Dynamics of Algorithmic Fairness

HARVARD UNIVERSITY

Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Harvard John A. Paulson School of Engineering and Applied Sciences and Philosophy
have examined a dissertation entitled:

"Dynamics of Algorithmic Fairness"

presented by: Lily Hu

Signature ______________________________
Typed name: Professor Y. Chen

Signature ______________________________
Typed name: Professor B. Grosz

Signature ______________________________
Typed name: Professor C. Dwork

April 12, 2022

Dynamics of Algorithmic Fairness

a dissertation presented
by
Lily Hu
to
The Department of The School of Engineering and Applied Sciences

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Applied Mathematics

Harvard University
Cambridge, Massachusetts
April 2022

Thesis advisor: Professor Yiling Chen Lily Hu

Dynamics of Algorithmic Fairness

Abstract

The rise of machine learning-based predictive models in making decisions of profound social impact has spurred study of those technical properties that may bear on the moral and political character of their deployment. One area of normative concern that has received particularly heightened scrutiny is the *fairness* of the outcomes that data-based classification tools issue. Much computer science and mathematics-based research in the fields of algorithmic fairness and fair machine learning looks to diagnose when classifiers may be engaging in discrimination or otherwise generating unfair outputs and to prevent such outcomes using a variety of methods that seeks to alter the classifier's behavior and thus the outcomes it produces.

This dissertation comprises contributions to the burgeoning field of algorithmic fairness that, rather than focusing on the internal workings of an algorithmic system itself, centers instead the interaction between machine classifications and the broader societal contexts within which data-based predictive tools are embedded. Each of these works thus conceive of algorithmic tools as only one component of a larger sociotechnical system that distributes key social benefits and burdens. Over the span of the three projects contained within—"Disparate Effects of Strategic Manipulation," "A Short-term Intervention for Long-term Fairness," and "Fair Classification and Social Welfare"—it considers changes to institutional incentive structures that data-based classification introduces, the strategic responses of agents who interact with such systems, and the welfare impacts of various fairness constraints that have been proffered in the field. Approaching the fairness problem with this wider lens of analysis builds in a broader and longer-term perspective from the start and necessarily draws on methods and insights beyond that of applied mathematics and computer science. In so doing, this research makes distinctive contributions to matters that are central in the scholarly discourse in algorithmic fairness, such as debate about fairness-accuracy trade-offs in algorithmic decision-making and the strategic interplay between machine classifications and agent behaviors. This dissertation therefore both advocates for and itself exemplifies a reorientation to questions of fairness by shifting focus from the machine as the central object of interest in favor of a broader vantage that addresses the broader social *dynamics* of algorithmic fairness. This approach not only challenges the standard methodological tacks taken in the field of algorithmic fairness but also generates insights that track more closely to how these tools actually operate in the world to effect key social outcomes. It thus is better suited to guiding work in algorithmic fairness towards the kinds of interventions we will need to construct a more equitable society.

Contents

Acknowledgments

Thank you to my collaborators on these projects: Yiling Chen, Nicole Immorlica, and Jenn Wortman Vaughan.

To my committee:

To Cynthia and Barbara, who taught me through their examples what it is to embody research and personal excellence, how to pursue research with great technical skill but also grounded in a deep scholarly sensibility that is all too rare. Your own work and the way you approach research is excellent by the standards of computer science and observing you both, I think, part of how you manage that excellence is by being so open-minded and receptive to the contributions of other disciplines. Seeing you both engage deeply beyond computer science assured me that my choice to pursue philosophy would not be taken by either of you as a failure. You are each absolute paragons of trailblazing and most remarkably, you supported me even when the trail that I wanted to blaze was different from your own.

To Yiling, who took in a student who did not know an ounce of real "computer science," whose experience coding was zilch, and through patience and a delicate balance of directed guidance alongside an attitude that encouraged exploration, transformed her into someone who found passion in a growing area of computer science and, surprisingly!, had something to contribute to it. Then when she make the crazy decision to take on a Philosophy degree, you supported the choice nonetheless—also in part on your dime. It took immense faith in me and a selflessness that few advisers have. Even fewer would have gone down all these rabbit holes with me, jumped through so many administrative hoops, into this great unknown, uncharted territory. I would be in a completely different place without your guidance. Your support in the past six years has quite literally transformed and continues to transform my academic career and my life.

Thank you to all my interlocutors, friends, and family all these years who have carried me through so many triumphs and even more defeats. I cannot express my gratitude with sufficient eloquence, so I'll just leave it at that: immense thank you truly from the bottom of my heart.

1

Introduction

Before there was anything nearing a fully-fledged discipline or an established research community dedicated to studying the so-called fairness properties of algorithmic systems, a small group of scholars identified a risk of using machine classification tools in decision procedures that distribute key social benefits and burdens. Works such as Dwork et al.'s "Fairness Through Awareness"[30] and Kamishima et al.'s "Fairness-aware Learning through Regularization Approach"[53] drew out a potential problem inherent to the task of machine-based classification. The aim of these classification systems is to draw distinctions among individuals, sorting them into types vis-à-vis some outcome of interest, and labeling them accordingly. But not all ways of differentiating individuals are permissible; some are morally wrongful and even legally prohibited. This raises a challenge: How can we make sure that data-based classifiers, tools that are optimized to be highly discriminating instruments of classification, do not discriminate in the wrongful sense on the basis of salient social groups like sex and race? How can those who design machine learning systems ensure that when they are deployed in systems that distribute important social goods such as access to loans, employment opportunities, and second-chances at public life, that they do not systematically exclude individuals on account of their race, sex, religion, and so on? How would we even know whether and when the highly complex patterns and distinctions that data-based algorithms trace out constitute unfair or discriminatory classification? And most importantly, how can we prevent classifiers from engaging in such discrimination?

These are the 10,000 foot questions of algorithmic fairness. Early works in computer science formalized this problem in distinctive ways by defining mathematical or computational notions of "fairness" or "unfairness" or "discrimination," and then providing approaches to resolving fairness problems or describing conditions under which solutions could or could not be found. The framework set by these early papers for how to approach technical questions of algorithmic fairness has remained highly influential to this day. These works also set the foundation upon which, in the following several years, what I will call the *first wave* of research on algorithmic fairness took place. This wave of works investigated formal properties of the classification system as the primary site of fairness or discrimination concerns. Seminal works such as Hardt et al.'s "Equality of Opportunity in Supervised Learning" and Zafar et al.'s "Fairness Constraints: Mechanisms for Fair Classification" probe the internal dynamics of classification and ask questions such as: How does the classifier treat the data inputs it receives? How well does a classifier's output meet various mathematical criteria of fairness? Research in this vein typically centers on a machine's optimization problem or a model of classifier behavior and then devises methods to ensure some formal fairness definition is met given a set of assumptions about the classifier setup. This overall approach has contributed greatly to our understanding of the many ways that algorithmic systems can raise concerns of discrimination. Such work has offered up many mathematical formalizations of fairness that have intuitive appeal and can be implemented in practice, illuminated the different sources of bias in models constructed from data, and shown the prospects and limitations of various de-biasing methods, among other key insights.

These formalizations of the problem, however, abstract from an important feature of algorithmic decision-making in practice. Socially impactful algorithms do not do anything on their own. They are embedded within existing institutions, plug into other pipelines of decision-making and into larger systems of rules and procedures. This integration with other features of social life is crucial to explaining how numbers and code in a machine can reach out into the world to bestow benefits or inflict harms on real people. Work that approaches fairness by looking to explicitly account for these factors outside of the machine make for what I will call a *second wave* of research in algorithmic fairness. And I take the works contained in this dissertation to be a small part of this overall effort.

I have titled this dissertation "Dynamics of Algorithmic Fairness" because the approach that I take to questions of fairness is centrally concerned with developing a perspective on fair machine learning that centers the interplay between machines, humans, institutions, and social structures. Importantly, however, my research project is still rooted in a mathematical perspective despite its emphasis on the so-called "sociotechnical" nature of algorithmic systems. I take sociotechnical features as a starting point of my analyses but proceed with a primarily mathematical and computational orientation, using tools from, as examples, learning theory and game theory. Still, while the works in this dissertation are in conversation with computer scientists working in the first wave of research in algorithmic fairness, rather than taking their setups and questions as given, I develop a distinctive formulation of longer-standing problems by highlighting the definite forms that algorithmic decision-making takes when such systems are embedded in a dynamic social world.

This shifted orientation gives rise to two themes that run through my research projects in algorithmic fairness, and I will now discuss each in turn. The first conveys a distinctive methodological tack taken in the chapters that follow. Since I am centrally interested in what arises out of the interplay between our algorithmic systems and the social systems within which they are embedded, the lens I take to the problem of fairness has a notably wider scope. I analyze classifier behavior and outputs by considering how agents interact with machine classification against a background social structure and set of institutions. This *wider lens of analysis* comprises of three key features:

1. Agents who interact with machines are strategic and heterogeneous.

2. Ours is a world characterized by persistent social inequalities.

3. Classification outcomes plug into and affect other features of social systems.

The first point reminds us that machine classifiers draw on data that are produced by individuals who are not just random draws from a probability distribution. They are reactive; they are strategic. Their behaviors are a function of their interests and their environment. Insofar as agents have different environments and often different interests, these data are heterogeneous in a deep sense. Models of machine behavior must explicitly account for the fundamentally social nature of data.

Second, if the world is indeed characterized by social inequality, then that fact must have some origin, some story that explains how it is so, which must be accounted for explicitly in one's model of it. One cannot address social inequality, in my view, without having a theory for how it arises and why it continues to be such a seemingly permanent fixture in our world.

The third point ties all this together: the social system is a system that links together agents, institutions, actions, beliefs, across time. Presently observed data are the product of a long lineage of previous social choices and conditions, and by the same token, a classification system's issued outcomes will serve to have ripple effects into the future. A dynamic model incorporates these features of temporality to string together a continuous picture of the world to explain why data look the way they do now, why people and in turn, why data-based machines behave the way they do, and what inertial forces exist to lock us into various steady state equilibria. My claim is that we can only make sense of these crucial facts if we pursue a dynamic and historical understanding of our social systems.

Two projects within this dissertation are illustrative of this wider lens of analysis. In Chapter 2, titled “The Disparate Effects of Strategic Manipulation,” Nicole Immorlica, Jenn Wortman Vaughan, and I probe the fairness properties of data-based classifiers that are optimized for a setting of strategic classification. We develop a model of strategic interaction between, on the one hand, those *candidates* who are being classified, and on the other, the *learner* who is doing the classifying. Candidates in our model are heterogeneous in their strategic behaviors; their ability to respond strategically to a classifier is a function of the costs they face when looking to “trick” a learner by manipulating their features. In cases of real world classification, an agent’s costs are not simply a function of their personal interest in receiving a positive classification but is bound up in a complex web of social factors that affect her ability to pursue certain action responses. In a setting of social inequality, those in disadvantaged groups face systematically higher costs than those in advantaged groups. Our results show that whenever one group’s costs are higher than the other’s in the strategic classification game, the learner’s equilibrium strategy exhibits an inequality-reinforcing phenomenon wherein the learner erroneously admits some candidates who are members of the advantaged group, while erroneously excluding some candidates who are members of the disadvantaged group.

The fact that interplay between a classifier and the broader strategic environment within which it is embedded may result in inequality-reinforcing feedback loops is a theme that runs also through the following chapter, titled “A Short-term Intervention for Long-term Fairness in the Labor Market.” In this chapter, Yiling Chen and I build a dynamic reputational model of firms and agents contracting in a labor market, which shows how unequal group outcomes may be immovable even when employers' hiring decisions are bound by an input-output notion of "individual fairness." This is precisely because the interaction between heterogeneous agents belonging to different social groups and firms generates feedback effects resulting from groups' divergent accesses to resources and as a result, investment choices, which then serve to reinforce asymmetric outcomes over time. To counter this outcome, we construct a dual labor market composed of a Temporary Labor Market (TLM), in which firms' hiring strategies are constrained to ensure statistical parity of workers granted entry into the pipeline, and a Permanent Labor Market (PLM), in which firms hire top performers as desired. Individual worker reputations produce externalities for their group; the corresponding feedback loop raises the collective reputation of the initially disadvantaged group via a TLM fairness intervention that need not be permanent. The feedback mechanism of the dynamic system is thus co-opted to bring about a regime of group-equitable outcomes.

The second theme that is characteristic of my works in this dissertation concerns a set of questions about *trade-offs* that have been central to the field of algorithmic fairness since its inception and on which my approach sheds, I think, distinctive and illuminating light. Questions of trade-offs concern how different desiderata of predictive algorithms should be weighed against each other. How should an interest in accuracy be traded-off for a concern for fairness? Some researchers have taken the existence of such trade-offs—the fact that one cannot simultaneously make progress on a system's fairness properties without losing progress on the accuracy front—to be "inevitable" features of machine classification[24,36,61]. Others have taken trade-offs to be an artifact of label bias or some other contingent assumption about the data that are not true in many cases[91,29]. More broadly, the debate centers the question of whether in the realm of algorithmic decision-making, we are essentially in a game of compromises, or whether, under certain circumstances, a win-win is possible. This matter has been taken to be one of the most fundamental disputes in the field.

An analysis that takes a dynamic approach and embeds algorithms in particular social and economic contexts shows the limitations of the standard framings of the trade-offs question, guides towards more interesting dimensions of the question, and also lends new insight into the problem more generally. My works enter the debate by bringing to light two facts and from there, analyzing questions of trade-offs from quite a different angle. First, once we take the problem of fairness to be a problem that emerges out of the interaction between algorithms and the network of social institutions within which they are embedded, the field's standard notions of fairness and accuracy no longer appear as the only central values at stake. Other values—values such as welfare and efficiency—emerge as normatively significant as well and how these values trade-off against fairness and accuracy is not well understood. For example, as I show in an extension of the strategic classification model in Chapter 2, even when members of a disadvantaged group have their higher costs subsidized by the learner, they are not necessarily made better-off by their greater ability to manipulate. Here we prove a rather paradoxical result about trade-offs: there exist cases in which providing a subsidy improves only the learner's utility while actually making both candidate groups worse-off—even the disadvantaged group that receives the subsidy.

