hat{Y}=0|Y=1) + (1-\pi_Y)c_Y p(\hat{Y}=1|Y=0)$$

where $c_Y \in [0,1]$, $p(\hat{Y}=0|Y=1)$ *is known as the* false negative rate *and* $p(\hat{Y}=1|Y=0)$ *as the* false positive rate*.*

We also wish to ensure that the hypothesis we learn is fair. Two common fairness measures are statistical parity and disparate impact, which compare outcomes for different sensitive variable groups using their difference and ratio respectively. In the analysis that follows we focus on the case where statistical parity and disparate impact are computed on the joint distribution $\mu_{XYS\hat{Y}}$. However, computing these measures only on part of the distribution yields other variants of interest, such as conditioning on $Y=1$ for quantified versions of equality of opportunity, or conditioning separately on $Y=1$ and $Y=0$ for quantified versions of equalized odds.

Definition 4.2 (Statistical parity [Calders and Verwer, 2010; Dwork et al., 2012])**.** *The statistical parity of an hypothesis h is*

$$SP(h) := p(\hat{Y}=1|S=1) - p(\hat{Y}=1|S=0).$$

Definition 4.3 (Disparate impact [United States Equal Opportunity Employment Commission, 1978; Feldman et al., 2015])**.** *The disparate impact of an hypothesis h is*

$$DI(h) := \frac{p(\hat{Y}=1|S=0)}{p(\hat{Y}=1|S=1)}.$$

Notice that $SP(h) \in [-1, 1]$, with equality of outcome corresponding to 0, while $DI(h) \in [0, \infty)$, with equality of outcome corresponding to 1. In both cases we want a value that is neither too low nor too high. It has been shown that this is equivalent to requiring that h and the 'anti-classifier' $1 - h$ both have values that are not too low (see Appendix C of [Menon and Williamson, 2018]).

The fair classification problem then takes the form, for some $R_{\text{fair}} \in \{SP, DI\}$:

$$\min_{h \in H} R_Y(h) \text{ subject to } \min[R_{\text{fair}}(h), R_{\text{fair}}(1-h)] \geq \tau, \tag{4.1}$$

where $H := [0,1]^{\mathcal{X}}$ and τ is a constant measuring the required level of fairness. For DI, $\tau \in [0, \infty)$, while for SP, $\tau \in [-1, 0]$ since $SP(1-h) = -SP(h)$.

It has been shown that a constraint on SP or DI of the type in (4.1) is equivalent to a constraint on a cost sensitive risk with respect to S (see Lemmas 1 and 2 of [Menon and Williamson, 2018]). Using Definition 4.1, this cost sensitive risk is written as:

$$R_S(h) := \pi_S(1 - c_S)p(\hat{Y} = 0|S = 1) + (1 - \pi_S)c_S p(\hat{Y} = 1|S = 0), \tag{4.2}$$

where $c_S \in [0, 1]$.

It is more convenient to work with an unconstrained variant of the fair classification problem:

$$\min_{h \in H} [R_Y(h) - \lambda R_S(h)], \tag{4.3}$$

where λ is a constant (not necessarily non-negative) controlling the trade-off between accuracy with respect to Y and fairness with respect to S. It has been shown [Menon and Williamson, 2018] that for some choice of λ, some solution to (4.3) is also a solution to (4.1).

Definition 4.4 (Optimal fair classification)**.** *Let the combined risk*

$$R_{YS}(h) := R_Y(h) - \lambda R_S(h).$$

Let $R_{YS}(h^*)$ be the value of (4.3) *and h^* be a corresponding hypothesis.*

Subsequently we will compare optimal fair classification to the case where we instead use fair representation learning as an intermediate step in fair classification.

4.3.2 Fair Representation Learning

In fair representation learning (Figure 4.1(b)), the data producer trains on samples of X, S and Y (we also examine the case where the data producer does not access Y), and learns the representation function $f : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z}$ is the set of possible cleaned variable values. The data producer samples X and applies f to each sample to produce cleaned variable $Z := f(X)$. The data producer learns f so that Z is still useful for predicting Y but suppresses information about S.

Let $\eta_Y^f(z) := p(Y = 1|Z = z)$ and $\eta_S^f(z) := p(S = 1|Z = z)$ be conditional probabilities of the positive classes of Y and S induced by f. The data user trains on samples of Z and Y and learns a stochastic hypothesis $g : \mathcal{Z} \to [0,1]$, which is used to construct modified decision variable $\hat{Y}^f \in \{0,1\}$ where $g(z) := p(\hat{Y}^f = 1|Z = z)$. At test time, the data producer samples X and passes it through f to produce a sample of Z, from which the data user makes a decision.

When the data user is not trusted, we are interested in constraining how unfair an *adversarial* user can be with the cleaned data. As in the fair classification case, this is equivalent to a constraint on an adversary's cost-sensitive risk with respect to S. We are also interested in ensuring that the cleaned data is still useful for predicting the target. We are therefore interested in the following problem:

$$\min_{f \in F} R_Y(g_Y^* \circ f) \text{ subject to } R_S(g_S^* \circ f) \geq \tau, \tag{4.4}$$

where τ is a constant measuring the required level of fairness, $\circ$ is function composition, $g_Y^* \in \arg\min_{g \in G} R_Y(g \circ f)$ is an optimal indifferent user of the cleaned data, $g_S^* \in \arg\min_{g \in G} R_S(g \circ f)$ is an optimal adversary using the cleaned data, $G := [0,1]^{\mathcal{Z}}$ and $F := \mathcal{Z}^{\mathcal{X}}$.

It is more convenient to work with the following unconstrained problem variant:

$$\min_{f \in F} [R_Y(g_Y^* \circ f) - \lambda R_S(g_S^* \circ f)]. \tag{4.5}$$

Using the form of the minimum cost-sensitive risk from [Zhao et al., 2013], we may express the terms in (4.5) as follows:

$$R_Y(g_Y^* \circ f) = \mathbb{E}_Z[\min((1 - c_Y)\eta_Y^f(Z), c_Y(1 - \eta_Y^f(Z)))] \tag{4.6}$$

$$R_S(g_S^* \circ f) = \mathbb{E}_Z[\min((1 - c_S)\eta_S^f(Z), c_S(1 - \eta_S^f(Z)))]. \tag{4.7}$$

Adversarial neural networks have previously been used to estimate g_Y^* and g_S^* [Edwards and Storkey, 2016; Beutel et al., 2017; Madras et al., 2018]. We observe that (4.6) and (4.7) simplify the fair representation learning cost function (4.5) by removing the two inner minimizations. Of course, there remains the task of estimating the underlying distribution and computing the outer minimization.

We focus on the case where the data producer learns a representation without using the target variable, i.e. we use unsupervised representation learning as in Chapter 2. This allows a single fair representation to be learned that can be used for multiple tasks. It also covers the situation where the data producer does not have access to the target variable. For example, Y contains commercially confidential information (e.g. defaults on a specific type of loan) known to the data user (e.g. a bank) but not the data producer (e.g. a credit bureau). Furthermore, we focus on the case $\mathcal{Z} = \mathcal{X}$ is a Euclidean space, which facilitates our analysis and covers many practical applications. In this case, we define *average reconstruction error* and show its use as a proxy for task performance.

Definition 4.5 (Average reconstruction error)**.** *Suppose $\mathcal{Z} = \mathcal{X}$ is a Euclidean space. Let $\mathbb{E}_X \|X - f(X)\|_2$ be the average reconstruction error of f with respect to X, where $\|\cdot\|_2$ is the Euclidean vector norm.*

Assuming the data producer does not access the target variable, we propose the following variant of the fair representation learning problem:

$$\min_{f \in F} [\mathbb{E}_X \|X - f(X)\|_2 - \lambda R_S(g_S^* \circ f)]. \tag{4.8}$$

We relate (4.8) and (4.5) as follows. This result allows us to select a principled objective function for the data producer.

Theorem 4.6 (Fair representation learning without accessing target variable)**.** *Suppose $\mathcal{Z} = \mathcal{X}$ and we have the Lipschitz condition that for some non-negative constant l_Y*

$$\forall x, x' \in \mathcal{X}, |\eta_Y(x) - \eta_Y(x')| \leq l_Y \|x - x'\|_2. \tag{4.9}$$

Then any $f \in F$ minimizing

$$\mathbb{E}_X \|X - f(X)\|_2 - \lambda R_S(g_S^* \circ f)$$

also minimizes an upper bound on

$$R_Y(g_Y^* \circ f) - l_Y \lambda R_S(g_S^* \circ f).$$

Proof idea. We upper bound $R_Y(g_Y^* \circ f) - l_Y \lambda R_S(g_S^* \circ f)$ by re-expressing the risks using Lemma 9 from [Menon and Williamson, 2018], and making use of the Lipschitz condition. We then observe that the f minimizing this upper bound also minimizes $\mathbb{E}_X \|X - f(X)\|_2 - \lambda R_S(g_S^* \circ f)$. See Section 4.8.2.1 for complete proof. □

4.4 Costs of Fair Representation Learning

We identify and quantify two costs of using fair representation learning rather than entrusting a single trusted data user to make decisions. These costs are incurred by decision-makers, as well as individuals about whom decisions are made. The first cost, which we refer to as the *cost of mistrust*, is the difference in the optimal fairness-accuracy trade-off available with the cleaned data produced by a representation function f compared to the original input. This cost is of interest to the data user – as well as potentially the data regulator. The second cost quantifies the extent to which individual fairness is violated by using a representation function f, which is primarily of interest to the data regulator. We show that both of these costs can be estimated by a data producer without accessing the target variable.

4.4.1 Cost of Mistrust

Suppose that after cleaning the data with the representation function f, we solve the following fair classification problem, which is equivalent to (4.3) but using the cleaned data.

$$\min_{g \in G}[R_Y(g \circ f) - \lambda R_S(g \circ f)] \tag{4.10}$$

Definition 4.7 (Cost of mistrust)**.** *Let g^* and h^* be hypotheses minimizing* (4.10) *and* (4.3) *respectively, where the value of λ is the same in both equations. The cost of mistrust for a representation function f is defined as*

$$R_{YS}(g^* \circ f) - R_{YS}(h^*).$$

The cost of mistrust is non-negative because f restricts the hypothesis class to a subset of H. If $\lambda = 0$ in (4.10) and (4.3), f may incur a cost for the target accuracy of the indifferent user, which seems unsurprising. However, for general λ we see that f may also incur a cost for fair classification. Without access to the sensitive variable S the data user has no way to estimate $R_S(g \circ f)$ in (4.10). However, even if they could somehow guess this quantity, f may create a suboptimal trade-off between fairness and accuracy compared to the trade-off available to a trusted data user using the original input. See Section 4.8.3 for examples where the cost of mistrust is either zero or positive.

We now show in Theorem 4.8 that we can express the cost of mistrust in analytical form. In our result, we use the expressions

$$h^*(x) = \mathbf{1}(\eta_Y(x) - c_Y \geq \lambda(\eta_S(x) - c_S)) \tag{4.11}$$

and

$$g^*(z) = \mathbf{1}(\eta_Y^f(z) - c_Y \geq \lambda(\eta_S^f(z) - c_S)), \tag{4.12}$$

obtained from Proposition 4 of [Menon and Williamson, 2018].

Theorem 4.8 (Analytical form of cost of mistrust)**.** *The cost of mistrust may be expressed as*

$$\begin{aligned} &R_{YS}(g^* \circ f) - R_{YS}(h^*) \\ &= \mathbb{E}_X[\min(\eta_Y^f(f(X)) - c_Y, \lambda(\eta_S^f(f(X)) - c_S)) - \min(\eta_Y(X) - c_Y, \lambda(\eta_S(X) - c_S))]. \end{aligned} \tag{4.13}$$

The cost of mistrust may be decomposed into accuracy and fairness differences, where the accuracy difference is

$$R_Y(g^* \circ f) - R_Y(h^*) = \mathbb{E}_X[h^*(X)(\eta_Y(X) - c_Y) - g^*(f(X))(\eta_Y^f(f(X)) - c_Y)], \tag{4.14}$$

and the fairness difference is

$$R_S(g^* \circ f) - R_S(h^*) = \mathbb{E}_X[h^*(X)(\eta_S(X) - c_S) - g^*(f(X))(\eta_S^f(f(X)) - c_S)], \tag{4.15}$$

which are combined in the overall cost of mistrust

$$R_{YS}(g^* \circ f) - R_{YS}(h^*) = R_Y(g^* \circ f) - R_Y(h^*) - \lambda(R_S(g^* \circ f) - R_S(h^*)).$$

Proof idea. We apply Lemma 9 of [Menon and Williamson, 2018] to express each of $R_Y(g^* \circ f)$, $R_Y(h^*)$, $R_S(g^* \circ f)$ and $R_S(h^*)$. Combining these yields a compact expression for $R_{YS}(g^* \circ f) - R_{YS}(h^*)$. See Section 4.8.2.2 for complete proof. □

The expression (4.13) for the cost of mistrust allows us to measure the quality of the fairness-accuracy trade-off available using f compared to using the original input. The decomposition reveals that the signs of the accuracy and fairness differences may vary. However, since the cost of mistrust is non-negative, for a fixed value of R_S we incur a value of R_Y that is at least as large using f as with the original input.

For intuition about the expression (4.13) for the cost of mistrust in Theorem 4.8, consider some point $z \in \mathcal{Z}$ and its preimage $\mathcal{X}_z := \{x \in \mathcal{X} | f(x) = z\}$. If for all $x \in \mathcal{X}_z$, we have the same value of $\mathbf{1}(\eta_Y(x) - c_Y \geq \lambda(\eta_S(x) - c_S))$, then the expectation conditioned on $x \in \mathcal{X}_z$ will be zero, otherwise it will be positive. Hence the cost of mistrust will be small when points mapped to the same value of z tend to have the same value of $\mathbf{1}(\eta_Y(x) - c_Y \geq \lambda(\eta_S(x) - c_S))$.

We are interested in situations where the data producer can guarantee that the cost of mistrust is small without accessing Y. When $\mathcal{Z} = \mathcal{X}$ and the conditional distributions $\eta_Y(x)$ and $\eta_S(x)$ are smooth, the cost of mistrust can be upper bounded in terms of average reconstruction error. This result, shown in Theorem 4.9, allows the data producer to bound the cost of mistrust using only X and Z.

Theorem 4.9 (Upper bound on cost of mistrust with smooth conditional distributions)**.** *Suppose $\mathcal{Z} = \mathcal{X}$ is a Euclidean space and we have the Lipschitz conditions that for some non-negative constants l_Y and l_S*

$$\forall x, x' \in \mathcal{X}, |\eta_Y(x) - \eta_Y(x')| \leq l_Y \|x - x'\|_2 \tag{4.16}$$

and

$$\forall x, x' \in \mathcal{X}, |\eta_S(x) - \eta_S(x')| \leq l_S \|x - x'\|_2. \tag{4.17}$$

Then

$$R_{YS}(g^* \circ f) - R_{YS}(h^*) \leq (l_Y + \lambda l_S)\mathbb{E}_X \|X - f(X)\|_2.$$

Proof idea. We observe that $R_{YS}(h^* \circ f)$ is an upper bound on $R_{YS}(g^* \circ f)$. We use Lemma 9 of [Menon and Williamson, 2018] to re-express R_{YS}. We then use the Lipschitz conditions to upper bound $R_{YS}(h^* \circ f) - R_{YS}(h^*)$. See Section 4.8.2.3 for complete proof. □

4.4.2 Cost for Individual Fairness

We investigate the cost of using a given representation in terms of *individual fairness* [Dwork et al., 2012]. This notion requires that similar decisions should be made for similar individuals, i.e. decisions are smooth. It is possible that a representation function maps points that are nearby in the input space to points that are distant from each other in the representation space. Therefore, smooth hypotheses may not be individually fair when applied to the cleaned data. We wish to quantify this cost for individual fairness by upper bounding the individual unfairness of an arbitrary smooth hypothesis applied to the cleaned data. We show that it is possible for a data user to provide this kind of certification to a data regulator by inspecting Z and X.

First, we restate a previous definition of individual fairness.

Definition 4.10 (Individual fairness [Dwork et al., 2012]). *Let D and d be subadditive functions. Hypothesis h is $D, d-$individually fair if*

$$\forall x, x' \in \mathcal{X}, D(h(x), h(x')) \leq d(x, x').$$

We also give a novel quantitative notion of individual *unfairness* by measuring the probability that a pair of randomly selected individuals will be treated unfairly according to Definition 4.10.

Definition 4.11 (Individual unfairness). *Hypothesis h has $D, d-$individual unfairness with respect to X defined as*

$$IU_{D,d}(h) := p(D(h(x), h(x')) > d(x, x')),$$

where x and x' are independent random samples of X.

In order to bound the level of individual unfairness induced by a representation, we introduce the following definition.

Definition 4.12 (Large reconstruction error rate). *Suppose $\mathcal{Z} = \mathcal{X}$. Let ϵ be a non-negative constant. Let $p(d(X, f(X)) > \epsilon)$ be the large reconstruction error rate of f.*

In Theorem 4.13 we show that if the large reconstruction error rate is small, then any hypothesis that is smooth (i.e. individually fair when applied to the original input) will not be too individually unfair when applied to the cleaned data. We observe that there is a tension between guaranteeing group fairness, which involves removing information to protect an adversary from inferring the sensitive variable, and individual fairness, which requires preserving information from the original input.

Theorem 4.13 (Upper bound on individual unfairness). *Suppose $\mathcal{Z} = \mathcal{X}$. Let*

$$d_\epsilon(x, x') := d(x, x') + 2\epsilon$$

and let h be any individually fair hypothesis. Then the $D, d_\epsilon-$individual unfairness of $h \circ f$ is upper bounded as follows:

$$IU_{D,d_\epsilon}(h \circ f) \leq 2p(d(X, f(X)) > \epsilon).$$

Proof idea. Let $\delta := p(d(X, f(X)) > \epsilon)$. For randomly drawn x and x', $d(x, f(x)) \leq \epsilon$ and $d(x', f(x')) \leq \epsilon$ with probability at least $1 - 2\delta$ by the union bound. If these statements hold, by the triangle inequality $D(h(f(x), h(f(x')) \leq d(x,x') + 2\epsilon$. See Section 4.8.2.4 for complete proof. □

4.5 Benefits of Fair Representation Learning

We quantify the benefits of some representation function f by measuring the discrimination achieved by an optimal adversary using Z, the representation variable induced by f. We show that a data producer can do this for both statistical parity and disparate impact. We can compute these two quantities directly for a given f, so that unlike in the optimization problems we considered earlier there is no need to use a cost-sensitive risk. The quantities we obtain can be given to a data regulator to certify that any use of the cleaned data will not be too unfair. If the data producer has access to the target variable, these quantities can also be evaluated on subsets of the data with the same value of the target, to measure quantified versions of equality of opportunity (conditioning on $Y = 1$) and equalized odds (conditioning separately on $Y = 1$ and $Y = 0$) [Hardt et al., 2016].

4.5.1 Benefit for Statistical Parity

We certify that any decision using the cleaned data has statistical parity (Definition 4.2) that is neither too small nor too large. In Theorem 4.14, we show that the maximum and minimum statistical parity of an adversary using Z can be expressed in closed form. The maximum and minimum will be closer if the induced conditional probability $\eta_S^f(z)$ does not deviate too much on average from the prior π_S. If $\eta_S^f(z) = \pi_S$ everywhere, we have statistical parity of zero, i.e. exact equality of outcome.

Theorem 4.14 (Statistical parity of optimal adversary)**.** *An adversarial user of Z achieves maximum and minimum statistical parity*

$$\max_{g \in G} SP(g \circ f) = 1 - \mathbb{E}_Z[\min(\frac{\eta_S^f(Z)}{\pi_S}, \frac{1-\eta_S^f(Z)}{1-\pi_S})]$$

$$\min_{g \in G} SP(g \circ f) = -1 + \mathbb{E}_Z[\min(\frac{\eta_S^f(Z)}{\pi_S}, \frac{1-\eta_S^f(Z)}{1-\pi_S})].$$

Proof idea. Observe that statistical parity is a linear transformation of balanced error rate. Apply the minimum balanced error rate from Equation 32 of [Zhao et al., 2013]. See Section 4.8.2.5 for complete proof. □

4.5.2 Benefit for Disparate Impact

We certify that any decision using the cleaned data has disparate impact (Definition 4.3) that is neither too small nor too large. In Theorem 4.15, we show that the maximum and minimum disparate impact of an adversary using Z can be expressed in closed form. The maximum and minimum will be closer if the induced conditional probability $\eta_S^f(z)$ never deviates too much from the prior π_S. If $\eta_S^f(z) = \pi_S$ everywhere, we have disparate impact of one, i.e. exact equality of outcome. Observe how disparate impact is more sensitive than statistical parity, since it requires $\eta_S^f(z)$ to be close to π_S *everywhere* rather than only in expectation.

Theorem 4.15 (Disparate impact of optimal adversary)**.** *Let* $\overline{\eta}_S^f := \max_{z \in \mathcal{Z}} \eta_S^f(z)$ *and* $\underline{\eta}_S^f := \min_{z \in \mathcal{Z}} \eta_S^f(z)$. *An adversarial user of* Z *achieves maximum and minimum disparate impact*

$$\max_{g \in G} DI(g \circ f) = \frac{\pi_S(1 - \underline{\eta}_S^f)}{\underline{\eta}_S^f(1 - \pi_S)}$$

$$\min_{g \in G} DI(g \circ f) = \frac{\pi_S(1 - \overline{\eta}_S^f)}{\overline{\eta}_S^f(1 - \pi_S)}.$$

Proof idea. Re-express $DI(g \circ f)$ using the law of total probability, the fact that $\hat{Y}^f$ and S are conditionally independent given Z, and Bayes' rule. Using this form we obtain the maximum and minimum values of $DI(g \circ f)$ and the corresponding choices of g. See Section 4.8.2.6 for complete proof. □

4.6 Experiments

We conducted experiments with two objectives in mind. First, to show that the formalization of the fair representation learning problem we suggested in Section 4.3 can be used in practice. Second, to illustrate how the costs and benefits identified in Sections 4.4 and 4.5 can be estimated and interpreted without requiring access to the target variable.

4.6.1 Datasets

We used the UCI Adult and ProPublica recidivism datasets, which are both well-known in the fair machine learning literature (e.g. [Calmon et al., 2017]).[3] We selected S to be gender for Adult and whether the person is of African-American ethnicity for ProPublica. Our experiments do not depend on a particular choice of Y. We learn f using 70% of the data and report results on the remaining 30%.

The Adult dataset contains financial and demographic information compiled from a census of about 32561 people, and contains 110 input columns once categorical features are represented as a one-hot encoding. We selected S as gender, while a possible choice of Y is whether the person's income is at least $50,000. This setting is similar to a situation where a financial institution uses an algorithm to decide whether to grant an individual a loan based on a prediction of their income.

[3] These datasets are located at https://archive.ics.uci.edu/ml/datasets/adult and https://github.com/propublica/compas-analysis respectively.

The ProPublica dataset contains information about 7214 criminal offences committed in Broward County, Florida and contains 79 input columns once categorical features are represented as a one-hot encoding. We processed the free text crime description column by converting it to a categorical variable where descriptions occurring at least 20 times have their own category (covering 82.9% of all offences) and all other descriptions are marked as 'other', and then using a one-hot encoding. We selected S as whether the person is of African-American ethnicity, while a possible choice of Y is whether the person reoffended within two years. This setting is similar to a situation where a court decides whether to grant a defendant a pre-trial release from custody, based on an algorithmic assessment of the individual's likelihood of reoffending.

4.6.2 Method

We approximated (4.8) by estimating f with an *encoder* neural network, testing several values of λ. We used a finite sample to estimate the average reconstruction error component of the cost function. To estimate the $R_S(g^* \circ f)$ component of the cost function we used the form given by (4.7) with $c_S = 0.5$, trained another *evaluator* neural network to estimate $\eta_S^f(z)$, and used a finite sample to approximate the expectation. The evaluator is comparable to the 'adversary' in [Edwards and Storkey, 2016; Beutel et al., 2017; Madras et al., 2018] since we alternated updating its weights with those of the encoder. However, the evaluator was used to estimate (4.7) which was used to evaluate f, rather than its performance directly being used to evaluate f. This approach is motivated by the fact that (4.7) gives us the performance of the optimal adversary.

Our training set consisted of N points, where the input and sensitive variable values for the nth point are given by x_n and s_n respectively. We estimated f using a fully-connected *encoder* neural network with one softplus ($\text{softplus}(x) := \ln(1+e^x)$) hidden layer of 100 units and a linear output layer with the same number of units as the input layer. To approximate (4.8) we updated f to minimize the following cost function:

$$\frac{1}{N}\sum_{n=1}^{N}\|x_n - f(x_n)\|_2 - \frac{\lambda}{2N}\sum_{n=1}^{N}\min[\widetilde{\eta}_S^f(f(x_n)), 1-\widetilde{\eta}_S^f(f(x_n))]. \tag{4.18}$$

We computed $\widetilde{\eta}_S^f(z)$ to estimate $\eta_S^f(z)$ using a fully-connected *evaluator* neural network with one softplus hidden layer of 100 units and a single sigmoidal output unit. The output layer of the encoder, which corresponds to the variable Z, is the input layer for the evaluator. For the evaluator network we updated $\widetilde{\eta}_S^f$ to minimize the following cost function:

$$-\frac{1}{N}\sum_{n=1}^{N}[s_n \ln \widetilde{\eta}_S^f(f(x_n)) + (1-s_n)\ln(1-\widetilde{\eta}_S^f(f(x_n)))].$$

We alternated updates of the weights in the encoder and evaluator networks, as in adversarial methods [Edwards and Storkey, 2016; Beutel et al., 2017; Madras et al., 2018]. We used the Adam Optimizer with a learning rate of 0.0001, a batch size of 100 and set the training set epochs to 100. We implemented the model in Python using the TensorFlow library.

We set the large reconstruction error rate threshold

$$\epsilon := 0.1 \times \frac{1}{N} \sum_{n=1}^{N} \|x_n\|_2$$

and set the individual fairness distance function[4]

$$d(x, x') := \|x - x'\|_2.$$

We evaluated the costs and benefits of a representation function f using the following empirical estimates computed over a test set of N' points:

$$\mathbb{E}_X \|X - f(X)\|_2 \approx \frac{1}{N'} \sum_{n=1}^{N'} \|x_n - f(x_n)\|_2$$

$$p(\|X - f(X)\|_2 > \epsilon) \approx \frac{1}{N'} \sum_{n=1}^{N'} \mathbf{1}(\|x_n - f(x_n)\|_2 > \epsilon)$$

$$\max_{g \in G} SP(g \circ f) \approx 1 - \frac{1}{N'} \sum_{n=1}^{N'} \min[\frac{\widetilde{\eta}_S^f(f(x_n))}{\widetilde{\pi}_S}, \frac{1 - \widetilde{\eta}_S^f(f(x_n))}{1 - \widetilde{\pi}_S}]$$

$$\max_{g \in G} DI(g \circ f) \approx \frac{\widetilde{\pi}_S (1 - \underline{\widetilde{\eta}}_S^f)}{\underline{\widetilde{\eta}}_S^f (1 - \widetilde{\pi}_S)}$$

where $\widetilde{\pi}_S := \frac{1}{N'} \sum_{n=1}^{N'} s_n$ and $\underline{\widetilde{\eta}}_S^f := \min_{n \leq N'} \widetilde{\eta}_S^f(f(x_n))$.

4.6.3 Results

We show our results in Figure 4.2. For several values of λ in (4.18) (shown on the horizontal axes of the plots), we estimated the costs and benefits of the learned representation function f (shown on the vertical axes of the plots). The trends for both datasets are similar. A subtlety is that we report proxies for the costs motivated by our theoretical results, which can be estimated by the data user without access to the target variable.

Recall that we may use average reconstruction error to upper bound the cost of mistrust (Theorem 4.9) and the large reconstruction error rate to upper bound the

[4] Euclidean distance is used to illustrate the effect of fair representation learning on individual fairness. However, the relationship between large reconstruction error rate and individual fairness holds for any choice of subadditive distance function, as shown in Theorem 4.13. See Dwork et al. [2012] for a discussion of approaches to selecting the individual fairness distance function.

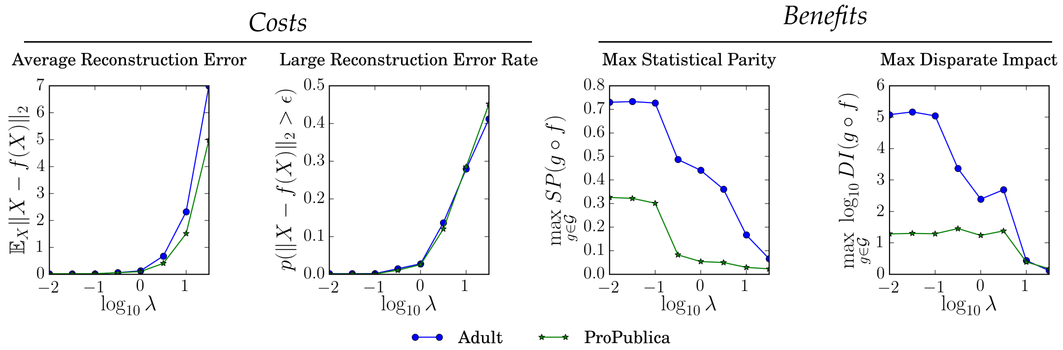


Figure 4.2: Estimates of costs and benefits of fair representation learning on Adult and ProPublica datasets, varying the parameter λ in (4.18). As λ increases, the cleaned data differs more from the original input as more information about the sensitive variable is suppressed. Lower is better on the vertical axes of all plots. See text for discussion.

cost for individual fairness (Theorem 4.13). Estimates of both of these cost proxies increase with λ, as the cleaned data becomes more distorted by f.

Furthermore, recall that we have the closed form of an adversary's maximum statistical parity (Theorem 4.14) and disparate impact (Theorem 4.15). Estimates of both of these quantities decline as λ increases, indicating benefits from using f. For disparate impact we use a log scale on the vertical axis for clarity, and observe that its empirical estimates appear noisier than those of statistical parity, since it requires us to estimate the minimum rather than the expectation of $\eta_S^f(z)$.

Our experiments, when combined with our theoretical results, reveal that the choice of λ in (4.8) starkly determines the relative costs and benefits of fair representation learning.

4.7 Conclusion

We have quantified the costs – an inferior fairness-accuracy trade-off and an increase in individual unfairness – incurred by a given representation. We have also quantified the benefits – narrower bands of statistical parity and disparate impact achievable by an adversary – of such a representation. The benefits result from restricting the decisions of adversarial data users, while the costs are due to applying those same restrictions to other data users. We showed how a data producer can estimate these costs and benefits, even without access to the target variable, to support a novel three-party governance model entailing a separation of concerns between fairness and accuracy. Future directions of interest include extending our results to finite samples, stochastic representation functions, multiple sensitive groups and variables, more general representation spaces, and other fairness definitions.

4.8 Appendix

We provide a summary of the problems, costs and benefits we consider, complete proofs of our results, and examples of the cost of mistrust.

4.8.1 Summary of Problems, Costs and Benefits Considered

We summarize the problems we have considered in Table 4.1. In Tables 4.2 and 4.3, we summarize the costs and benefits respectively of a given representation function f. In each table we distinguish between cases where access to the target variable Y is required, and cases where it is not required. Observe that if access to the target variable is available, the benefits described in Table 4.3 can be computed for specific subsets of the data based on the target variable (e.g. conditioning on $Y = 1$ or $Y = 0$). Definitions of all terms can be found in the main text.

Table 4.1: Problems

Problem	**Reference**	**Optimization Problem**
Access to target variable required		
Fair classification	(4.3)	$\min_{h\in H}[R_Y(h) - \lambda R_S(h)]$
Fair representation learning	(4.5)	$\min_{f\in F}[R_Y(g_Y^* \circ f) - \lambda R_S(g_S^* \circ f)]$
Access to target variable not required		
Fair representation learning without accessing target variable	(4.8)	$\min_{f\in F}[\mathbb{E}_X\|X - f(X)\|_2 - \lambda R_S(g_S^* \circ f)]$

Table 4.2: Costs of a representation function f

Cost	**Reference**	**Analytical Form**
Access to target variable required		
Cost of mistrust	Theorem 4.8	$\mathbb{E}_X[\min(\eta_Y^f(f(X))-c_Y, \lambda(\eta_S^f(f(X)) - c_S)) - \min(\eta_Y(X) - c_Y, \lambda(\eta_S(X) - c_S))]$
Access to target variable not required		
Upper bound on cost of mistrust using average reconstruction error	Theorem 4.9	$(l_Y + \lambda l_S)\mathbb{E}_X\|X - f(X)\|_2$
Upper bound on individual unfairness using large reconstruction error rate	Theorem 4.13	$2p(d(X, f(X)) > \epsilon)$

Table 4.3: Benefits of a representation function f

Benefit	**Reference**	**Analytical Form**
Access to target variable not required		
Maximum and minimum statistical parity	Theorem 4.14	$\max_{g \in G} SP(g \circ f) = 1 - \mathbb{E}_Z[\min(\frac{\eta_S^f(Z)}{\pi_S}, \frac{1-\eta_S^f(Z)}{1-\pi_S})]$ $\min_{g \in G} SP(g \circ f) = -1 + \mathbb{E}_Z[\min(\frac{\eta_S^f(Z)}{\pi_S}, \frac{1-\eta_S^f(Z)}{1-\pi_S})]$
Maximum and minimum disparate impact	Theorem 4.15	$\max_{g \in G} DI(g \circ f) = \frac{\pi_S(1-\underline{\eta}_S^f)}{\underline{\eta}_S^f(1-\pi_S)}$ $\min_{g \in G} DI(g \circ f) = \frac{\pi_S(1-\overline{\eta}_S^f)}{\overline{\eta}_S^f(1-\pi_S)}$

4.8.2 Theorem Proofs

We present complete proofs of our theoretical results.

4.8.2.1 Proof of Theorem 4.6 (Fair Representation Learning Without Accessing Target Variable)

Proof. We derive an upper bound on

$$R_Y(g_Y^* \circ f) - l_Y \lambda R_S(g_S^* \circ f).$$

Let $h^*_Y \in \arg\min_{h \in H} R_Y(h)$, which takes the form $h^*_Y(x) = \mathbf{1}(\eta_Y(x) \geq c_Y)$ [Zhao et al., 2013].

$$
\begin{aligned}
&R_Y(g^*_Y \circ f) - l_Y \lambda R_S(g^*_S \circ f) \\
&\leq R_Y(h^*_Y \circ f) - l_Y \lambda R_S(g^*_S \circ f) \\
&= R_Y(h^*_Y \circ f) - R_Y(h^*_Y) + R_Y(h^*_Y) - l_Y \lambda R_S(g^*_S \circ f) \\
&= \mathbb{E}_X[(c_Y - \eta_Y(X))h^*_Y(f(X))] - \mathbb{E}_X[(c_Y - \eta_Y(X))h^*_Y(X)] + R_Y(h^*_Y) - l_Y \lambda R_S(g^*_S \circ f) && (4.19) \\
&= \mathbb{E}_X[(c_Y - \eta_Y(X))(h^*_Y(f(X)) - h^*_Y(X))] + R_Y(h^*_Y) - l_Y \lambda R_S(g^*_S \circ f) && (4.20) \\
&\leq l_Y \mathbb{E}_X \|X - f(X)\|_2 + R_Y(h^*_Y) - l_Y \lambda R_S(g^*_S \circ f). && (4.21)
\end{aligned}
$$

For (4.19) we apply Lemma 9 from [Menon and Williamson, 2018]. For (4.20) we apply linearity of expectation.

For (4.21), for any x where $h^*_Y(x) \neq h^*_Y(f(x))$, there must exist some x' on the decision boundary of h^* such that

$$c_Y - \eta_Y(x') = 0 \tag{4.22}$$

and

$$\|x - x'\|_2 \leq \|x - f(x)\|_2. \tag{4.23}$$

Combining (4.22) and (4.23) with the Lipchitz condition (4.9) yields

$$
\begin{aligned}
&c_Y - \eta_Y(x) \\
&\leq c_Y - \eta_Y(x') + l_Y \|x - x'\|_2 \\
&\leq l_Y \|x - f(x)\|_2.
\end{aligned}
$$

Since this is true for every x it is also true in expectation.

We then observe

$$
\begin{aligned}
&\arg\min_{f \in F} [l_Y \mathbb{E}_X \|X - f(X)\|_2 + R_Y(h^*_Y) - l_Y \lambda R_S(g^*_S \circ f)] \\
&= \arg\min_{f \in F} [\mathbb{E}_X \|X - f(X)\|_2 - \lambda R_S(g^*_S \circ f)].
\end{aligned}
$$

□

4.8.2.2 Proof of Theorem 4.8 (Analytical Form of Cost of Mistrust)

Proof. First we show the analytical expression for the cost of mistrust (4.13). Applying Proposition 4 of [Menon and Williamson, 2018], we have that (4.11) and (4.12) are hypotheses h^* and g^* corresponding to solutions to (4.3) and (4.10) respectively.

$$
\begin{aligned}
&\mathbb{E}_X[\min(\eta_Y(X) - c_Y, \lambda(\eta_S(X) - c_S))] && (4.24)\\
&= \mathbb{E}_X[(1 - h^*(X))(\eta_Y(X) - c_Y)] + \lambda\mathbb{E}_X[h^*(X)(\eta_S(X) - c_S)] && (4.25)\\
&= \mathbb{E}_X[\eta_Y(X) - c_Y] - \mathbb{E}_X[h^*(X)(\eta_Y(X) - c_Y)] + \lambda\mathbb{E}_X[h^*(X)(\eta_S(X) - c_S)]\\
&= \pi_Y - c_Y - \mathbb{E}_X[h^*(X)(\eta_Y(X) - c_Y)] + \lambda\mathbb{E}_X[h^*(X)(\eta_S(X) - c_S)]\\
&= \pi_Y - c_Y + R_Y(h^*) - (1 - c_Y)\pi_Y + \lambda\mathbb{E}_X[h^*(X)(\eta_S(X) - c_S)] && (4.26)\\
&= \pi_Y - c_Y + R_Y(h^*) - (1 - c_Y)\pi_Y - \lambda R_S(h^*) + \lambda(1 - c_S)\pi_S && (4.27)\\
&= R_Y(h^*) - \lambda R_S(h^*) - c_Y(1 - \pi_Y) + \lambda(1 - c_S)\pi_S.
\end{aligned}
$$

(4.25) follows from the form of h^* given in (4.11). (4.26) and (4.27) both involve substitutions using Lemma 9 from [Menon and Williamson, 2018].

Similarly, using the form of g^* from (4.12) we conclude that

$$
\begin{aligned}
&\mathbb{E}_X[\min(\eta^f_Y(f(X)) - c_Y, \lambda(\eta^f_S(f(X)) - c_S))] && (4.28)\\
&= \mathbb{E}_Z[\min(\eta^f_Y(Z) - c_Y, \lambda(\eta^f_S(Z) - c_S))]\\
&= R_Y(g^* \circ f) - \lambda R_S(g^* \circ f) - c_Y(1 - \pi_Y) + \lambda(1 - c_S)\pi_S. && (4.29)
\end{aligned}
$$

The result (4.13) follows by substituting (4.24) from (4.28) and applying linearity of expectation.

The decomposed form follows from applying Lemma 9 of [Menon and Williamson, 2018] to each of $R_Y(g^* \circ f)$, $R_Y(h^*)$, $R_S(g^* \circ f)$ and $R_S(h^*)$, then applying linearity of expectation to express $R_Y(g^* \circ f) - R_Y(h^*)$ as in (4.14) and $R_S(g^* \circ f) - R_S(h^*)$ as in (4.15). □

4.8.2.3 Proof of Theorem 4.9 (Upper Bound on Cost of Mistrust with Smooth Conditional Distributions)

Proof.

$$
\begin{aligned}
&R_{YS}(g^* \circ f) - R_{YS}(h^*)\\
&\le R_{YS}(h^* \circ f) - R_{YS}(h^*)\\
&= R_Y(h^* \circ f) - R_Y(h^*) - \lambda(R_S(h^* \circ f) - R_S(h^*))\\
&= \mathbb{E}_X[(c_Y - \eta_Y(X))(h^*(f(X)) - h^*(X))] - \lambda\mathbb{E}_X[(c_S - \eta_S(X))(h^*(f(X)) - h^*(X))]\\
& && (4.30)\\
&= \mathbb{E}_X[(c_Y - \eta_Y(X) - \lambda(c_S - \eta_S(X)))(h^*(f(X)) - h^*(X))] && (4.31)\\
&\le (l_Y + \lambda l_S)\mathbb{E}_X\|X - f(X)\|_2. && (4.32)
\end{aligned}
$$

(4.30) is by Lemma 9 from [Menon and Williamson, 2018] and linearity of expectation. (4.31) is by linearity of expectation.

For (4.32), using the form of h^* from (4.11), for any x where $h^*(x) \neq h^*(f(x))$, there must exist some x' on the decision boundary of h^* such that

$$c_Y - \eta_Y(x') - \lambda(c_S - \eta_S(x')) = 0 \tag{4.33}$$

and

$$\|x - x'\|_2 \leq \|x - f(x)\|_2. \tag{4.34}$$

Combining (4.33) and (4.34) with the Lipchitz conditions (4.16) and (4.17) yields

$$\begin{aligned} & c_Y - \eta_Y(x) - \lambda(c_S - \eta_S(x)) \\ \leq\; & c_Y - \eta_Y(x') + l_Y\|x - x'\|_2 - \lambda(c_S - \eta_S(x') - l_S\|x - x'\|_2) \\ \leq\; & (l_Y + \lambda l_S)\|x - f(x)\|_2. \end{aligned}$$

Since this is true for every x it is also true in expectation. □

4.8.2.4 Proof of Theorem 4.13 (Upper Bound on Individual Unfairness)

Proof. Let $\delta := p(d(X, f(X)) > \epsilon)$. Let h be a $D, d-$individually fair hypothesis (see Definition 4.10).

Consider points x and x' drawn independently at random using the input X. With probability $1 - \delta$,

$$d(x, f(x)) \leq \epsilon. \tag{4.35}$$

Similarly, with probability $1 - \delta$,

$$d(x', f(x')) \leq \epsilon. \tag{4.36}$$

By the union bound, both statements hold with probability at least $1 - 2\delta$. In that case, the following statements also hold:

$$\begin{aligned} & D(h(f(x), h(f(x')) \\ \leq\; & D(h(f(x)), h(x)) + D(h(x), h(f(x'))) && (4.37) \\ \leq\; & \epsilon + D(h(x), h(f(x'))) && (4.38) \\ \leq\; & \epsilon + D(h(x), h(x')) + D(h(x'), h(f(x'))) && (4.39) \\ \leq\; & 2\epsilon + D(h(x), h(x')) && (4.40) \\ \leq\; & 2\epsilon + d(x, x'). && (4.41) \end{aligned}$$

(4.37) and (4.39) apply the triangle inequality since D is subadditive. (4.38) and (4.40) hold due to (4.35) and (4.36) respectively, along with Definition 4.10. (4.41) applies Definition 4.10. Therefore $IU_{D,d_\epsilon}(h \circ f) \leq 2\delta$. □

4.8.2.5 Proof of Theorem 4.14 (Statistical Parity of Optimal Adversary)

Proof. Let

$$BER(h) := \frac{1}{2}p(\hat{Y} = 0|S = 1) + \frac{1}{2}p(\hat{Y} = 1|S = 0)$$

be the balanced error rate of an hypothesis h. Observe that $SP(h) = 1 - 2BER(h)$ for all h.

Therefore

$$\begin{aligned}
&\max_{g \in G} SP(g \circ f) \\
&= 1 - 2\min_{g \in G} BER(g \circ f) \\
&= 1 - \mathbb{E}_Z[\min(\frac{\eta_S^f(Z)}{\pi_S}, \frac{1 - \eta_S^f(Z)}{1 - \pi_S})], \qquad (4.42)
\end{aligned}$$

where (4.42) uses Equation 32 from [Zhao et al., 2013].

Similarly,

$$\begin{aligned}
&\min_{g \in G} SP(g \circ f) \\
&= 1 - 2\max_{g \in G} BER(g \circ f) \\
&= 1 - 2\max_{g \in G}[1 - BER(1 - g \circ f)] \qquad (4.43) \\
&= -1 + \min_{g \in G} BER(1 - g \circ f) \\
&= -1 + \min_{g \in G} BER(g \circ f) \\
&= -1 + \mathbb{E}_Z[\min(\frac{\eta_S^f(Z)}{\pi_S}, \frac{1 - \eta_S^f(Z)}{1 - \pi_S})],
\end{aligned}$$

where for (4.43) we used the fact that $BER(h) = 1 - BER(1 - h)$ for all h. □

4.8.2.6 Proof of Theorem 4.15 (Disparate Impact of Optimal Adversary)

Proof. Disparate impact can be expressed as follows:

$$
\begin{aligned}
& DI(g \circ f) \\
&= \frac{p(\hat{Y}^f = 1|S = 0)}{p(\hat{Y}^f = 1|S = 1)} \\
&= \frac{\int_z p(Z = z|S = 0)p(\hat{Y}^f = 1|S = 0, Z = z)dz}{\int_z p(Z = z|S = 1)p(\hat{Y}^f = 1|S = 1, Z = z)dz} \\
&= \frac{\int_z p(Z = z|S = 0)g(z)dz}{\int_z p(Z = z|S = 1)g(z)dz} \qquad (4.44) \\
&= \frac{\pi_S \int_z p(Z = z)(1 - \eta_S^f(z))g(z)dz}{(1 - \pi_S) \int_z p(Z = z)\eta_S^f(z)g(z)dz} \qquad (4.45) \\
&= \frac{\pi_S \mathbb{E}_Z[(1 - \eta_S^f(Z))g(Z)]}{(1 - \pi_S)\mathbb{E}_Z[\eta_S^f(Z)g(Z)]}. \qquad (4.46)
\end{aligned}
$$

For (4.44) we used the fact that $\hat{Y}^f$ and S are conditionally independent given Z. For (4.45) we used Bayes' rule.

Recall that $\overline{\eta}_S^f := \max\limits_{z \in \mathcal{Z}} \eta_S^f(z)$ and $\underline{\eta}_S^f := \min\limits_{z \in \mathcal{Z}} \eta_S^f(z)$. Let γ be an arbitrary constant in the range $(0, 1]$. Using the form of $DI(g \circ f)$ in (4.46), we have:

$$
\max_{g \in G} DI(g \circ f) = \frac{\pi_S(1 - \underline{\eta}_S^f)}{\underline{\eta}_S^f(1 - \pi_S)}
$$

where the maximum is obtained for

$$
g(z) = \begin{cases} \gamma & \text{if } \eta_S^f(z) = \underline{\eta}_S^f \\ 0 & \text{otherwise.} \end{cases}
$$

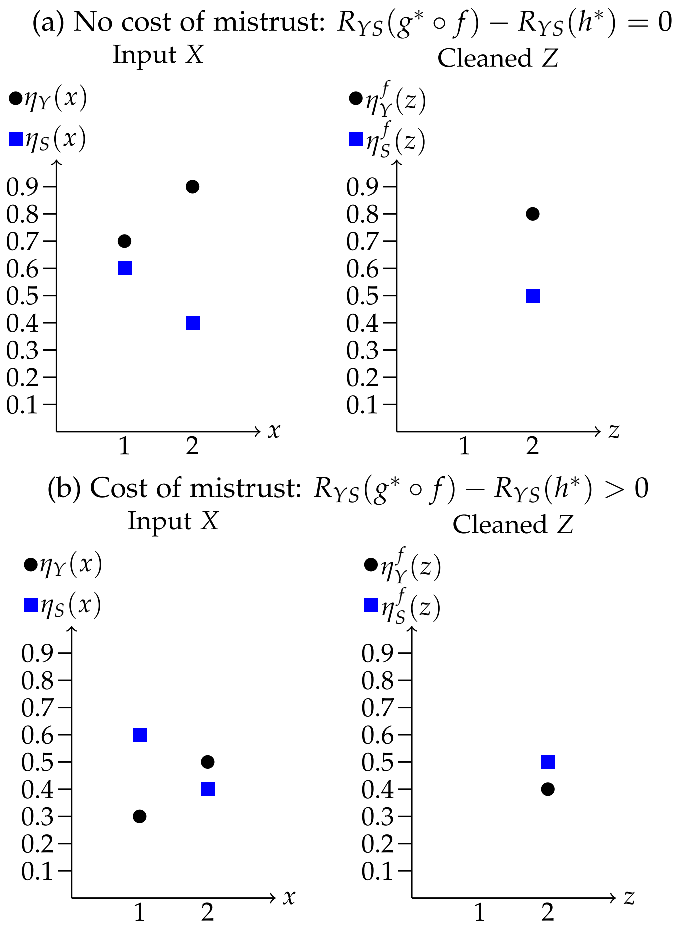
Similarly,

$$
\min_{g \in G} DI(g \circ f) = \frac{\pi_S(1 - \overline{\eta}_S^f)}{\overline{\eta}_S^f(1 - \pi_S)}
$$

where the minimum is obtained for

$$
g(z) = \begin{cases} \gamma & \text{if } \eta_S^f(z) = \overline{\eta}_S^f \\ 0 & \text{otherwise.} \end{cases}
$$

□


(a) No cost of mistrust: $R_{YS}(g^* \circ f) - R_{YS}(h^*) = 0$

(b) Cost of mistrust: $R_{YS}(g^* \circ f) - R_{YS}(h^*) > 0$

Figure 4.3: Two examples illustrating the cost of mistrust. See text for discussion.

4.8.3 Examples of the Cost of Mistrust

We use examples to demonstrate that the cost of mistrust may be either zero or positive, as depicted in Figure 4.3.

Let $\mathcal{X} = \mathcal{Z} = \{1,2\}$, $c_Y = c_S = 0.5$, $p(X = 1) = p(X = 2) = 0.5$, $\lambda = 1$ in (4.3), (4.5) and (4.10). In both examples, $\eta_S(1) = 0.6$ and $\eta_S(2) = 0.4$. In (a) $\eta_Y(1) = 0.7$ and $\eta_Y(2) = 0.9$, while in (b) $\eta_Y(1) = 0.3$ and $\eta_Y(2) = 0.5$. While setting f to map all points to a constant is a crude example, it suffices for our illustration. In (a) the cost of mistrust is 0, while in (b) it is 0.05. This is because in (a), h^* predicts the same value for the two points combined by f and is hence unaffected by f, while in (b), h^* predicts different values and is hence affected by f.

We compare the representation function $f(x) = 2$ to the identity representation function $f_I(x) = x$, which makes our analysis sufficiently general to cover all choices of representation function with this distribution. First we show that in both examples f is a solution to (4.5). In (a), we may show

$$R_Y(g_Y^* \circ f) = R_Y(g_Y^* \circ f_I) = 0.1$$

by applying (4.6). However, by applying (4.7) we may show $R_S(g_S^* \circ f) = 0.25$, while $R_S(g_S^* \circ f_I) = 0.2$. Similarly in (b),

$$R_Y(g_Y^* \circ f) = R_Y(g_Y^* \circ f_I) = 0.2,$$

while $R_S(g_S^* \circ f) = 0.25$ and $R_S(g_S^* \circ f_I) = 0.2$.

Now we compute the cost of mistrust for both cases by applying Theorem 4.8. In (a),

$$R_{YS}(g^* \circ f) - R_{YS}(h^*) = (1 \times 0) - (0.5 \times 0.1 + 0.5 \times -0.1) = 0 - 0 = 0.$$

In (b),

$$R_{YS}(g^* \circ f) - R_{YS}(h^*) = (1 \times -0.1) - (0.5 \times -0.2 + 0.5 \times -0.1) = 0.05.$$

Thus we observe that it is straightforward to construct examples where the cost of mistrust is both zero as well as those where the cost of mistrust is positive.

The examples in Figure 4.3 can be used as intuition for interpreting the expression for the cost of mistrust in Theorem 4.8. For some point $z \in \mathcal{Z}$, define its preimage

$$\mathcal{X}_z := \{x \in \mathcal{X} | f(x) = z\}.$$

If for all $x \in \mathcal{X}_z$, we have the same value of

$$\mathbf{1}(\eta_Y(x) - c_Y \geq \lambda(\eta_S(x) - c_S)),$$

as in (a), then the expectation (4.13) conditioned on $x \in \mathcal{X}_z$ will be zero. Otherwise, as in (b), the conditional expectation will be positive.

Chapter 5

Equalized Odds Implies Partially Equalized Outcomes Under Realistic Assumptions

5.1 Introduction

Definitions of fairness – and conflicts between them – are an important topic in recent quantitative fairness literature [Barocas et al., 2018]. Such definitions often involve avoiding discrimination on the basis of a particular kind of group membership, such as race or gender. In a particular situation, different definitions may be invoked by different stakeholders [Nayaranan, 2018].

The controversy created by the COMPAS recidivism prediction system showed this in practice. The system provided risk assessments about the likelihood that individuals would reoffend within a fixed period, in order to inform decisions in the criminal justice system such as whether to grant pre-trial release [Northpointe Inc., 2012]. The news organization ProPublica tested the system on past data containing both risk assessments and observed reoffences. ProPublica claimed that COMPAS was unfair towards African-Americans based on analysis showing that among observed non-reoffenders, African-Americans were more likely to be marked high risk than whites, and among observed reoffenders, whites were more likely to be marked low risk than African-Americans [Angwin et al., 2016] – i.e. the algorithm violated *equalized odds* (see Definition 5.1). The COMPAS response was that the algorithm was not unfair because among those marked high risk, African-Americans were not less likely to reoffend than whites [Dieterich et al., 2016] – i.e. the algorithm satisfied *test-fairness* (see Definition 2.1 of [Chouldechova, 2017]).

Subsequently it was shown that no algorithm can simultaneously satisfy both equalized odds and test-fairness under realistic assumptions [Chouldechova, 2017]. A similar result was shown in the more general setting of continuous rather than binary risk scores [Kleinberg et al., 2017b], replacing test-fairness with a related concept known as *calibration* (see Definition 5.10). Our work generalizes this latter result by exploring the relationship between *equalized outcomes* (see Definition 5.2) and equalized odds, as summarized in Table 5.1.

Table 5.1: Summary of main definitions and results.

Definitions
Equalized Odds
True positive rates same for each group
False positive rates same for each group
Partially Equalized Outcomes
Predicted difference between groups less than observed difference between groups
Calibration
Predicted probability equals observed probability for each group and each probability value
Results
Existing
Equalized Odds $\implies$ Not Calibration [Kleinberg et al., 2017b]
New
Equalized Odds $\implies$ Partially Equalized Outcomes
Partially Equalized Outcomes $\implies$ Not Calibration

We saw in Chapter 4 that judging the success of fair representation learning critically depends upon our choice of fairness definition. This chapter shows that conflicts between such definitions provide hard limits on what can be achieved by fair representation learning, or indeed by any technique which aims to achieve fairness in machine learning.

5.1.1 Motivation for Equalized Odds

We motivate *equalized odds*, which we define formally in Definition 5.1, using recidivism prediction as a running example (see Section 5.6.2 for a more extended discussion). Among observed non-reoffenders, we may want to ensure that those from one group are not marked higher risk on average than those from another group, i.e. our false positive rates for both groups are equal. This has been dubbed *equality of opportunity* [Hardt et al., 2016]. If we also ensure that among observed reoffenders, those from one group are not marked higher risk on average than those from another group, we have *equalized odds* [Hardt et al., 2016], i.e. our true positive rates for both groups are also equal.[1] It has been observed that in order to make the relative utility of different groups more equal, absolute utility may be reduced [Corbett-Davies and Goel, 2018; Corbett-Davies et al., 2017; Menon and Williamson, 2018]. However, equalized odds has some intuitive appeal as a fairness measure since it ensures that incorrect predictions do not disproportionately impact any group.

[1] Equality of true positive rates between groups is equivalent to equality of false negative rates between groups.

5.1.2 The Debate about Equalized Outcomes

Equality of outcomes between groups is a well-known fairness criterion [Phillips, 2004]. As we saw in Chapter 4, it can be mathematically formalized through concepts such as *statistical parity* [Calders and Verwer, 2010; Dwork et al., 2012], avoiding *disparate impact* [United States Equal Opportunity Employment Commission, 1978; Feldman et al., 2015], and achieving *independence* between outcomes and group membership [Barocas et al., 2018]. These technical definitions have prompted debate about whether they are suitable measures of fairness.

A critique of equalized outcomes is that if the observed rates (e.g. of recidivism) are different across the two groups in the training data, then an algorithm that reflects this difference is not 'unfair' but is rather a reflection of real underlying differences [Hardt et al., 2016; Zafar et al., 2017a]. The argument goes: surely we would not want to label 'unfair' a prediction algorithm which is perfectly accurate! The job of the algorithm is to predict the world as it is; changing the world is out of scope.

However, *not* equalizing outcomes across groups creates the risk of discrimination in situations where the data collection process systematically disadvantages one group [Barocas and Selbst, 2016; Zafar et al., 2017a; O'Neil, 2017]. For example, increased policing of particular populations based on pre-existing risk assessments can distort trends in reoffence data. Equalized outcomes may help algorithms to avoid perpetuating this structural inequality. More generally, the question of whether redistribution should be used to reduce inequality is at the core of the left-right political divide [Jaeger, 2008]. As such, the debate on equalized outcomes is unlikely to be definitively won or lost by either side.

5.1.3 Contribution of this Chapter

The core contribution of this chapter is to formalize and quantify the relationship between equalized odds and equalized outcomes, two important but seemingly distinct notions of fairness. We quantify the extent to which outcomes are equalized in an intuitive way, via a comparison between the predicted and observed differences between groups (Section 5.2). We prove that if we want to satisfy equalized odds, we must partially equalize outcomes – even if we only want approximately equalized odds (Section 5.3). In addition, we generalize a well-known existing result about the incompatibility of equalized odds and a different fairness measure known as *calibration* [Kleinberg et al., 2017b], using a simpler proof technique (Section 5.4). Our conclusion (Section 5.5) highlights why we should accept the reality that algorithmic decisions are imperfect when defining measures of fairness. Our technical results can be interpreted as an example of the problem of group-to-individual inference [Fisher et al., 2018]: learning from trends across groups may lead to incorrect inferences about individuals, and those incorrect inferences may disproportionately affect certain groups.