Second, conceiving of algorithms as a part of a dynamic social system that inherits data from the past and produces results that play a hand in influencing outcomes well into the future reveals the standard trade-offs conversation to be besot with a troubling ambiguity. When precisely are we to evaluate an algorithm's ”fairness” and ”accuracy”? What does it mean to claim that an algorithm is fair? At what time horizon are we even evaluating an algorithm's fairness or accuracy? If, as dynamic models show, the accuracy and fairness features of some system change over time, then there is no sense in which an algorithm just is or is not fair, or is or is not accurate. Fairness and accuracy claims that may hold at a given point in time do not hold necessarily for any later point in time. This is exemplified in our model of the labor market in Chapter 3 within which the fairness constraint that we propose does not immediately realize fair outcomes; it only generates group-equitable outcomes at steady-state in the long-term. Still, we show that the move to the group-fair regime may be desirable even for those who care only about the efficiency of the labor market or the welfare of employers. We prove that there exist market conditions under which the group-equitable equilibrium Pareto-dominates the group-inequitable ones arising from strategies that statistically discriminate or employ a “group-blind” criterion. Hence, the trade-off story in this setting is certainly more complicated but also more optimistic.

I tackle the relationship between fairness and welfare head on in Chapter 4’s “Fair Classification and Social Welfare,” a project in which Yiling Chen and I present a welfare-based analysis of fair classification regimes. We ask about the broad set of works that adopt formal parity-based definitions of fairness the following question: How do leading notions of fairness as defined by computer scientists map onto longer-standing notions of social welfare? Our main findings assess the welfare impact of fairness-constrained empirical risk minimization programs on the individuals and groups who are subject to their outputs. Our method of analysis, which maps changes in “fairness” space into changes in “welfare” space, assesses whether and which fair learning procedures result in classification outcomes that make groups better-off welfare-wise. Do gains in fairness always result in gains in the welfare of disadvantaged groups and losses in the welfare of advantaged group? In a surprising result, we show that applying stricter fairness criteria codified as parity constraints can worsen welfare outcomes for *both* groups. More generally, always preferring “more fair” classifiers does not abide by the Pareto Principle—a fundamental axiom of social choice theory and welfare economics. This makes for another complication in the standard ”trade-offs” narrative: improving along the axis of “fairness” may not necessarily translate into actually improved outcomes for any individuals or groups. This raises an important question: what exactly are gains in fairness *for*?

The works in this dissertation build upon while also challenging paradigms in the algorithmic fairness literature. Even in the short time that has passed since the field’s first significant wave of work, many scholars in the field have recognized the need to consider classifier behaviors as only one aspect of sociotechnical systems, which must as a whole be the target of inquiry into matters of fairness and discrimination. I take this to be a healthy sign of a growing consensus within the community about the limitations of an approach to fairness that considers only the computational features of some classifier or algorithm and an indication of the field's continued development. I thus close this dissertation with some reflections on the field's progression in the time I have been fortunate to contribute to it, and some suggestions for fruitful next steps that the community of scholars working on algorithmic fairness can take on in the coming years to both build on past strands of research as well as grow in new directions.

2

Disparate Effects of Strategic Manipulation

2.1 Introduction

The expanding realm of algorithmic decision-making has not only altered the ways that institutions conduct their day-to-day operations, but has also had a profound impact on how individuals interface with these institutions. It has changed the ways we communicate with each other, receive crucial resources, and are granted important social and economic opportunities. In theory, algorithms have great potential to reform existing systems to become both more efficient and equitable, but as exposed by various high-profile investigations[87,77,6,34], prediction-based models that make or assist with consequential decisions are, in practice, highly prone to reproducing past and current patterns of social inequality.

While few algorithmic systems are explicitly designed to be discriminatory, there are many underlying forces that drive socially biased outcomes. For one, since most of the features used in these models are based on proxy, rather than causal, variables, outputs often reflect the various structural factors that bear on a person's life opportunities rather than the individualized characteristics that decision-makers often seek. Much of the previous work in algorithmic fairness has examined a particular undesirable proxy effect in which a classifier's features may be linked to socially significant and legally protected attributes like race and gender, interpreting correlations that have arisen due to centuries of accumulated disadvantage as genuine attributes of a group of people marked as members of some social category.[51,80,60,43]

But algorithmic models do not only generate outcomes that passively correlate with social advantages or disadvantages. These tools also provoke a certain type of reactivity, in which agents see a classifier as a guide to action and actively change their behavior to accord with the algorithm's preferences. On this view, classifiers both *evaluate* and *animate* their subjects, transforming static data into strategic responses. Just as an algorithm's use of certain features differentially advantages some populations over others, the room for strategic response that is inherent in many automated systems also naturally favors social groups of privilege. Admissions procedures that heavily weight SAT scores motivate students who have the means to take advantage of test prep courses and even take the exam multiple times. Loan approval systems that rely on existing lines of credit as an indication of creditworthiness encourage those who can to apply for more credit in their name.

Thus an algorithm that scores applicants to determine how a resource should be allocated sets a standard for what an ideal candidate's features ought to be. A responsive subject would look to alter how she appears to a classifier in order to increase her likelihood of gaining the system's approval. But since reactivity typically requires informational and material resources that are not equally accessible to all. Thus, even when an algorithm draws on features that seem to arise out of individual effort, these metrics can be skewed to favor those who are more readily able to alter their features.

In the machine learning literature, agent reactivity to a classifier is termed "strategic manipulation." Since previous work in strategic classification has typically depicted interactions between the agent who selects the classifier, the so-called *learner*, and those candidates who are being classified as antagonistic, such actions are usually viewed as distortions that aim to undermine the published classifier.[15,44] As shown in Hardt et al.,[44] a learner who anticipates these responses can, under certain formulations of agent costs, adapt to protect against the misclassification errors that would have resulted from manipulation, recovering an accuracy level that is arbitrarily close to the theoretical maximum. These results are welcome news for a learner who correctly assesses agents' best-responses. Indeed in most strategic manipulation models, agents are depicted as equally able to pursue manipulation, allowing the learner who knows their costs to accurately preempt strategic responses. While there are occasions in which agents do largely face homogeneous costs—an even playing field, as it were—in many other social uses of machine learning tools, agents encounter differing costs of altering the attributes that are ultimately observed and assessed by the classifier. As such, in this chapter we ask, *"What are the effects of strategic classification and manipulation in a world of social stratification?"*

As in previous work in strategic classification, we cast the problem as a Stackelberg game in which the learner moves first and publishes her classifier before candidates best-respond and manipulate their features.[15,44,4,25] But in contrast with the models in previous work by Brückner and Scheffer[15] and Hardt et al.,[44] we formalize the setting of a society comprised of social groups that not only may differ in terms of distributions over unmanipulated features and true labeling functions but also face different costs to manipulation. This extra set of differences brings to light questions that favor an analysis that focuses on the welfares of the candidates who must contend with these classifiers: Do classifiers formulated with strategic behavior in mind impose disparate burdens on different groups? If so, how can a learner mitigate these adverse effects? The altered gameplay and outcomes of strategic classification raise questions of fairness that are intertwined with those of optimality.

Though our model is quite general, we obtain technical results that reveal important social ramifications of using classification in systems marked by inequality and a potential for manipulation. Our analysis shows that, under our model, even when the learner knows the costs faced by different groups, her equilibrium classifier will always act to reinforce existing inequalities by mistakenly excluding qualified candidates who are less able to manipulate their features while also mistakenly admitting those candidates for whom manipulation is less costly, perpetuating the relative advantage of group already advantaged in the social structure. We delve into the cost disparities that generate such inevitable classification errors.

Next, we consider the impact of providing subsidies to lighten the burden of manipulation for individuals who are in the disadvantaged group that faces higher costs. We find that such an intervention can improve the learner's classification performance as well as mitigate the extent to which her errors are inequality-reinforcing. However, we show that there exist cases in which providing subsidies enforces an equilibrium learner strategy that actually makes some individual candidates worse-off without making any better-off. Paradoxically, in these cases, paying a subsidy to the disadvantaged group actually benefits only the learner while both candidate groups experience a welfare decline! Further analysis of these scenarios reveals that, in many cases, all parties would have preferred a world in which manipulation of features was not possible for any candidates.

Our chapter's agent-centric analysis views data points as representing individuals and classifications as impacting those individuals' welfares. This orientation departs from the dominant perspective in learning theory, which privileges a vendor's predictive accuracy, and instead evaluates classification regimes in light of the social consequences of the outcomes they issue. By incorporating insights and techniques from game theory and economics, domains that consider deeply the effects of various policies on agents' behaviors and outcomes, we hope to broaden the perspective that machine learning takes on socially-oriented tools. Presenting more democratically-inclined analysis has been central to the field of algorithmic fairness, and we hope our work sheds new light on this generic setting of classification with strategic agents.

2.1.1 Related Work

While many earlier approaches to strategic classification in the machine learning literature have tended to view learner-agent interactions as adversarial,[55,9] our work does not assume inherently antagonistic relationships, and instead, shares the Stackelberg game-theoretic perspective akin to that presented in Brückner and Scheffer[15] and built upon by Hardt et al.[44] Departing from these models' focus on static prediction and homogeneous manipulation costs, Dong et al.[27] propose an online setting of strategic classification in which agents appear sequentially and have individual costs for manipulation that are unknown to the learner. Unlike our work, they take a traditional learner-centric view, whereas our concerns are with the welfare of the candidates.

Agent features and potential manipulations in the face of a learner classifier can also be interpreted as serving *informational* purposes. In the economics literature on signaling theory, agents interact with a principal—the counterpart to our learner—via signals that convey important information relevant to a particular task at hand. Classic works, such as Spence's paper on job-market signaling, focus their analysis on the varying quality of information that signals provide at equilibrium.[85] The emphasis in our analysis on different group costs shares features with a recent update to the signaling literature by Frankel and Kartik,[39] who also distinguish between natural actions, corresponding to unmanipulated features in our model, and "gaming" ability, which operate similarly to our cost functions. The connection between gaming capacity and social advantage is also explicitly discussed in work by Esteban and Ray[33] who consider the effects of wealth and lobbying on governmental resource allocation. While most works in the economics signaling literature center on the decay of the informativeness of signals as gaming and natural actions become indistinguishable, some recent work in computer science has also considered the effect of costly signaling on mechanism design.[58,59] In contrast to both of these perspectives, our work highlights the effect of manipulation on a learner's action and as a consequence, on the agents' welfares.

In independent, concurrent work by Milli et al.[73] also consider the social impacts of strategic classification. Whereas our model highlights the interplay between a learner's Stackelberg equilibrium classifier and agents' best-response manipulations at the feature level, their work traces the relationship between the learner's utility and the social burden, a measure of agents' manipulation costs. They show that an institution must select a point on the outcome curve that trades off its predictive accuracy with the social burden it imposes. In their model, an agent with an unmanipulated feature vector $\mathbf{x}$ has a likelihood $\ell(\mathbf{x})$ of having a positive label and can manipulate to change her original feature vector to any vector $\mathbf{y}$ with $\ell(\mathbf{y}) \leqslant \ell(\mathbf{x})$ at zero cost, or to $\mathbf{y}$ with $\ell(\mathbf{y}) > \ell(\mathbf{x})$ for a positive cost. This assumption, which the authors call "outcome monotonicity," allows them to reason about manipulations in (one-dimensional) likelihood space rather than feature space, since the optimal learner strategies amount to thresholds on likelihoods. In contrast, we allow features to be differently manipulable (perhaps a student can boost her SAT score via test prep courses, but can do nothing to change her grades from the previous year, and cannot freely obtain a higher SAT score in exchange for a worse record of extracurricular activities), which affects the forms of both the learner's equilibrium classifier and agents' best-response manipulations. Despite these differences in model and focus, their analysis yields results that are qualitatively similar to ours. Highlighting the differential impact of classifiers on social groups, they also find that overcoming stringent thresholds is more burdensome on the disadvantaged group.

2.2 Model Formalization

As in Brückner and Scheffer[15] and Hardt et al.,[44] we formalize the Strategic Classification Game as a Stackelberg competition in which the learner moves first by committing to and publishing a binary classifier f. Candidates, who are endowed with "innate" features, best respond by manipulating their feature inputs into the classifier. Formally, a candidate is defined by her d-dimensional feature vector $\mathbf{x} \in X = [0,1]^d$ and group membership A or B, with A signifying the advantaged group and B the disadvantaged. Group membership bears on manipulation costs such that a candidate from group m who wishes to move from a feature vector $\mathbf{x}$ to a feature vector $\mathbf{y}$ must pay a cost of $c_m(\mathbf{y}) - c_m(\mathbf{x})$. We note that these cost function forms are similar to the class of separable cost functions considered in Hardt et al.[44] We assume that higher feature values indicate higher quality to the learner, and thus restrict our attention to manipulations such that $\mathbf{y} \geqslant \mathbf{x}$, where the symbol $\geqslant$ signifies a component-wise comparison such that $\mathbf{y} \geqslant \mathbf{x}$ if and only if $\forall i \in [d], y_i \geqslant x_i$. Throughout this chapter, we study non-negative monotone cost functions such that the cost of manipulating from a feature vector $\mathbf{x}$ to a feature vector $\mathbf{y}$ increases as $\mathbf{x}$ and $\mathbf{y}$ get further apart.

To motivate this distinction between features and costs, consider the use of SAT scores as a signal of academic preparedness in the U.S. college admissions process. The high-stakes nature of the SAT has encouraged the growth of a test prep industry dedicated to helping students perform better on the exam. Test preparation books and courses, while also exposing students to content knowledge and skills that are covered on the SAT, promise to "hack" the exam by training students to internalize test-taking strategies based on the format, structure, and style of its questions. One can view SAT scores as a feature used by a learner building a classifier to select candidates with sufficient academic success according to some chosen standard. The existence of test prep resources then presents an opportunity for some applicants to inflate their scores, which might "trick" the tool into classifying the candidates as more highly qualified than they are in fact. In this example, a candidate's strategic manipulation move refers to her investment in these resources, which despite improving her exam score, do not confer any genuine benefits to her level of academic preparation for college.

Just as access to test prep resources tends to fall along income and race lines, we view candidates' different abilities to manipulate as tied to their group membership. We model these group differences with respect to availability of resources and opportunity by enforcing a *cost condition* that orders the two groups. We suppose that for all $\mathbf{x} \in [0,1]^d$ and $\mathbf{y} \geqslant \mathbf{x}$,

$$c_A(\mathbf{y}) - c_A(\mathbf{x}) \leqslant c_B(\mathbf{y}) - c_B(\mathbf{x}). \tag{2.1}$$

Manipulating from a feature vector $\mathbf{x}$ to $\mathbf{y}$ is always at least as costly for a member of group B as it is for a member of group A. We believe our model's inclusion of this cost condition reflects an authentic aspect of our social world wherein one group is systematically disadvantaged with respect to a task in comparison to another.