5.2 Problem Formalization

We mathematically formalize the setting we have informally described above. Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{S}$ be sets corresponding to the input variable X, target variable Y and sensitive variable S respectively. The sensitive variable encodes some form of group membership. We focus on the case where $\mathcal{S} = \{0, 1\}$ (e.g. race coded as 1 for African-American or 0 for non African-American) and $\mathcal{Y} = \{0, 1\}$ (e.g. ground truth of whether the person reoffended). The choice of $\mathcal{X}$ is arbitrary in our analysis (e.g. a person's criminal record expressed as a real-valued vector). Let $h : \mathcal{X} \times \mathcal{S} \to [0, 1]$ be a stochastic hypothesis, which can also be interpreted as a scoring function.[2] Let decision variable $\hat{Y}$ be constructed such that $p(\hat{Y} = 1 | X = x, S = s) := h(x, s)$. While setting S, Y and $\hat{Y}$ to be binary variables is an assumption, this allows us to cover many cases of interest – such as the recidivism prediction example – and facilitates our analysis and interpretation.

Drawing X, S and Y and making decision $\hat{Y}$, we have a joint distribution μ of all four variables. We may also derive marginal distributions over one or more variables, such as the marginal distribution of Y:

$$\mu_Y(Y = y) := \int_{x \in \mathcal{X}} \sum_{s \in \{0,1\}} \sum_{\hat{y} \in \{0,1\}} \mu(X = x, S = s, Y = y, \hat{Y} = \hat{y}) dx.$$

Similarly, we may derive conditional distributions, such as the marginal distribution of $\hat{Y}$ conditioned on $Y = 1$:

$$\mu_{\hat{Y}|Y=1}(\hat{Y} = \hat{y}) := \frac{\mu_{Y,\hat{Y}}(Y = 1, \hat{Y} = \hat{y})}{\mu_Y(Y = 1)}.$$

We use notation of the form $p(Y = y) := \mu_Y(Y = y)$ for marginal distributions and $p(\hat{Y} = \hat{y} | Y = 1) := \mu_{\hat{Y}|Y=1}(\hat{Y} = \hat{y})$ for conditional distributions. For example, $p(\hat{Y} = 1 | Y = 1)$ is known as the true positive rate (e.g. predicted reoffence rate for reoffenders) and $p(\hat{Y} = 1 | Y = 0)$ is known as the false positive rate (e.g. predicted reoffence rate for non-reoffenders). We use the symbol $\perp$ to denote probabilistic independence between variables.

5.2.1 Impossibility Results with respect to Fairness

An impossibility result states several candidate properties of a joint distribution, and shows that *no* distribution can simultaneously satisfy all of these properties. We briefly review several impossibility results with respect to fairness that have been established in prior works.

A well-known impossibility result (Theorem 1.1 of [Kleinberg et al., 2017b], restated in Theorem 5.11) considered the relationship between calibration – which requires that for both groups, each risk score accurately reflects the true risk associated

[2] This underpins our comparisons with [Kleinberg et al., 2017b], which analyzes risk scores. Interpreting such scores as decision probabilities facilitates our analysis.

with individuals assigned that score (see Definition 5.10) – and equalized odds. The result showed that it is impossible to simultaneously satisfy both fairness criteria and other realistic assumptions (average value of target variable differs between groups, imperfect decisions).

Variants exist involving approximate versions of equalized odds (Theorem 1 of [Pleiss et al., 2017]), calibration or both (Theorem 1.2 of [Kleinberg et al., 2017b]). We mentioned earlier the incompatibility of equalized odds and *test-fairness* – where the risk scores are binary and the true risk of individuals with a given score must be the same for both groups [Chouldechova, 2017]. Simple rules of conditional probability may be used to show that $\hat{Y} \perp S | Y$ – corresponding to equalized odds – and $Y \perp S | \hat{Y}$ – which is closely related to calibration – cannot both simultaneously hold under realistic assumptions [Barocas et al., 2018]. The incompatibility of equalized odds and the independence relationship $\hat{Y} \perp S$ (i.e. perfect statistical parity) has also been shown [Kleinberg et al., 2017b; Barocas et al., 2018].

In our work we derive impossibility results involving equalized outcomes and equalized odds, which are of interest given the debates about these fairness criteria described in the Section 5.1. As we shall see in Section 5.4, our analysis also allows us to generalize Theorem 1.1 of [Kleinberg et al., 2017b], by exploiting the relationship between equalized outcomes and calibration.

5.2.2 Fairness Definitions

We formalize the definition of equalized odds.

Definition 5.1 (Equalized odds [Hardt et al., 2016; Zafar et al., 2017a])**.** *Equalized odds is satisfied if both of the following hold:*

$$p(\hat{Y} = 1 | S = 1, Y = 1) = p(\hat{Y} = 1 | S = 0, Y = 1) \tag{5.1}$$

i.e. the true positive rate is the same for both groups, and

$$p(\hat{Y} = 1 | S = 1, Y = 0) = p(\hat{Y} = 1 | S = 0, Y = 0) \tag{5.2}$$

i.e. the false positive rate is the same for both groups.

We present a novel formalization of equalized outcomes.

Definition 5.2 (Equalized outcomes)**.** *Let*

$$p(\hat{Y} = 1 | S = 1) - p(\hat{Y} = 1 | S = 0) = \alpha(p(Y = 1 | S = 1) - p(Y = 1 | S = 0)), \tag{5.3}$$

where α is a constant we refer to as the equalized outcomes coefficient. If (5.3) *holds for $\alpha = 0$ we have* fully equalized outcomes*. If* (5.3) *holds for some $\alpha \in (0, 1)$ we have* partially equalized outcomes*. If* (5.3) *holds for $\alpha = 1$ we have* non-equalized outcomes*.*

Fully equalized outcomes corresponds to the well-known definition of perfect *statistical parity* [Calders and Verwer, 2010; Dwork et al., 2012], or equivalently the *independence* $\hat{Y} \perp S$ [Barocas et al., 2018]. The value of introducing the parameter α is that we quantify the extent to which outcomes are equalized in an intuitive way, via a comparison with the observed difference between groups. Under partially equalized outcomes, the predicted difference between groups is smaller than the observed difference between groups. Non-equalized outcomes means that predicted outcomes are *faithful* to the observed difference in outcomes between groups. If $\alpha > 1$ the predicted difference amplifies the observed difference, while if $\alpha < 0$ the predicted difference flips the sign of the observed difference. These options do not appear advantageous in terms of either fairness or accuracy, and we do not focus on them.

5.3 The Relationship Between Equalized Odds and Equalized Outcomes

Assuming equalized odds is satisfied, we show there is a quantifiable trade-off between accuracy and the extent to which outcomes are equalized. As a corollary, we show that equalized odds implies partially equalized outcomes under realistic assumptions. We consider the cases where equalized odds either exactly or approximately holds.[3]

5.3.1 Perfectly Equalized Odds

We show in Theorem 5.3 that given perfectly equalized odds, the extent to which we equalize outcomes is given by the difference α between the true positive rate and false positive rate. This novel result is of interest because it precisely quantifies the relationship between the well-known but seemingly distinct notions of equalized odds and equalized outcomes.

As we shall see shortly in Corollary 5.7, we may use Theorem 5.3 to show that, under mild assumptions, equalized odds implies partially equalized outcomes. This implies that if we have non-equalized outcomes then we cannot satisfy equalized odds. If there is an observed difference between groups (for example, in average recidivism rates), then faithfully retaining this difference in our predictions (i.e. non-equalized outcomes) might seem fair since it reflects a trend present in the world. However, we show that this would imply the violation of equalized odds, which may be perceived as unfair – for example, this was the major critique made by ProPublica of the COMPAS recidivism prediction system [Angwin et al., 2016]. To avoid violating equalized odds, the predicted difference between groups must be less than the observed difference between groups, which can be seen as a form of algorithmic 'affirmative action' [Chander, 2016], itself a controversial notion. Non-equalized outcomes (i.e. avoiding affirmative action) and equalized odds might both seem fair, but in most realistic situations we must choose one at the expense of the other.

[3] While the exact version is a special case of the approximate version, we consider the exact case first as it makes the presentation of the results more intuitive.

Theorem 5.3 (Equalized outcomes given equalized odds)**.** *Let*

$$\alpha := p(\hat{Y} = 1|Y = 1) - p(\hat{Y} = 1|Y = 0).$$

Suppose (5.1) *and* (5.2) *hold, i.e. equalized odds is satisfied. Then* (5.3) *is satisfied, i.e. α is the equalized outcomes coefficient satisfying*

$$p(\hat{Y} = 1|S = 1) - p(\hat{Y} = 1|S = 0) = \alpha(p(Y = 1|S = 1) - p(Y = 1|S = 0)).$$

Proof. We have

$$\begin{aligned} &p(\hat{Y} = 1|S = 1) \\ &= p(Y = 1|S = 1)p(\hat{Y} = 1|S = 1, Y = 1) + p(Y = 0|S = 1)p(\hat{Y} = 1|S = 1, Y = 0) \end{aligned} \tag{5.4}$$

and

$$\begin{aligned} &p(\hat{Y} = 1|S = 0) \\ &= p(Y = 1|S = 0)p(\hat{Y} = 1|S = 0, Y = 1) + p(Y = 0|S = 0)p(\hat{Y} = 1|S = 0, Y = 0) \end{aligned} \tag{5.5}$$

by the law of total probability.

Applying (5.1) and (5.2) to (5.4) yields

$$\begin{aligned} &p(\hat{Y} = 1|S = 1) \\ &\qquad = p(Y = 1|S = 1)p(\hat{Y} = 1|Y = 1) + p(Y = 0|S = 1)p(\hat{Y} = 1|Y = 0) \end{aligned} \tag{5.6}$$

and similarly, applying (5.1) and (5.2) to (5.5) yields

$$\begin{aligned} &p(\hat{Y} = 1|S = 0) = \\ &\qquad p(Y = 1|S = 0)p(\hat{Y} = 1|Y = 1) + p(Y = 0|S = 0)p(\hat{Y} = 1|Y = 0). \end{aligned} \tag{5.7}$$

Subtracting (5.7) from (5.6) and using the definition of α, we have

$$p(\hat{Y} = 1|S = 1) - p(\hat{Y} = 1|S = 0) = \alpha(p(Y = 1|S = 1) - p(Y = 1|S = 0)). \tag{5.8}$$

□

5.3.2 Realistic Assumptions

We now introduce three realistic assumptions, which help to illuminate the relationship between equalized odds and equalized outcomes. In the subsequent results in this chapter, we flag whether one or more of the assumptions is used.

The first assumption is that the observed rates (e.g. of recidivism) are different across groups, which is true for most cases of interest.

Assumption 5.4 (Different observed rates)**.**

$$p(Y = 1|S = 1) \neq p(Y = 1|S = 0) \tag{5.9}$$

The other two assumptions are that our decisions are *imperfect* (i.e. they are not always accurate) and *non-vacuous* (i.e. they have some predictive power). This covers the bulk of realistic situations in which algorithmic decisions are used. We observe that the imperfect decisions assumption will hold if Y cannot be expressed as a deterministic function of X and S. In this case, changing $\hat{Y}$ will not help. This is typically the case when we are making predictions about the future actions of individuals.

Assumption 5.5 (Imperfect decisions)**.** *At least one of the following holds:*

$$p(\hat{Y} = 1|Y = 0) > 0, \tag{5.10}$$

i.e. some negative examples are misclassified, or

$$p(\hat{Y} = 1|Y = 1) < 1, \tag{5.11}$$

i.e. some positive examples are misclassified.

Assumption 5.6 (Non-vacuous decisions)**.**

$$p(\hat{Y} = 1|Y = 1) > p(\hat{Y} = 1|Y = 0), \tag{5.12}$$

i.e. the decision is more likely to be positive for positive examples than for negative examples.

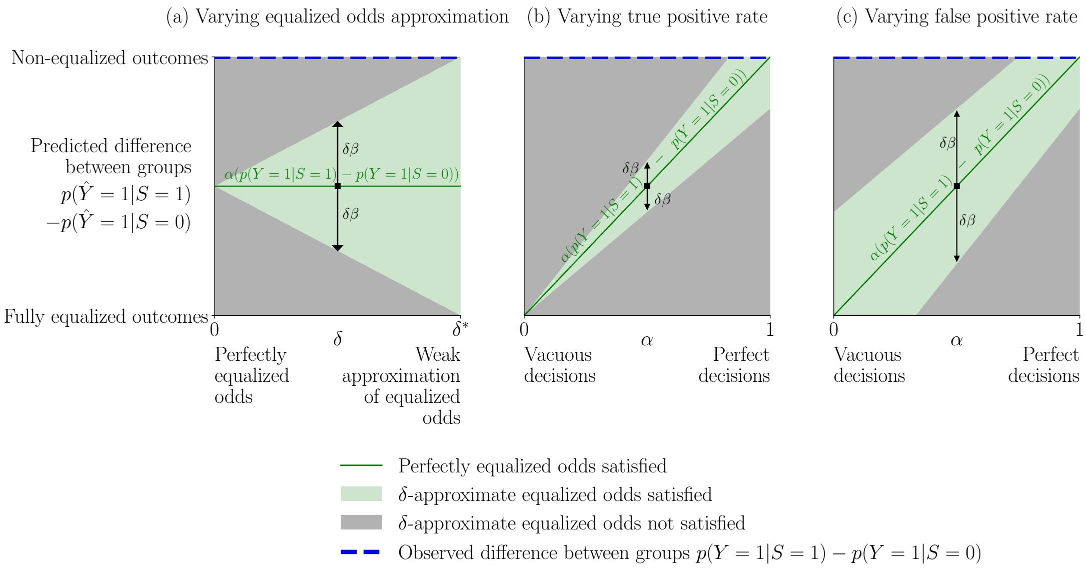
As a consequence of Theorem 5.3 and our realistic assumptions, if we have perfectly equalized odds then we have partially equalized outcomes, as shown in Corollary 5.7. While as we mentioned above the incompatibility of equalized odds and fully equalized outcomes (i.e. $\alpha = 0$, perfect statistical parity) was already known, we are the first to show that equalized odds is also incompatible with non-equalized outcomes ($\alpha = 1$) or indeed any value of α outside the interval $(0, 1)$ under our realistic assumptions.

Corollary 5.7 (Equalized odds implies partially equalized outcomes under realistic assumptions)**.** *Suppose* (5.1) *and* (5.2) *hold, i.e. we have equalized odds. Suppose also that Assumptions 5.4, 5.5 and 5.6 hold. Then satisfying* (5.3) *requires $\alpha \in (0, 1)$, i.e. we have partially equalized outcomes.*

Proof. By Theorem 5.3 we know that given equalized odds, the equation (5.3) is satisfied for

$$\alpha = p(\hat{Y} = 1|Y = 1) - p(\hat{Y} = 1|Y = 0).$$

Applying Assumption 5.4 (different observed rates), this is the *only* value of α satisfying (5.3). Applying Assumption 5.5 (imperfect decisions) we have $\alpha < 1$. Applying Assumption 5.6 (non-vacuous decisions) we have $\alpha > 0$. The result follows. □



(a) Varying equalized odds approximation (b) Varying true positive rate (c) Varying false positive rate

Figure 5.1: Visualization of key results. Certain combinations of equalized outcomes, equalized odds and accuracy are possible (light green regions), while other combinations are impossible (dark gray regions). In (a) we vary equalized odds approximation parameter δ, fixing accuracy parameter $\alpha := p(\hat{Y}=1|Y=1) - p(\hat{Y}=1|Y=0)$. In (b) we vary the $p(\hat{Y}=1|Y=1)$ term in α and in (c) we vary the $p(\hat{Y}=1|Y=0)$ term in α, fixing δ. β is a distribution-dependent parameter (see Theorem 5.9).

5.3.3 Approximately Equalized Odds

We consider a relaxation of the equalized odds condition, allowing the false positive rates to slightly differ across groups and the false negative rates to likewise slightly differ across groups. The parameter δ quantifies the degree of this relaxation, with $\delta = 0$ corresponding to perfectly equalized odds.

Definition 5.8 (Approximately equalized odds)**.** *For some constant $\delta \geq 0$, δ-approximately equalized odds holds if*

$$\begin{aligned} p(\hat{Y}=1|S=1,Y=1), p(\hat{Y}=1|S=0,Y=1) \in \\ [(1-\delta)p(\hat{Y}=1|Y=1), (1+\delta)p(\hat{Y}=1|Y=1)], \end{aligned} \tag{5.13}$$

i.e. the true positive rate is approximately the same for both groups, and

$$\begin{aligned} p(\hat{Y}=1|S=1,Y=0), p(\hat{Y}=1|S=0,Y=0) \in \\ [(1-\delta)p(\hat{Y}=1|Y=0), (1+\delta)p(\hat{Y}=1|Y=0)], \end{aligned} \tag{5.14}$$

i.e. the false positive rate is approximately the same for both groups.

In Theorem 5.9 we show that if δ-approximately equalized odds is satisfied, then the extent to which we equalize outcomes is given by an interval. This midpoint of the interval is determined by the difference α between the true positive rate and false positive rate. The size of the interval is determined by δ and a distribution-dependent parameter β. Section 5.3.4 provides interpretation of the result, by visualizing how the approximately equalized odds constraint creates limited achievable combinations of equalized outcomes and accuracy.

Theorem 5.9 (Equalized outcomes given approximately equalized odds)**.** *Let*

$$\alpha := p(\hat{Y} = 1|Y = 1) - p(\hat{Y} = 1|Y = 0),$$

$$\epsilon := p(Y = 1|S = 1) + p(Y = 1|S = 0)$$

and

$$\beta := \epsilon p(\hat{Y} = 1|Y = 1) + (2 - \epsilon)p(\hat{Y} = 1|Y = 0).$$

Observe that $\beta \geq 0$. Suppose (5.13) *and* (5.14) *hold, i.e. δ-approximately equalized odds is satisfied. Then*

$$p(\hat{Y} = 1|S = 1) - p(\hat{Y} = 1|S = 0) \in$$
$$[\alpha(p(Y = 1|S = 1) - p(Y = 1|S = 0)) - \delta\beta, \alpha(p(Y = 1|S = 1) - p(Y = 1|S = 0)) + \delta\beta].$$

Proof idea. As in the proof of Theorem 5.3, express $p(\hat{Y} = 1|S = 1)$ and $p(\hat{Y} = 1|S = 0)$ using the law of total probability. Then apply the δ-approximately equalized odds assumption to upper and lower bound their difference. See Section 5.6.1.1 for complete proof. □

5.3.4 Interpretation

We visualize our results in Figure 5.1. In each plot the vertical axis shows the predicted difference between groups, i.e. the extent to which outcomes are equalized, on a scale from zero (bottom) to the observed difference between groups (top). We vary other parameters along the horizontal axes of the plots.

If perfectly equalized odds is satisfied there is an exact relationship between equalized outcomes and α (see Theorem 5.3, green line on plots). If δ-approximate equalized odds is satisfied there is a region of permissible combinations of equalized outcomes and α values (see Theorem 5.9, light green region on plots). Combinations outside this region violate δ-approximate equalized odds (dark gray region on plots).

In Figure 5.1(a), we see that if we relax the constraint on equalized odds by increasing the parameter δ (see Definition 5.8), we have a larger region of possible combinations.[4] The size of this region is quantified by the slack term $\delta\beta$. The region is an interval centered on the product of α and the observed difference between

[4] When δ equals $\delta^* := \frac{\alpha}{\beta}(p(Y = 1|S = 1) - p(Y = 1|S = 0))$, the edge of this region intersects the horizontal axis. Figure 5.1(a) uses the fixed parameters $\alpha := 0.5$, $\beta := 1$ and $\epsilon := 1$.

groups. We see visually why for $\alpha \in (0,1)$, i.e. for decisions that are imperfect and non-vacuous, we have partially equalized outcomes.

Figures 5.1(b) and 5.1(c) show that if we have equalized odds, then increasing accuracy (measured by α) moves towards non-equalized outcomes.[5] We may increase α by increasing the true positive rate, as in Figure 5.1(b), where we assume no false positives. We may also increase α by decreasing the false positive rate, as in Figure 5.1(c), where we assume no false negatives. Under perfectly equalized odds the effect on equalized outcomes is the same, while under approximately equalized odds the permissible regions differ because β depends on the false positive rate and the true positive rate.

5.4 Generalization of Calibration-Equalized Odds Impossibility Result

The relationship between equalized odds and equalized outcomes, in addition to its intrinsic interest, allows us to generalize a well-known result about the impossibility of simultaneously satisfying calibration and equalized odds (Theorem 1.1 of [Kleinberg et al., 2017b]). We use a proof technique involving elementary probabilities, which also provides a simpler proof of Kleinberg's result.

5.4.1 Review of Existing Result

We first introduce the definition of group-conditional calibration proposed in previous work [Kleinberg et al., 2017b; Pleiss et al., 2017]. This means that for both groups, each risk score equals the observed risk associated with individuals assigned that score.

Definition 5.10 (Group-conditional calibration [Kleinberg et al., 2017b; Pleiss et al., 2017]). *Both of the following statements hold* $\forall c \in [0,1]$*:*

$$p(Y=1|h(x,s)=c, S=1) = c \tag{5.15}$$

$$p(Y=1|h(x,s)=c, S=0) = c \tag{5.16}$$

We now state the well-known calibration-equalized odds impossibility result (Theorem 1.1 of [Kleinberg et al., 2017b], restated to align with our definitions).

Theorem 5.11 (Calibration-equalized odds impossibility [Kleinberg et al., 2017b]). *Suppose* (5.1), (5.2), (5.15) *and* (5.16) *hold, i.e. equalized odds and group-conditional calibration are both satisfied. Then at least one of Assumption 5.4 or Assumption 5.5 is violated, i.e. the observed rates are the same for both groups and/or the decision is perfect.*

In other words, equalized odds implies not calibration under realistic assumptions, as stated in Table 5.1.

[5] Figures 5.1(b) and 5.1(c) use $\epsilon := 1$ and $\delta := 0.2(p(Y=1|S=1) - p(Y=1|S=0))$.

5.4.2 Group-Conditional Calibration Implies Non-Equalized Outcomes

In preparation for generalizing Theorem 5.11, we show in Lemma 5.12 that group-conditional calibration implies non-equalized outcomes but not vice versa.

Lemma 5.12 (Group-conditional calibration implies non-equalized outcomes but not vice versa)**.** *If* (5.15) *and* (5.16) *hold, then* (5.3) *holds for* $\alpha = 1$*, i.e. group-conditional calibration implies non-equalized outcomes. However, if* (5.3) *holds for* $\alpha = 1$*, then it is not the case that* (5.15) *and* (5.16) *must hold, i.e. non-equalized outcomes does not imply group-conditional calibration.*

Proof idea. Use laws of probability to show that group-conditional calibration implies non-equalized outcomes. Then construct a counterexample to show that non-equalized outcomes does not imply group-conditional calibration. See Section 5.6.1.2 for complete proof. □

Using the contrapositive of the fact that group-conditional calibration implies non-equalized outcomes, partially equalized outcomes implies not calibration as stated in Table 5.1. We observe that in contrast to group-conditional calibration, test-fairness as proposed in [Chouldechova, 2017] does not in general imply non-equalized outcomes.

5.4.3 The Generalized Result

The existing result stated in Theorem 5.11 shows that if group-conditional calibration and equalized odds hold, realistic assumptions are violated. Our new result in Theorem 5.13 shows that if non-equalized outcomes and equalized odds hold, the same realistic assumptions are violated.

As we just showed in Lemma 5.12, group-conditional calibration implies non-equalized outcomes but not vice versa, i.e. non-equalized outcomes is a weaker condition than group-conditional calibration. Therefore Theorem 5.13 is more general than Theorem 5.11, since with a weaker condition we arrive at the same conclusion. It is straightforward to see that Lemma 5.12 and Theorem 5.13 together imply Theorem 5.11. We observe that our proof technique appears simpler, since it relies only on elementary manipulation of probabilities.

Theorem 5.13 (Generalization of calibration-equalized odds impossibility result)**.** *Suppose* (5.1) *and* (5.2) *hold, and* (5.3) *holds for* $\alpha = 1$*, i.e. equalized odds and non-equalized outcomes are both satisfied. Then at least one of Assumption 5.4 or Assumption 5.5 is violated, i.e. the observed rates are the same for both groups and/or the decision is perfect.*

Proof. Suppose (5.3) holds for $\alpha = 1$, i.e. non-equalized outcomes is satisfied, and (5.1) and (5.2) hold, i.e. equalized odds is satisfied. Applying Theorem 5.3,

$$\begin{aligned} &p(\hat{Y} = 1|S = 1) - p(\hat{Y} = 1|S = 0) \\ &= (p(\hat{Y} = 1|Y = 1) - p(\hat{Y} = 1|Y = 0))(p(Y = 1|S = 1) - p(Y = 1|S = 0)). \end{aligned} \tag{5.17}$$

Combining (5.3) and (5.17), we have

$$\begin{aligned}&p(Y=1|S=1)-p(Y=1|S=0)\\&=(p(\hat{Y}=1|Y=1)-p(\hat{Y}=1|Y=0))(p(Y=1|S=1)-p(Y=1|S=0)).\end{aligned}\tag{5.18}$$

We conclude from (5.18) that at least one of the following holds:

$$p(Y=1|S=1)=p(Y=1|S=0)\tag{5.19}$$

$$p(\hat{Y}=1|Y=1)-p(\hat{Y}=1|Y=0)=1\tag{5.20}$$

If (5.19) holds then Assumption 5.4 is violated, i.e. the observed rates are the same for both groups. If (5.20) holds, then $p(\hat{Y}=1|Y=1)=1$ and $p(\hat{Y}=1|Y=0)=0$. Therefore Assumption 5.5 is violated, i.e. the decision is perfect. □

5.5 Conclusion

When algorithms make predictions of the future actions of individuals, a certain degree of inaccuracy seems inevitable. In this context, naively using trends observed across groups to make predictions about individuals – a problem known as group-to-individual inference [Fisher et al., 2018] – creates the risk that incorrect inferences may disproportionately affect certain groups, in the legal system and beyond. We have formalized the intuition that when algorithms conduct group-to-individual inference – or in other words, *stereotype* – they tend to be unfair to individuals who are 'atypical' (e.g. non-reoffenders from a group with higher reoffence rates). In particular, we have seen that an imperfect algorithm for which the predicted and observed differences between groups are equal will violate equalized odds. Avoiding this requires partially equalized outcomes, which can be seen as an instantiation of 'algorithmic affirmative action' [Chander, 2016].

5.6 Appendix

We present the remaining proofs of the chapter's theoretical results, as well as motivating examples of equalized odds and its relationship to equalized outcomes.

5.6.1 Supplementary Proofs

We present the proofs of Theorem 5.9 and Lemma 5.12.

5.6.1.1 Proof of Theorem 5.9 (Equalized Outcomes Given Approximately Equalized Odds)

Proof. We have

$$\begin{aligned}&p(\hat{Y}=1|S=1)\\&=p(Y=1|S=1)p(\hat{Y}=1|S=1,Y=1)+p(Y=0|S=1)p(\hat{Y}=1|S=1,Y=0)\end{aligned}$$

and

$$\begin{aligned}&p(\hat{Y}=1|S=0)\\&=p(Y=1|S=0)p(\hat{Y}=1|S=0,Y=1)+p(Y=0|S=0)p(\hat{Y}=1|S=0,Y=0)\end{aligned}$$

by the law of total probability.

Assuming δ-approximately equalized odds, we have

$$\begin{aligned}&p(\hat{Y}=1|S=1)-p(\hat{Y}=1|S=0)\\&\leq p(Y=1|S=1)(1+\delta)p(\hat{Y}=1|Y=1)+p(Y=0|S=1)(1+\delta)p(\hat{Y}=1|Y=0)\\&\quad-p(Y=1|S=0)(1-\delta)p(\hat{Y}=1|Y=1)-p(Y=0|S=0)(1-\delta)p(\hat{Y}=1|Y=0)\\&=\alpha(p(Y=1|S=1)-p(Y=1|S=0))+\delta\beta.\end{aligned}$$

The equality follows by rearranging the terms and using the definitions of α and β.

Similarly,

$$\begin{aligned}&p(\hat{Y}=1|S=1)-p(\hat{Y}=1|S=0)\\&\geq p(Y=1|S=1)(1-\delta)p(\hat{Y}=1|Y=1)+p(Y=0|S=1)(1-\delta)p(\hat{Y}=1|Y=0)\\&\quad-p(Y=1|S=0)(1+\delta)p(\hat{Y}=1|Y=1)-p(Y=0|S=0)(1+\delta)p(\hat{Y}=1|Y=0)\\&=\alpha(p(Y=1|S=1)-p(Y=1|S=0))-\delta\beta.\end{aligned}$$

□

5.6.1.2 Proof of Lemma 5.12 (Group-Conditional Calibration Implies Non-Equalized Outcomes but Not Vice Versa)

Proof. Suppose (5.15) and (5.16) hold, i.e. group-conditional calibration is satisfied. Then

$$\begin{aligned}&p(Y=1|S=1)-p(Y=1|S=0)\\&=\int_0^1 p(h(x,s)=c|S=1)p(Y=1|h(x,s)=c,S=1)\,dc\\&\quad-\int_0^1 p(h(x,s)=c|S=0)p(Y=1|h(x,s)=c,S=0)\,dc\end{aligned}$$

by the law of total probability

$$= \int_0^1 p(h(x,s) = c|S = 1)c\,dc - \int_0^1 p(h(x,s) = c|S = 0)c\,dc$$

by group-conditional calibration, substituting in (5.15) and (5.16)

$$\begin{aligned}&= \int_0^1 p(h(x,s) = c|S = 1)p(\hat{Y} = 1|h(x,s) = c, S = 1)\,dc \\ &- \int_0^1 p(h(x,s) = c|S = 0)p(\hat{Y} = 1|h(x,s) = c, S = 0)\,dc\end{aligned}$$

by the definition $p(\hat{Y} = 1|X = x, S = s) := h(x,s)$

$$= p(\hat{Y} = 1|S = 1) - p(\hat{Y} = 1|S = 0)$$

by the law of total probability. Hence (5.3) holds for $\alpha = 1$ and we have shown that group-conditional calibration implies non-equalized outcomes.

However, we may have non-equalized outcomes without group-conditional calibration. For example, consider the case that

$$h(x,s) = p(Y = 1|S = s) + \eta$$

where η is generated by random noise with range

$$[-p(Y = 1|S = s), 1 - p(Y = 1|S = s)]$$

and mean zero.

Therefore

$$\begin{aligned}&p(\hat{Y} = 1|S = s) \\ &= \int_x p(X = x|S = s)p(\hat{Y} = 1|X = x, S = s)dx \\ &= \int_x p(X = x|S = s)h(x,s)dx \\ &= \int_x p(X = x|S = s)[p(Y = 1|S = s) + \eta]dx \\ &= p(Y = 1|S = s).\end{aligned}$$

Hence (5.3) holds for $\alpha = 1$, i.e. we have non-equalized outcomes.

We also have $\forall c \in [0,1]$

$$p(Y = 1|h(x,s) = c, S = 1) = p(Y = 1|S = 1)$$

and

$$p(Y = 1|h(x,s) = c, S = 0) = p(Y = 1|S = 0).$$

Hence (5.15) and (5.16) do not in general hold and we have shown that non-equalized outcomes does not imply group-conditional calibration. □

5.6.2 Motivating Examples of Equalized Odds and its Relationship to Equalized Outcomes

We present a motivating example for equalized odds using recidivism prediction. We also present an example which motivates the relationship between equalized odds and equalized outcomes.

5.6.2.1 Equalized Odds

We explore the definition of and rationale behind *equalized odds*, using recidivism prediction with the ProPublica dataset, which contains information about 7214 criminal offences committed in Broward County, Florida. We used the individual's age, gender, race and criminal history to predict whether they would reoffend within two years.[6] We applied a 70/30 training/test split of the data, trained a logistic regression model[7] on the training set, and used this model to predict the probability that each individual in the test set would reoffend.

The model achieved an area under the curve (AUC) of 0.72 on the test set, indicating that the model is far from perfect but a lot better than a random guess.[8] The results are shown in Table 5.2. We note there is a difference in the observed reoffence rates between African-American and non African-American individuals in the data. The predicted reoffence rates were close to the observed reoffence rates for both groups, and thus showed a difference of a similar magnitude, i.e. non-equalized outcomes was approximately satisfied. The model rated African-American individuals as higher risk on average, but one could justify this by arguing that the model simply reflects trends in the data.

However, looking separately at those individuals who were observed as non-reoffenders, and those who were observed as reoffenders, we find that the predictions were far from satisfying equalized odds. Looking at the non-reoffenders, for African-Americans the predicted reoffence rate was 47.8% while for non African-Americans it was 36.2%. In other words, the *false positive rate* was much higher for African-Americans than for non African-Americans. Now looking only at the reoffenders, we notice a difference in the *true positive rate* across racial groups – for African-Americans the predicted reoffence rate was 61.5% while for non African-Americans it was 47.4%. Equivalently, the *false negative rate* for non African-Americans (52.6%) was much higher than for African-Americans (38.5%).

Among non-reoffenders, non African-Americans would be better off with this model since they are less likely to be incorrectly classified as high risk. Among reoffenders, non African-Americans would also be better off since they are more

[6] The dataset is available at https://github.com/propublica/compas-analysis/blob/master/compas-scores.csv. We predicted the column `is_recid` using `sex`, `age_cat`, `juv_fel_count`, `juv_misd_count`, `juv_other_count` `priors_count` and `c_charge_degree`, representing categorical variables as a one-hot encoding.

[7] Implemented in Python using the sklearn package.

[8] AUC can be interpreted as the probability that a randomly selected positive example will receive a higher score than a randomly selected negative example. A perfect classifier achieves an AUC of 1, while a random classifier achieves an AUC of 0.5.

Table 5.2: Test set results for a recidivism prediction model on the ProPublica dataset. The example motivates equalized odds.

Metric	**African-American**	**Non African-American**	**Overall**
Observed reoffence rate	54.1%	39.9%	47.2%
Predicted reoffence rate	55.2%	40.7%	48.1%
Predicted reoffence rate among non-reoffenders	47.8%	36.2%	41.4%
Predicted reoffence rate among reoffenders	61.5%	47.4%	55.7%

likely to be incorrectly classified as low risk. These two types of discrimination are precisely what ProPublica reported about the COMPAS algorithm [Angwin et al., 2016]. Our example has shown how easily this can occur, even if on the face of it the model seems to just reflect differences between two groups in its training data. It also shows how individuals are impacted by inferences made from past observations of others who appear similar to them – in effect they are stereotyped by the algorithm.

In summary, our example has shown how a model's true positive rates and false positive rates may differ across groups, which may disadvantage a particular group. This observation motivates the definition of equalized odds – requiring that the true positive rates and false positive rates are equal across groups – which, if satisfied, prevents this form of disadvantage [Hardt et al., 2016].

5.6.2.2 The Relationship between Equalized Odds and Equalized Outcomes

Continuing with our ProPublica dataset example, we ask whether our findings – that our observed and predicted reoffence rates were close for both groups, and that we violated equalized odds – are quirks of this particular algorithm or dataset? As our theoretical results have shown, this is far from a coincidence – in fact, under realistic assumptions this combination is inevitable!

The core contribution of our work is to formalize the relationship between equalized odds and equalized outcomes. To provide further intuition on this relationship, we pre-processed the ProPublica data to suppress information about race using a technique proposed in [Edwards and Storkey, 2016]. The technique is governed by a parameter λ – increasing this parameter changes the data to make it harder to distinguish between the records of African-Americans and non African-Americans.[9] We then ran logistic regression (as in Section 5.6.2.1) on the pre-processed data and reported results on the test set, as shown in Figure 5.2.

[9] We learned a representation function f, which is applied to each input, by approximating (4.8) via generator and adversary neural networks trained in tandem, i.e. *learning fair representations with an adversary*. The objective function jointly depends on how well the generator approximates the input, and how well an adversary can estimate a particular sensitive variable (in this case race) from it. The latter is more important for larger λ. Simply omitting the sensitive variable is not sufficient, since it may be possible to infer this variable from other columns.

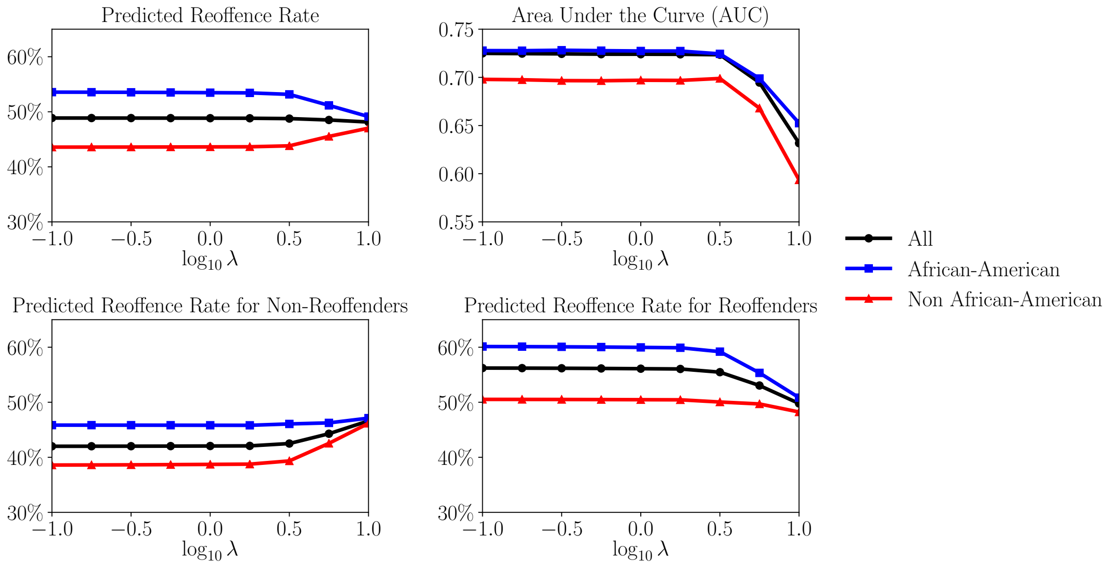


Figure 5.2: Motivating example: equalized odds appears related to equalized outcomes. The horizontal axis shows the parameter λ used in pre-processing (see text) on a log scale, while the vertical axes show several performance measures of interest.

This technique yielded more *equalized outcomes* with increasing λ, i.e. the predicted reoffence rates for African-Americans and non African-Americans became closer (top left). The accuracy of the model as measured by AUC declined somewhat with increasing λ (top right). The predicted reoffence rates for non-reoffenders became closer for the two groups with increasing λ (bottom left). The predicted reoffence rates for reoffenders for the two groups also became closer (bottom right). In other words, we achieved a tighter approximation of equalized odds by increasing λ.

In summary, our example showed anecdotal evidence of a relationship between equalized odds and equalized outcomes, and raised questions about whether this relationship has a mathematical foundation. The technical results in this chapter have formalized the mathematical relationship between these two notions of fairness.

Part III

Applying Representation Learning to Fairness: Two Case Studies

Chapter 6

Trade-offs in Algorithmic Risk Assessment: an Australian Domestic Violence Case Study

6.1 Introduction

Actuarial methods have been part of criminal law and its enforcement in jurisdictions around the world for nearly a century [Harcourt, 2006]. These methods employ probability theory to shape risk management tools designed to help humans make decisions about who to search, what geographical areas to police, eligibility for bail, eligibility for parole, the length of a criminal sentence and the kind of prison a convicted offender should be incarcerated in [Harcourt, 2006]. The criminal justice system can be said to have been employing algorithms and crunching 'big' data for decision-making long before these words became part of the popular lexicon surrounding automated decisions.

More recently, a range of commercial and government providers have developed software that embeds actuarial methods in code, using machine learning methods on large bodies of data and marketed under the umbrella of artificial intelligence (AI) [Berk, 2012]. While the effects of using these kinds of probabilistic methods in criminal justice contexts – such as higher incarceration rates among certain racial groups and distorted future predictions – have been critiqued by legal and social science scholars for several years [Rice and Harris, 1995], they have also become issues for the computer scientists and engineers developing these software solutions.

In-depth investigations of commercial criminal recidivism algorithms, like the COMPAS software developed by US-based company Equivant (formerly known as Northpointe), have become flashpoints in discussions of bias and prejudice in AI (see Chapter 5 and [Angwin et al., 2016]). Within the computer science community, developing quantitative methods to build fairer, more transparent decision-making systems is an increasingly important research area [Nayaranan, 2018].

We trial one quantitative approach designed to address potential discrimination in the outputs of a pre-existing case study predicting domestic violence recidivism in the Australian context. This chapter is a case study that considers the practical potential and consequences of fair representation learning in the criminal justice system, and complements the theoretical perspective on fair representation learning we presented in Chapter 4.

As we have already seen in Chapters 4 and 5, there is no one authoritative definition of fairness, in computer science or in any other discipline. 'Fairness' as a word carries significant cultural heritage [Wierzbicka, 2006]. John Rawls' famed "veil of ignorance" proposes an approach to fairness akin to an impartial observer, who does not know what status they will have in society and how the definition of fairness is agreed on [Rawls, 1971]. Other scholars have noted this abstract approach of fairness, when put into practice, does not reduce perceptions of unfair outcomes [Trautmann and van de Kuilen, 2016]. Previous explorations of varied definitions of fairness in disciplines as diverse as philosophy, law, neuroscience and information theory have concluded there is no single foundation on which to rest for the purposes of fair algorithms [Menon and Williamson, 2018].

To paraphrase the science fiction author Margaret Atwood: "Fair never means fairer for everyone. It always means worse, for some" [Atwood, 1985]. This chapter does not assert its approach to fairness as the 'right' one. What is 'fair' is not a technical consideration, but a moral one [Nayaranan, 2018]. We are interested in the insights that quantitative methods for fairness give human decision makers, allowing us to make explicit certain implicit trade-offs that have long been part of how humans make decisions. Efforts to quantify what is 'fair' allow us to measure the impact of these trade-offs.

Used effectively in a criminal justice context, machine learning methods could help human decision makers make more transparent, informed decisions about a person's likelihood of recidivism. Whatever definition of 'fairness' is employed, there are real world consequences. The impact of varying trade-offs in 'fair' decision-making on victims and offenders should be carefully considered in a domestic violence context.

6.2 Algorithmic Risk Assessment in an Australian Domestic Violence Context

In a 2016 paper [Fitzgerald and Graham, 2016], Australian researchers evaluated the potential of using existing administrative data drawn from the NSW Bureau of Crime Statistics and Research (BOCSAR) Re-offending Database (ROD) to predict domestic violence-related recidivism [NSW Bureau of Crime Statistics and Research, 2018]. Being able to reliably and accurately assess which offenders, in which contexts, are likely to recommit domestic violence is a priority for law enforcement, victim support services and of course, for victims themselves.

Domestic violence (DV), also referred to as family violence or domestic abuse, is defined as a pattern of violence, intimidation or abuse between individuals in a current or former intimate relationship. A World Health Organization study found that within each of dozens of studies conducted around the world, between 10% and

69% of women reported having experienced physical abuse by an intimate partner, and between 5% and 52% reported having experienced sexual violence by an intimate partner [Krug et al., 2002].

In Australia, one in six women and one in twenty men have experienced at least one instance of domestic violence since the age of 15 [Australian Bureau of Statistics, 2017b; Cox, 2012]. On average, police in Australia respond to a domestic violence matter every two minutes [Bulmer, 2015]. These statistics emphasize the scale and the gendered nature of this issue. Indeed, aggregate prevalence rates further highlight the negative impact of DV and family violence more broadly. DV is one of the top ten risk factors contributing to disease burden among adult women [Australian Institute of Health and Welfare and Australia's National Research Organisation for Women's Safety, 2016; Australian Institute of Health and Welfare, 2018], and the economic costs of violence against women and children in Australia (including both domestic and non-domestic violence) are estimated at around $13.6 billion per year [Australian Government Department of Social Services, 2009]. Existing statistics and surveys suggest that Indigenous communities face domestic violence issues at much greater rates than the rest of the population.[1]

6.2.1 The Evolution of Algorithmic Risk Assessments

Actuarial methods and probability theory have been employed to help humans make decisions in a criminal justice context for many years [Harcourt, 2006]. It's only recently that they've been embedded in software [Desmarais and Singh, 2013]. While these longstanding methods could be said to be 'algorithmic' in nature – taking a rule-based approach to predictions – for the purposes of this chapter we use the term "algorithmic risk assessment" to refer to the more recent automated, software-driven systems. An example is the Public Safety Assessment [Laura and John Arnold Foundation, 2017], which has been used in the U.S. states of Kentucky, Arizona and New Jersey and several other U.S. counties [Laura and John Arnold Foundation].

Algorithmic risk assessment systems have several potential advantages. They offer a mechanism to improve the accuracy of decisions made in the criminal justice system.[2] They are readily scalable, offering greater consistency than human judgment [Kleinberg et al., 2017a]. They offer increased transparency of decisions, if the system's code, methodology and input data are accessible [Zeng et al., 2017]. And they often have adjustable parameters (as in this work), which render trade-offs explicit in decision-making and allow them to be managed.

However, investigations of existing algorithmic risk assessment systems have demonstrated that these systems can – by choice – also be shrouded in secrecy, un-

[1] In NSW in 2016, 2.9% of the population were Indigenous [Australian Bureau of Statistics, 2017a] while 65% of victims of family and domestic violence overall were Indigenous [Australian Bureau of Statistics, 2017d].

[2] For example, a recent study using data from more than 750,000 pre-trial release decisions made by New York City judges found that, at the same jailing rate as human judges, an algorithm could reduce crime by 14.4-24.7%. Alternatively, without any increase in crime, an algorithm could reduce jail rates by 18.5-41.9% [Kleinberg et al., 2017a].

necessarily complex, and disadvantageous to particular groups [Angwin et al., 2016]. It has been shown that COMPAS – which used over a hundred variables for predictions – performs no better than a logistic regression classifier using age and total number of previous convictions [Dressel and Farid, 2018]. A controversial recent example of a risk assessment system in the Australian context is the Suspect Targeting Management Plan (STMP) [NSW Police Force, 2016]. In the cases of both COMPAS [Angwin et al., 2016] and STMP [Sentas and Pandolfini, 2017], concerns have been raised that the systems are unfair, in the former case towards African-Americans and in the latter case towards Indigenous Australians.

6.2.2 Predicting Domestic Violence Recidivism using Administrative Data

A primary aim of any recidivism prediction is accuracy.[3] This allows law enforcement agencies to identify which offenders are most likely to recommit a crime and subsequently (1) adjust their access to bail or parole, or period of incarceration accordingly; and (2) understand the risk factors associated with recidivism in order to better target resources and programs aimed at crime prevention. But what is considered an 'accurate' prediction is complicated by risk-based, profiling approaches to policing that inevitably see certain populations overrepresented in data about past offenders, which is then used for making future predictions. In what senses are predictions based on this past data 'fair', and to whom are they 'fair'? Answering this question depends on identifying and managing the trade-offs involved in the design of recidivism assessments.

Although domestic violence (DV) is a serious problem in Australia, to date there has been relatively little research on the risks associated with family violence and DV recidivism in the Australian context [Boxall et al., 2015; Fitzgerald and Graham, 2016]. Recidivism in this chapter refers to reoffending following conviction for an offence. Broadly speaking, a 'reoffender' is an individual who is a repeat or chronic offender. In the context of DV recidivism, national and state-based agencies have begun to develop and implement computerized decision support systems (DSS) and risk assessment tools that draw on standardized data (within and/or across agencies) to help understand the risk of DV recidivism for sub-groups within the population. There is increasing interest in evidence-based crime and social welfare governance that draw on data science and big data, perhaps due to a perception that these kinds of DSS and risk assessment tools are more efficient, objective and less costly than existing approaches [Gillingham and Graham, 2017].

The point of these DSS and risk assessment tools is to enhance, refine and better target programs and resources to prevent DV, rather than simply punishment and control. While computer-based DSS have been criticized in, for example, child welfare and protection [Gillingham, 2006], recent studies suggest that DV-related risk assessment tools can be effective, particularly to assist under-resourced front-line agencies to make informed and speedy decisions about detention, bail and victim

[3] The term 'accuracy' is used here in a broad sense. In practice, a cost-sensitive risk may be appropriate given that false positives and false negatives may carry different social costs.

assistance [Mason and Julian, 2009; Messing et al., 2017]. A standard practice is to measure the accuracy of risk assessment tools using Area Under the Curve (AUC) [Fawcett, 2004]. Predictive risk assessment tools for DV recidivism have been shown to provide reasonably high levels of predictive performance, with AUC scores in the high 0.6 to low 0.7 range [Rice et al., 2010].[4]

6.2.3 Findings from Previous Studies

Fitzgerald and Graham [2016] applied statistical methods to existing administrative data on NSW offenders who had committed domestic violence, to examine the kinds of factors – for example, socioeconomic status, history of past offences, Indigenous or non-Indigenous status – which were predictive of future domestic violence offences. They used logistic regression to examine the future risk of violent DV offending among a cohort of individuals convicted of any DV offence (regardless of whether it is violent or not) over a specific time period. They found that applying their models to unseen data achieved AUC of 0.69, indicating a reasonable level of predictive accuracy, on par with other risk assessment tools in other countries and contexts. A follow-up study explored using a decision tree induction approach on the same dataset [Wijenayake et al., 2018]. Although these prior works showed the potential for such models to be deployed to enhance targeted programs and resources for DV prevention, Fitzgerald and Graham's study also highlighted a significant problem: the authors found that the use of their model could disadvantage the Indigenous population in the justice system.

Fitzgerald and Graham argued that whilst DSS that incorporate logistic regression might offer a satisfactory tool for predicting the risk of domestic violence recidivism in the *overall population*, the efficacy is reduced for making predictions for particular sub-groups, particularly for individuals who identify as Indigenous. In their study, Indigenous individuals were *more than twice as likely* to be predicted as reoffenders (29.4%) by the model compared to the observed rate (13.7%), whereas non-Indigenous individuals were less than *half as likely* to be predicted as reoffenders (2.3%) compared to the observed rate (6.1%).[5]

In other words, when it came to predicting DV recidivism for the Indigenous subgroup, Fitzgerald and Graham found that the model was biased on two fronts: over-predicting Indigenous reoffenders and under-predicting non-Indigenous reoffenders. If deployed as a risk assessment tool, this model could have serious negative consequences that may reinforce existing inequalities that have resulted from historical and contemporary injustices and oppression of Indigenous Australians. The output of the model not only reflects but also potentially *amplifies and reinforces* these inequalities. Indeed, the fact that Indigenous status (as an independent variable) appears at

[4] AUC can be interpreted as the probability that a randomly selected reoffender will receive a higher risk score than a randomly selected non-reoffender. A random guess has expected AUC of 0.5 while the perfect prediction has AUC of 1.

[5] Looking at the entire population the predicted (7.4%) and observed (7.6%) recidivism rates are relatively well-aligned. The large differences between predicted and observed recidivism rates only become visible looking separately at the Indigenous and non-Indigenous cohorts.

all in the dataset brings to light the politics of data collection and statistical forms of reasoning. The data provided through the BOCSAR ROD, and subsequently used in the study by Fitzgerald and Graham, reflects a "practical politics" that involves negotiating and deciding what to render visible (and invisible) in an information system context [Bowker and Star, 1996]. This example shows the importance of the issue of fairness in algorithmic decision-making as we move towards computerized risk assessment tools in criminal justice and social welfare. At the same time, caution needs to be taken in how such fairness is defined and achieved.

6.3 Designing Fair Algorithmic Risk Assessments

The impact of an algorithmic risk assessment is determined by both its design and the context in which it is used. This context – which includes human judgment, policy settings and broader social trends – will remain an important determinant of outcomes in the justice system and elsewhere. No algorithm can rectify all of the past and present structural disadvantage faced by a particular social group. However, algorithmic risk assessments influence human decisions, which in turn determine the extent to which structural disadvantage is entrenched. As we have already seen in Chapters 4 and 5, considerable research is underway to incorporate fairness into the design of algorithmic systems. This approach requires clear definitions of fairness, and modifications to algorithm design to accommodate these definitions.

6.3.1 Defining Fairness in the Australian DV Recidivism Context

We must be precise about what we mean if we are to embed fairness in computer code – a definition that seems simplistic or reductionist may still be preferable to none at all. Therefore we necessarily consider a narrow subset of the possible meanings of 'fairness'. We consider applying parity-based definitions of fairness, which we introduced in Chapter 4, to the context of recidivism prediction. Parity-based definitions may be used to assess the fairness of a recidivism risk assessment model which generates a probability that an individual will reoffend. Given the issues associated with the context of DV in Australia, parity between Indigenous and non-Indigenous populations in the criminal justice system is of special interest.

An important design choice is selecting a subset of the population to which a particular parity-based definition of fairness is applied. We then ask for parity of average predictions between groups only within this subset. For example, in the recidivism context we might consider all individuals, or only those who reoffended, or only those who did not reoffend. Consider the difference between Indigenous and non-Indigenous populations for each of the following:

- **Predicted reoffence rate**: the average probability of reoffence predicted by the model.

- **Predicted reoffence rate for non-reoffenders**: the average probability of reoffence predicted by the model, for those individuals who were not observed to reoffend.

- **Predicted reoffence rate for reoffenders**: the average probability of reoffence predicted by the model, for those individuals who were observed to reoffend.

We recall several possible parity-based definitions of fairness introduced in Chapter 4, and discuss them in the context of predicting DV recidivism in Australia. Parity between groups of predicted reoffence rates among non-reoffenders is referred to as *equality of opportunity* [Hardt et al., 2016] in the quantitative fairness literature. If we also have parity of predicted reoffence rates among reoffenders, this is referred to as *equalized odds* [Hardt et al., 2016] (also known as avoiding *disparate mistreatment* [Zafar et al., 2017a]). Enforcing these parity measures between Indigenous and non-Indigenous populations has some intuitive appeal, since it ensures that disagreements between the algorithm's predictions and the subsequently observed data do not disproportionately impact one racial group. However, these measures are sensitive to the way in which the reoffence data was collected. Profiling of particular populations, based on pre-existing risk assessments, can distort trends observed in reoffending. A feedback loop may be created, where this reoffence data in turn influences future risk assessments [O'Neil, 2017].

Overall parity between groups of predicted reoffence rate is referred to in the quantitative fairness literature as *statistical parity* [Dwork et al., 2012] or avoiding *disparate impact* [Zafar et al., 2017a]. We may not want overall parity of predicted reoffence rate if the observed rates of reoffence for Indigenous and non-Indigenous populations are different. However, overall parity has the advantage that it does not depend on the way that reoffence data was collected, which may systematically disadvantage one group [Barocas and Selbst, 2016]. Furthermore, an actual difference in reoffence rates between groups may be the result of a complex historical process. In the case of Indigenous Australians this includes founding violence, structural violence, cultural breakdown, intergenerational trauma, disempowerment, and alcohol and drugs [The Healing Foundation and White Ribbon Australia, 2017]. Legal decision-makers may wish to intervene in this process by reducing the discrepancy between incarceration rates for Indigenous and non-Indigenous populations.[6] To support this intervention, it may be appropriate for the design of a risk assessment system to incorporate greater parity in predicted reoffence rates. By contrast, other fairness definitions may be used to justify and perpetuate current rates of Indigenous incarceration.

A risk assessment model should also be accurate, subject to the previous caveat that reoffence data is likely to be imperfect and is possibly biased. While AUC does not consider fairness with respect to group membership, it is certainly related to fairness insofar as it measures the extent to which observed reoffenders are assessed as higher risk than observed non-reoffenders.

[6] As of 2017, the incarceration rate of Australia's Aboriginal and Torres Strait Islander population stood at 2434 per 100,000 people, versus 160 per 100,000 people for the non-Indigenous population [Australian Bureau of Statistics, 2017c].

6.3.2 Learning Fair Representations with an Adversary

We are interested in using a modified recidivism prediction algorithm to achieve greater parity in predicted outcomes for Indigenous and non-Indigenous populations. We saw in Chapter 4 that there are several possible approaches to achieving this. For a review of using techniques for algorithmic fairness in the context of recidivism prediction, see [Berk et al., 2017].

In this chapter we explore using an approach known as *learning fair representations with an adversary*, which was proposed in [Edwards and Storkey, 2016] and we considered in Chapter 4. While this is not the only possible method for this problem setting, as we shall see it provides a proof of concept that it is possible to achieve the objective of greater parity in predicted outcomes between Indigenous and non-Indigenous populations. A *data producer* pre-processes the data to remove information about the sensitive variable (in this case race). This means that the *data user* making decisions with the data does not need to incorporate fairness into their algorithm design. This approach enables certain fairness criteria to be met even in the case where the data user is not trusted to be fair. It may also be more convenient for the data user since they can continue to use whatever prediction algorithm they choose, in the knowledge that fairness concerns have already been addressed at the data pre-processing stage.

We describe how this approach works in the context of ensuring that recidivism predictions do not discriminate on the basis of race. A data producer learns a cleaned variable (Z) such that an adversary is unable to predict race (S) from it, while also trying to make the cleaned variable similar to the original input (X). In our case we assume that the data producer does not have access to whether the person has reoffended (Y), which means that it is not affected by any bias in the way that reoffence data is collected.

We introduce a parameter λ, a non-negative constant, to control the trade-off between the two objectives involved in the construction of the cleaned variable (Z). When λ is large, the algorithm focuses more on making the adversary unable to predict race (S). When λ approaches zero, the algorithm focuses more on making the inputs and cleaned data similar. The algorithm does not provide any guidance as to how to select λ. Rather, this depends on a decision about the relative importance assigned to inter-group parity and accuracy in the design of the algorithmic risk assessment. Such a decision is a social, political and regulatory one – the algorithm simply provides an implementation for whatever decision is made.

The learning steps of the algorithm are summarized in Figure 6.1.[7] The data producer learns a neural network parameterized by weights θ_1, which produces cleaned records from input records. The adversary learns a neural network parameterized by weights θ_2, which predicts race from the cleaned records. Four steps are repeated for each batch of examples from the training data:

[7] See Chapter 4 and [Edwards and Storkey, 2016] for further details. We also considered a variant of the adversary training objective proposed in [Madras et al., 2018] but found it did not substantively change the results.