In our setup, we also allow groups to have distinct probability distributions $\mathcal{D}_A$ and $\mathcal{D}_B$ over unmanipulated features and to be subject to different true labeling functions h_A and h_B defined as

$$h_A(\mathbf{x}) = \begin{cases} 1, & \forall \mathbf{x} \text{ such that } \sum_{i=1}^d w_{A,i}x_i \geqslant \tau_A, \\ 0, & \forall \mathbf{x} \text{ such that } \sum_{i=1}^d w_{A,i}x_i < \tau_A, \end{cases} \tag{2.2}$$

$$h_B(\mathbf{x}) = \begin{cases} 1, & \forall \mathbf{x} \text{ such that } \sum_{i=1}^d w_{B,i}x_i \geqslant \tau_B, \\ 0, & \forall \mathbf{x} \text{ such that } \sum_{i=1}^d w_{B,i}x_i < \tau_B. \end{cases} \tag{2.3}$$

We assume that $h_A(\mathbf{x}) = 1 \implies h_B(\mathbf{x}) = 1$ for all $\mathbf{x} \in [0,1]$. Returning to the SAT example, research has shown that scores are skewed by race even before factoring in additional considerations such as access to manipulation.[18] In such cases, the true threshold for the disadvantaged group is lower than that for the advantaged group. We leave this generality in our model to acknowledge and account for the influence that various social and historical factors have on candidates' unmanipulated features and not, we emphasize, as an endorsement of a view that groups are fundamentally different in ability. A formal description of the Strategic Classification Game with Groups is given in the following definition.

Definition 1 (Strategic Classification Game with Groups). *In the Strategic Classification Game with Groups, candidates with features* $\mathbf{x} \in [0,1]^d$ *and group memberships A or B are drawn from distributions* $\mathcal{D}_A$ *and* $\mathcal{D}_B$*. The population proportion of each group is given by* p_A *and* p_B *where* $p_A + p_B = 1$*. A candidate from group m pays cost* $c_m(\mathbf{y}) - c_m(\mathbf{x})$ *to move from her original features* $\mathbf{x}$ *to* $\mathbf{y} \geqslant \mathbf{x}$*. There exist true binary classifiers* h_A *and* h_B*, for candidates of each group. Probability distributions, cost functions, and true binary classifiers are all common knowledge. Gameplay proceeds in the following manner:*

1. *The learner issues a classifier f generating outcomes* $\{0,1\}$*.*
2. *Each candidate observes f and manipulates her features* $\mathbf{x}$ *to* $\mathbf{y} \geqslant \mathbf{x}$*.*

A group m candidate with features $\mathbf{x}$ *who moves to* $\mathbf{y}$ *earns a payoff*

$$f(\mathbf{y}) - (c_m(\mathbf{y}) - c_m(\mathbf{x})).$$

The learner incurs a penalty of

$$C_{FP} \sum_{m \in \{A,B\}} p_m P_{\mathbf{x} \sim \mathcal{D}_m}[h_m(\mathbf{x}) = 0, f(\mathbf{y}) = 1] + C_{FN} \sum_{m \in \{A,B\}} p_m P_{\mathbf{x} \sim \mathcal{D}_m}[h_m(\mathbf{x}) = 1, f(\mathbf{y}) = 0],$$

where C_{FP} *and* C_{FN} *denote the cost of a false positive and a false negative respectively.*

The learner looks to correctly classify candidates with respect to their original features $\mathbf{x}$, whereas each candidate hopes to manipulate her features to attain a positive classification, expending as little cost as possible in the process. Under this setup, candidates are only willing to manipulate their features if it flips their classification from 0 to 1 and if the cost of the manipulation is less than 1. We note that defining the utility of a positive classification to be 1 can be considered a scaling and thus is without loss of generality.

This learner-candidate interaction is very similar to that studied in Hardt et al.[44] However, our inclusion of groups with distinct manipulation costs leads to an ambiguity regarding a candidate's initial features that does not exist when all candidates have an equal opportunity to manipulate. In very few cases can a vendor distinguish among candidates based on their group membership for the explicit purpose of issuing distinct classification policies, especially if that group category is a protected class attribute. As such, in our setup, we require that a learner publish a classifier that is not adaptive to different agents based on their group memberships.

It is important to note that the positive results in Hardt et al.'s[44] formulation of the Strategic Classification Game, wherein for separable cost functions, the learner can attain a classification error at test-time that is arbitrarily close to the optimal payoff attainable, do not carry over into this setting of heterogeneous groups and costs. Even when $h_A = h_B$, the existence of different costs of agent manipulation, even when separable as in our model, introduces a base uncertainty to the learning problem that generates errors that cannot be extricated so long as the learner must publish a classifier that does not distinguish candidates based on their group memberships. Second, an analysis of the learner's strategy and performance, the perspective typically taken in most learning theory papers, contributes only a partial view of the total welfare effect of using classification in strategic settings. The main objective of this chapter is to offer a more thorough and holistic inspection of all agents' outcomes, paying special heed to the different outcomes experienced by candidates of the two groups. Insofar as all social behaviors are impelled by goals, interests, and purposes, we should view data that is strategically generated to be the rule rather than the exception in social machine learning settings.

Remark on the assumption that h_A and h_B are known.

Our assumption that the learner has knowledge of groups' true labeling functions is not central to our analysis. We make such an assumption to highlight the pure effect of groups' differential costs of manipulation on equilibrium gameplay and consequent welfares rather than the potential side effects due to a learner's noisy estimation of the true classifiers. Our general findings do not substantially rely on this feature of the model, and the overall results carry through into a setting in which the learner optimizes from samples.

Remark on unequal group costs

The differences in costs c_A and c_B encoded by the cost condition is not restricted to referring only to differences in the monetary cost of manipulation. Instead, as is common in information economics and especially signaling theory, "cost" reflects the multiplicity of factors that bear on the effort exertion required by feature manipulation.[85,86,66,10] To demonstrate the generality of our formulation of distinct group costs, we show that the cost condition given in (2.1) is equivalent to a more explicit derivation of the choice that an agent faces when deciding whether to manipulate her feature.

A rational agent with feature $\mathbf{x}$ will only pursue manipulation if her value for a positive classification minus her cost of manipulation exceeds her value for a negative classification:

$$v(f(\mathbf{x}) = 0) \leqslant v(f(\mathbf{y}) = 1) - u(c(\mathbf{y}) - c(\mathbf{x})). \tag{2.4}$$

The monotone function u translates the costs borne by a candidate to manipulate from $\mathbf{x}$ to $\mathbf{y}$ into her "utility space," i.e., it reflects the value that she places on that expenditure. We can rewrite the previous inequality to be

$$c(\mathbf{y}) - c(\mathbf{x}) \leqslant u^{-1}\big(v(f(\mathbf{y}) = 1) - v(f(\mathbf{x}) = 0)\big). \tag{2.5}$$

Substituting in $k = u^{-1}\big(v(f(\mathbf{y}) = 1) - v(f(\mathbf{x}) = 0)\big)$, we have $c(\mathbf{y}) - c(\mathbf{x}) \leqslant k$. Since the same cost expenditure is valued more highly by the disadvantaged group than by the advantaged group, the function u is more convex for group B than for group A. Thus all else equal, we have $c_A(\mathbf{y}) - c_A(\mathbf{x}) \leqslant c_B(\mathbf{y}) - c_B(\mathbf{x})$ as desired. More generally, the functions v, c, and u may each be different for the groups. As such, the disadvantage encoded in the cost condition can arise due to differences in valuations of classifications (v), differences in costs (c), or differences in valuations of those costs (u).

2.3 Equilibrium Analysis

We begin by studying agents' best-response strategies in the basic Strategic Manipulation Game with Groups in which candidates belong to one of two groups A and B, and the cost condition holds so that group B members face greater costs to manipulation than group A members. To build intuition, we first consider best-response strategies in the one-dimensional case in which candidates have features $x \in [0, 1]$ and group cost functions are of any non-negative monotone form. We then move on to consider the d-dimensional case in which candidate features are given as vectors $\mathbf{x} \in [0, 1]^d$ and manipulation costs are assumed to be linear.

2.3.1 One-dimensional Features

In the $d = 1$ case, the cost condition given in (2.1) may be written as $c'_A(x) \leqslant c'_B(x)$ for all $x \in [0, 1]$. Since the true decision boundaries are linear, in the one-dimensional case, they may be written as threshold functions where thresholds τ_A and τ_B are constants in $[0, 1]$ and for agents in group m,

$h_m(x) = 1$ if and only if $x \geqslant \tau_m$. A university admissions decision based on a single score is an example of such a classifier. Although the SAT does not act as the sole determinant of admissions in the U.S., in countries such as Australia, Brazil, and China, a single exam score is often the only factor of applicant quality that is considered for admissions.

When the learner has access to τ_A and τ_B, and group costs c_A and c_B satisfy the cost condition, the following proposition characterizes the space of undominated strategies for the learner who seeks to minimize any error-penalizing cost function.

Proposition 1 (One-D Undominated Learner Strategies). *Given group cost functions c_A and c_B and true label thresholds τ_A and τ_B where $\tau_B \leqslant \tau_A$, there exists a space of undominated learner threshold strategies $[\sigma_B, \sigma_A] \subset [0, 1]$ where $\sigma_A = c_A^{-1}(c_A(\tau_A) + 1)$ and $\sigma_B = c_B^{-1}(c_B(\tau_B) + 1)$. That is, for any error penalties C_{FP} and C_{FN}, the learner's equilibrium classifier f is based on a threshold $\sigma \in [\sigma_B, \sigma_A]$ such that for all manipulated features y,*

$$f(y) = \begin{cases} 1, & \forall y \geqslant \sigma, \\ 0, & \forall y < \sigma. \end{cases} \tag{2.6}$$

To understand this result, first notice that if the learner were to face only those candidates from group A, she would achieve perfect classification by labeling as 1 only those candidates with unmanipulated feature $x \geqslant \tau_A$. This strategy is enacted by considering candidates' best-response manipulations. A rational candidate would only be willing to manipulate her feature if the gain she receives in her classification exceeds her costs of manipulation. The learner would like to guard against manipulations by candidates with $x < \tau_A$ but still admit candidates with $x \geqslant \tau_A$, so she considers the maximum manipulated feature y that is attainable by a rational candidate with $x = \tau_A$ who is willing to spend up to a cost of one in order to secure a better classification, as illustrated in Figure 2.1. The maximum such y value is σ_A, and thus, the learner sets a threshold at σ_A, admitting all those with $y \geqslant \sigma_A$ and rejecting all those with $y < \sigma_A$. The same reasoning applies to a learner facing only group B candidates, and the learner sets a threshold at σ_B, admitting all those candidates with $y \geqslant \sigma_B$ and rejecting all those with $y < \sigma_B$.

It can be shown that for all valid values of τ_A, τ_B, c_A, and c_B, necessarily $\sigma_B \leqslant \sigma_A$. Then all classifiers with threshold $\sigma < \sigma_B$ are dominated by σ_B, in the sense that for any arbitrary error penalties C_{FP} and C_{FN}, the learner would suffer higher costs by setting her threshold to be σ rather than σ_B. In the same way, all thresholds $\sigma > \sigma_A$ are dominated by σ_A, thus leaving $[\sigma_B, \sigma_A]$ to be the space of undominated thresholds. For an account of the full proof of this result (and all omitted proofs), see the appendix.

Even without committing to a particular learner cost function, the space of optimal strategies characterized in Proposition 1 leads to an important consequence. A rational learner in the Strategic Classification Game always selects a classifier that exhibits the following phenomenon: it mistakenly admits unqualified candidates from the group with lower costs and mistakenly excludes qualified candidates from the group with higher costs. This result is formalized in Proposition 2.

To state the proposition, the following definition is instructive. Whereas the true thresholds τ_A and τ_B are a function of unmanipulated features, the learner only faces candidate features that may have been manipulated. In order to make these observed features commensurable with τ_A and τ_B, it is helpful for the learner to "translate" a candidate's possibly manipulated feature y to its minimum corresponding original unmanipulated value.

Definition 2 (Correspondence with unmanipulated features). *For any observed candidate feature $y \in [0, 1]$, the minimum corresponding unmanipulated feature is defined as*

$$\begin{aligned} \ell_A(y) &= \max\{0, c_A^{-1}(c_A(y) - 1)\}, \\ \ell_B(y) &= \max\{0, c_B^{-1}(c_B(y) - 1)\} \end{aligned} \tag{2.7}$$

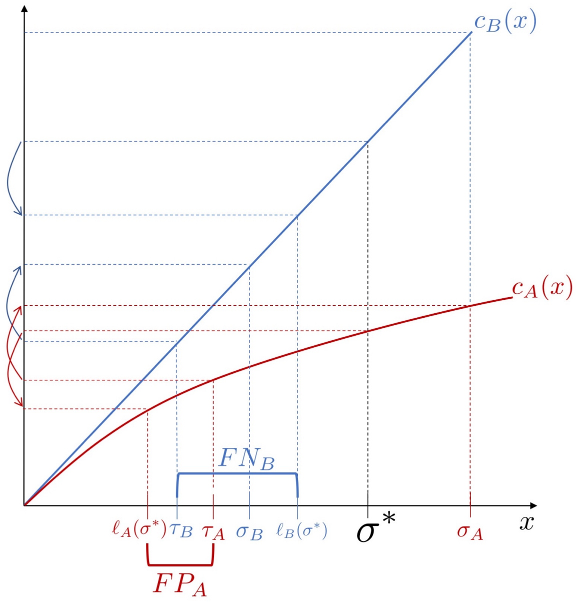


Figure 2.1: Group cost functions for a one-dimensional feature x. τ_A and τ_B signify true thresholds on unmanipulated features for group A and B, but a learner must issue a classifier on manipulated features. The threshold σ_A perfectly classifies group A candidates; σ_B perfectly classifies group B candidates. A learner selects an equilibrium threshold $\sigma^* \in [\sigma_B, \sigma_A]$, committing false positives on group A (red bracket) and false negatives on group B (blue bracket).

for a candidate belonging to group A and group B respectively.

The corresponding values $\ell_A(y)$ and $\ell_B(y)$ are defined such that a candidate who presents feature y must have as her true unmanipulated feature $x \geqslant \ell_A(y)$ if she is a group A member and $x \geqslant \ell_B(y)$ if she is a group B member.

Proposition 2 (Learner's Cost in 1 Dimension). *A learner who employs a classifier f based on a threshold strategy $\sigma \in [\sigma_B, \sigma_A]$ only commits false positives errors on group A and false negatives errors on group B. The cost $C(\sigma)$ of such a classifier is*

$$C_{FN} p_B P_{x\sim\mathcal{D}_B}\big[x \in [\tau_B, \ell_B(\sigma))\big] + C_{FP} p_A P_{x\sim\mathcal{D}_A}\big[x \in [\ell_A(\sigma), \tau_A)\big],$$

where false negative errors entail penalty C_{FN}, and false positive errors entail penalty C_{FP}.