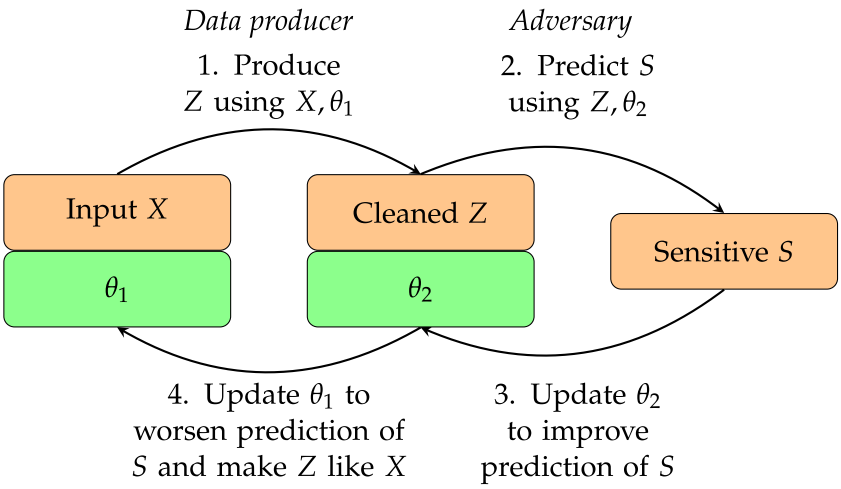

Data producer
1. Produce Z using X, θ_1
Adversary
2. Predict S using Z, θ_2
Input X
Cleaned Z
Sensitive S
θ_1
θ_2
4. Update θ_1 to worsen prediction of S and make Z like X
3. Update θ_2 to improve prediction of S

Figure 6.1: Learning fair representations with an adversary. In the text we use the example of X=criminal record, Z=the cleaned version of the criminal record, S=race, Y=whether the person reoffended. θ_1 and θ_2 are parameters of the learning algorithm.

1. On receiving examples of X, the data producer passes them through a neural network with weights θ_1 to produce examples of Z

2. On receiving examples of Z, the adversary passes them through a neural network with weights θ_2 to predict the values of S

3. By comparing the true values of S to its predictions for these examples, the adversary updates θ_2 to improve its prediction of S in future

4. By comparing the true values of S to the adversary's predictions for these examples, the data producer updates θ_1 to worsen the adversary's prediction of S in future while also trying make Z similar to X. The trade-off between these two objectives is governed by the parameter λ.

Once learning is complete, for each individual the data producer passes their input record through a neural network with weights θ_1. This cleaned record is then provided to the data user, who uses it to make a prediction about whether the individual will reoffend.

6.4 Predicting DV Recidivism with the BOCSAR Dataset

We applied learning fair representations with an adversary to the prediction of DV recidivism in Australia with the BOCSAR ROD used in Fitzgerald and Graham [2016]. As a result, we achieved improved fairness compared to Fitzgerald and Graham's study on several measures. However, this case study also highlights the inevitable trade-offs involved. Our proposed approach allows us to reduce the relative

Table 6.1: Independent features in the BOCSAR dataset.

Feature	**Description**
	Offender demographic characteristics
Gender	Whether the offender was recorded in ROD as male or female.
Age	The age category of the offender at the court appearance, derived from the date of birth of the offender and the date of finalization for the court appearance.
Indigenous status	Recorded in ROD as 'Indigenous' if the offender had ever identified as being of Aboriginal or Torres Strait Islander descent, otherwise 'non-Indigenous'.
Disadvantaged areas index quartile	Measures disadvantage of an offender's residential postcode at the time of the offence. Based on the Socio-Economic Index for Areas (SEIFA) score produced by the Australian Bureau of Statistics.
	Conviction characteristics
Concurrent offences	Number of concurrent proven offences, including the principal offence, at the offender's court appearance.
AVO breaches	Number of proven breaches of Apprehended Violence Order (AVO) at the court appearance.
	Criminal history characteristics
Prior juvenile or adult convictions	Number of Youth Justice Conferences or finalized court appearances with any proven offences as a juvenile or adult prior to the court appearance.
Prior serious violent offence conviction past 5 years	Number of Youth Justice Conferences or finalized court appearances in the 5 years prior with any proven homicide or serious assault.
Prior DV-related property damage offence conviction past 2 years	Number of Youth Justice Conferences or finalized court appearances in the 2 years prior with any proven DV-related property damage offence.
Prior bonds past 5 years	Number of finalized court appearances in the 5 years prior at which given a bond.
Prior prison or custodial order	Number of previous finalized court appearances at which given a full-time prison sentence or custodial order.

disadvantage faced by Indigenous defendants incurred by using the original input data, but at the cost of predictive accuracy.

6.4.1 BOCSAR Dataset Experiments

The BOCSAR ROD contains 14776 examples and 11 categorical and ordinal input features for each example, as shown in Table 6.1. The input features are grouped to represent the offender demographic characteristics, conviction characteristics, and criminal history characteristics for each case. The target variable is whether or not an individual re-committed a DV-related offence within a duration of 24 months since the first court appearance finalization date. DV-related offences include any physical, verbal, emotional, and/or psychological violence or intimidation between domestic partners. We used a random 50% sample for training and the remaining 50% for testing, as in some experiments in [Fitzgerald and Graham, 2016].

Our baseline experiments used the original data, including the Indigenous status variable. We also tested the pre-processing method described in Section 6.3.2, applied to the original data without the Indigenous status variable, for several values of the

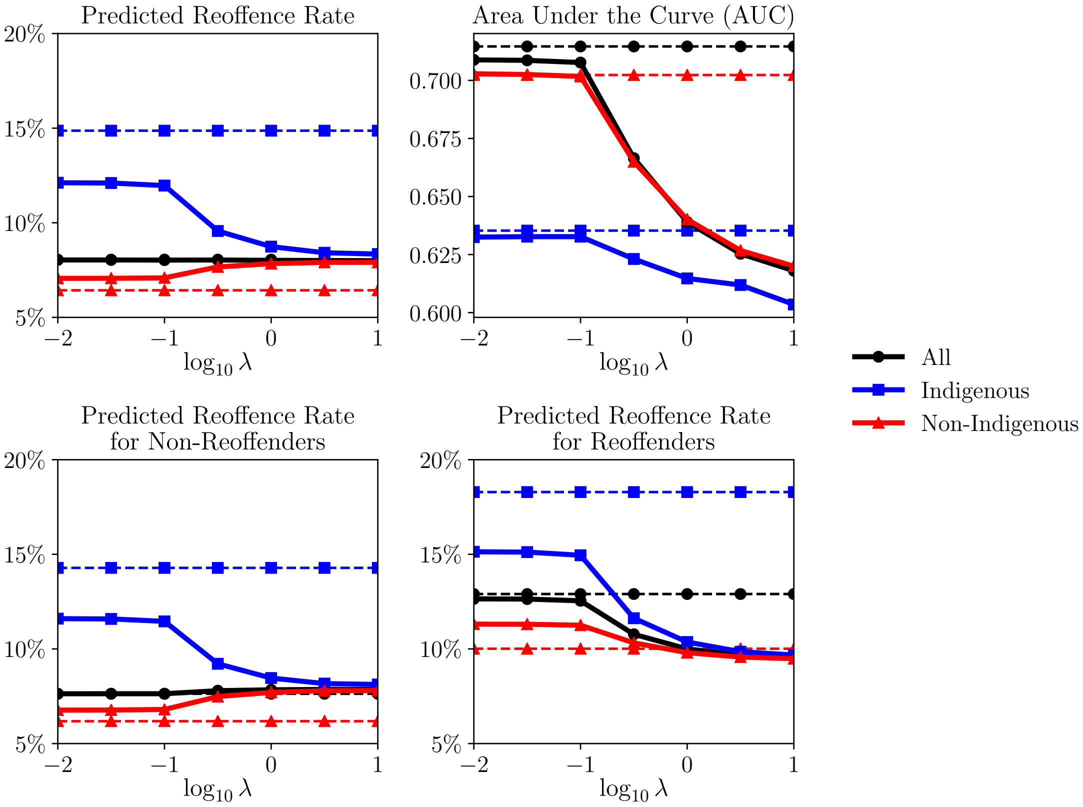


Figure 6.2: Results of applying pre-processing to the BOCSAR dataset, followed by logistic regression, to predict DV reoffences. Baselines using logistic regression without pre-processing are shown as dashed lines, and experiments using logistic regression with pre-processing are shown as solid lines. The vertical axes show several measures of interest on the test data. The horizontal axes show the parameter λ (see text) used in pre-processing on a logarithmic scale.

parameter λ. We predicted recidivism from the data – the original data in the baseline experiments and the pre-processed data in the other experiments – using logistic regression as in Fitzgerald and Graham's study, which predicts the probability of reoffence for each individual. We computed the metrics described in Section 6.3.1, as shown in Figure 6.2. We computed each of these metrics for all individuals, for Indigenous individuals and for non-Indigenous individuals.

6.4.2 Discussion of the BOCSAR Dataset Results

We discuss our results by comparing the performance of the baseline method with our proposed pre-processing method.

6.4.2.1 Baseline Results using Original Data

Using the baseline, there are significant differences in the average predicted reoffence rates for Indigenous and non-Indigenous individuals. These predicted rates are close to the observed rates in the test set: for Indigenous 14.9% predicted vs 14.6% observed, and for non-Indigenous 6.4% predicted vs 6.5% observed. Our baseline does not display the severe overestimation of Indigenous reoffence observed in the Fitzgerald and Graham's model. Furthermore, the baseline test set AUC was 0.71 (slightly superior to the 0.69 previously reported by Fitzgerald and Graham), indicating that the model has some predictive accuracy.

Interestingly, the baseline AUC was higher for the overall population than for either Indigenous or non-Indigenous subpopulations, which is likely because comparing Indigenous and non-Indigenous individuals contributed positively to the overall population AUC whereas these comparisons do not occur within the subpopulation AUC scores. Furthermore, the AUC was lowest for the Indigenous subpopulation, indicating that the model found it harder to make accurate predictions within this group relative to the non-Indigenous subpopulation. However, separately conditioning only on observed reoffenders and on observed non-reoffenders revealed that the *types* of inaccurate predictions made for Indigenous and non-Indigenous subpopulations differed – an issue which is not evident from the AUC figures alone.

There are several other potential issues with the baseline:

- variations in the way that reoffence data is collected among Indigenous and non-Indigenous populations may influence and be reinforced by predictions made by the model
- among observed non-reoffenders the average predicted reoffence rate was 14.3% for Indigenous vs 6.2% for non-Indigenous populations, indicating that Indigenous non-reoffenders were rated more than twice as risky as a non-Indigenous non-reoffenders
- among observed reoffenders, the average predicted reoffence rate was 18.3% for Indigenous vs 10.0% for non-Indigenous populations, indicating that non-Indigenous reoffenders were rated only just over half as risky as Indigenous reoffenders[8]
- from a process perspective, it may be viewed as unfair that a person's Indigenous status is considered by the model.

Removing the Indigenous status column in the data is a possible step towards remediating these issues. It would address the final concern around fair process; but

[8] By the contrapositive of Corollary 5.7, if predicted reoffence rates are equal to observed reoffence rates for both Indigenous and non-Indigenous populations, and the observed Indigenous and non-Indigenous reoffence rates are different from each other, and the model is not perfectly accurate, then the predicted reoffence rate for non-reoffenders is different between Indigenous and non-Indigenous populations and/or the predicted reoffence rate for reoffenders is different between Indigenous and non-Indigenous populations.

as has previously been observed [Dwork et al., 2012] and is reinforced by our results, this "fairness through blindness" approach is insufficient to address the first three concerns, which remain even without the presence of this column.

6.4.2.2 Results using Pre-processed Data

Recall that the results using pre-processed data depend on the parameter λ, which controls the extent to which information about Indigenous status is removed by the pre-processing. The solid lines on the left hand side of the plots, where λ approaches zero and the data is effectively left untouched except for the exclusion of the Indigenous status column, indicate that while the discrepancies between the Indigenous and non-Indigenous populations are not as acute as in the baseline case, they are still very much present. Information contained in the other columns still results in different outcomes for Indigenous and non-Indigenous populations, a phenomenon known as *redundant encoding* [Dwork et al., 2012].

Applying pre-processing with increasing values of λ, the issues associated with the baseline described in Section 6.4.2.1 are addressed:

- the predicted reoffence rate for non-reoffenders is more similar for Indigenous and non-Indigenous populations (for $\lambda = 10$, 8.1% for Indigenous vs 7.8% for non-Indigenous)

- the predicted reoffence rate for reoffenders is more similar for Indigenous and non-Indigenous populations (for $\lambda = 10$, 9.7% for Indigenous vs 9.5% for non-Indigenous)

- the predicted reoffence rate overall is more similar for Indigenous and non-Indigenous populations (for $\lambda = 10$, 8.3% for Indigenous vs 7.9% for non-Indigenous).

There was a cost to pre-processing in terms of accurately predicting reoffence. The AUC dropped to 0.62, so that the predictions were less accurate than the baseline (AUC 0.71), while still significantly more accurate than a random prediction (AUC 0.5).[9] Overall predicted reoffence rates for non-reoffenders were higher compared to the baseline: 7.9% for $\lambda = 10$ vs 7.6% for the baseline, a 10.2% increase. Overall predicted reoffence rates for reoffenders were lower compared to the baseline: 9.6% for $\lambda = 10$ vs 12.9% for the baseline, a 26.0% decrease. This reduced accuracy is not surprising as the pre-processing removed information from the dataset. The decrease in predicted reoffence rates for reoffenders caused by the pre-processing is undesirable from the perspective of potential victims of domestic violence. Furthermore, this decrease was greater for Indigenous individuals, whose potential victims are more likely to also be Indigenous.

[9] By Theorem 5.3, given equal Indigenous and non-Indigenous predicted reoffence rates among reoffenders, among non-reoffenders and overall, the predicted reoffence rates for reoffenders and non-reoffenders must be equal (assuming that the observed Indigenous and non-Indigenous reoffence rates are unequal).

6.4.2.3 Summary of the BOCSAR Dataset Results

In summary, our approach improved on several measures of fairness compared to Fitzgerald and Graham's study. The baseline approach of learning from the original input data resulted in a prediction indicating that the average risk associated with Indigenous individuals was more than twice that of their non-Indigenous counterparts, even among non-reoffenders, while using pre-processing with a value of $\lambda = 10$ these average risks were comparable. As discussed previously, this could not have been achieved simply by removing the Indigenous status column from the data. However, achieving comparable risks came at the cost of overall predictive accuracy (AUC 0.71 to AUC 0.62). It is worth repeating that our approach does not prescribe a particular value of the trade-off parameter λ, but rather provides a quantitative tool to estimate the effect of this trade-off. We discuss further implications of fairness trade-offs in our conclusion.

6.5 Conclusion

The Australian DV case study shows that without incorporating an explicit fairness criterion into algorithm design, individuals from one racial group may be marked higher risk than another, even when separately considering only observed reoffenders or only observed non-reoffenders. This is still true when race is simply dropped from the input data: blindness is not enough. An alternative is the use of the fair representation learning, which we analyzed in Chapter 4, as a data pre-processing technique. This approach yielded more equal predicted reoffence rates for different racial groups: among reoffenders, among non-reoffenders and overall.

The case study also reveals an important trade-off involved in the design of algorithmic risk assessments. From the perspective of Indigenous defendants who in the baseline scenario were considered higher risk than non-Indigenous defendants, both among reoffenders and among non-reoffenders, this pre-processing makes the system fairer. The flipside is that non-Indigenous non-reoffenders are judged to be more risky. And all reoffenders – particularly Indigenous reoffenders – are judged to be less risky, which is not in the interests of potential victims.

The trade-off between the interests of different stakeholders is equally a part of human decision-making in the criminal justice system. The advantage of our approach is making this trade-off explicit and precisely controllable through a model parameter, which may be set according to whatever weighting is deemed appropriate by society. The approach we propose – involving an explicit trade-off between certain quantitative definitions of accuracy and fairness – also applies to other contexts where prediction algorithms are used to support decisions about individuals such as the provision of credit or insurance, and to other demographic groups besides racial groups.

There is a second trade-off involved here: between explicit and implicit explanations for decisions. Transparency allows individuals to better understand the social systems – including the criminal justice system – that make decisions about their lives. However, when the rationale for these decisions is laid bare, they may be less palatable than when they are opaque. Algorithms – with their stark rules implemented in code – have the effect of illuminating the myriad forms of inclusion and exclusion that invisibly form our social fabric. Perhaps the more profound trade-off is determining to what extent we are willing to shine that light.

Chapter 7

Using Cohort Analysis and Predictive Modeling to Inform Targeted Student Support

7.1 Introduction

Universities, students and society at large have a shared interest in students achieving outcomes at university. These outcomes include program completion, passing individual courses, academic grades, student satisfaction and relevant graduate employment. In the Australian context, outcomes are reported by universities through the Higher Education Information Management System (HEIMS) and monitored by the federal government [Australian Government Department of Education and Training, 2015; Australian Government Tertiary Education Quality and Standards Agency, 2017].

To help students achieve outcomes, universities frequently offer support services such as accommodation, financial aid, academic skills development programs, peer learning communities, counselling and mentoring. Data-driven approaches can help universities to target their student support to maximize its effectiveness. Descriptive statistics allow universities to understand the performance of various student cohorts across several outcomes, and hence plan future interventions to support particular student cohorts [Norton and Cherastidtham, 2018; Edwards and McMillan, 2015]. Individualized predictions of student outcomes offer universities the chance to be even more targeted, by directing their assistance to students who will most benefit [Jia and Maloney, 2015; Korhonen and Rautopuro, 2018].

Targeting student support becomes even more critical when a university's admissions process changes. This is the situation currently faced by The Australian National University (ANU), which recently announced plans to move to a new national undergraduate admissions model starting in 2020 [Australian National University, 2018; Hughes-Warrington et al., 2019]. The new model is aligned to the mandate of ANU as Australia's national university and is intended to democratize access to undergraduate places at the university. It is expected that the new model will significantly increase the diversity – including in terms of geographic spread and socio-economic status – of the student population. To make the new model a success, ANU must adapt its existing support services to enable previously underrepresented cohorts to achieve across a range of student outcomes. This process starts with understanding student outcomes for particular cohorts and identifying individuals who are likely to require support, the focus of this work.

A risk in predicting student outcomes is that the resulting interventions may lead to unequal outcomes for particular demographic groups, particularly in the realistic case where the predictions are not always correct. It is possible to quantify these effects – for example, by analyzing the extent to which different demographic groups are subject to incorrect predictions [Hardt et al., 2016; Zafar et al., 2017a]. Recent scholarship on fairness in machine learning allows such risks to be managed by incorporating equity considerations into predictive model design, including via *fair representation learning* [Edwards and Storkey, 2016].[1] We show how such an approach can be applied to the context of predicting ANU student outcomes. This provides another case study of the practical use of fair representation learning, which addresses a different set of issues compared to the recidivism prediction case study presented in Chapter 6.

This chapter focuses on topics of interest for the Australian National University administration[2] and the higher education sector, and also includes topics that are more closely connected to the earlier chapters of this thesis. In Section 7.2 we introduce the data inputs and student outcomes from ANU on which our analysis is based. In Section 7.3 we analyze the performance of several cohorts across multiple student outcomes. In Section 7.4 we make and evaluate individualized predictions of student outcomes. In Section 7.5 we examine the equity considerations of our predictions and present approaches to incorporating equity into predictive model design via fair representation learning. We conclude in Section 7.6.

7.2 Data Inputs and Student Outcomes

We present the ANU data inputs and student outcomes used in the analysis. Given the focus of ANU on reforming its undergraduate admissions model for domestic school-leavers, we limit our analysis to this student group. While some findings are specific to this group, our methodology as well as certain findings may apply to other student groups and higher education institutions.

7.2.1 Dataset and Feature Extraction

This work uses a dataset of 8498 domestic undergraduate current school-leavers admitted between 2011-17 at ANU. The dataset incorporates data from a range of ANU databases as well as data licensed from the Universities Admissions Centre (UAC).

[1] See Chapter 4 for an introduction to fair representation learning.

[2] This project was sponsored by the Australian National University Deputy Vice-Chancellor (Academic). The Planning and Performance Measurement Division of the Australian National University collaborated on this project, including providing source data.

Table 7.1: Features and outcomes in the analysis.

Feature	**Options**	**Source**
	Prior studies	
Australian Tertiary Admissions Rank (ATAR)	Continuous	UAC
ATAR top 3 in school	Yes/No	UAC
ATAR top 2% in school	Yes/No	UAC
High school English attempted	Yes/No	UAC
High school Maths attempted	Yes/No	UAC
High school subject attempts	268 Yes/No variables	UAC
High school subject grades	264 continuous variables	UAC
	Demographics	
Gender	Male/Female	UAC
Age at application date	Continuous	UAC
Home state	ACT/NSW/NT/QLD/SA/TAS/VIC/WA/Other	UAC
Home address socio-economic status (SES)	High/Medium/Low	UAC
Home address remoteness	Major cities / Inner regional / Outer regional	UAC
Financial difficulties	Yes/No	UAC
Medical disadvantage	Yes/No	UAC
Geographic disadvantage	Yes/No	UAC
Youth Allowance	Yes/No	UAC
Family Tax Benefit Part A	Yes/No	UAC
Non-English Speaking Background	Yes/No	UAC
School Index of Community Socio-Educational Advantage (ICSEA) band	<900 / 900-1100 / >1100	ANU
Indigenous	Yes/No	ANU
Disabled	Yes/No	ANU
	University enrolment	
Combined Program	Yes/No	ANU
Bachelor of Philosophy – Honours (PhB)	Yes/No	ANU
Attendance type	Full-time/Part-time	ANU
Basis of Admission	Higher Education/Secondary Education/Other	ANU
Program Primary Broad Field of Education	7 fields	ANU
Program Primary Narrow Field of Education	15 fields	ANU
	University outcomes	
Attrition year 1	Yes/No	ANU
Attrition year 2 or later	Yes/No	ANU
Failed at least one course year 1	Yes/No	ANU
Failed at least one course year 2 or later	Yes/No	ANU
Grade point average year 1	Continuous (fail=0, passing grades=4-7 scale)	ANU
Grade point average year 2 or later	Continuous (fail=0, passing grades=4-7 scale)	ANU

The features cover the student's prior studies, demographics and university enrolment – all of which are known before the student commences.[3] There are also several university outcomes of interest. The features and outcomes are shown in Table 7.1. For each feature or outcome we describe its options, if it is a categorical variable, or else note that it is a continuous variable. We also show whether the source of the feature is ANU or UAC.

ATAR is a national percentile rank based on the student's academic performance at high school.[4] For the categorical variables, we only included options that applied to at least 20 students. The Program Primary Broad and Narrow Fields of Education

[3] The home state and address refer to the student's permanent home location when applying to ANU.

[4] The ATAR top 2% in school feature is crudely computed by dividing the within-school ATAR rank of the student by the total number of graduating students from the school, and checking if this quantity is no greater than 0.02. This method has shortcomings, however, notably that in schools of less than 50 graduating students, even the top-ranked student is determined not to be in the top 2%!

(FoE) conform to the Australian Standard Classification of Education [Australian Bureau of Statistics, 2001]. In the case of high school subjects, we focused on year 12 subjects across all Australian states and territories, including subjects offered through the International Baccalaureate (IB). We included records of subject attempts for those subjects taken by at least 20 students in the dataset. We also included the individual grades for those subjects for which a grade was awarded. For a small minority of high school subjects for which ordinal grades were given (Queensland and some IB subjects), we applied a simple heuristic method to convert these to numerical grades.

7.2.2 Student Outcomes

When describing student outcomes we use the term *course* to refer to a single unit of study, and *program* to refer to a qualification such as a bachelor degree comprised of several courses. We focused on describing and predicting the following student outcomes:

1. Attrition, defined as leaving ANU without completing a program[5]
2. Failure of at least one attempted course
3. Grade Point Average (GPA), where the average is weighted by the credits allocated to a course.

Each of these outcomes (or in the case of attrition and course failure, their converse) is an important indicator of a successful student experience at ANU. For each outcome we separately considered:

- The outcome in the student's first year, using features known before they commence their first year
- The outcome in the student's second and later years, using features known before they commence their second year.

Separately analyzing first year outcomes was motivated by a recognition that this can be a particularly challenging time for students, given the change in circumstances they have recently experienced. It is also a period where the university has fewer data points about a student on which to make informed interventions, making targeting support particularly challenging.

7.3 Analysis of Student Cohorts

For each outcome of interest, we investigated how different cohorts of students performed. We focused our attention on first year outcomes, given the importance of this period in a student's time at university. We describe the methodology for the analysis of student cohorts, followed by our results for each outcome.

[5] Transferring to another institution and ceasing studies altogether are both counted as attrition. It was not possible to distinguish between these two cases based on the dataset used in this study.

7.3.1 Methodology

We analyzed the relationship of each outcome to a set of *standard features*, comprised of the prior studies, demographic and university enrolment features from Table 7.1, except for the high school subject attempts and grades. For each outcome and categorical feature, we computed the difference between the average outcome when the feature was present and the average outcome when the feature was absent. For continuous features we computed the difference between the average outcome when the feature was above average and the average outcome when the feature was below average. For example, we computed the difference between first year attrition rates among students with ATARs above the dataset average and among students with ATARs below the dataset average.

When analyzing each outcome, we:

- identified a subset of the standard features which had a statistically significant relationship with the outcome

- investigated the relationships between the outcome and membership of several *equity cohorts*, comprised of students facing various forms of potential educational disadvantage.

Given the large number of the high school subject attempt and grade features, we excluded these from the set of standard features to allow us to focus our attention on the remainder of the features. For GPA, we separately analyzed the relationship of this outcome to high school subject attempts and grades.

For each feature and outcome, we conducted a two sample test for the difference in outcome when the feature is present vs absent (above vs below average for continuous features). For attrition and failing at least one course, which are binary, the statistical test was a two-tailed Pearson's chi-squared test implemented via the prop.test function in R. For GPA, which is continuous, the statistical test was a two-tailed Student's t-test implemented via the t.test function in R.

Each test produced an estimate of the difference in outcome, a 95% confidence interval for the difference, and a p-value. The p-value indicates the probability of a difference at least as large as that observed, given a null hypothesis that there is no difference in average outcomes between the two samples. Because we simultaneously tested multiple features for each outcome, we applied the Bonferroni correction, which required a smaller p-value for statistical significance compared to testing a single feature. This ensured that the family-wise error rate – the probability that at least one of the statistical significance tests passed given null hypotheses for all features – was at most 0.05. The result was a more conservative subset of features marked as statistically significant.

7.3.2 Attrition

The results of the cohort analysis for first year attrition are shown in Figure 7.1. Data was available for all 8498 students. The results showed that the intensity and duration

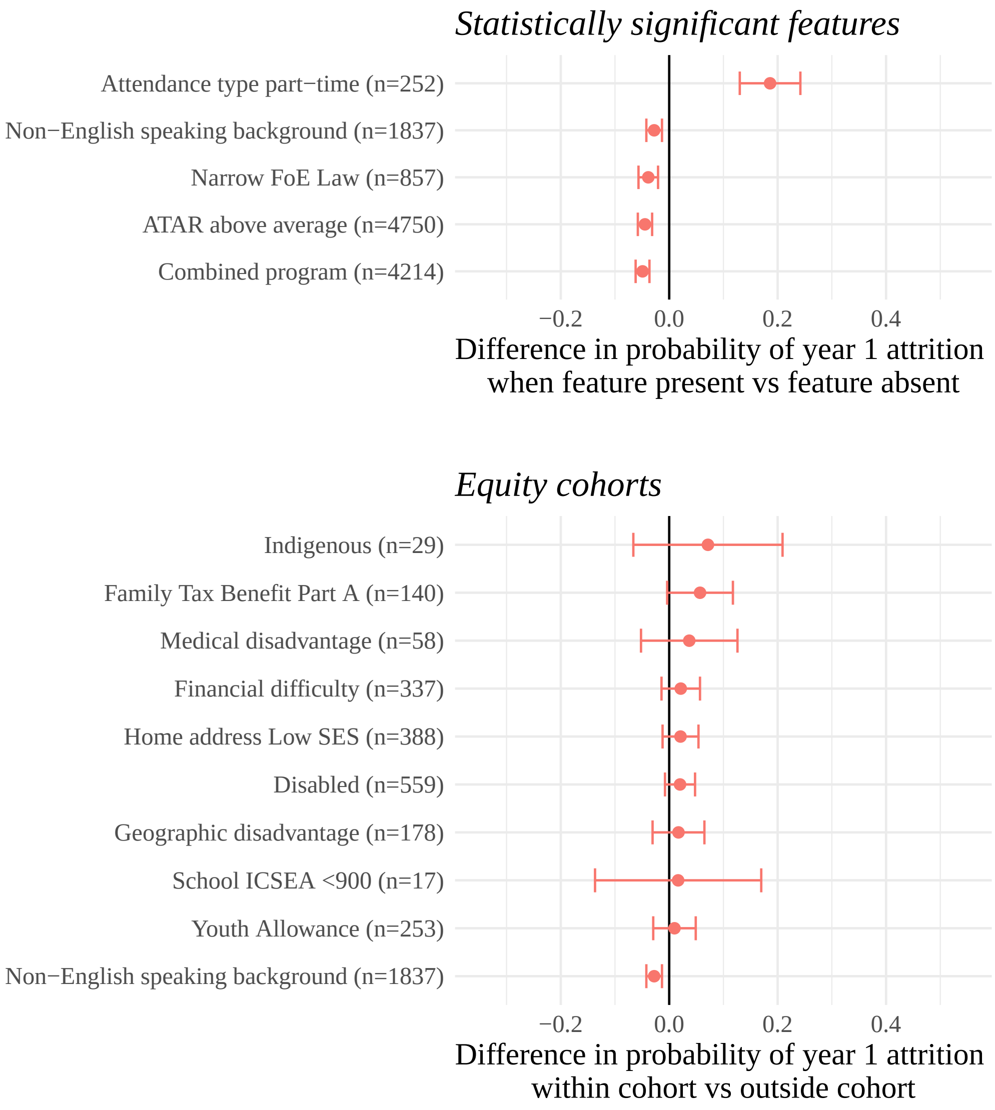


Figure 7.1: Cohort analysis of first year attrition, showing statistically significant features (top) and equity cohorts (bottom).

of a student's program were predictors of attrition. Part-time students showed higher rates of attrition, while combined program students showed lower rates of attrition. Among the equity cohorts considered, the only statistically significant difference was among non-English speaking background students who tended to attrit less in first year. While students with an above average ATAR showed lower rates of attrition, attrition is often a personal decision of which academic performance is only one part.