A learner who commits to classifying only one of the groups correctly bears costs given by the following corollaries.

Corollary 1. *A classifier based on σ_A perfectly classifies group A candidates and bears cost* $C(\sigma_A) = C_{FN}p_B P_{x\sim\mathcal{D}_B}\big[x \in [\tau_B, \ell_B(\sigma))\big]$.

Corollary 2. *A classifier based on σ_B perfectly classifies group B candidates and bears cost* $C(\sigma_B) = C_{FP}p_A P_{x\sim\mathcal{D}_A}\big[x \in [\ell_A(\sigma), \tau_A)\big]$.

Notice that the learner's errors always cut in the same direction—by unduly benefiting group A candidates and unduly rejecting group B candidates, these errors act to reinforce the existing social inequality that had generated the unequal group cost conditions in the first place. Since these errors arise out of the asymmetric group costs of manipulation, the Strategic Classification Game can be viewed as an interactive model that itself perpetuates the relative advantage of group A over group B candidates.

Within the undominated region $[\sigma_B, \sigma_A]$, the equilibrium learner threshold σ^* is attained as the solution to the optimization problem

$$\sigma^* = \underset{\sigma\in[\sigma_B,\sigma_A]}{\arg\min}\, C(\sigma). \tag{2.8}$$

In the game's greatest generality where candidates are drawn from arbitrary probability distributions, groups bear any costs that abide by the cost condition, and the learner has arbitrary error penalties, the equilibrium learner threshold σ^* cannot be specified any further. However, under some special cases of candidate cost functions and probability distributions, the equilibrium threshold can be characterized more precisely. Specifically, when candidates from both groups are assumed to be drawn from a uniform distribution over unmanipulated features in $[0, 1]$, an error-minimizing learner seeks a threshold value σ^* that minimizes the length of the interval of errors, given by the following quantity:

$$\sigma^* = \underset{\sigma \in [\sigma_B, \sigma_A]}{\arg\min}\, \ell_B(\sigma) - \ell_A(\sigma).$$

From here, one natural assumption of candidate cost functions would have that groups A and B bear costs that are proportional to each other. In this case, the curvature of the cost functions is determinative of a learner's equilibrium threshold.

Proposition 3. *Suppose group cost functions are proportional such that $c_A(x) = qc_B(x)$ for $q \in (0,1)$, that $\mathcal{D}_A$ and $\mathcal{D}_B$ are uniform on $[0,1]$, and that $C_{FN} = C_{FP}$ and $p_A = p_B = \frac{1}{2}$. Let σ^* be the learner's equilibrium threshold.*

When cost functions are strictly concave, $\sigma^* = \sigma_B$. When cost functions are strictly convex, $\sigma^* = \sigma_A$. When cost functions are affine, the learner is indifferent between all $\sigma^* \in [\sigma_B, \sigma_A]$.

2.3.2 General d-Dimensional Feature Vectors

In the general d-dimensional case of the Strategic Classification Game, candidates are endowed with features that are given by a vector $\mathbf{x} \in [0,1]^d$ and can choose to manipulate and present any feature $\mathbf{y} \geqslant \mathbf{x}$ to the learner. In this section, we consider optimal learner and candidate strategies when group costs are linear such that they may be written as

$$c_A(\mathbf{x}) = \sum_{i=1}^{d} c_{A,i}x_i; \quad c_B(\mathbf{x}) = \sum_{i=1}^{d} c_{B,i}x_i \tag{2.9}$$

for groups A and B respectively. Now, the cost condition $c_A(\mathbf{y}) - c_A(\mathbf{x}) \leqslant c_B(\mathbf{y}) - c_B(\mathbf{x})$ for all $\mathbf{y} \geqslant \mathbf{x}$—defined component-wise as before—implies that $\forall i \in [d]$, $c_{A,i} \leqslant c_{B,i}$. In d dimensions, the true classifiers h_A and h_B have linear decision boundaries such that for a group A candidate with feature **x**,

$$h_A(\mathbf{x}) = \begin{cases} 1 & \sum_{i=1}^{d} w_{A,i} x_i \geqslant \tau_A, \\ 0 & \sum_{i=1}^{d} w_{A,i} x_i < \tau_A, \end{cases} \tag{2.10}$$

and for a group B candidate with feature **x**,

$$h_B(\mathbf{x}) = \begin{cases} 1 & \sum_{i=1}^{d} w_{B,i} x_i \geqslant \tau_B, \\ 0 & \sum_{i=1}^{d} w_{B,i} x_i < \tau_B. \end{cases} \tag{2.11}$$

We assume that all components x_i contribute positively to an agent's likelihood of being classified as 1 so that $w_{A,i}, w_{B,i} \geqslant 0$ for all i. To ensure that the cost of manipulation is always non-negative, all cost coefficients are positive: $c_{B,i}, c_{A,i} \geqslant 0$ for all $i \in [d]$.

A candidate may now manipulate any combination of the d components of her initial feature **x** to reach the final feature **y** that she presents to the learner. Despite this increased flexibility on the part of the candidate, we are still able to characterize the performance of undominated learner classifiers, generalizing the result in Proposition 2. All potentially optimal classifiers exhibit the same inequality-reinforcing property inherent within the one-dimensional interval of undominated threshold strategies, trading off false positives on group A candidates with false negatives on group B candidates. Before we formally present this result, we first describe candidates' best-response strategies. Here, a geometric view of the space of potential manipulations is informative.

Suppose a candidate endowed with a feature vector **x** faces costs $\sum_{i=1}^{d} c_i x_i$ and is willing to expend a total cost of 1 for manipulation. Then she can move to any $\mathbf{y} \geqslant \mathbf{x}$ contained within the d-simplex with orthogonal corner at **x** and remaining vertices at $\mathbf{x} + \frac{1}{c_i}\mathbf{e}_i$ where $\mathbf{e}_i$ is the ith standard basis vector. This region is given by

$$\Delta(\mathbf{x}) = \left\{ \mathbf{x} + \sum_{i=1}^{d} \frac{t_i}{c_i} \mathbf{e}_i \in [0,1]^d \middle| \sum_{i=1}^{d} t_i \leqslant 1\,; t_i \geqslant 0 \; \forall i \right\}. \tag{2.12}$$

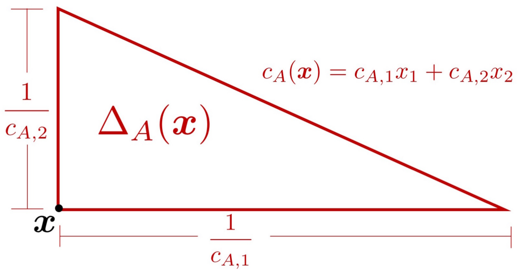

$c_A(\boldsymbol{x}) = c_{A,1}x_1 + c_{A,2}x_2$
$\frac{1}{c_{A,2}}$
$\Delta_A(\boldsymbol{x})$
$\boldsymbol{x}$
$\frac{1}{c_{A,1}}$

Figure 2.2: The forward simplex. A candidate in group A with unmanipulated feature vector $\mathbf{x}$ can manipulate to reach any feature vector $\mathbf{y} \in \Delta_A(\mathbf{x})$ at a cost of at most 1.

$\Delta(\mathbf{x})$, depicted in Figure 2.2, gives the space of potential movement for a candidate with unmanipulated feature $\mathbf{x}$ who is willing to expend a total cost of 1. Notice that t_i can be interpreted as the cost that a candidate expends on movement in the ith direction. Thus $\sum_{i=1}^d t_i$ gives the total cost of manipulation. Moving beyond the range of possible moves, in order to describe how a rational candidate will best-respond to a learner, we must consider the published classifier.

Suppose a learner publishes a classifier f based on a hyperplane $\sum_{i=1}^d g_i y_i = g_0$, so that $f(\mathbf{y}) = 1$ if and only if $\sum_{i=1}^d g_i y_i \geqslant g_0$. A best-response manipulation occurs along the direction that generates the greatest increase in the value $\sum_{i=1}^d g_i(y_i - x_i)$ for the least cost. As such, a candidate will move in any directions $i \in \arg\max_{i\in[d]} \frac{g_i}{c_i}$. This result is formalized in the following lemma.

Lemma 1 (d-D Candidate Best Response). *Suppose a learner publishes the classifier $f(\mathbf{y}) = 1$ if and only if $\sum_{i=1}^d g_i y_i \geqslant g_0$. Consider a candidate with unmanipulated feature vector $\mathbf{x}$ and linear costs $\sum_{i=1}^d c_i x_i$. If $f(\mathbf{x}) = 1$ or if for all $i \in [d], f(\mathbf{x} + \frac{1}{c_i}\mathbf{e}_i) = 0$, the candidate's best response is to set $\mathbf{y} = \mathbf{x}$. Otherwise, letting $K = \arg\max_{i\in[d]} \frac{g_i}{c_i}$, her manipulation takes the form*

$$y = \mathbf{x} + \sum_{i=1}^d \frac{t_i}{c_i}\mathbf{e}_i$$

for any $\mathbf{t}$ such that $t_i \geqslant 0$ for all $i \in [d]$, $t_i = 0$ for all $i \notin K$, and $\sum_{i=1}^d g_i(x_i + \frac{t_i}{c_i}) = g_0$.

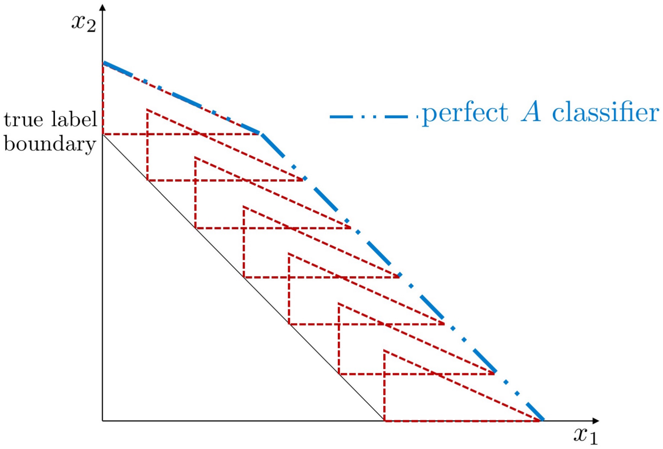


Figure 2.3: A perfect classifier for group A. Every candidate with unmanipulated feature vector $\mathbf{x}$ on or above the true decision boundary for group A is able to manipulate to a point $\mathbf{y} \in \Delta_A(\mathbf{x})$ on or above the blue decision boundary depicted here. No candidate with an unmanipulated feature vector below the true decision boundary is able to do so. The kink in the blue decision boundary arises due to the restriction of features to $[0, 1]^d$. A perfect classifier for group A does not need to have this kink; for example, a more lenient perfect classifier can be formed by "straightening" it out.

While in the d-dimensional case, a candidate has many more choices of manipulation directions to pursue, a best response strategy will always lead her to increase her feature in those components that are most valued by the learner and least costly for manipulation. That is, she behaves according to a "bang for your buck" principle, in which the optimal manipulations are in the direction or directions where the ratio $\frac{g_i}{c_i}$ is highest.

Despite the fact that the optimal manipulation may not be unique, as in the cases where there are multiple equivalently good directions for a candidate to move in, a learner who knows candidates' costs can still anticipate best-response manipulations and avoid errors on that group. As such, we are once again able to construct a perfect classifier for candidates of group A and a perfect classifier for candidates of group B.

Theorem 1 (d-D Space of Dominant Learner Strategies). *In the general d-dimensional Strategic Classification Game with linear costs, there exists a classifier that perfectly classifies group A and a classifier that perfectly classifies group B. All undominated classifiers commit no false positive errors on* *group A and no false negative errors on group B.*

A full exposition of the proof appears in the appendix, but here we present an abbreviated explanation of the result.

For each group m, the learner computes an optimal boundary that perfectly classifies all of its members by considering the set of simplices $\{\Delta_m(\mathbf{x})\}$ anchored at the vectors $\bar{\mathbf{x}}$ that satisfy $\mathbf{w}_m^\intercal \bar{\mathbf{x}} = \tau_m$ and drawing the strictest hyperplane that intersects each simplex. That is for all hyperplanes $g_i : \sum_{j=1}^d g_{i,j} x_j = g_{i,0}$ that are constructed to intersect each simplex, then $g_1 : \sum_{j=1}^d g_{1,j} x_j = g_{1,0}$ is the strictest if for all $\mathbf{x} \in [0,1]^d$,

$$\sum_{j=1}^d g_{1,j} x_j = g_{1,0} \implies \sum_{j=1}^d g_{i,j} x_j = g_{i,0} \geqslant g_{j,0}$$

for all g_i. Due to the cost ordering, for any $\mathbf{x} \in [0,1]^d$, $\Delta_B(\mathbf{x}) \subseteq \Delta_A(\mathbf{x})$, and thus wherever a comparison is possible, the group A boundary is at least as strict as the group B boundary. Figure 2.3 gives a visualization of a boundary formed by connecting the simplices $\Delta(\bar{\mathbf{x}})$; the corresponding classifier perfectly classifies the group.

As in the one-dimensional general costs case, learner strategies necessarily entail inequality-reinforcing classifiers: a rational learner equipped with any error-penalizing cost function will select an equilibrium strategy that trades off undue optimism with respect to group A for undue pessimism with respect to group B. We note that except in the extreme case in which there exists a perfect classifier for all candidates in the population, this result implies that the classifier for group A issues false negatives on group B, and the classifier for group B issues false positives on group A. In order to formalize this result, we would like to generalize the idea behind the minimum correspondence unmanipulated features given by $\ell_A(\cdot)$ and $\ell_B(\cdot)$ in (2.7) for general d-dimensions and linear costs.

A learner who observes a possibly manipulated feature vector $\mathbf{y}$ must consider the space of unmanipulated feature vectors that the candidate could have had. Thus we can make use of the simplex idea of potential manipulation; however in this case, the learner seeks to project a simplex "backward" to "undo" the potential candidate manipulation. Since groups are subject to different costs, simplices $\Delta_A^{-1}(\mathbf{y})$ and $\Delta_B^{-1}(\mathbf{y})$—a depiction is given in Figure 2.4—which represent the region from where a candidate could have manipulated, will differ based on the candidate's group membership, with

$$\Delta_A^{-1}(\mathbf{y}) = \Big\{\mathbf{y} - \sum_{i=1}^{d} \frac{t_i}{c_{A,i}}\mathbf{e}_i \in [0,1]^d \Big| \sum_{i=1}^{d} t_i \leqslant 1\,; t_i \geqslant 0\;\forall i\Big\}, \tag{2.13}$$

$$\Delta_B^{-1}(\mathbf{y}) = \Big\{\mathbf{y} - \sum_{i=1}^{d} \frac{t_i}{c_{B,i}}\mathbf{e}_i \in [0,1]^d \Big| \sum_{i=1}^{d} t_i \leqslant 1\,; t_i \geqslant 0\;\forall i\Big\}. \tag{2.14}$$

We can now use these constructs in order to define d-dimensional generalizations of $\ell_A(\mathbf{y})$ and $\ell_B(\mathbf{y})$.