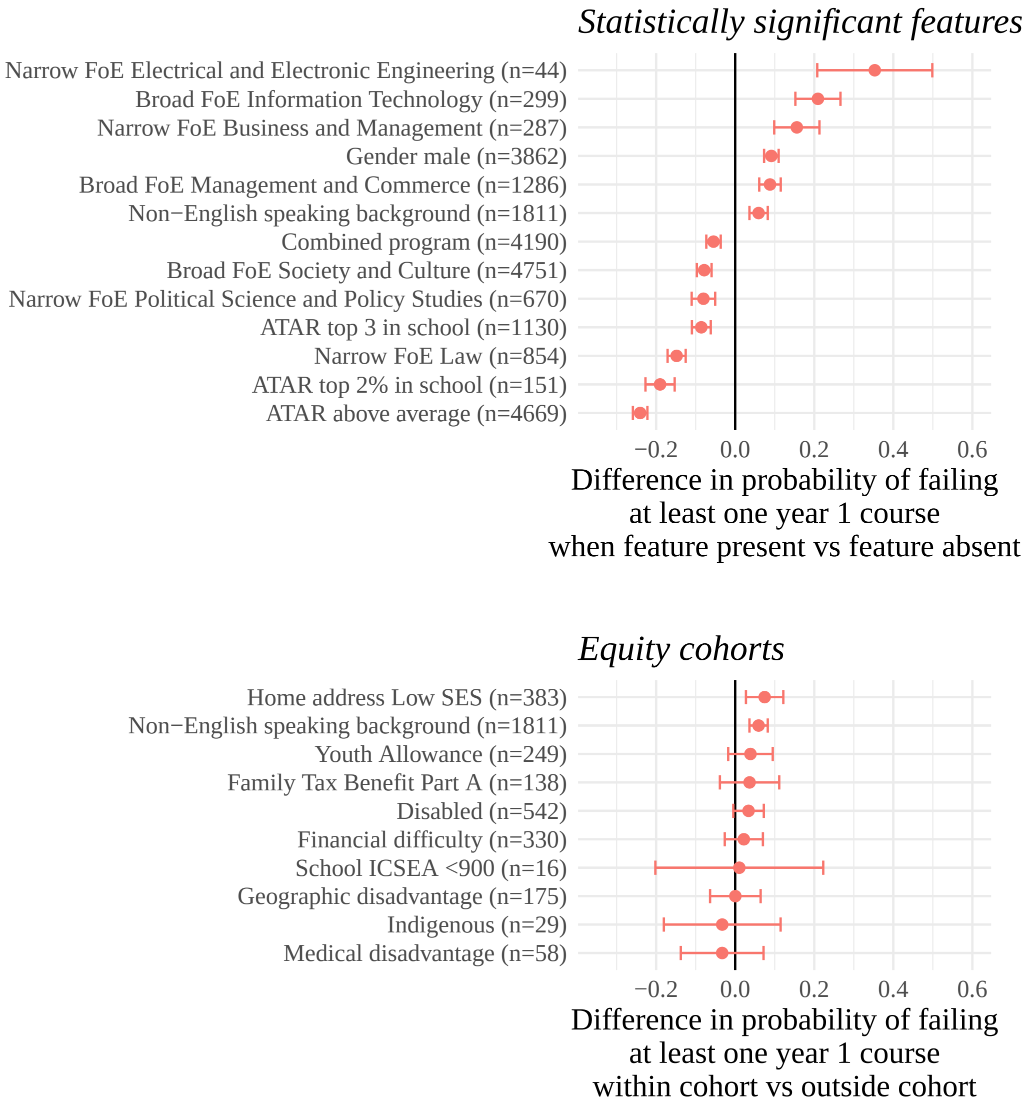


Figure 7.2: Cohort analysis of failing at least one first year course, showing statistically significant features (top) and equity cohorts (bottom).

7.3.3 Failing At Least One Course

The results for the cohort analysis of failing at least one first year course are shown in Figure 7.2. Data was available for 8384 students (98.6%). The results show that a student's high school academic performance, as measured by their ATAR, is closely related to their probability of failing a course – students with higher ATARs were less likely to fail. Students enrolled in programs within certain primary fields of education – like information technology and commerce – had higher rates of failure, while those in law and social sciences had lower rates. We observe that in some cases the courses failed may have been outside of their primary field of education.

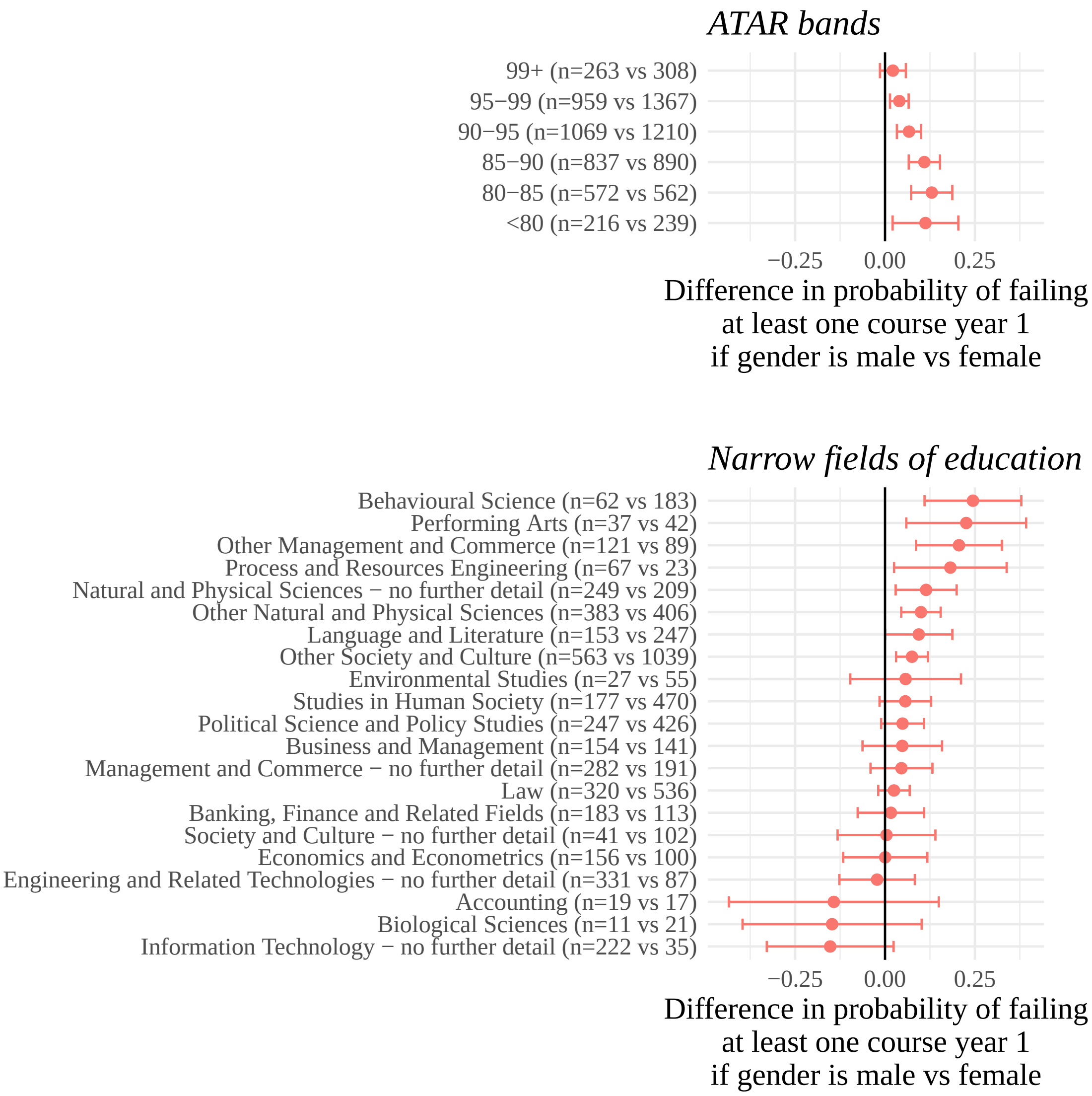


Figure 7.3: Relationship of gender to probability of failing at least one first year course, broken down by ATAR band (top) and Program Primary Narrow Field of Education (bottom). Sample sizes are shown for males vs females.

Among the equity cohorts, students with low SES home addresses and from non-English speaking backgrounds had the highest rates of failure. The non-English speaking background group is particularly interesting, given that they are less likely to attrit but more likely to fail courses than other students. Failure and attrition are more tightly coupled with respect to other features such as ATAR. Further analysis is required to explain this trend.

Another cohort of students with higher rates of failure is males. To better understand the relationship between failing courses and gender, we disaggregated our analysis by ATAR bands and Program Primary Narrow FoE as shown in Figure 7.3.

For each ATAR band and Narrow FoE we report the number of male vs female students. We omitted cases where either of these numbers was below 10, motivated both by privacy and the fact that with very small sample sizes it is difficult to draw any meaningful conclusions.

The trend for males to have a higher probability of course failure holds up across ATAR bands. While it is less pronounced in the 99+ cohort compared to the cohorts below 90, this is not surprising since the low rate of failure across the board in the 99+ cohort leaves less room for variability between genders. The results provide evidence that academic support services need to be appropriately designed to support male students, in the same way that boys' education in the school system is a specific topic in educational research [Epstein, 1998].

The picture is more mixed once we disaggregate by gender and narrow field of education. Females outperform males in passing courses in the majority of fields, and markedly so in certain fields across both the sciences (e.g. behavioural science, process and resources engineering, other natural and physical sciences) and the humanities (e.g. performing arts, other society and culture). However, in several other fields (e.g. law, banking and finance, economics) there is no marked difference between genders. Information technology is something of an outlier in the opposite direction, with males tending to outperform females. These results indicate that efforts to deliver academic student support services may need to be tailored to students of different genders across different fields of education.

7.3.4 Grade Point Average

The results for the cohort analysis of first year GPA are shown in Figure 7.4. Data was available for 8384 students (98.6%). The trends are broadly similar to those for failing at least one first year course. This is expected since a failed course incurs a zero mark (pass marks are in the range 4-7), which in turn affects GPA. Students who had lower ATARs, were male, came from a non-English speaking background, came from a low SES home address, and were in disciplines such as information technology and commerce had higher rates of failure and lower GPAs. The analysis shows that part-time students and those from outside the ACT also lagged their peers on GPA to a statistically significant degree. Indigenous and disabled students both appeared to have lower GPA, although not to statistical significance. Further analysis of lower GPA cohorts could help to design and target academic support services.

Separately, we looked at the relationship between high school subject attempts and grades and first year GPA, as shown in Figure 7.5. This illustrates that looking beyond ATAR gives a more detailed picture of student preparedness for the university academic environment. Above average grades in physics and advanced mathematics subjects were prominent at the top of the list of high school subjects associated with high first year GPA. However, above average subjects across a diverse range of disciplines including accounting, biology, politics, economics and chemistry were also associated with high first year GPA. Above average grades from jurisdictions far from ANU were particularly strongly associated with high GPA. A possible expla-

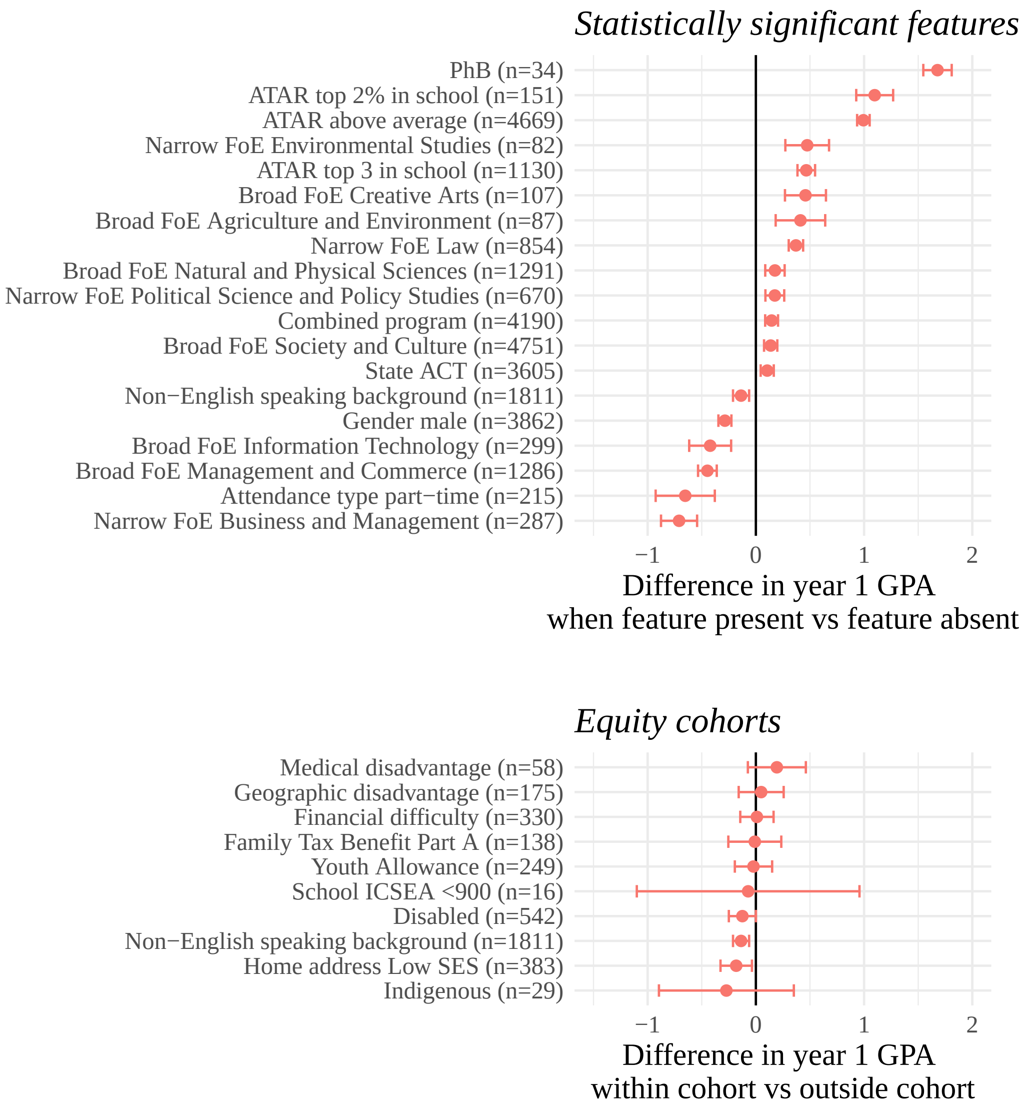


Figure 7.4: Cohort analysis of first year GPA, showing statistically significant features (top) and equity cohorts (bottom).

nation is that these students had previously demonstrated both academic aptitude based on their grades and a commitment to their studies based on their willingness to relocate a significant distance to Canberra. Attempts at some subjects were associated with lower first year GPA, including general mathematics, business and information technology.

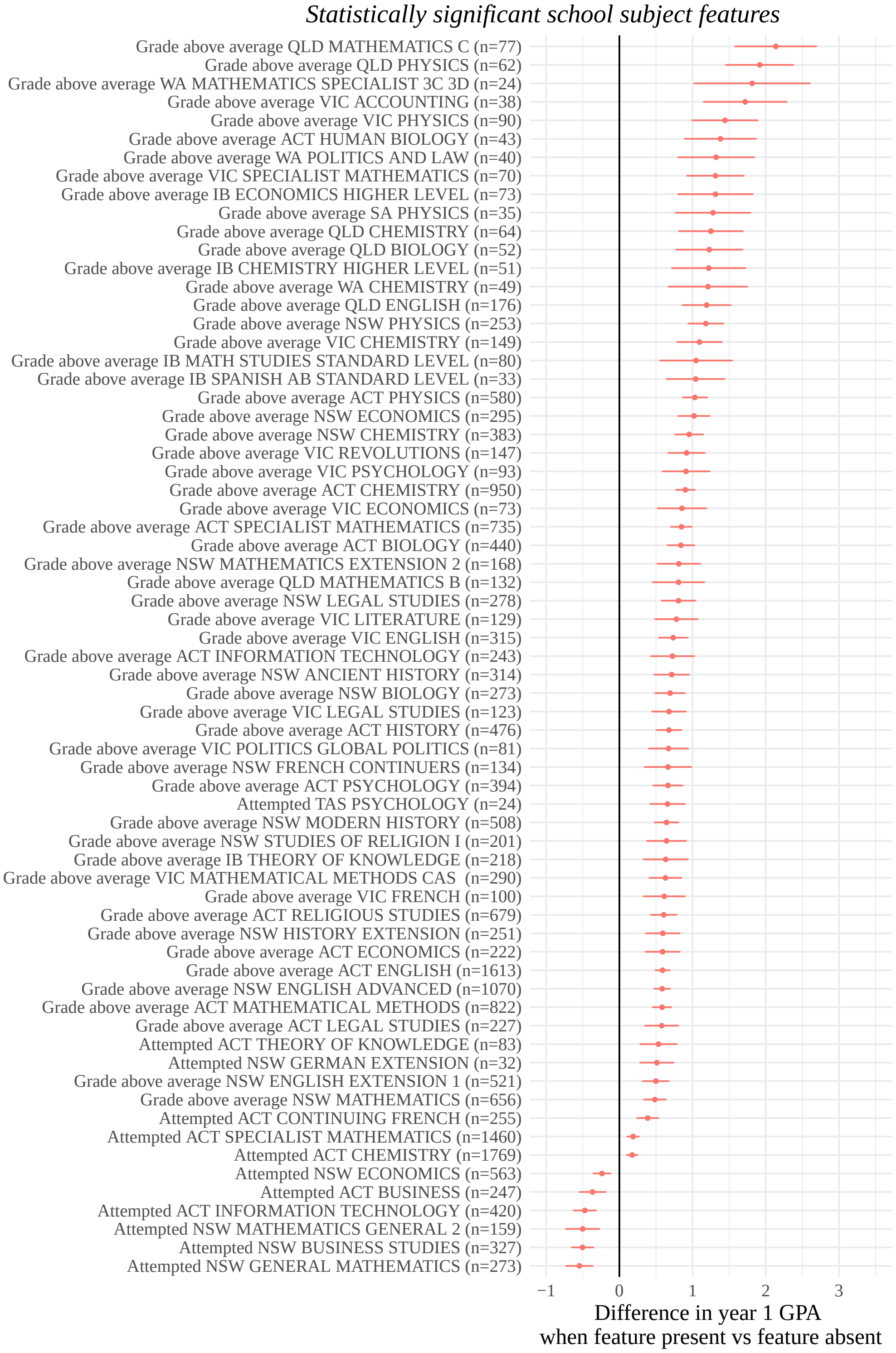


Figure 7.5: Descriptive results for first year GPA, showing statistically significant high school subject features. For the 'grade above average' features, 'feature absent' applies to students who attempted a subject and achieved a grade below the dataset average.

Table 7.2: Feature sets used for individualized predictions.

Feature Set	**Features used in predicting year 1 outcomes**	**Features used in predicting year 2+ outcomes**
One feature	ATAR	Year 1 GPA
Standard features	Features listed in Table 7.1 except high school subject grades and attempts	Features listed in Table 7.1 except high school subject grades and attempts *plus* year 1 information including GPA, whether any courses were failed, and satisfaction with student support services
Standard features + high school subjects	Features listed in Table 7.1 including high school subject grades and attempts	Features listed in Table 7.1 including high school subject grades and attempts *plus* year 1 information including GPA, whether any courses were failed, and satisfaction with student support services

7.4 Individualized Predictions of Student Outcomes

We developed and applied a methodology for making individualized predictions of student outcomes. This involved extracting features from the dataset and testing several predictive models that took these features as inputs.

7.4.1 Feature Extraction

For our predictive modeling, we used features about an individual's prior studies, demographics and university enrolment to predict several outcomes of interest, as described in Table 7.1. We conducted the following additional processing of the dataset using standard techniques, to enable us to make predictions using all rows and columns of the dataset:

- replaced missing values with the feature mean for continuous variables and with an 'unknown' category for categorical variables
- applied min-max normalization to the continuous variables, so that all values fell into the range $[0, 1]$
- converted categorical variables to dummy variables, using one-hot encoding for all categories but one to avoid redundancy.

We then considered three types of feature sets, as shown in Table 7.2. We also considered using only features found to be statistically significant in Section 7.3, but preliminary investigation indicated that this did not substantively change the results.

7.4.2 Predictive Models

We considered the tasks of predicting each student outcome – both in first year and in second and later years – and investigated the difficulty of each of these tasks separately. Attrition and failing at least one course are binary, yielding two *classification* tasks. GPA is continuous, yielding a *regression* task.

For each outcome, we randomly selected 70% of the students to form the training set, trained several predictive models on the training set, and then tested the models on the remaining 30% of the of the data. We tested several linear and non-linear models. In each case we attempted to select model parameters that produced reasonable performance, but did not exhaustively investigate parameter selection. The models were:

- *linear/logistic regression* – a simple linear model whose outputs were a weighted sum of the input features. In regression tasks the outputs were taken as is (*linear regression*), while in classification tasks the outputs were converted to probabilities via the logistic function (*logistic regression*). The weights were learned from the training set, using lasso regularization to make the model more robust to noise from features with limited predictive value [Tibshirani, 1996]. We implemented the model using the glmnet R package.
- *decision tree* – a simple non-linear model which assigned each data point to a leaf node in a tree based on its input features, and associated predictions with each leaf node in a tree [Breiman et al., 1984]. The tree structure and leaf node predictions were learned from the training set, with the procedures varying between classification and regression tasks. We implemented the model using the rpart R package.
- *random forest* – a more complex and resource-intensive non-linear model which learned a set of decision trees, each based on a random subset of the rows and columns of the training set [Breiman, 2001]. Predictions were made for new data points by aggregating the predictions of each decision tree, with the aggregation varying between regression and classification. We implemented the model using the randomForest R package.

We used standard evaluation measures of model performance over the test set for both regression and classification tasks. For regression, we evaluated the models' predictions using root mean squared error (RMSE), including RMSE-specific 95% confidence intervals implemented using R. Lower RMSE indicates better performance. For the classification tasks, we evaluated the models' predictions using area under the curve (AUC), including AUC-specific 95% confidence intervals implemented using the pROC R package. AUC can be interpreted as the probability that a randomly selected positive example will receive a higher predicted probability than a randomly selected negative example [Fawcett, 2004]. Higher AUC indicates better performance. A random guess is expected to achieve AUC of 0.5, while a perfect rank ordering achieves AUC of 1.

7.4.3 Results

The results of the predictive models on the classification tasks are shown in Figure 7.6. There are several interesting aspects of the results:

- In the case of all outcomes, there was evidence that they were somewhat predictable (some models achieved AUC above 0.5), but not entirely predictable (no models achieved AUC of 1).

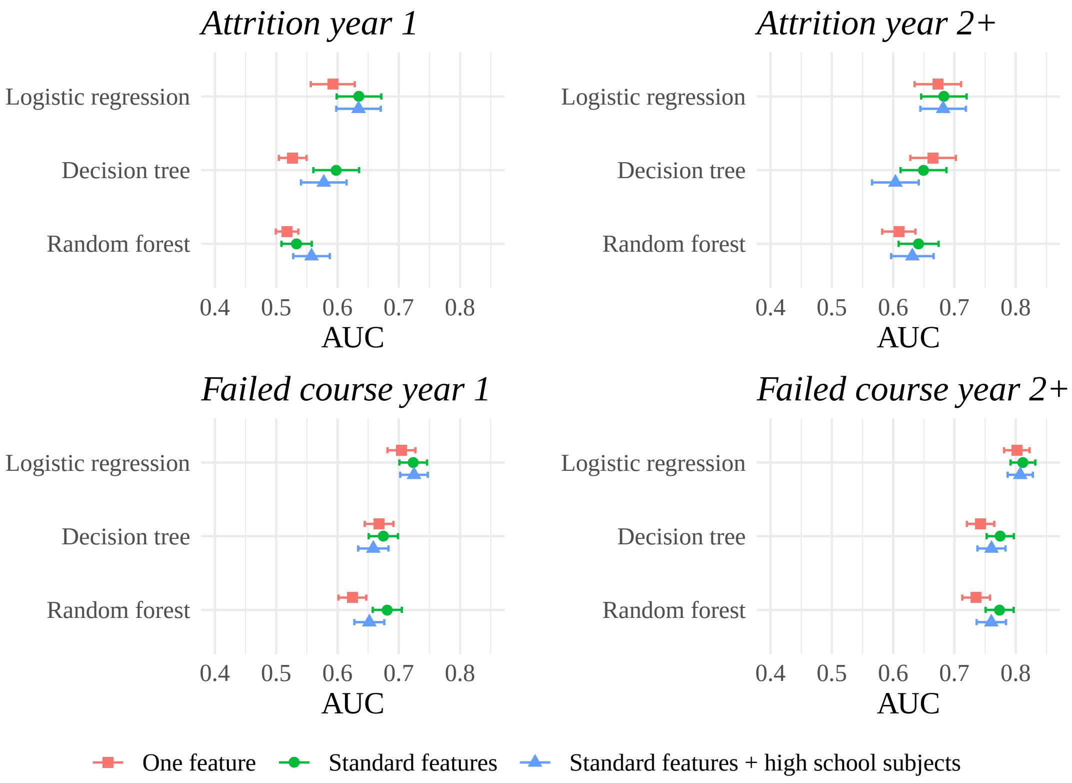


Figure 7.6: Performance of several models and feature sets across predicting attrition and failing at least one course. Higher AUC is better.

- The AUC scores were higher for course failure than for attrition. It is perhaps not surprising that academic performance was easier to predict than attrition, which may be a personal decision.

- It was easier making predictions in years 2 and later – using additional information about student outcomes in first year – compared to making predictions of first year outcomes using only pre-university predictors. It is not surprising that a student's performance and experience at university becomes more predictable once data is available about their time at university to date. There also appeared to be a survivorship bias, where there was less variability in outcomes among students who made it through first year.

- Logistic regression mostly outperformed both of the non-linear models. This suggests that for prediction purposes it does not hurt too much to treat the interactions between different features as primarily linear (despite the existence of some non-linear effects, as we saw with the relationship of gender, ATAR bands and fields of education). The lasso regularization used in logistic regression was an effective way to focus the model's efforts on the most robust predictors. The non-linear models may have been more 'distracted' by the other

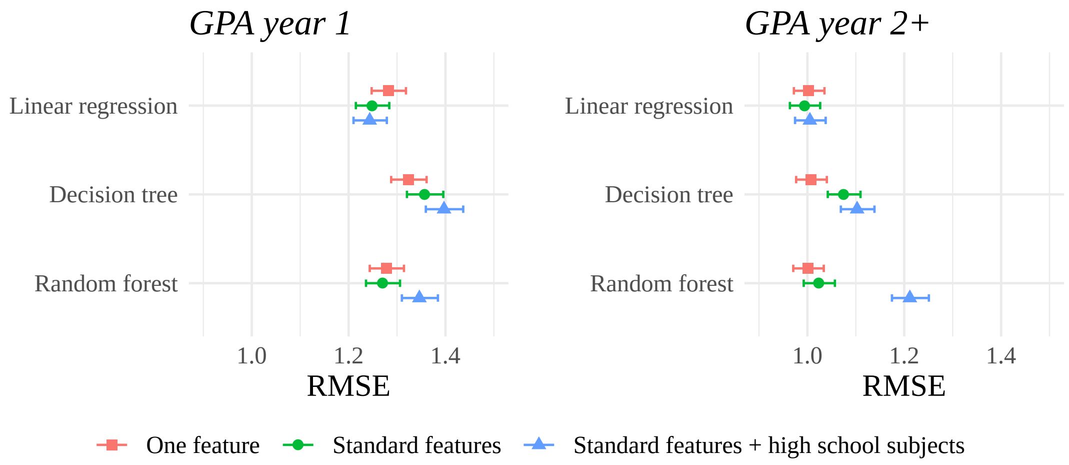


Figure 7.7: Performance of several models and feature sets in predicting GPA. Lower RMSE is better.

less predictive features. It is possible that parameter tuning in the non-linear models could improve performance, but there is no evidence to suggest they would significantly outperform linear models. This is not too surprising given that these tasks do not fall into the categories where non-linear models tend to be most dominant: very high-dimensional tasks with very large training sets.

- The standard features combined with logistic regression performed strongly on most tasks. The addition of the high school subject features made little difference (and in the case of the non-linear models, sometimes worsened performance). Using just ATAR to predict first year outcomes was not far behind the standard set of features. Using just year 1 GPA to predict later year outcomes yielded comparable performance to using more features. This indicates that a single number measuring past academic performance can be an extremely powerful predictor of student outcomes, and even more so when it measures past academic performance at university. The other features appear to offer some predictive benefit – but this benefit is surprisingly small.

The results of the models on the regression task of predicting GPA are shown in Figure 7.7. The trends in the results are broadly comparable with those of the classification tasks, particularly for students failing at least one course. We observe that even the most successful models had considerable variability – an RMSE of around 1 – which is relatively large considering that of the 7 point GPA scale, passing grades lie within a 3 point range. We should remember that we are predicting the future behavior of human beings, which to a certain extent will always be unpredictable.

7.5 Incorporating Fairness into Predictive Model Design

We consider the implications for fairness of using our predictive models, and explore changes to their design to incorporate fairness. We examine a case study of predicting students who will fail at least one course, looking at the fairness implications with respect to gender. A potential concern is that certain types of errors made by a predictive model disproportionately affect students of one gender, which may be viewed as discriminatory. We measure model error rates for both male and female students, and show a proof of concept approach to reducing the difference in these error rates between genders. The definitions and techniques we present have wider applications beyond this particular problem setting.

7.5.1 Defining Quantitative Fairness

A concern arising from the use of predictive models is the risk that decisions arising from these models may discriminate on the basis of group membership, such as gender, language background or socio-economic status [O'Neil, 2017]. Researchers in the field of machine learning have in recent years developed a quantitative perspective on the question of whether a predictive model is unfair. This has included the development of several quantitative definitions of fairness [Mitchell and Shadlen, 2018; Nayaranan, 2018] and approaches to incorporating fairness into the design of predictive models [Dwork et al., 2012; Menon and Williamson, 2018] (see Section 4.2).

We investigate whether the potential for unfairness exists in the case of predicting ANU student outcomes, and explore possible solutions to manage this risk. As we observed earlier, in our dataset the male students tended to be more likely to fail at least one course compared to the female students. As a case study, we consider the effects for both genders of the use of our predictive models of subject failure.

We focus on one possible definition of quantitative fairness, which has variously been referred to in the literature as *equalized odds* [Hardt et al., 2016] or the avoidance of *disparate mistreatment* [Zafar et al., 2017a] (see Definition 5.1). Applying this definition to our context, first we consider all students who did in fact fail at least one first year course, and measure the difference in average predicted probabilities of failure among males and among females. This difference is small (ideally zero) for a model satisfying equalized odds. Second, we consider all students who did *not* in fact fail at least one first year course, and again measure the difference in average predicted probabilities of failure among males and among females. Once more, this difference is small (ideally zero) for a model satisfying equalized odds.

The motivation for this definition of fairness is that conditioned on the individual's actual behavior, the model's predictions are the same on average for both groups, i.e. they are not determined by group membership. A related motivation is that the mistakes made by the model are evenly distributed across both groups. False positives – where the model predicts that someone will fail when in fact they do not – occur at the same rate for both genders. Similarly, false negatives – where the model predicts that someone will not fail when in fact they do – also occur at the same rate for both genders.

7.5.2 Achieving Quantitative Fairness

We might hope that simply excluding group membership – for example, the column encoding a person's gender – from the data on which the model is trained might be sufficient to achieve equalized odds. However, as we saw in the recidivism case study in Chapter 6, there may be correlations between group membership and the other features. For example, it may be possible to infer a student's gender with some accuracy from their program field of education or other demographic characteristics. In that case, the model may still have different effects on different groups even if it does not explicitly consider group membership. The shortcomings of this 'fairness through blindness' approach are well known [Dwork et al., 2012].

While there are several approaches to ensuring that equalized odds is (at least approximately) satisfied [Hardt et al., 2016; Zafar et al., 2017a], we explore the use of *learning fair representations with an adversary* [Edwards and Storkey, 2016] (see Chapters 4 and 6 for further details). In this approach, we prepare a *cleaned* version of the input data such that group membership – in this case, gender – cannot be inferred from the cleaned data. The cleaned data can be seen as an alternative *representation* of the original input.

We briefly recap the technique used in learning fair representations with an adversary. The cleaned data is prepared by passing the original input through a neural network model. The weights of the neural network are learned by optimizing a combination of two objectives on the training set: preventing an adversary (itself another neural network model) from inferring the group membership of individuals in the data, and otherwise making the cleaned data as similar as possible to the original input. In our experiments we selected parameters for the neural network learning which delivered reasonable performance, without exhaustively exploring the issue of parameter selection. We implemented the neural network models using the TensorFlow library in Python.

An advantage of *learning fair representations with an adversary* is that we may retain the same approach to predictive modeling, only applied to the cleaned data instead of the original input. This separation of concerns between achieving fairness in data pre-processing, without otherwise altering our approach to predictive modeling, is convenient. It also has governance benefits in that the party conducting predictive modeling need not be trusted to be fair, since this has already been guaranteed by the data pre-processing, which may be conducted by another party (see Chapter 4).

7.5.3 Results on the ANU Dataset

Recall that logistic regression using standard features was an approach which performed strongly in predicting at least one first year course failure (see Figure 7.6). We explored the effect of using logistic regression with:

- *original input* – standard features from Table 7.1
- *original input without gender* – standard features from Table 7.1, except gender

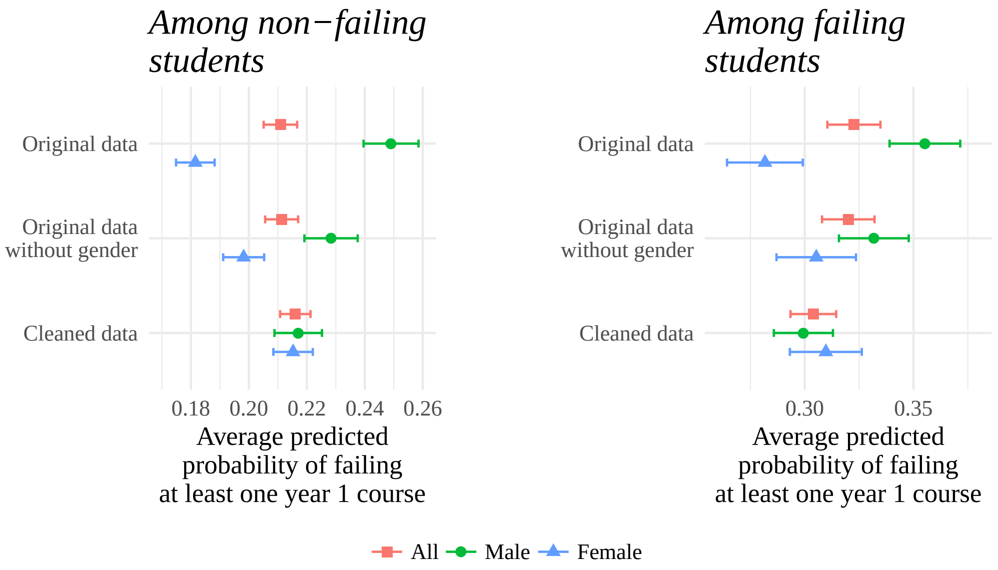


Figure 7.8: Addressing potential for gender discrimination when predicting at least one year 1 course failure.

- *cleaned data* – learning fair representations with an adversary applied to the original input.

The results are shown in Figure 7.8. In the left-hand plot we consider only students who did not in fact fail at least one first year course (*non-failing students*), showing the average predicted probability of failing for all students, male students and female students. In the right-hand plot we consider only students who did in fact fail at least one first year course (*failing students*), again showing the average predicted probability of failing for all students, male students and female students. 95% confidence intervals are shown, which were computed using a Student's t-test in R.

Using the original input, males were predicted to be more likely to fail than females to a statistically significant degree, both among non-failing and failing students. If such predictions were used to make decisions about which students were offered academic support, two fairness concerns may arise. Among the non-failing students, males may be disproportionately targeted for academic support programs that they do not in fact need. Among the failing students, females may disproportionately miss out on academic support programs that they do in fact need.

Using the original input without gender, the gap between genders reduces somewhat among both non-failing and failing students. This is because the predictive model can no longer explicitly consider gender. However, the gap between genders persists due to correlations between other features and gender.

Using the cleaned data, the predicted probabilities of failing converge for males and females among both non-failing and failing students. This convergence is particularly apparent among non-failing students, who make up a sizeable majority of students. If such predictions were used to make decisions about which students were offered academic support, the probability of non-failing students being targeted for an academic support program would be similar both males and females. Furthermore, the probability of failing students being omitted from an academic support program would be similar for both males and females. Hence, the fairness concerns we identified using the original input would be addressed.

However, among non-failing students the overall average predicted probability of failing appears somewhat higher using the cleaned data compared to the original input. Among failing students the overall average predicted probability of failing appears somewhat lower using the cleaned data compared to the original input. Using the cleaned data, the model is less accurate in *absolute* terms, while in *relative* terms its inaccuracies are more evenly spread between genders. This is not surprising given that the cleaned data has removed information present in the original input. The cost for overall utility of requiring such quantitative definitions of fairness has previously been observed in the literature [Corbett-Davies et al., 2017; Corbett-Davies and Goel, 2018; Menon and Williamson, 2018].

7.6 Conclusion

Cohort analysis and predictive modeling are both useful tools to assist higher education institutions in understanding and shaping student outcomes at university. Cohort analysis can help to identify particular characteristics of students who frequently achieve – or do not achieve – positive outcomes. This can assist universities in planning which groups of students to focus their student support efforts on. Predictive modeling offers the ability to be even more targeted, identifying which individuals would benefit most from assistance from the university. Inevitably, however, the trajectory of each individual student is to a degree unpredictable, and it is important not to overstate the accuracy of such models. This work covered only some possible outcomes – notably, it would be interesting to also consider graduate employment outcomes in future work.

Universities should expect and aspire to increasingly diverse student populations over time. The move to a national undergraduate admissions model at ANU is just one example of this. To make this transition a success, universities will need to ensure that they are providing support that is most needed by particular students and cohorts. They will also need to consider the equity implications of such interventions. Understanding the value – and limitations – of data-driven approaches will be essential in enabling institutions to help students get the most out of their time at university.

An interesting question is determining when it is appropriate for predictive models to be used. For example, is there a difference between using predictive models to target student support services and using them to inform admissions decisions? When predictive models are used, universities should be mindful of the risk that decisions made based on such models may have discriminatory effects against particular groups.

Fair representation learning is one mechanism to adjust models to minimize the risks of these discriminatory effects. This chapter demonstrated how this technique could be applied in the context of predicting higher education student outcomes, building upon our theoretical analysis in Chapter 4 and the recidivism case study in Chapter 6. However, the student outcomes case study again highlighted the tradeoff between maximizing the absolute utility of a predictive model and equalizing its relative utility for different groups.

In reality, individual identities are not binary and often cut across multiple group memberships, a phenomenon known as *intersectionality*. The small sample sizes in these intersections makes cohort analysis more difficult. An interesting area of active research is designing fairness definitions and interventions to address situations involving multiple intersecting groups [Kearns et al., 2018].

Chapter 8

Conclusion

This thesis has examined the question: when is representation learning provably useful? While we have provided specific examples to this question in Chapters 2, 3 and 4, it is possible to draw upon these cases to answer the question in a more holistic way (see Section 8.1). Our work on several problem variants has also yielded a common set of methodological insights (see Section 8.2).

This thesis has also examined a second question: how can representation learning help to achieve fairness in machine learning, and what are its limitations? Chapters 4 and 5 have addressed technical aspects of this question using mathematics, while Chapters 6 and 7 have examined case studies illuminating this issue. We step back and consider the relationship between representation learning and fairness with greater critical distance (see Section 8.3).

In the course of writing this thesis many opportunities for interesting future work have been identified, with respect to representation learning, fairness in machine learning, and their intersection. We preview a few such directions (see Section 8.4).

8.1 When Representation Learning is Provably Useful

Deep learning techniques, which transform raw data into successively more abstract representations, are at the core of contemporary machine learning. Such techniques have been successful in computer vision, natural language processing and other domains with high dimensional sensory input [Goodfellow et al., 2016]. Intuitively, such techniques learn structure from raw data, in the same way that a baby learns to make sense of the 'blooming, buzzing confusion' [James, 1890] their senses confront them with.

The term *representation learning* – while synonymous in the machine learning context with *feature learning* – evokes the broader notion of *representation* which stands in for an original, i.e. a *signifier* replacing a *signified* [Saussure, 2011]. A photo represents a scene in the physical world; a spoken or written word represents an abstract concept; a member of parliament represents their constituents. In each case, the representation synthesizes something more complex. It is easy to see what is lost – the detail of the original that is not present in the representation – as formalized in results such as the *data processing inequality* (Theorem 2.8.1 of [Cover and Thomas,

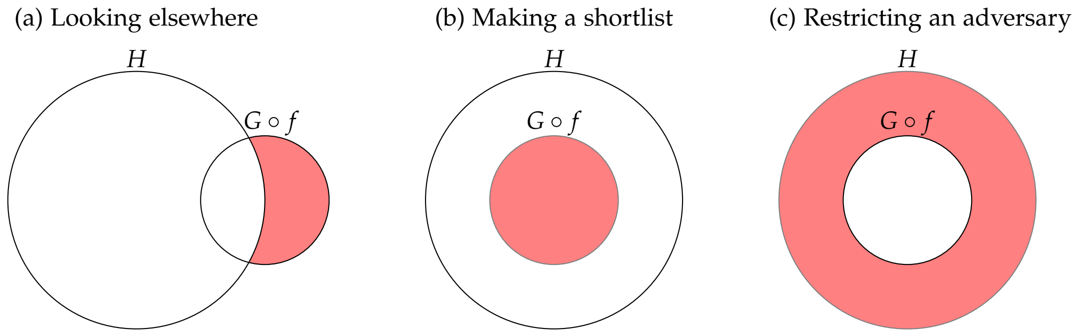


Figure 8.1: Three situations where representation learning is provably useful. In each case, we would like the hypothesis of interest to lie in the region shaded red.

2012]). We feel that we have gained something in return, but what exactly? The paradox is that while the downside of moving from the original to the representation can be crisply quantified, the upside can be intuited but is harder to pin down.

This thesis has considered the question of why a representation might be useful in the context of supervised learning. In particular, we have compared using an original hypothesis class H to an hypothesis class induced by representation function f, $G \circ f := \{g \circ f | g \in G\}$. This thesis has provided three possible answers to the question of when representation learning is provably useful, as summarized in Figure 8.1.

(a) Looking elsewhere: The original hypothesis class H does not contain the target hypothesis, whereas $G \circ f$ does. In Chapter 2 we investigated how to find such an f using unlabeled data.

(b) Making a shortlist: $G \circ f$ is a subset of H, both of which contain the target hypothesis. However, because of the smaller size of $G \circ f$, we can bound generalization error with fewer samples. In Chapter 3 we investigated how to find one or more such choices of f using labeled data from another task.

(c) Restricting an adversary: $G \circ f$ is a subset of H. While H contains an hypothesis of interest to an adversary, $G \circ f$ does not. In Chapter 4 we investigated how to verify that for some f, no hypothesis in $G \circ f$ can be too unfair to a particular group.

Inspecting Figure 8.1 it becomes clear that what makes a good representation function f depends on the hypothesis classes H and G. In (a), if H is all possible hypotheses, we will not be able to 'look elsewhere'. The representation function f is only useful insofar as $G \circ f$ *compensates for the deficiencies* of H. In (b), G must not be too large, or else our 'shortlist' $G \circ f$ will not be very short after all. In (c) by contrast, it makes sense for H and G to be the sets of all possible hypotheses corresponding to their respective type signatures, since there may still be unfair hypotheses available to an adversary in H but not in $G \circ f$. Our results are stronger if we are able to restrict an adversary who is optimally unfair.

Our analysis has revealed that a good representation function f also depends on the probability distribution and loss function used in the supervised learning problem. We conclude that there is no such thing as a universally useful representation function. This context dependence is reminiscent of the notion of compatibility between an hypothesis class and a probability distribution, which was introduced in the analysis of semi-supervised learning [Balcan and Blum, 2010].

8.2 Methodological Insights

This thesis has considered the theoretical foundations of representation learning across several distinct settings. There were several common aspects of our methodology in each case. In particular, *problem specification*, *dependence on assumptions* and *modularity* each played a role.

Problem specification: A wide range of techniques are used in machine learning. But it is important to first clearly state the problem that is being solved. In the cases of unsupervised representation learning (Chapter 2), learning transferrable representations (Chapter 3) and fair representation learning (Chapter 4), we formalized the definitions of successful representation learning before examining the conditions under which such definitions could be satisfied.

Dependence on assumptions: In constructing examples where particular forms of representation learning were provably useful, we had to introduce assumptions. In our example where unsupervised representation learning is useful we required a set of relationships between the representation function class, the hypothesis class, and the joint and unlabeled probability distributions (Chapter 2). Similarly, in our example where transferring representations is useful we considered particular classes of two-layer neural networks, and source and target tasks with related distributions (Chapter 3). We cannot expect unsupervised or transfer representation learning to always be useful. However, it is interesting to note that in practice these techniques have been found to be effective even when it is not clear that such assumptions have been met [Hinton and Salakhutdinov, 2006; Mikolov et al., 2013; Yosinski et al., 2014; Donahue et al., 2014]. In these cases, there are other aspects of the problem setting which we have not considered in our analysis – for example, the use of gradient-based optimization techniques, and regularities in particular kinds of domain-specific data – which may play a role in the success of unsupervised and transfer representation learning techniques, but about which we do not yet have a theoretical understanding.

Modularity: Modularity is an important principle in systems design, including in software engineering [Pressman, 2005]. Rather than solving a complex problem in one attempt, we instead decompose it into several components which are more manageable. The motivation for representation learning is similar – while our ultimate objective is solving a supervised learning problem, learning a new representation of the input is a first step. However, modular optimization is suboptimal compared to global optimization, in the same way that a greedy algorithm is suboptimal compared to the global solution. A security benefit of modularity is that user access can be limited to certain required system components, as we saw in the case of using representation learning to restrict an adversary in Chapter 4. The perspective of modularity helps us to understand both the advantages and disadvantages of representation learning as an approach to supervised learning problems.

8.3 Issues in Applying Representation Learning to Fairness

During the course of writing this PhD, fairness in machine learning dramatically increased its profile as an issue in public debate, in the academic community and in industry. Symptomatic of this trend, several of the world's largest technology companies released open source fairness toolkits, including IBM's AI Fairness 360 and Google's What-If Tool.[1] This thesis has considered applying representation learning to fairness, both from a theoretical perspective (Chapter 4), and in practice in the contexts of predicting recidivism in the criminal justice system (Chapter 6) and student outcomes at university (Chapter 7). It is of interest to consider issues associated with implementing such techniques in practice.

A central issue is how fairness is quantitatively defined. Chapter 4 showed that fair representation learning may improve *group fairness*, while simultaneously negatively impacting *individual fairness*. Furthermore, there are multiple possible definitions of group fairness, including several parity-based definitions. We may look at the difference in outcomes between groups (*statistical parity*) or the ratio of outcomes between groups (*disparate impact*). We also face a choice whether to condition these measures on some variable – for example, conditioning on the target variable we have *equalized odds*. Chapter 5 showed that there are hard trade-offs between equalized outcomes and equalized odds. We are unable to satisfy certain combinations of fairness definitions, regardless of whether we use fair representation learning. In practice, we must pick particular fairness definitions and recognize that they may represent the interests of certain stakeholders.

Representation learning for fairness raises issues of interpretability and transparency. In machine learning, it is common to convert an input vector into a feature vector via a representation function, and then use the feature vector to make predictions. A data point may be represented as a new point in a representation space that may be different from the original space, or even as a distribution over points if the representation function is stochastic. This method, which as we have seen may help to achieve some fairness objective, may also make it harder to interpret the reasons behind the algorithm's decisions. This kind of interpretability is beginning to be demanded by regulators, such as the 'right to an explanation' included in the European Union's General Data Protection Regulation (Recital 71 and Article 13, 2f of [European Union, 2016]). It is unclear what the public reaction will be to the identity of an individual being represented inside an algorithm as a vector which is not readily interpretable. This public reaction will have to be considered in the practical roll-out of such approaches.

The role of trust – and its absence – was highlighted in our analysis of applying representation learning to fairness. In Chapter 4, we saw that there are different routes to achieving fairness in the context of a trusted data user, i.e. *fair classification*, and in the context of an untrusted and potentially adversarial data user, i.e. *fair representation learning*. We introduced the notion of the *cost of mistrust*, which formalizes

[1] Accessible at https://developer.ibm.com/code/open/projects/ai-fairness-360/ and https://pair-code.github.io/what-if-tool/ respectively.

the price paid in terms of the accuracy-fairness trade-off available if the data user is not trusted, relative to the case of a trusted data user. This echoes findings about the role of trust in economics: trust reduces transaction costs, and without it, economic activity may be stifled [Algan and Cahuc, 2010; Gould and Hijzen, 2016]. While mistrust incurs a cost, in the scenario of data release to many parties, the assumption that a data user cannot be trusted to be fair may be inevitable. This would bring fairness in machine learning in line with other machine learning problem settings where users are not trusted, such as privacy-aware learning [Wainwright et al., 2012] and adversarial machine learning [Huang et al., 2011].

8.4 Future Work

In the course of conducting the research contained in this thesis, many interesting directions for future work have emerged. These are summarized below within the subject areas of representation learning and fairness in machine learning.

8.4.1 Representation Learning

We have considered representation learning as a tool in supervised learning problems. Representation learning promises a more principled alternative to feature extraction by humans. Feature extraction is particularly common with structured data which is not in the matrix form expected by a standard classifier, such as data from sequences, graphs, or several joined tables in a relational database. Representation learning has been applied to such structured data, such as time series data [Längkvist et al., 2014; Keogh and Pazzani, 1998]. An interesting direction is to formally investigate the role of representation learning in such cases.

Of particular interest is the formal analysis of learning representations of discrete *entities*, from data about the relationships between those entities. Entities are naturally represented using a one-hot code. Hinton [1984] proposed moving to 'distributed' vector representations, where aspects of the representation of an entity are shared with other entities. Where the entities are words, for example, embeddings can be found which place words that frequently co-occur nearby in the vector space [Mikolov et al., 2013; Turian et al., 2010]. The success of such an approach rests on the *distributional hypothesis*, i.e. that co-occurring words tend to be semantically similar, or in other words 'you shall know a word by the company it keeps' [Firth, 1957]. In social network analysis, the entities are people and their embeddings are learned from a graph of their relationships – here the analog of the distributional hypothesis is *homophily*, i.e. people connected to each other tend to have more similar tastes [Tang and Liu, 2011; Perozzi et al., 2014]. A comparable approach can be used to learn representations of entities from a bipartite graph – such as the graph of user ratings about movies – and is common in recommender systems [Menon and Elkan, 2010].

Given the ubiquity of representation learning of entities from relational data,[2] it would be interesting to formalize the conditions under which such techniques work well. We speculate that it may be possible to formalize the distributional hypothesis and quantify it along a spectrum between 'public' and 'private' behaviors, i.e. different types of behaviors may be more or less influenced by the behaviors of neighboring entities. Such a formalization would be of interest both from technical and sociological perspectives. A related question is the value of learning representations of entities, as compared to a distance metric between entities, and the extent to which the two notions are equivalent.

Representation learning is also at the heart of broader topics in machine learning and artificial intelligence. We intuit that the brain is capable of forming abstract representations of the world, including from unlabeled data. Should we be learning more about how [Fong et al., 2018], in order to guide research on machine learning? An example of this approach is research which identified how faces are encoded in the brains of macaques [Chang and Tsao, 2017]. Representation learning is an important part of the move towards automated machine learning (*AutoML*[3]), where the design choices involved in setting up a machine learning system are automated, in addition to standard learning of parameters. AutoML systems aim to make decisions of the kind that human designers typically make at present: choosing not only the representation of the data, but also the learning algorithm, the data collection process, and the infrastructure on which the system runs. Can we provide a mathematical formalization of AutoML, and the conditions under which it works well, more generally?

8.4.2 Fairness in Machine Learning

Fairness in machine learning is a growing field with many open research questions of interest. Among these, there are questions about *fairness definitions*, about the relationship between *privacy and fair representation learning*, and about the *fairness implications of generalizing from data*.

Fairness definitions: Definitions of fairness have been studied well before the current surge of interest in fairness in machine learning [Hutchinson and Mitchell, 2019]. Engaging with a breadth of issues associated with defining fairness – from a range of disciplines including law and philosophy – will be central to the next steps in the field of fair machine learning. Here are four particular directions of interest.

1. **Multiple groups.** The analysis in this thesis has focused on the case of ensuring fairness where there are two distinct groups, i.e. the sensitive variable is binary. However, in many realistic applications there are more than two groups, or group membership may be continuous or probabilistic. Furthermore, a person may be a member of multiple groups, as analyzed via the notion of intersection-

[2] An example of their popularity is the NeurIPS 2018 workshop on Relational Representation Learning (see https://r2learning.github.io/).

[3] See https://www.automl.org/ and https://cloud.google.com/automl/ for examples.

ality. Researchers have begun to consider this problem setting [Williamson and Menon, 2019; Kearns et al., 2018; Hebert-Johnson et al., 2018] and we expect that considering both fair representation learning and fairness impossibility results would be interesting in these cases.

2. **Conditioning variable.** The difference between equalized outcomes and equalized odds is that the former involves parity between groups overall, while the latter involves parity between groups conditioned on the target variable. It is possible that past data about the target variable is either not available (e.g. in the context of a hiring decision, it is not clear what the definition of the target should be), or affected by historic injustice or structural inequality. What about the case where it is possible to quantify the extent to which the target variable is observed or corrupted? Can we develop a principled interpolation between equalized outcomes and equalized odds?

3. **Relative vs absolute errors.** Equalized odds is far from the only possible measure of fairness, and has the downside that it is concerned only with relative error rates between groups rather than absolute error rates [Corbett-Davies and Goel, 2018]. What are the properties of possible variants, such as the absolute error rate among the group for which the algorithm has the highest error [Williamson and Menon, 2019]? Is it possible to apply concepts from finance – such as the Sharpe ratio for measuring risk adjusted returns [Sharpe, 1994], which trades off absolute returns and variation in returns over time – to the problem of trading off absolute error and variation in error across groups?

4. **Process fairness.** If a sensitive variable can be discerned from other variables, could it be acceptable to consider those other variables but not acceptable to consider the sensitive variable in a decision-making process? The two may be equivalent from a mathematical perspective, but perceived as distinct by society. More generally, can the legal idea of certain kinds of evidence being *admissible* and others not [Murphy, 2007] be extended to the context of algorithmic decisions? It would be interesting to incorporate human perceptions of process fairness into quantitative definitions of fairness, building on emerging work on this topic [Grgic-Hlaca et al., 2018].

Privacy and fair representation learning: Methods which alter data to preserve privacy have much in common with fair representation learning. In both cases the intent is to obscure some aspect of the data from an adversary, while preserving other useful information. It would be interesting to investigate the relationship between fairness metrics and differential privacy, which requires that summary statistics from a dataset are not too different, whether or not an individual record is included in the dataset [Dwork, 2008]. Techniques from *privacy preserving data publishing* [Fung et al., 2010] – where data is published with noise added, or is synthetically generated – may be applicable to fairness. Definitions such as k-anonymity (each individual's record is indistinguishable from k other records) and l-diversity (each released group of records is aggregated from a diverse range of records) may also have interpretations as fairness metrics. There is scope to bring insights from the fields of privacy and fair representation learning together, allowing techniques and definitions developed in one research area to be applied to the other.

Fairness implications of generalizing from data: A final topic of interest is understanding the implications for fairness of generalizing from past data. The process of generalizing from past data is at the heart of what machine learning systems do: as such, it cannot readily be excised. When it is acceptable to make predictions or decisions about someone on the basis of generalizations from the past behaviors of others? For example, recalling the case study on predicting student outcomes at university from Chapter 7, is there a difference between using this approach for targeted support as compared to university admissions decisions? Will the generalization process on which machine learning algorithms depend be perceived as objectionable to broader society in some contexts, and lead to a popular backlash? As machine learning is deployed to make or inform an increasing number of important decisions about people's lives, a key challenge – for researchers, industry practitioners, regulators, and society more broadly – is determining in a principled way the contexts where the algorithmic process of generalization can appropriately be used.

To understand the fairness implications of algorithmic generalizations from data, it is useful to consider similar generalizations made by humans. The concept of a *stereotype* has been defined as "an individual's set of beliefs about the characteristics or attributes of a group" [Judd and Park, 1993], which in many cases is informed by past data. Stereotypes can help individuals to explain trends they observe across groups [Campbell, 1967; McGarty et al., 2002] and are not necessarily inaccurate [Lee et al., 1995] – although they can be, and the term 'stereotype' carries a connotation of potential inaccuracy.[4] Analogously, we may view an algorithm's beliefs about the characteristics or attributes of a group – often based on generalizations from past data – as an *algorithmic stereotype*. These beliefs can be probabilistic in nature and measured in terms of observable behavior: for example, "members of group A are twice as likely on average to exhibit a certain behavior compared to members of group B".