Definition 3 (Correspondence with Unmanipulated Features in d-D). *For any observed candidate feature* $\mathbf{y} \in [0,1]^d$*, the minimum corresponding unmanipulated feature vectors are given by*

$$\ell_A(\mathbf{y}) = \big\{\mathbf{x} \in \Delta_A^{-1}(\mathbf{y}) \cap [0,1]^d | \nexists \hat{\mathbf{x}} \in \Delta_A^{-1}(\mathbf{y}) \textit{ such that } \hat{\mathbf{x}} < \mathbf{x}\big\}, \tag{2.15}$$

$$\ell_B(\mathbf{y}) = \big\{\mathbf{x} \in \Delta_B^{-1}(\mathbf{y}) \cap [0,1]^d | \nexists \hat{\mathbf{x}} \in \Delta_B^{-1}(\mathbf{y}) \textit{ such that } \hat{\mathbf{x}} < \mathbf{x}\big\} \tag{2.16}$$

for a candidate belonging to group A and group B respectively.

The corresponding values $\ell_A(\mathbf{y})$ and $\ell_B(\mathbf{y})$ are defined such that a candidate who presents feature $\mathbf{y}$ must have had a true unmanipulated feature vector $\mathbf{x} \geqslant \bar{\mathbf{x}}$ for some $\bar{\mathbf{x}} \in \ell_A(\mathbf{y})$ if she is a group A member and $\mathbf{x} \geqslant \bar{\mathbf{x}}$ for some $\bar{\mathbf{x}} \in \ell_B(\mathbf{y})$ if she is a group B member.

For any hyperplane decision boundary g containing vectors $\mathbf{y}$, the minimum corresponding feature vectors given by $\ell_A(\mathbf{y})$ and $\ell_B(\mathbf{y})$ are helpful for determining the effective thresholds that g

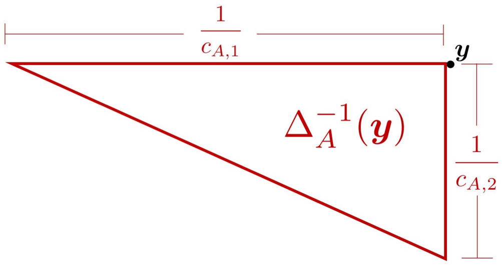


Figure 2.4: The backward simplex. A candidate in group A with manipulated feature vector $\mathbf{y}$ could have started with any feature vector $\mathbf{x} \in \Delta_A^{-1}(\mathbf{y})$ and paid a cost of at most 1.

generates on unmanipulated features for groups A and B.

Lemma 2. *Suppose a learner classifier f is based on a hyperplane $g : \sum_{i=1}^{d} g_i x_i = g_0$. Construct the set*

$$\mathcal{L}_m(g) = \left\{ \underset{\mathbf{x}\in \ell_m(\mathbf{y})}{\arg\min} \sum_{i=1}^{d} g_i x_i \middle| \forall \mathbf{y} \; s.\ t.\ \sum_{i=1}^{d} g_i y_i = g_0 \right\} \tag{2.17}$$

Then a group m agent with feature $\mathbf{x}$ can move to some $\mathbf{y}$ with $f(\mathbf{y}) = 1$ and $c_m(\mathbf{y}) - c_m(\mathbf{x}) \leqslant 1$ if and only if $\mathbf{x} \geqslant \ell$ for some $\ell \in \mathcal{L}_m(g)$.

By definition, for any two $\ell_1, \ell_2 \in \mathcal{L}_m(g)$,

$$\sum_{i=1} g_i \ell_{1,i} = \sum_{i=1} g_i \ell_{2,i} = g_0 - \frac{g_{k_m}}{c_{m,k_m}},$$

where $k_m \in \arg\max_{i=[d]} \frac{g_i}{c_{m,i}}$. Thus a learner who cares only about the true label of presented features, will construct her decision boundary g such that all $\ell \in \mathcal{L}_m(g)$ have the same true label.

A cost-minimizing learner who publishes a classifier f based on a hyperplane g on manipulated features will commit errors on those candidates with unmanipulated features $\mathbf{x} \in [0,1]^d$ contained within the boundaries given by $\mathcal{L}_A(g)$ and $\mathcal{L}_B(g)$. This space can be understood as the d-dimensional generalization of the $[\ell_A(\sigma), \ell_B(\sigma)]$ error interval in one-dimension.

Proposition 4 (Learner's Cost in d Dimensions). *A learner who publishes an undominated classifier f based on a hyperplane $\mathbf{g}^\intercal \mathbf{x} = g_0$ can only commit false positives on group A candidates and false negatives on group B candidates. The cost of such a classifier is*

$$C_{FN} P_{x \sim \mathcal{D}_B}\Big[\mathbf{x} \in \big(\mathbf{g}^\intercal \mathbf{x} < g_0 - \frac{g_{k_B}}{c_{k_B}} \bigcap \mathbf{w}_B^\intercal \mathbf{x} \geqslant \tau_B\,\big)\Big] \\ + C_{FP} P_{x \sim \mathcal{D}_A}\Big[\mathbf{x} \in \big(\mathbf{w}_A^\intercal \mathbf{x} < \tau_A \bigcap \mathbf{g}^\intercal \mathbf{x} \geqslant g_0 - \frac{g_{k_A}}{c_{k_A}}\big)\Big],$$

where $k_B \in \arg\max_{i \in [d]} \frac{g_i}{c_{B,i}}$ and $k_A \in \arg\max_{i \in [d]} \frac{g_i}{c_{A,i}}$.

2.4 Learner Subsidy Strategies

Since in our setting, the learner's classification errors are directly tied to unequal group costs, we ask whether she would be willing to subsidize group B candidates in order to shrink the manipulation gap between the two groups and as a result, reduce the number of errors she commits. In this section, we formalize subsidies as interventions that a learner can undertake to improve her classification performance. Although in many high-stakes classification settings, the barriers that make manipulation differentially accessible are non-monetary—such as time, information, and social access—in this section, we consider subsidies that are monetary in nature to alleviate the financial burdens of manipulation.

We introduce these subsidies for the purpose of analyzing their effects on not only the learner's classification performance but also candidate groups' outcomes. Since subsidies mitigate the inherent disparities in groups' costs and increase access to manipulation, one might expect that their implementation would surely improve group B's overall welfare. In this section, we show that in some cases, optimal subsidy interventions can surprisingly have the effect of lowering the welfare of candidates from *both* groups without improving the welfare of even a single candidate.

2.4.1 Subsidy Formalization

There are different ways in which a learner might choose to subsidize candidates costs. In the main text of this chapter, we focus on subsidies that reduce each group B candidate's costs such that the agent need only pay a β fraction of her original manipulation cost.

Definition 4 (Proportional subsidy). *Under a proportional subsidy plan, the learner pays a proportion* $1 - \beta$ *of each group B candidate's cost of manipulation for some* $\beta \in [0, 1]$. *As such, a group B candidate who manipulates from an initial feature vector* $\mathbf{x}$ *to a final feature vector* $\mathbf{y}$ *bears a cost of* $\beta\big(c_B(\mathbf{y}) - c_B(\mathbf{x})\big)$.

In the appendix, we also introduce flat subsidies in which the learner absorbs up to a flat α amount from each group B candidate's costs, leaving the candidate to pay $\max\{0, c_B(\mathbf{y}) - c_B(\mathbf{x}) - \alpha\}$. Similar results to those shown in this section hold for flat subsidies.

When considering proportional subsidies, the learner's strategy now consists of both a choice of β and a choice of classifier f to issue. The learner's goal is to minimize her penalty

$$C_{FP} \sum_{m \in \{A,B\}} p_m P_{\mathbf{x} \sim \mathcal{D}_m}\big[h_m(\mathbf{x}) = 0, f(\mathbf{y}) = 1\big] + C_{FN} \sum_{m \in \{A,B\}} p_m P_{\mathbf{x} \sim \mathcal{D}_m}\big[h_m(\mathbf{x}) = 1, f(\mathbf{y}) = 0\big] + \lambda \mathit{cost}(f, \beta),$$

where $\mathit{cost}(f, \beta)$ is the monetary cost of the subsidy, C_{FP} and C_{FN} denote the cost of a false positive and a false negative respectively as before, and $\lambda \geqslant 0$ is some constant that determines the relative weight of misclassification errors and subsidy costs for the learner.

For ease of exposition, the remainder of the section is presented in terms of one-dimensional features. In Section A.1.3 of the appendix, we show that in many cases, the d-dimensional linear costs setting can be reduced to this one-dimensional setting.

As an analog of (2.7), we define $\ell_B^{\beta}(y) = (\beta c_B)^{-1}(\beta c_B(y) - 1)$, giving the minimum corresponding unmanipulated feature x for any observed feature y. Under the proportional subsidy, for a given y, the group B candidate must have $x \geqslant \ell_B^\beta(y)$. From this, we define σ_B^β such that $\ell_B^\beta(\sigma_B^\beta) = \tau_B$.

In order to compute the cost of a subsidy plan, we must determine the number of group B candidates who will take advantage of a given subsidy benefit. Since manipulation brings no benefit in itself, candidates will only choose to manipulate and use the subsidy if it will lead to a positive classification. For a published classifier f with threshold σ, we then have

$$cost(f, \beta) = \big(1 - \beta\big) \int_{\ell_B^\beta(\sigma)}^{\sigma} \big(c_B(\sigma) - c_B(x)\big) P_{x \sim \mathcal{D}_B}(x) dx.$$

Although the learner's optimization problem can be solved analytically for various values of λ, we are primarily interested in taking a welfare-based perspective on the effects of various classification regimes on both the learner and candidate groups. In the following section, we analyze how the implementation of a subsidy plan can alter a learner's classification strategy and consider the potential impacts of such policies on candidate groups.

2.4.2 Group Welfare Under Subsidy Plans

While a learner would choose to adopt a subsidy strategy primarily in order to reduce her error rate, offering cost subsidies can also be seen as an intervention that might equalize opportunities in an environment that by default favors those who face lower costs. That is, if costs are keeping group B down, then one might believe that reducing costs will surely allow group B a fairer shot at manipulation, and, as a result, a fairer shot at positive classification. Alas we find that mitigating cost disparities by way of subsidies does not necessarily lead to better outcomes for group B candidates. In fact, an optimal subsidy plan can actually reduce the welfares of *both* groups. Paradoxically, in some cases, the subsidy plan boosts only the learner's utility, whereas every individual candidate from both groups would have preferred that she offer no subsidies at all.

The following theorem captures the surprising result that subsidies can be harmful to all candidates, even those from the group that would appear to benefit.

Theorem 2 (Subsidies can harm both groups). *There exist cost functions c_A and c_B satisfying the cost conditions, learner distributions $\mathcal{D}_A$ and $\mathcal{D}_B$, true classifiers with threshold τ_A and τ_B, population proportions p_A and p_B, and learner penalty parameters C_{FN}, C_{FP}, and λ, such that no candidate in either group has higher payoff at the equilibrium of the Strategic Classification Game with proportional subsidies compared with the equilibrium of the Strategic Classification Game with no subsidies, and some candidates from both group A and group B are strictly worse off.*

We note that a slightly weaker version of the theorem holds for flat subsidies. In particular, there exist cases in which some individual candidates have higher payoff at the equilibrium of the Strategic Classification Game with flat subsidies compared with the equilibrium with no subsidies, but both group A and group B candidates have lower payoffs on average with the subsidies.

To prove the theorem, it suffices to give a single case in which both candidate groups are harmed by the use of subsidies. However, to illustrate that this phenomenon does not arise only as a rare corner case, we provide one such example here plus two in the appendix, and discuss general conditions under which this occurs. In each example, we consider a particular instance of the Strategic Classification Game and compare the welfares of candidates at equilibrium when the learner is able to select a proportional subsidy with their welfares at equilibrium when no subsidy is allowed.

Example 1. *Suppose that a learner is error-minimizing such that $C_{FN} = C_{FP} = 1$ and $\lambda = \frac{3}{4}$. Suppose that unmanipulated features for both groups are uniformly distributed with $p_A = p_B = \frac{1}{2}$. Let group cost functions be given by $c_A(x) = 8\sqrt{x} + x$ and $c_B(x) = 12\sqrt{x}$; note that the cost condition $c'_A(x) < c'_B(x)$ holds for $x \in [0, 1]$. Let the true group thresholds be given by $\tau_A = 0.4$ and $\tau_B = 0.3$.*

When subsidies are not allowed, the learner chooses a classifier with threshold $\sigma^* = \sigma_B \approx 0.398$ at equilibrium. This threshold perfectly classifies all candidates from group B, while permitting false positives on candidates from group A with features $x \in [0.272, 0.4)$.

If the learner decides to implement a proportional subsidies plan, at equilibrium the learner chooses a classifier with threshold $\sigma^*_{prop} = \sigma_A \approx 0.546$ *and a subsidy parameter* $\beta^* = 0.558$. *Her new threshold now correctly classifies all members of group A, while committing false negatives on group B members with features* $x \in [0.3, 0.348)$.

Some candidates in group B are thus strictly worse-off, while none improve. Without the subsidy offering, group B members had been perfectly classified, but now there exist some candidates who are mistakenly excluded. Further, one can show that candidates who are positively classified must pay more to manipulate to the new threshold in spite of receiving the subsidy benefit. This increased cost is due to the fact that the higher classification threshold imposes greater burdens on manipulation than the β *subsidy alleviates.*

Group A candidates are also strictly worse-off since the threshold increase eliminates false positive benefits that some members had previously been granted in the no-subsidy regime. Moreover, all candidates who manipulate must expend more to do so, since these candidates do not receive a subsidy payment. Only the learner is strictly better off with the implementation of this subsidy plan.

Additional examples in the appendix show cases in which both groups experience diminished welfare when they bear linear costs. Even when the learner has an error function that penalizes false negatives twice as harshly as false positives and thus is explicitly concerned with mistakenly excluding group B candidates, an equilibrium subsidy strategy can still make both groups worse-off.

We thus highlight two consequences of subsidy interventions: On the one hand, with reduced cost burdens, more candidates from the disadvantaged group should be able to manipulate to reach a positive classification. However, subsidy payments also allow a learner to select a classifier that is at least as strict as the one issued without offering subsidies. These are opposing forces, and these examples show that without needing to distort underlying group probability distributions or the learner's penalty function in extreme ways, the effect of mitigating manipulation costs may be outweighed by the overall impact of a stricter classifier.