Existing typologies of stereotype inaccuracy can help to analyze the effect of algorithmic stereotypes on particular groups. One possible distinction [Judd and Park, 1993] is between *stereotypic inaccuracy* – where the belief is inaccurate with respect to the average behavior of individuals in the group – and *dispersion inaccuracy* – where the belief does not take account of within-group differences. Our results in Chapter 5 showed that if, for each group, the average outcomes predicted by an algorithm equal the average observed outcomes – i.e. stereotypic inaccuracy is not present – the algorithm also tends to make less accurate predictions about individuals who are

[4] An alternative definition of 'stereotype', taken from the Cambridge Dictionary, captures this potential inaccuracy explicitly: "a set idea that people have about what someone or something is like, especially an idea that is wrong" [O'Shea and Waterhouse, 2012]. A proposed definition of 'stereotype' from an influential earlier work [Allport et al., 1954] – "an exaggerated belief associated with a category" – insists upon the inaccuracy of stereotypes, as opposed to objective "group traits".

atypical of their group – i.e. dispersion inaccuracy *is* present. Drawing upon the literature on stereotypes is a promising direction for understanding the implications for fairness of algorithmic decision-making.

Bibliography

Abu-Mostafa, Y., 2012. *Learning From Data*. AMLBook.

Algan, Y. and Cahuc, P., 2010. Inherited Trust and Growth. *American Economic Review*, 100, 5 (2010), 2060–92.

Allport, G. W.; Clark, K.; and Pettigrew, T., 1954. *The Nature of Prejudice*. Addison-Wesley.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L., 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Argyriou, A.; Evgeniou, T.; and Pontil, M., 2008. Convex Multi-task Feature Learning. *Machine Learning*, 73, 3 (2008), 243–272.

Arora, S.; Khandeparkar, H.; Khodak, M.; Plevrakis, O.; and Saunshi, N., 2019. A Theoretical Analysis of Contrastive Unsupervised Representation Learning. In *International Conference on Machine Learning*.

Arora, S. and Risteski, A., 2017. Provable Benefits of Representation Learning. arXiv 1706.04601.

Atwood, M., 1985. *The Handmaid's Tale*. McClelland and Stewart.

Australian Bureau of Statistics, 2001. Australian Standard Classification of Education.

Australian Bureau of Statistics, 2017a. Aboriginal and Torres Strait Islander Population Census 2016 Data Summary.

Australian Bureau of Statistics, 2017b. Personal Safety Survey 2016.

Australian Bureau of Statistics, 2017c. Prisoners in Australia 2017.

Australian Bureau of Statistics, 2017d. Recorded Crime – Victims, Australia 2016.

Australian Government Department of Education and Training, 2015. Completion Rates of Higher Education Students – Cohort Analysis, 2005-2014. https://docs.education.gov.au/system/files/doc/other/cohort_analysis_2005-2014_0.pdf.

Australian Government Department of Social Services, 2009. The Cost of Violence against Women and their Children. Report of the National Council to Reduce Violence against Women and their Children.

Australian Government Tertiary Education Quality and Standards Agency, 2017. Characteristics of Australian Higher Education Providers and their Relation to First-year Student Attrition. https://www.teqsa.gov.au/for-providers/resources/characteristics-australian-higher-education-providers-and-their-relation.

Australian Institute of Health and Welfare, 2018. Family, Domestic and Sexual Violence in Australia.

Australian Institute of Health and Welfare and Australia's National Research Organisation for Women's Safety, 2016. Examination of the Health Outcomes of Intimate Partner Violence against Women: State of Knowledge Paper.

Australian National University, 2018. New Admissions for 2020. http://www.anu.edu.au/study/apply/new-admissions-for-2020.

Balcan, M.-F. and Blum, A., 2010. A Discriminative Model for Semi-Supervised Learning. *Journal of the ACM*, 57, 3 (2010), 19.

Balcan, M.-F.; Blum, A.; and Vempala, S., 2015. Efficient Representations for Lifelong Learning and Autoencoding. In *Conference on Learning Theory*.

Bansal, M.; Gimpel, K.; and Livescu, K., 2014. Tailoring Continuous Word Representations for Dependency Parsing. In *Association for Computational Linguistics*.

Barocas, S.; Hardt, M.; and Narayanan, A., 2018. *Fairness and Machine Learning*. fairmlbook.org.

Barocas, S. and Selbst, A., 2016. Big Data's Disparate Impact. *California Law Review*, 104 (2016).

Baxter, J., 2000. A Model of Inductive Bias Learning. *Journal of Artificial Intelligence Research*, 12, 3 (2000), 149–198.

Bechavod, Y. and Ligett, K., 2017. Penalizing Unfairness in Binary Classification. arXiv 1707.00044.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W., 2010. A Theory of Learning from Different Domains. *Machine Learning*, 79, 1 (2010), 151–175.

Ben-David, S. and Borbely, R. S., 2008. A Notion of Task Relatedness Yielding Provable Multiple-task Learning Guarantees. *Machine Learning*, 73, 3 (2008), 273–287.

Bengio, Y., 2012. Deep Learning of Representations for Unsupervised and Transfer Learning. In *ICML Workshop on Unsupervised and Transfer Learning*.

Bengio, Y.; Courville, A.; and Vincent, P., 2013. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35, 8 (2013), 1798–1828.

Bengio, Y.; Paiement, J.-F.; Vincent, P.; Delalleau, O.; Le Roux, N.; and Ouimet, M., 2004. Out-of-sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering. In *Advances in Neural Information Processing Systems*.

Berk, R., 2012. *Criminal Justice Forecasts of Risk: A Machine Learning Approach*. Springer.

Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; and Roth, A., 2017. Fairness in Criminal Justice Risk Assessments: the State of the Art. arXiv 1703.09207.

Beutel, A.; Chen, J.; Zhao, Z.; and Chi, E., 2017. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. In *FAT/ML Workshop*.

Bowker, G. and Star, S., 1996. How Things (Actor-Net)Work: Classification, Magic and the Ubiquity of Standards. *Philosophia*, 25, 3 (1996), 195–220.

Boxall, H.; Rosevear, L.; and Payne, J., 2015. Identifying First Time Family Violence Perpetrators: The Usefulness and Utility of Categorisations Based on Police Offence Records. *Trends and Issues in Crime and Criminal Justice*, 487 (2015).

Breiman, L., 2001. Random Forests. *Machine Learning*, 45, 1 (2001), 5–32.

Breiman, L.; Friedman, J.; Olshen, R.; and Stone, C., 1984. *Classification and Regression Trees*. Wadsworth & Brooks/Cole Advanced Books & Software.

Bulmer, C., 2015. 'Australian Police Deal with a Domestic Violence Matter Every Two Minutes'. http://www.abc.net.au/news/2015-05-29/domestic-violence-data/6503734.

Calders, T. and Verwer, S., 2010. Three Naive Bayes Approaches for Discrimination-Free Classification. *Data Mining and Knowledge Discovery*, 21, 2 (2010), 277–292.

Calmon, F.; Wei, D.; Ramamurthy, K. N.; and Varshney, K., 2017. Optimized Data Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems*.

Campbell, D., 1967. Stereotypes and the Perception of Group Differences. *American Psychologist*, 22, 10 (1967), 817–829.

Chander, A., 2016. The Racist Algorithm. *Michigan Law Review*, 115 (2016), 1023–1045.

Chang, L. and Tsao, D., 2017. The Code for Facial Identity in the Primate Brain. *Cell*, 169, 6 (2017), 1013–1028.

Chouldechova, A., 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5, 2 (2017).

Corbett-Davies, S. and Goel, S., 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. arXiv 1808.00023.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A., 2017. Algorithmic Decision Making and the Cost of Fairness. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Cover, T. and Thomas, J., 2012. *Elements of Information Theory*. Wiley.

Cox, P., 2012. Violence Against Women in Australia: Additional Analysis of the Australian Bureau of Statistics' Personal Safety Survey. Horizons Research Report, Australia's National Research Organisation for Women's Safety.

Datta, A.; Tschantz, M. C.; and Datta, A., 2015. Automated Experiments on Ad Privacy Settings. *Proceedings on Privacy Enhancing Technologies*, 2015, 1 (2015), 92–112.

Deerwester, S.; Dumais, S.; Furnas, G.; Landauer, T.; and Harshman, R., 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41, 6 (1990), 391–407.

Desmarais, S. and Singh, J., 2013. Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States.

Dieterich, W.; Mendoza, C.; and Brennan, T., 2016. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Northpointe Inc.

Doersch, C.; Gupta, A.; and Efros, A., 2015. Unsupervised Visual Representation Learning by Context Prediction. In *IEEE International Conference on Computer Vision*.

Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; and Darrell, T., 2014. DeCAF: a Deep Convolutional Activation Feature for Generic Visual Recognition. In *International Conference on Machine Learning*.

Donini, M.; Oneto, L.; Ben-David, S.; Shawe-Taylor, J.; and Pontil, M., 2018. Empirical Risk Minimization Under Fairness Constraints. In *Advances in Neural Information Processing Systems*.

Dressel, J. and Farid, H., 2018. The Accuracy, Fairness, and Limits of Predicting Recidivism. *Science Advances*, 4, 1 (2018).

Dwork, C., 2008. Differential Privacy: A Survey of Results. In *International Conference on Theory and Applications of Models of Computation*.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R., 2012. Fairness Through Awareness. In *Innovations in Theoretical Computer Science Conference*.

Dwork, C.; Immorlica, N.; Kalai, A.; and Leiserson, M., 2018. Decoupled Classifiers for Group-Fair and Efficient Machine Learning. In *Conference on Fairness, Accountability and Transparency*.

Edwards, D. and McMillan, J., 2015. Completing University in Australia: A Cohort Analysis Exploring Equity Group Outcomes. Joining the Dots Research Briefings.

Edwards, H. and Storkey, A., 2016. Censoring Representations with an Adversary. In *International Conference on Learning Representations*.

Elkan, C., 2001. The Foundations of Cost-Sensitive Learning. In *International Joint Conference on Artificial Intelligence*.

Epstein, D., 1998. *Failing Boys?: Issues in Gender and Achievement*. McGraw-Hill Education.

Erhan, D.; Bengio, Y.; Courville, A.; Manzagol, P.-A.; Vincent, P.; and Bengio, S., 2010. Why Does Unsupervised Pre-training Help Deep Learning? *Journal of Machine Learning Research*, 11 (2010), 625–660.

European Union, 2016. General Data Protection Regulation. https://eur-lex.europa.eu/eli/reg/2016/679/oj.

Evgeniou, T. and Pontil, M., 2004. Regularized Multi-task Learning. In *International Conference on Knowledge Discovery and Data Mining*.

Fawcett, T., 2004. ROC Graphs: Notes and Practical Considerations for Researchers. *Machine Learning*, 31, 1 (2004), 1–38.

Feldman, M.; Friedler, S.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S., 2015. Certifying and Removing Disparate Impact. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Feurer, M.; Klein, A.; Eggensperger, K.; Springenberg, J.; Blum, M.; and Hutter, F., 2015. Efficient and Robust Automated Machine Learning. In *Advances in Neural Information Processing Systems*.

Firth, J., 1957. A Synopsis of Linguistic Theory, 1930-1955. In *Studies in Linguistic Analysis*. Basil Blackwell.

Fisher, A.; Medaglia, J.; and Jeronimus, B., 2018. Lack of Group-to-Individual Generalizability is a Threat to Human Subjects Research. *Proceedings of the National Academy of Sciences*, 115, 27 (2018), E6106–E6115.

Fitzgerald, R. and Graham, T., 2016. Assessing the Risk of Domestic Violence Recidivism. *Crime and Justice Bulletin*, 189 (2016).

FLORIDI, L., 2010. *Information: A Very Short Introduction*. Oxford University Press.

FONG, R.; SCHEIRER, W.; AND COX, D., 2018. Using Human Brain Activity to Guide Machine Learning. *Scientific Reports*, 8, 1 (2018), 5397.

FUNG, B.; WANG, K.; CHEN, R.; AND YU, P., 2010. Privacy-preserving Data Publishing: A Survey of Recent Developments. *ACM Computing Surveys*, 42, 4 (2010).

GALANTI, T.; WOLF, L.; AND HAZAN, T., 2016. A Theoretical Framework for Deep Transfer Learning. *Information and Inference*, 5, 2 (2016), 159–209.

GANIN, Y.; USTINOVA, E.; AJAKAN, H.; GERMAIN, P.; LAROCHELLE, H.; LAVIOLETTE, F.; MARCHAND, M.; AND LEMPITSKY, V., 2016. Domain-adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17, 59 (2016), 1–35.

GERMAIN, P.; HABRARD, A.; LAVIOLETTE, F.; AND MORVANT, E., 2013. A PAC-Bayesian Approach for Domain Adaptation with Specialization to Linear Classifiers. In *International Conference on Machine Learning*.

GHASSAMI, A.; KHODADADIAN, S.; AND KIYAVASH, N., 2018. Fairness in Supervised Learning: An Information Theoretic Approach. arXiv 1801.04378.

GHIFARY, M.; KLEIJN, B.; AND ZHANG, M., 2014. Domain Adaptive Neural Networks for Object Recognition. In *Pacific Rim International Conference on Artificial Intelligence*.

GILLINGHAM, P., 2006. Risk Assessment in Child Protection: Problem Rather than Solution? *Australian Social Work*, 59, 1 (2006), 86–98.

GILLINGHAM, P. AND GRAHAM, T., 2017. Big Data in Social Welfare: The Development of a Critical Perspective on Social Work's Latest Electronic Turn. *Australian Social Work*, 70, 2 (2017), 135–147.

GIRSHICK, R.; DONAHUE, J.; DARRELL, T.; AND MALIK, J., 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

GITELMAN, L., 2013. *Raw Data is an Oxymoron*. MIT Press.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A.; AND BENGIO, Y., 2016. *Deep Learning*. MIT Press.

GOULD, E. AND HIJZEN, A., 2016. *Growing Apart, Losing Trust? The Impact of Inequality on Social Capital*. International Monetary Fund.

GRGIC-HLACA, N.; REDMILES, E.; GUMMADI, K.; AND WELLER, A., 2018. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In *World Wide Web Conference*.

HARCOURT, B., 2006. *Against Prediction: Profiling, Policing and Punishing in an Actuarial Age*. University of Chicago Press.

Hardt, M.; Price, E.; and Srebro, N., 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*.

Hebert-Johnson, U.; Kim, M.; Reingold, O.; and Rothblum, G., 2018. Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In *International Conference on Machine Learning*.

Hinton, G., 1984. Distributed Representations. Technical Report, Carnegie Mellon University.

Hinton, G.; Osindero, S.; and Teh, Y.-W., 2006. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18, 7 (2006), 1527–1554.

Hinton, G. and Salakhutdinov, R., 2006. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313, 5786 (2006), 504–507.

Hoeffding, W., 1963. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58, 301 (1963), 13–30.

Hoffman, J.; Guadarrama, S.; Tzeng, E.; Hu, R.; Donahue, J.; Girshick, R.; Darrell, T.; and Saenko, K., 2014. LSDA: Large Scale Detection Through Adaptation. In *Advances in Neural Information Processing Systems*.

Hornik, K., 1991. Approximation Capabilities of Multilayer Feedforward Networks. *Neural Networks*, 4, 2 (1991), 251–257.

Huang, L.; Joseph, A.; Nelson, B.; Rubinstein, B.; and Tygar, J., 2011. Adversarial Machine Learning. In *ACM Workshop on Security and Artificial Intelligence*.

Hughes-Warrington, M.; Nurbasari, L.; Evans, E.; Hawkins, S.; and Hilton, R., 2019. Admissions, Education's Double: Building a National Admissions Model for a National University. Publication forthcoming.

Hutchinson, B. and Mitchell, M., 2019. 50 Years of Test (Un) fairness: Lessons for Machine Learning. In *Fairness, Accountability and Transparency Conference*.

Jaeger, M. M., 2008. Does Left–Right Orientation have a Causal Effect on Support for Redistribution? Causal Analysis with Cross-Sectional Data using Instrumental Variables. *International Journal of Public Opinion Research*, 20, 3 (2008), 363–374.

James, W., 1890. *The Principles of Psychology*. Henry Holt and Company.

Jia, P. and Maloney, T., 2015. Using Predictive Modelling to Identify Students at Risk of Poor University Outcomes. *Higher Education*, 70, 1 (2015), 127–149.

Johndrow, J. and Lum, K., 2017. An Algorithm for Removing Sensitive Information: Application to Race-Independent Recidivism Prediction. arXiv 1703.04957.

Johnson, W. and Lindenstrauss, J., 1984. Extensions of Lipschitz Mappings into a Hilbert Space. *Contemporary Mathematics*, 26 (1984), 189–206.

Judd, C. and Park, B., 1993. Definition and Assessment of Accuracy in Social Stereotypes. *Psychological Review*, 100, 1 (1993), 109–128.

Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S., 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *International Conference on Machine Learning*.

Keogh, E. and Pazzani, M., 1998. An Enhanced Representation of Time Series Which Allows Fast and Accurate Classification, Clustering and Relevance Feedback. In *KDD*.

Kilbertus, N.; Carulla, M. R.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B., 2017. Avoiding Discrimination through Causal Reasoning. In *Advances in Neural Information Processing Systems*.

Kleinberg, J.; Lakkaraju, H.; Leskovec, J.; Ludwig, J.; and Mullainathan, S., 2017a. Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 133, 1 (2017), 237–293.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M., 2017b. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Innovations in Theoretical Computer Science*.

Korhonen, V. and Rautopuro, J., 2018. Identifying Problematic Study Progression and "At-Risk" Students in Higher Education in Finland. *Scandinavian Journal of Educational Research*, (2018), 1–14.

Krug, E.; Dahlberg, L.; Mercy, J.; Zwi, A.; and Lozano, R., 2002. The World Report on Violence and Health. World Health Organization.

Kusner, M.; Loftus, J.; Russell, C.; and Silva, R., 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*.

Längkvist, M.; Karlsson, L.; and Loutfi, A., 2014. A Review of Unsupervised Feature Learning and Deep Learning for Time-series Modeling. *Pattern Recognition Letters*, 42 (2014), 11–24.

Laura and John Arnold Foundation. Public Safety Assessment Frequently Asked Questions . https://www.psapretrial.org/about/faqs.

Laura and John Arnold Foundation, 2017. Public Safety Assessment: Risk Factors and Formula. https://www.arnoldfoundation.org/wp-content/uploads/PSA-Risk-Factors-and-Formula.pdf.

LeCun, Y.; Bengio, Y.; and Hinton, G., 2015. Deep Learning. *Nature*, 521, 7553 (2015), 436–444.

Lee, Y.-T.; Jussim, L.; and McCauley, C., 1995. *Stereotype Accuracy: Toward Appreciating Group Differences.* American Psychological Association.

Lipton, Z.; Chouldechova, A.; and McAuley, J., 2018. Does Mitigating ML's Impact Disparity Require Treatment Disparity? In *Advances in Neural Information Processing Systems*.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M., 2015. Learning Transferable Features with Deep Adaptation Networks. In *International Conference on Machine Learning*.

Louizos, C.; Swersky, K.; Li, Y.; Zemel, R.; and Welling, M., 2016. The Variational Fair Autoencoder. In *International Conference on Learning Representations*.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R., 2018. Learning Adversariially Fair and Transferable Representations. In *International Conference on Machine Learning*.

Mansour, Y.; Mohri, M.; and Rostamizadeh, A., 2009. Domain Adaptation: Learning Bounds and Algorithms. In *Conference on Learning Theory*.

Mason, R. and Julian, R., 2009. Analysis of the Tasmania Police Risk Assessment Screening Tool (RAST), Final Report. Tasmanian Institute of Law Enforcement Studies, University of Tasmania.

Maurer, A.; Pontil, M.; and Romera-Paredes, B., 2016. The Benefit of Multitask Representation Learning. *Journal of Machine Learning Research*, 17, 81 (2016), 1–32.

McGarty, C.; Yzerbyt, V.; and Spears, R., 2002. *Stereotypes as Explanations: The Formation of Meaningful Beliefs about Social Groups*. Cambridge University Press.

Menon, A. and Elkan, C., 2010. Predicting Labels for Dyadic Data. *Data Mining and Knowledge Discovery*, 21, 2 (2010), 327–343.

Menon, A. and Williamson, R., 2018. The Cost of Fairness in Binary Classification. In *Conference on Fairness, Accountability and Transparency*.

Messing, J. T.; Campbell, J.; Sullivan Wilson, J.; Brown, S.; and Patchell, B., 2017. The Lethality Screen: the Predictive Validity of an Intimate Partner Violence Risk Assessment for Use by First Responders. *Journal of Interpersonal Violence*, 32, 2 (2017), 225–226.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J., 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*.

Mitchell, S. and Shadlen, J., 2018. Mirror Mirror: Reflections on Quantitative Fairness. https://speak-statistics-to-power.github.io/fairness.

Mohri, M.; Rostamizadeh, A.; and Talwalkar, A., 2012. *Foundations of Machine Learning*. MIT Press.

Murphy, P., 2007. *Murphy on Evidence*. Oxford University Press.

Nayaranan, A., 2018. Tutorial: 21 Fairness Definitions and their Politics. https://www.youtube.com/watch?v=jIXIuYdnyyk.

Northpointe Inc., 2012. Practitioners guide to COMPAS.

Norton, A. and Cherastidtham, I., 2018. Dropping Out: The Benefits and Costs of Trying University. https://grattan.edu.au/wp-content/uploads/2018/04/904-dropping-out-the-benefits-and-costs-of-trying-university.pdf.

NSW Bureau of Crime Statistics and Research, 2018. Re-offending Statistics for NSW.

NSW Police Force, 2016. NSW Police Force Corporate Plan 2016-18. (2016).

Olshausen, B. and Field, D., 1996. Emergence of Simple-cell Receptive Field Properties by Learning a Sparse Code for Natural Images. *Nature*, 381, 6583 (1996), 607–609.

O'Neil, C., 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books.

Ong, C. S.; Smola, A.; and Williamson, R., 2005. Learning the Kernel with Hyperkernels. *Journal of Machine Learning Research*, 6 (2005), 1043–1071.

O'Shea, S. and Waterhouse, H. (Eds.), 2012. *Cambridge Learner's Dictionary, 4th Edition*. Cambridge University Press.

Pan, S. J. and Yang, Q., 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22, 10 (2010), 1345–1359.

Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; and Efros, A., 2016. Context Encoders: Feature Learning by Inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Pentina, A. and Lampert, C. H., 2014. A PAC-Bayesian bound for Lifelong Learning. In *International Conference on Machine Learning*.

Perozzi, B.; Al-Rfou, R.; and Skiena, S., 2014. Deepwalk: Online Learning of Social Representations. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Phillips, A., 2004. Defending Equality of Outcome. *Journal of Political Philosophy*, 12, 1 (2004), 1–19.

Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; and Weinberger, K., 2017. On Fairness and Calibration. In *Advances in Neural Information Processing Systems*.

Pressman, R., 2005. *Software Engineering: A Practitioner's Approach*. Palgrave Macmillan.

Qu, L.; Ferraro, G.; Zhou, L.; Hou, W.; Schneider, N.; and Baldwin, T., 2015. Big Data Small Data, In domain Out-of Domain, Known Word Unknown Word: the Impact of Word Representation on Sequence Labelling Tasks. In *Conference on Computational Natural Language Learning*.

Raina, R.; Ng, A.; and Koller, D., 2006. Constructing Informative Priors using Transfer Learning. In *International Conference on Machine Learning*.

Rawls, J., 1971. *A Theory of Justice*. Harvard University Press.

Rice, M. and Harris, G., 1995. Violent Recidivism: Assessing Predictive Validity. *Journal of Consulting and Clinical Psychology*, 63 (1995), 737–748.

Rice, M.; Harris, G.; and Hilton, Z., 2010. The Violence Risk Appraisal Guide and Sex Offender Risk Appraisal Guide for Violence Risk Assessment. In *Handbook of Violence Risk Assessment*. Routledge.

Rigollet, P., 2007. Generalization Error Bounds in Semi-supervised Classification Under the Cluster Assumption. *Journal of Machine Learning Research*, 8 (2007), 1369–1392.

Romei, A. and Ruggieri, S., 2014. A Multidisciplinary Survey on Discrimination Analysis. *The Knowledge Engineering Review*, 29, 5 (2014), 582–638.

Sarwar, B.; Karypis, G.; Konstan, J.; and Riedl, J., 2000. Application of Dimensionality Reduction in Recommender System – A Case Study. Technical report, University of Minnesota.

Saussure, F. D., 2011. *Course in General Linguistics*. Columbia University Press.

Saxe, A.; McClelland, J.; and Ganguli, S., 2014. Exact Solutions to the Nonlinear Dynamics of Learning in Deep Linear Neural Networks. In *International Conference on Learning Representations*.

Sentas, V. and Pandolfini, C., 2017. Policing Young People in NSW: A Study of the Suspect Targeting Management Plan. Youth Justice Coalition.

Shalev-Shwartz, S. and Ben-David, S., 2014. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

Sharpe, W., 1994. The Sharpe Ratio. *Journal of Portfolio Management*, 21, 1 (1994), 49–58.

Shmueli, G., 2010. To Explain or to Predict? *Statistical Science*, 25, 3 (2010), 289–310.

Singh, A.; Nowak, R.; and Zhu, X., 2009. Unlabeled Data: Now it Helps, Now it Doesn't. In *Advances in Neural Information Processing Systems*.

Socher, R.; Ganjoo, M.; Manning, C.; and Ng, A., 2013. Zero-shot Learning through Cross-modal Transfer. In *Advances in Neural Information Processing Systems*.

Sutskever, I.; Jozefowicz, R.; Gregor, K.; Rezende, D.; Lillicrap, T.; and Vinyals, O., 2015. Towards Principled Unsupervised Learning. arXiv 1511.06440.

Sutskever, I.; Vinyals, O.; and Le, Q., 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*.

Tang, L. and Liu, H., 2011. Leveraging Social Media Networks for Classification. *Data Mining and Knowledge Discovery*, 23, 3 (2011), 447–478.

The Healing Foundation and White Ribbon Australia, 2017. Towards an Aboriginal and Torres Strait Islander Violence Prevention Framework for Men and Boys.

Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 1 (1996), 267–288.

Trautmann, S. and van de Kuilen, G., 2016. Process Fairness, Outcome Fairness, and Dynamic Consistency: Experimental Evidence for Risk and Ambiguity. *Journal of Risk and Uncertainty*, 53, 2/3 (2016), 75–88.

Turian, J.; Ratinov, L.; and Bengio, Y., 2010. Word Representations: A Simple and General Method for Semi-supervised Learning. In *Association for Computational Linguistics*.

United States Equal Opportunity Employment Commission, 1978. Uniform Guidelines on Employee Selection Procedures.

Van Rooyen, B. and Williamson, R., 2015. A Theory of Feature Learning. arXiv 1504.00083.

Wainwright, M.; Jordan, M.; and Duchi, J., 2012. Privacy Aware Learning. In *Advances in Neural Information Processing Systems*.

Wheeler, J., 1990. Information, Physics, Quantum: The Search for Links. In *Complexity, Entropy, and the Physics of Information*. Addison-Wesley.

Wierzbicka, A., 2006. *English: Meaning and Culture*. Oxford University Press.

Wijenayake, S.; Graham, T.; and Christen, P., 2018. A Decision Tree Approach to Predicting Recidivism in Domestic Violence. arXiv 1803.09862.

Williamson, R. and Menon, A., 2019. Fairness Risk Measures. In *International Conference on Machine Learning*.

Yosinski, J.; Clune, J.; Bengio, Y.; and Lipson, H., 2014. How Transferable are Features in Deep Neural Networks? In *Advances in Neural Information Processing Systems*.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K., 2017a. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment. In *International Conference on World Wide Web*.

ZAFAR, M. B.; VALERA, I.; GOMEZ RODRIGUEZ, M.; AND GUMMADI, K., 2017b. Fairness Constraints: Mechanisms for Fair Classification. In *International Conference on Artificial Intelligence and Statistics*.

ZEMEL, R.; WU, Y.; SWERSKY, K.; PITASSI, T.; AND DWORK, C., 2013. Learning Fair Representations. In *International Conference on Machine Learning*.

ZENG, J.; USTUN, B.; AND RUDIN, C., 2017. Interpretable Classification Models for Recidivism Prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180, 3 (2017), 689–722.

ZHAO, M.-J.; EDAKUNNI, N.; POCOCK, A.; AND BROWN, G., 2013. Beyond Fano's Inequality: Bounds on the Optimal F-Score, BER, and Cost-Sensitive Risk and their Implications. *Journal of Machine Learning Research*, 14 (2013), 1033–1090.

ZHUANG, J.; WANG, J.; HOI, C. H.; AND LAN, X., 2011. Unsupervised Multiple Kernel Learning. *Journal of Machine Learning Research*, 20 (2011), 129–144.

ZLIOBAITE, I., 2015. A Survey on Measuring Indirect Discrimination in Machine Learning. arXiv 1511.00148.