This result can also be extended to show that a setup in which candidates are unable to manipulate their features at all can be preferred by all three parties—groups A and B as well as the learner—to both the manipulation and subsidy regimes. We provide an informal statement of this proposition below and defer the interested reader to its formal statement and demonstration in the appendix.

Proposition 5. *There exist general cost functions such that the outcomes issued by a learner's equilibrium classifier under a non-manipulation regime is preferred by all parties—the learner, group A, and group B—to outcomes that arise both under her equilibrium manipulation classifier and under her equilibrium subsidy strategy.*

2.5 Discussion

Social stratification is constituted by forms of privilege that exist along many different axes, weaving and overlapping to create an elaborate mesh of power relations. While our model of strategic manipulation does not attempt to capture this irreducible complexity, we believe this work highlights a likely consequence of the expansion of algorithmic decision-making in a world that is marked by deep social inequalities. We demonstrate that the design of classification systems can grant undue rewards to those who *appear* more meritorious under a particular conception of merit while justifying exclusions of those who have failed to meet those standards. These consequences serve to exacerbate existing inequalities.

Our work also shows that attempts to resolve these negative social repercussions of classification, such as implementing policies that help disadvantaged populations manipulate their features more easily, may actually have the opposite effect. A learner who has offered to mitigate the costs facing these candidates may be encouraged to set a higher classification standard, underestimating the deeper disadvantages that a group encounters, and thus serving to further exclude these populations. However, it is important to note that these unintended consequences do not always arise. A conscientious learner who offers subsidies to equalize the playing field can guard against such paradoxes by making sure to classify agents in the same way even when offering to mitigate costs.

Other research in signaling and strategic classification has considered models in which manipulation is desirable from the learner's point of view.[39,62] Though this perspective diverges from the one we consider here, we acknowledge that there do exist cases in which manipulation serves to improve a candidate's quality and thus leads a learner to encourage such behaviors. It is important to note, however, that although this account may accurately represent some social classification scenarios, differential group access to manipulation remains an issue, and in fact, cases in which manipulation genuinely improves candidate quality may present even more problematic scenarios for machine learning systems. As work in algorithmic fairness has shown, feedback effects of classification can lead to deepening inequalities that become "justified" on the basis of features both manipulated and "natural".[32]

The rapid adoption of algorithmic tools in social spheres calls for a range of perspectives and approaches that can address a variety of domain-specific concerns. Expertise from other disciplines ought to be imported into machine learning, informing and infusing our research in motivation, application, and technical content. As such, our work seeks to investigate, from a theoretical learning perspective, some of the potential adverse effects of what sociology has called "quantification," a world increasingly governed by metrics. In doing so, we bring in techniques from game theory and information economics to model the interaction between a classifier and its subjects. This chapter adopts a framework that tries to capture the genuine unfair aspects of our social reality by modeling group inequality in a population of agents. Although this perspective deviates from standard idealized settings of learner-agent interaction, we believe that so long as machine learning tools are designed for deployment in the imperfect social world, pursuing algorithmic fairness will require us to explicitly build models and theory to address critical issues such as social stratification and unequal access.

3

A Short-term Intervention for Long-term Fairness

3.1 Introduction

As algorithms are increasingly deployed to make social decisions that have previously been under the sole purview of humans, a growing body of work has challenged the reigning primacy of optimality and efficiency when issues of bias and discrimination are potentially at stake. Research in the growing field of algorithmic fairness has sought to address these concerns about the machine decision-making process by examining and manipulating standard tasks such as ranking or classification under generalized constraints of "fairness." Such computational notions of fairness have been varied but two broad opposing perspectives have proposed solutions that either defend fairness at the individual level (similar individuals are treated similarly)[30] or at the group level (groups are awarded proportional representation).[53,35] While this chapter similarly adopts a constraint-based intervention to achieve fairness, we depart from standard accounts of fairness that consider static domain-general algorithms and instead develop a dynamic model for the specific domain of decision-making in the labor market. Our work considers the role that firms' hiring practices play in perpetuating economic inequalities between social groups by way of the disparate outcomes that groups experience in their employment opportunities and wage prospects. We address the issue by building upon a dynamic model of worker and firm behavior that has been shown to generate the asymmetric group outcomes that are observed empirically between black and white workers in the United States[16,5,41] and appending a constraint on firms' hiring practices that successfully induces a group-equitable equilibrium.

As we focus on the particular domain of labor market dynamics, our chapter draws upon an extensive literature in economics. The theory of statistical discrimination, originally set forth in two seminal papers by Phelps[78] and Arrow,[8] explains disparate group outcomes as the result of rational agent behaviors that lock a system into an unfavorable equilibrium. In the basic model, workers compete for a skilled job with wage w. Skill acquisition requires workers to expend an investment cost of c, which is distributed according to a function F. A worker's investment decision is an assessment of her expected wage gain compared with her investment cost. Firms seek information about a worker's hidden *ability* level but can only base hiring decisions on observable attributes: her noisy *investment* signal and group membership. The firm's response to this missing information problem is to update its beliefs about a worker's qualifications by drawing on its prior for her group's ability levels. Therefore if a firm holds different priors for different groups, it will also set different group-specific hiring thresholds. Further, since these distinct thresholds are observed and internalized by workers, they adjust their own investment strategies accordingly—individuals within the unfavored group will lower their investment levels, and individuals in the favored group will continue to invest at a high level. Notably, even when the distribution of investment costs F is the same for each group[*], an asymmetric equilibrium can arise in which groups invest at different levels, further informing firms' distinct priors and reinforcing disparate employment prospects. In other words, rational workers and firms best respond in ways that exactly confirm the others' beliefs and strategies, and thus, the discriminatory outcome is "justified."

A proponent of "individual fairness" may diagnose the problem of statistical discrimination as a failure to treat candidates of similar investments similarly[†]. After all, the mistaken inference of unequal group ability levels indeed appears to be the origin of firms' inequitable hiring decisions. Moreover, when investment level is positively correlated with likelihood of being qualified, hiring based solely on investments is both rational and individually-fair. However, this group-blind solution fails to take into account a critical aspect of workers' investments—namely that they are *choices* rather than *givens*. Failure to recognize the upstream causes of observed data features brings to light the prickly notion of "ground truth" that has, from the start, plagued work on machine learning bias. Within a system as complex as the labor market, an input-output account of fairness that assesses the mapping of workers' investment levels to their hiring outcomes does not resolve the underlying source of inequalities that drives the differences in attributes between groups. Because both

[*] This has been the standard assumption in the economics literature since Arrow.[8]

[†] In the exposition of "individual fairness" proposed by Dwork et al.,[30] the built-in flexibility of the generic similarity metric between persons can include group membership and even be used to justify "fair affirmative action." However, within an economic signaling environment where firms' hiring standards affect workers' investments, a more flexible metric approach that compares quality within and across groups still fails to account for the strategy and incentive features of the labor market and thus the group coordination failure that characterizes many statistical discrimination equilibria.

statistical discrimination and machine learning rely on data that harbor historical inequalities, *local* fairness checks are often incapable of addressing the self-perpetuating nature of biases. Even without group biases, the paradox remains: the cyclic equilibrium ensures local procedural fairness—fairness with respect to investment choices—while maintaining global disparate outcomes.

The difficulty in pinpointing a particular cause of observed system-wide asymmetric outcomes challenges our mission in designing constraints to ensure fairness within the domain. If the outcomes themselves are trapped in a feedback loop, a successful fairness constraint should first jolt the system out of its current steady-state, and second, launch it on a path towards a preferable equilibrium. As such, a successful approach must consider fairness *in situ.* This chapter presents a *domain-specific* dynamic model with an intervention that effects *system-wide* impact, guaranteeing a group-equitable equilibrium that is stable and self-sustaining.

In our model, workers invest in human capital, enter first a Temporary Labor Market (TLM) and then transition into a Permanent Labor Market (PLM)[‡]. We use this partition to impose a constraint on TLM hiring practices that enforces group statistical parity representation. However, the restriction need not apply in the PLM where firms select natural best response hiring strategies. Our employment model is *reputational*—a worker carries an individual reputation, which is a summary of her past job performances and belongs to a group with a collective reputation, which is a measure of the proportion of its members producing "good" outcomes.

Working within this model, we show that by imposing this constraint on firms' hiring strategies in the TLM, the resulting steady-state in the PLM is symmetric such that an equal proportion of workers in the two groups produce good outcomes and are thus hired. The labor market at equilibrium, both procedurally and in outcomes, satisfies leading notions of "fairness"–group, individual, meritocratic[35,30,57]—discussed in the algorithmic fairness literature. Furthermore, we show that

[‡] Contracting in a segmented market is common in the labor economics literature. Of these, our work is most similar to Kim and Loury,[70] but notably they model the effects of statistical discrimination, while ours explicitly requires group-equitable outcomes.

under particular labor market conditions, it Pareto-dominates the asymmetric outcomes that arise under two unconstrained rational hiring strategies: group-blind hiring and statistical discriminatory hiring. Our fairness intervention exploits the complementary nature of individual and collective reputations such that the system produces its own feedback loop that incrementally addresses initial inequalities in group social standing. As such, the TLM intervention need not be permanent—statistical parity of hired workers becomes the natural result of firms' optimal hiring strategies once group equality is restored and the fairness constraint becomes obsolete.

This chapter's constraint-based approach to achieving equitable group outcomes in a reputational model of labor market interactions melds the perspectives and techniques of labor economics with the motivations of algorithmic fairness. However, our system-wide view also challenges a thread of work in the literature that characterizes notions of fairness as input-output-based properties of a decision-making function. By casting workers and firms as strategic agents in a dynamic game, we incorporate complexities of the labor market dynamic such as agents' expectations, incentives, and externalities that are otherwise difficult to encapsulate in a static classification setting. We advocate for an intervention that addresses the root of disparities between black and white workers' positions in the labor market and society—not only positions of unequal prospects and outcomes but as important, positions of unequal opportunities and, as a result, qualifications. Ensuring procedural fairness in the hiring decision alone is insufficient for this greater task. Our proposed constraint is designed to perturb a labor market at asymmetric equilibrium by co-opting the system's own cyclic effects to install group-equality that is self-sustaining in the long-term.

In Section 2, we present a standard model of labor market dynamics and introduce our fairness intervention. Section 3 contains an overview of the equilibria results of the constrained-hiring model along with a comparison against equilibria arising from two rational hiring strategies free from such a constraint. The chapter ends with a reflection on the equilibrium tendencies of discrimination and their implications on the design of fairness constraints. We also offer some comments on the dynamic feedback effects that are inherent features of persistent inequalities and the challenges they issue upon future work in algorithmic fairness.

3.1.1 Related Work

Within the algorithmic fairness literature, Zemel et al.[97] address group and individual notions of fairness by constructing a mapping of agent data to an intermediate layer of clusters that each preserve statistical parity while obfuscating protected attributes. A second map taking cluster assignments to their final classifications then allows "similar" agents to be treated similarly. This dual-map approach roughly corresponds to the roles of the TLM and PLM in our model. Related work has sought distance metrics to guide the initial mapping,[30] but since criteria for similarity vary by domain, general approaches often face obstacles of application. Our chapter's concentrated treatment of labor market dynamics aims to addresses this concern. We answer a call by Friedler et al.[40] to specify a particular world view of fairness within a domain and classification task. Our model starts with an assumption of inherent equality between groups. As such, differences in observable investment decisions or job outcomes are due to unequal societal standing, producing secondary effects of inequality, rather than fundamental differences in the nature of the individuals.

Labor market discrimination has been of long-standing interest in economics due to the persistent inequalities in employment prospects among groups of different race, gender, and other socially-salient attributes.[16,5,41] Since most explicit forms of wage discrimination are now illegal in the U.S. and genetic accounts of group differences have been largely discredited,[76] modern theories of labor market discrimination have updated the classical works—Becker's "taste-based" discrimination[12] and Phelps' model of exogenous group productivity differences[78]—by examining the *social* sources of asymmetric outcomes. Research in the field has produced models that consider temporal dynamics, utilize distinct group cost functions, and develop wages endogenously.[19,7] We follow in this line of work by incorporating a dynamic group reputation parameter into an individual's cost function, a modeling choice informed by the vast empirical literature showing the differential externalities produced by groups of differential social standing. Our model is not the first that makes explicit this linkage. In research examining the impact of neighborhood segregation on agents' accesses to resources for skill acquisition, Bowles, Loury, and Sethi[14] include a group "skill share" metric that functions similarly to our notion of group reputation in its effect on individuals' costs.

This chapter also frames the hiring process as reputational in nature, following a distinct literature on collective reputation.[88,92] Of these, our work shares most in common with the model proposed by Levin,[67] in which workers carry an individual reputation that contributes to their group's reputation. Levin shows that even when cost conditions evolve stochastically, reputations can produce a persistent feedback effect that leads to convergence to an asymmetric equilibrium in which groups occupy distinct social standings. Unlike in Levin, the notion of collective reputation in our model bears not only on workers' forward-looking expectations and incentives but also explicitly impacts future generations' investment costs. Additionally, since our work has in mind the information-processing capabilities of artificial intelligence agents, we formalize the concept of "individual reputation" as composed of a total history of previous outcomes. These additional "data," while potentially overwhelming for human decision-makers, can be handled by an algorithmic decision-maker. Since the functionality of machine learning in the hiring process is ultimately based in a form of "rational" statistical discrimination of worker data and job histories, this strand of economics literature is particularly relevant for considerations of algorithmic fairness in the labor market.

3.2 Model

We highlight the role of the fairness constraint within the rest of the standard labor market dynamics of the model by utilizing a dual labor market setup composed of a Temporary Labor Market

(TLM) and a Permanent Labor Market (PLM). In the former, a hiring constraint is established to ensure statistical parity, and in the latter, firms hire according to their best response hiring practices in a reputational model applied to the particular setting of employment.

This partition does little to impinge upon the standard dynamics of the labor market—workers flow from the TLM to the PLM, wages are labor-market-wide, and individual worker reputations in the PLM produce externalities for the collective group reputations that play a key role in individuals' pre-TLM investment decisions.

3.2.1 General Setup

Consider a society of n workers who pass through the labor market sequentially at times $t = 0, 1, \dots$. The labor markets maintain a constant relative size: m proportion of the workers reside in the TLM, and $1 - m$ reside in the PLM. Movement is governed by Poisson processes—workers immediately replace departing ones in the TLM, transition from the TLM to the PLM according to the parameter κ, and leave the PLM at rate λ.

Each worker belongs to one of two groups $\mu \in \{B, W\}$ with population share σ_B and $1 - \sigma_B$ respectively. We assume that these subpopulation proportions of workers are stable such that a worker of group μ who leaves the labor market is replaced via the birth of a new worker of the same group. The distribution of individual abilities, described by the CDF $F(\theta)$, is stable over time and identical across groups. In contrast, societal reputation varies with time and by group. A group's time t reputation π_t^μ gives the proportion of all individuals in group μ who are producing "good" outcomes in the labor market, over the interval timespan $[t - \tau, t]$, where the parameter $\tau \geqslant 0$ controls the time-lag effect of a group's previous generations' performance on its present reputation.

Prior to entering the labor market, workers select education investment levels η, weighing the cost of investment with its expected reward. Firms hire and pay workers based on expected performance, awarding wage $w(g_t)$ for a "good" worker, where g_t gives the proportion of "good" workers in the

PLM at time t. To prevent constant fluctuation at each time step, the wage $w_t = w(g_{t'})$ updates in a Poisson manner such that $t' < t$ gives the time of the last wage change. The hiring process is formalized by assigning workers to either skilled or unskilled tasks with distinct wages. For simplicity, workers who do not pass particular hiring thresholds may still be considered "hired," but they are assigned to an unskilled task and paid a wage normalized to 0.

As a function, the wage premium w_t is decreasing in g_t, since as the relative supply of "good" workers increases, imperfect worker substitutability lowers their marginal productivity, thus decreasing wage. We impose a minimum wage $\underline{w}$ such that $\lim_{g_t \to \infty} w(g_t) = \underline{w}$ and a maximum wage $\overline{w}$ such that $\lim_{g_t \to 0} w(g_t) = \overline{w}$. In the context of the model, minimum and maximum wages should not be considered as only products of labor laws, rather they also act to track the supply of "good" workers relative to firms' demand.

3.2.2 Temporary Labor Market

A worker i of group μ chooses to invest in human capital $\eta_i \geqslant 0$ according to her expected wage gain of being in the skilled labor market w_t[§] and her personal cost function for investment, $c_{\pi_t^\mu}(\theta_i, \eta_i)$, which is a function decreasing in her individual ability θ_i and increasing in her selected level of investment η_i. The incorporation of group reputation π^μ into an individual's cost function reflects the differential externalities produced by groups of differential social standing.[14] We posit that a worker belonging to a group with a superior societal reputation has improved cost conditions relative to her counterparts with equal ability in the lower reputation group. Formally, $\forall \pi_t^\mu < \pi_t^\nu, c_{\pi_t^\mu}(\theta_i, \eta_i)$ is a positive monotonic transformation of $c_{\pi_t^\nu}(\theta_i, \eta_i)$.

Investment in human capital operates as an imperfect signal, and workers have a hidden true type: *qualified* or *unqualified*, $\rho \in \{Q, U\}$. Let $\gamma : \mathbb{R}_{\geqslant 0} \to [0, 1]$ be a monotonically increasing function that maps a worker's investment level to her probability of being qualified. Unlike in

[§] Workers are boundedly rational and unable to anticipate future wage dynamics.

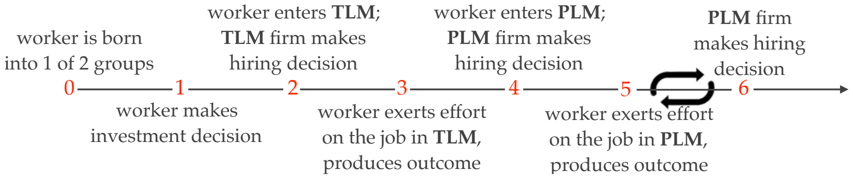


Figure 3.1: Timeline of worker and firm interactions throughout the labor market pipeline.

Spence's original work on education signaling[84] in which investment confers no productivity benefits and thus operates purely as a signal to employers, in our model, a worker's chosen investment level η has intrinsic value insofar as it is positively correlated with her likelihood of being qualified $\gamma(\eta)$.

Given this setup, a firm's *TLM hiring strategy* is a mapping $\mathcal{H}_T : \mathbb{R}_{\geqslant 0} \times \mu \to \{0, 1\}$ such that the hiring decision for worker i is based only her observable investment level $\eta_i \in \mathbb{R}_{\geqslant 0}$ and group membership μ. A worker who is hired into the TLM enters the pipeline and is eligible to compete for a PLM skilled job; a worker who does not pass the TLM hiring stage remains in the market but is permanently excluded from candidacy for the skilled wage. In this chapter, we mainly consider only those workers who successfully enter the skilled hiring pipeline, considering all others as "not hired." As such, we use the terms "skilled" and "hired" interchangeably.

3.2.3 Permanent Labor Market

Labor market dynamics follow in the style of repeated principal-agent interactions with hidden actions (effort exertion) but observable histories (reputation of outcomes). Once hired into the TLM, a worker i exerts on-the-job effort—choosing either *high* (H) or *low* (L) effort—which stochastically produces an observable *good* (G) or *bad* (B) outcome that affects her individual reputation and thus future reward. Exerting L is free, but exerting H bears cost $e_\rho(\theta_i)$, which is a function of qualification $\rho \in \{Q, U\}$ and ability level θ_i. Effort is more costly for unqualified individuals: $\forall \theta_i, e_U(\theta_i) > e_Q(\theta_i)$. We emphasize here that the notions of ability level θ and qualification status

ρ are distinct worker qualities. A high ability worker is one who has the general attributes that bear on success in the realms of education and work, whereas a qualified worker is one who has the appropriate training and skills for a given job. We may say, very crudely, that a worker is "born" with an ability level and "earns" a qualification status. In our model, a worker's ability level precedes her investment decision, which begets a qualification status.

High effort increases the probability of a good outcome G. If $p_{\rho,k}$ gives the probability of achieving outcome G with qualifications ρ and effort level k, then the following inequalities hold.

$$p_{Q,H} > p_{Q,L}; \quad p_{U,H} > p_{U,L}; \quad p_{Q,L} > p_{U,L}$$

Since the effect of qualifications on exerting high effort is already incorporated in its cost, $p_{Q,H} = p_{U,H}$, we write both quantities as p_H. We then simplify $p_{Q,L}$ and $p_{U,L}$ to p_Q and p_U respectively.

We emphasize the distinction between the *effort* exertion cost functions $e(\cdot)$ here and the previous *investment* cost functions $c(\cdot)$—the former are pertinent to workers already in the labor market and differ by qualification status, whereas the latter relate to pre-labor-market decisions and differ by group membership. Separate cost functions allow for a finer analysis of the salient factors that influence agent behavior at distinct points of the labor market pipeline. The inclusion of group membership into human-capital investment costs reflects the genuine differences in resources available to workers of different groups in their paths to education attainment[¶].

A worker keeps the same TLM job until the Poisson process with parameter κ selects her to move into the PLM, where at each time step, she cycles through jobs, exerting a chosen effort level, producing an observable outcome, and accumulating a history of past performances that includes her TLM outcome. At each time step, firms in the PLM want to hire all and only those workers who

[¶] We do not claim that group membership ceases to be a relevant factor impacting agent behavior once workers are in the labor market, but we note that a worker's qualifications, or the extent to which her skill investment proved to be successful, becomes an overriding determinant. Insofar as education investment bears on qualification status, a worker's group membership continues to impact her labor market outcomes.

consistently exert effort. To do so, firms distill a worker's history of observable outcomes into her "individual reputation" Π_i^t, which gives the proportion of outcomes G in her recent length-t history. In a labor market system of repeated worker-firm contracting, firms have the power to use these observable individual reputations to set self-enforcing relational contracts. A firm's *PLM hiring strategy* is a mapping $\mathcal{H}_P : [0,1] \to \{0,1\}$ such that the decision is solely a function of Π_i^t. Figure 3.1 depicts a timeline of how workers move through the labor market pipeline and interact with firms.

While "fairness" is a notoriously thorny ethical concept to define, the goal here of achieving long-term fairness is equivalent to attaining group equality in labor market outcomes. Since groups do not differ in fundamental or intrinsic ways, their job and wage prospects should also not systematically diverge at a fair steady-state.

Table 3.1: Table of notation

Notation	Significance
$F(\theta)$	CDF of ability levels θ
π^μ	group μ reputation
σ_μ	group μ population share
w_t	wage at time t
g_t^μ	proportion of group μ workers producing good outcomes at time t
η	investment level
p_H, p_Q, p_U	probability of producing G given effort level
$c_{\pi_t^\mu}(\theta, \eta)$	cost of investment
$\gamma(\eta)$	probability of being qualified
$\rho \in \{Q, U\}$	hidden qualification status
$e_\rho(\theta)$	cost of effort exertion
Π_i^t	individual reputation at time t

3.3 Results

Reputation-based labor market models, such as the one described in this chapter, can generate asymmetric group outcomes when firms utilize rational strategies such as statistical discrimination or group-blind hiring.[8,22,7,19] Since this chapter examines the effect of our proposed intervention on system-wide dynamics and outcomes, in the following section, we consider only those strategies and equilibria outcomes that arise in this fairness-constrained setting.

3.3.1 Equilibrium Strategies and Steady-States

We start by describing TLM strategies resulting from the fairness constraint, then move onto the PLM and analyze firms' and workers' best response strategies together. Gameplay in the PLM mirrors repeated principal-agent interactions wherein firms have the power to enforce contracts by monitoring individual reputations, and thus we consider strategies that constitute a sequential equilibrium.

Since a firm in the TLM prefers candidates who are more likely to be qualified, optimal hiring follows a threshold strategy: Given a hiring threshold $\hat{\eta}$, $\forall i$ such that $\eta_i \geqslant \hat{\eta}$, $\mathcal{H}_{\mathcal{T}}(i) = 1$, and inversely, $\forall i$ such that $\eta_i < \hat{\eta}$, $\mathcal{H}_{\mathcal{T}}(i) = 0$. However, since firms must abide by the statistical parity hiring rule, their optimal threshold strategy is uniquely determined: if a firm aims to hire a fraction ℓ of all workers, its investment thresholds will be implicitly defined and group-specific, so that in the TLM, skilled employees from groups μ and ν will constitute $\sigma_\mu \ell$ and $(1 - \sigma_\mu)\ell$ proportions of the full worker population respectively.

A worker of group μ, observing her group-specific TLM investment threshold $\widehat{\eta}_\mu$, will weigh her cost of investment with her expected wage gain w_t. All workers i with $c_{\pi_t^\mu}(\theta_i, \widehat{\eta}_\mu) \leqslant w_t$ will choose to invest exactly at the level $\eta_i = \widehat{\eta}_\mu$ and be hired for the skilled position in the TLM; all other workers will invest at level $\eta_i = 0$ and fail to enter the pipeline to compete for the skilled job. Workers who pass the first hiring stage know that their future PLM opportunities will depend on their observable outcome in the TLM, and as such they exert effort in a one-shot game. A worker i with qualification status ρ exerts high effort on the job if and only if $e_\rho(\theta_i) \leqslant w_t(p_H - p_\rho)$.

As previously shown, while the statistical parity constraint preserves the fundamental equality of ability distributions $F(\theta)$ between groups, the group-specific investment thresholds $\widehat{\eta}_\mu$ generate group-specific investment strategies. As consequence, since investment has positive returns on qualification status, groups may have differing proportions of qualified candidates in the PLM pool. We denote by γ_t^μ the proportion of candidates in group μ who are qualified at time t, leaving $1 - \gamma_t^\mu$ who are unqualified. Then the proportion of group μ workers in the TLM who produce good outcomes follows the recursive model

$$\begin{gathered} g_t^\mu = p_H[1 - F(\widehat{\theta}_Q)\gamma_t^\mu - F(\widehat{\theta}_U)(1 - \gamma_t^\mu)] + p_Q F(\widehat{\theta}_Q)\gamma_t^\mu + p_U F(\widehat{\theta}_U)(1 - \gamma_t^\mu) \qquad (3.1) \\ \text{where } \widehat{\theta}_\rho = e_\rho^{-1}(w_t(p_H - p_\rho)) \text{ and } g_{t'} = \sigma_\mu \ell g_{t'}^\mu + (1 - \sigma_\mu)\ell g_{t'}^\nu \end{gathered}$$

with $w_t = w(g_{t'})$ where t' gives the time of the last wage update.

It is important to note that g_t^μ gives the proportion of workers in the skilled labor market who at time t are producing good outcomes in their jobs. This quantity does not exactly coincide with group reputation, π_t^μ, which gives a (time-interval average) normalized metric that scales with the proportion of *all* members in group μ–including those who are not granted entry into the skilled job pipeline–who are producing good outcomes.

A PLM worker's future-anticipatory strategy is a selection of time, reputation, wage, and hiring threshold-dependent probabilities of effort exertion $\varepsilon(\Pi_i^{t'})$ with $\Pi_i^{t'} \in \{\Pi^{t'}\}$ where the index i of $\Pi_i^{t'}$ denotes a particular individual reputation level in the set of all possible reputation levels $\{\Pi\}$ and t' tracks the length of time that has passed since the last wage update. Supposing that workers engage in N-depth reasoning where $N \gg t'$, this quantity may be computed via backward induction on the continuation value for a given individual reputation, $V(\Pi_i^{t'})$. With this setup, the continuation value $V(\Pi^N) = 0$, and the agent with ability θ and qualification ρ solves the following dynamic programming problem

$$
\begin{aligned}
V(\Pi_i^{t'}, \hat{\Pi}^{t'}, w_t) = \sup_{\varepsilon(\Pi_i^{t'}) \in [0,1]} & \Big\{ (1-\lambda)[V(\Pi_i^{t'+1}, G)[\varepsilon(\Pi_i^{t'})(p_H - p_\rho) + p_\rho] \\
& + V(\Pi_i^{t'+1}, B)[(-\varepsilon(\Pi_i^{t'}))(p_H - p_\rho) + 1 - p_\rho]] + 1_{\Pi_i^{t'} \geqslant \hat{\Pi}^{t'}} w_t \Big\} \\
\text{where } V(\Pi_i^{t'}, G) = V(\frac{\Pi_i^{t'} t' + 1}{t' + 1}, \hat{\Pi}^{t'}, w_t) & \text{ and } V(\Pi_i^{t'}, B) = V(\frac{\Pi_i^{t'} t'}{t' + 1}, \hat{\Pi}^{t'}, w_t) \\
\text{and } \forall t, w_t = w_T & \text{ when the agent looks forward from time } T
\end{aligned}
$$

where the worker solves for optimal effort exertion probabilities $\varepsilon(\Pi_i^{t'})$ for each possible reputation $\Pi_i^{t'} \in \{\Pi^{t'}\}$, and high effort is only optimal at time t if $V(\Pi_i^{t'}, G)(p_H - p_\rho) \geqslant e_\rho(\theta)$.

If firms seek those workers who appear willing and able to exert high effort upon being hired, their equilibrium strategy is to select a reputation threshold $\hat{\Pi}^{t'} = p_H - \Delta_{t'}$ when facing a worker with history length t' since the last wage update. $\Delta_{t'} > 0$ acts as the firm's optimistic forgiveness buffer, permitting a worker's recent time t' reputation to be slightly under the p_H threshold, to ensure that it does not penalize workers who exert high effort but are unlucky and receive B outcomes. An optimal choice of $\Delta_{t'}$ monotonically decreases in t' toward 0 as the reputation of a worker consistently exerting high effort converges to p_H as $t' \rightarrow \infty$. Note that the firm must also take care not to decrease $\Delta_{t'}$ too slowly, lest workers are able to exert low effort and continue to be hired. Thus the firm optimizes its hiring threshold $\hat{\Pi}^{t'} = p_H - \Delta_{t'}$ by decreasing Δ just enough at each time step to motivate consistent high effort from workers who can afford it. All other workers exert low effort in each round. Thus given a firm's reputation threshold $\hat{\Pi}^{t'}$, its equilibrium PLM hiring strategy $\mathcal{H}_\mathcal{P}$ is a mapping such that if and only if the worker's accumulated reputation since the last wage update $\Pi_i^{t'}$ exceeds the threshold $\hat{\Pi}^{t'}$, $\mathcal{H}_\mathcal{P}(\Pi_i^{t'}) = 1$, and the worker is hired. Otherwise

$\mathcal{H}_{\mathcal{P}}(\Pi_i^{t'}) = 0$, and the worker does not earn the wage premium. This strategy is summarized in the following Proposition, and we defer the interested reader to the Appendix for its proof.

Proposition 6. *There exists a pair of PLM equilibrium strategies $(\mathcal{H}, \mathcal{E})$ of firm-hiring and worker-effort respectively such that*

(i) A firm's hiring strategy $\mathcal{H}$ is a selection of a reputation threshold function of the form $\hat{\Pi}^{t'} = p_H - \Delta_{t'}$, where $\Delta_{t'}$ is a monotonically decreasing function in t', such that $\mathcal{H}(\Pi^{t'}) = 1$ if and only if $\Pi_i^{t'} \geqslant \hat{\Pi}^{t'}$, otherwise $\mathcal{H}(i) = 0$.

(ii) A worker's effort strategy $\mathcal{E}$ is a selection of effort levels that considers only the wage w_t and cost of effort such that $\mathcal{E}(w_t) = H$ if and only if $e_\rho(\theta) \leqslant w_t(p_H - p_\rho)$, else $\mathcal{E}(w_t) = L$.

Interestingly, the strategies employed in the repeated worker-firm interactions in the PLM generate a recursive relationship of the proportion of "good" workers for each group that mirrors the structure of (3.1). PLM firms' stringent threshold reputation hiring strategy imposes the same type of "pressure" on workers at each round of employment as does the single-shot game in the TLM. In both labor markets, every outcome "counts."

Having elaborated upon the dynamics of both the TLM and PLM, we incorporate worker movement and combine the results to obtain a recursive relationship that governs the sequence of workers' performance results from an initial wage w_0. Note that the multiplicity of possible firm hiring strategies produces a multiplicity of dynamic paths of outcomes $\{(g_t^\mu, g_t^\nu)\}_0^\infty$ to steady-state, but given that in our model, firms are willing to hire only and all workers who consistently exert high effort, firm and worker equilibrium strategies are as described in Proposition 1, there is a unique sequence of group outcome pairs (g_t^μ, g_t^ν) such that there exists a time $t = T$ with the property that $\forall t \geqslant T, (g_T^\mu, g_T^\nu) = (g_t^\mu, g_t^\nu)$.

Theorem 3. *Under the described labor market conditions in which ℓ proportion of workers gain entry into the TLM and firms abide by the statistical parity hiring constraint, the proportion of all workers* *in group μ producing good outcomes at time t, g_t^μ in the full labor market follows the recursive system*

$$g_{t+1}^\mu = p_H[1 - F(\theta_Q)\gamma_t^\mu - F(\theta_U)(1-\gamma_t^\mu)] + p_Q F(\theta_Q)\gamma_t^\mu + p_U F(\theta_U)(1-\gamma_t^\mu) \tag{3.2}$$

$$\textit{where } \pi_t^\mu = \frac{\sigma_\mu \ell}{\tau}\sum_{j=t-\tau}^{t} g_j^\mu, \tag{3.3}$$

$$\gamma_t^\mu = \varphi(\widehat{\eta}_\mu(\pi_t^\mu)), \tag{3.4}$$

$$\theta_\rho = e_\rho^{-1}(w_t(p_H - p_\rho)), \tag{3.5}$$

$$g_t = \sigma_\mu \ell g_t^\mu + (1-\sigma_\mu)\ell g_t^\nu, \tag{3.6}$$

where φ and $\widehat{\eta}_\mu$ in Eq. 3.4 are monotonically increasing functions whose composition combines the labor market's reputational feedback effect with firms' TLM constrained group-investment thresholds. Then there exists a unique stable symmetric steady-state equilibrium and convergence time T, wherein $\tilde{\pi}_t^\mu = \tilde{\pi}_t^\nu = \tilde{\pi}, \forall t > T$, satisfying system-wide fairness, with a corresponding unique stable wage $\tilde{w}$.

To understand why the existence of this unique stable symmetric equilibrium is guaranteed when TLM firms are bound to the statistical parity requirement, consider the two variables that affect a group μ worker i's likelihood of producing a good outcome: her ability level θ_i and her probability of being qualified $P(Q|\widehat{\eta}_\mu) = \gamma^\mu$. Since there are positive returns to investment, γ^μ is increasing in π^μ: As her group μ social standing rises, cost conditions improve, and as a result, workers in future generations are more likely to be qualified. With the imposition of the TLM hiring constraint, firms recognize the groups' different costs of investment and hire in a manner that retains equality between the two groups' underlying ability distributions $F(\theta)$ within the labor market, which assures that the proportions of workers producing good outcomes in each group g^μ do not diverge within the skilled labor market pipeline. Moreover, the statistical parity hiring constraint requires that firms hire in a manner such that workers from a disadvantaged group μ are not inequitably blocked from entering the skilled labor market and always constitute $\sigma_\mu \ell$ of the TLM. As a result of maintaining both identical ability distributions $F(\theta)$ and proportional representation σ_μ in the TLM, statistical parity hiring ensures that as group outcomes in the skilled labor market converge, so do group reputations. Thus, the γ_t-generated positive feedback loop that pushes towards diverging group outcomes is always constrained, allowing the natural reputational feedback on group investment cost functions c_{π^μ} to drive the convergence of group outcomes and thus group reputations to a single steady-state value. Importantly, throughout the path of $\{(g_t^\mu, g_t^\nu)\}$ outcomes toward this symmetric steady-state, the "severity" of the TLM fairness constraint on firms' hiring strategies continually slackens until it recedes into disuse. For a full exposition of the proof, see the Appendix.

Under statistical parity hiring in the TLM, groups with unequal initial social standing will gradually approach the same reputation level according to time-lag τ. The constraint has the effect of co-opting the "self-confirming" loop for group reputation improvement—collective reputation produces a positive externality, lowering individual group members' cost functions, thus improving investment conditions for future workers, further raising individual and group reputation. We point out that the empirically-validated link between group reputations and members' investment costs makes a TLM statistical parity constraint a more efficient means of addressing group inequalities than a similar intervention in the PLM. Since the TLM represents the entry point into the market, enforcing statistical parity at the onset ensures that lower reputation workers are not disproportionately excluded from the pipeline as a whole.

We next compare this steady-state under the TLM constraint with long-term outcomes of other rational hiring strategies that are not bound by any fairness constraints and show that under particular market conditions, the fair steady-state is Pareto-dominant.

3.3.2 Comparative Statics with Unconstrained Hiring Strategies

In the absence of any constraint, firms are free to select any strategy that will maximize their probability of employing high-ability, qualified workers. Two such common strategies are group-blind, sometimes called “meritocratic,” and statistical discriminatory hiring. We provide an overview of each practice and then continue on to comparing their long-term equilibria outcomes with the symmetric steady-state that arises under our TLM hiring constraint.

Consider a *group-blind* TLM hiring strategy that is individual-based, operating under an equal-treatment philosophy. Without considering agent group membership— suppose again $\mu \in \{B, W\}$— the firm hires a proportion ℓ of workers by selecting a single investment level threshold $\tilde{\eta}$ for all workers, implicitly defined as

$$\ell = (1 - \sigma_B)\Big(1 - F(c^{-1}_{\pi^W}(\tilde{\eta}(p_H - p_\rho))\Big) + \sigma_B\Big(1 - F(c^{-1}_{\pi^B}(\tilde{\eta}(p_H - p_\rho))\Big)$$

where σ_B and $1 - \sigma_B$ give the proportion of individuals in groups B and W respectively, and the function $c_{\pi^\mu}(\cdot)$ determines the group μ investment level. Pragmatically under this strategy, the firm will examine the broad distribution of all investment levels and select a threshold above which it is willing to employ workers. This strategy is also rationalized by the fact that the threshold $\tilde{\eta}$ maximizes the expected number of hired workers who are qualified.

An alternative class of firm hiring strategies employ *statistical discrimination*, in which priors regarding a worker’s observable attributes, such as group membership, are used to infer a particular individual’s hidden attributes. In particular, if TLM firms hold priors ξ_B and ξ_W about the two groups’ capabilities, upon observing an applicant’s group μ and investment level η, they will update their beliefs of the prospective employee’s qualifications according to:

$$P(Q|\mu, \eta) = \frac{p_Q(\eta)\xi_\mu}{p_Q(\eta)\xi_\mu + (1 - \xi_\mu)p_U(\eta)}$$

where $p_Q(\eta)$ and $p_U(\eta)$ give the probability of a qualified and unqualified worker having investment level η respectively.

Theorem 4. *In a PLM with unsaturated demand ($w = \bar{w}$) for skilled workers, the TLM constraint leads to a symmetric steady-state equilibrium that Pareto-dominates the asymmetric equilibria that arise under group-blind and statistical discriminatory hiring.*

We present an abbreviated exposition of the underlying factors that drive unconstrained hiring strategies to Pareto-dominated outcomes. For the full account of the proof, see the Appendix.

Group-blind hiring satisfies neither of the two key constrained hiring guarantees described in the proof explanation for Theorem 3—namely, groups no longer share equal ability distributions $F(\theta)$ nor are they proportionally represented in the market according to their demographic shares σ_μ. The violation of both of these criteria contribute to group reputation divergence and thus the existence of persistent asymmetric outcomes between groups.

At the asymmetric steady-state, groups retain distinct investment costs that, under a group-blind investment threshold, generate group-specific ability level thresholds $\widetilde{\theta_B}$ and $\widetilde{\theta_W}$. If group reputation $\pi_B < \pi_W$, then these ability thresholds may be ranked with respect to the threshold $\bar{\theta}$ that arises under the fairness constraint: $\widetilde{\theta_W} < \bar{\theta} < \widetilde{\theta_B}$. These hiring strategies inequitably bound the proportion of able and qualified workers in group B who are eligible to compete for skilled jobs, leaving behind an untapped source of group B individuals who would have otherwise been hired. Under PLM conditions in which demand for skilled workers is unsaturated and the wage $w(g_t) = \bar{w}$, workers in group W who are barred from entering the labor market in the proposed fair regime are not hired at equilibrium under group-blind hiring anyway. With strictly better-off employment outcomes for group B workers and no worse outcomes for group W workers, the constrained-hiring equilibrium Pareto-dominates the group-blind hiring equilibrium.

Similarly, statistical discriminatory hiring leads to group-specific ability thresholds and does not guarantee statistical parity. As Coate and Loury[22] show, self-confirming asymmetric equilibria also exist under this regime, wherein lower investment levels within the group with lower social standing are justified by firms' more stringent hiring standards. These effects have consequences that mirror the Pareto-dominated results under group-blind hiring.

3.4 Discussion

Describing disparate outcomes in employment as caused by rational agent best response strategies suggests that the field of algorithmic fairness should consider the labor market's inherent dynamic setting in its approach to potential interventions. Fairness constraints that are conceived as isolated procedural checks have a limited capacity to install system-wide fairness that is self-sustaining and long-lasting. The problem of fairness in the labor market is fundamentally tied to historical factors. Within nearly all societal domains in which fairness is an issue, past and current social relations differentially impact subjects, producing distinct sets of resources, options, and opportunities that continue to mark agents' choices and outcomes today. Empirical evidence points to what economist and social theorist Glenn Loury has called "development bias," in which black members of society have reduced chances of realizing their potential, as the greater source of racial inequality in welfare outcomes than discriminatory hiring.[71] This perspective challenges the notion that assuring "individual fairness" of the actual procedure of hiring should be the primary concern in assuring a labor market that is unbiased as a whole.

Not only is the standard learning theory formulation of the problem, in which agent attributes are treated as *a priori* givens, inadequate to attend to development bias, it also neglects the (arguably) meritocratic goals of the labor market. In economic settings, rewarding merit primarily serves an instrumental purpose—to incentivize investment and effort—rather than existing simply to pass along desert-based awards to candidates. Framing the problem as one of clustering or classification fails to understand the labor market as an incentive-oriented system. Fairness criteria that solely assess an algorithm's treatment of workers' qualifications similarly fall into the trap of viewing hiring decisions only as rewards to meritorious individuals without considering the incentive purposes of the reward system at-large.

In contrast, a dynamic model recognizes the ripple effect of development bias in the past and calls for a fairness intervention with incentive features that carries momentum into the future. The labor market as a source of economic opportunity is an ideal setting for a notion of fairness that is oriented toward a future beyond the short timeline of firm hiring cycles. It is precisely our focus on steady-state outcomes that allows for this long-term conception of fairness. However, it should be noted that the employment outcomes along the path to the symmetric equilibrium are by no means guaranteed to satisfy any notions of fairness, neither individual nor group. But we claim that conceiving of fairness in this way—as a project that aims to achieve permanent societal group-egalitarianism—is an ambition that is not only a worthy goal in itself but also one that we show may be economically socially optimal.

Our model of individual reputations as a sequence of previous outcomes in the PLM fits within the hiring regime today, in which employers have increased access to worker data. Since algorithms will be largely responsible for making sense of this historical data, future work should consider how systems that sift through a worker's history should be designed to determine when group membership-related considerations, such as the ones embedded in the TLM constraint proposed here, should be taken into account. As machine decision-makers are deployed increasingly throughout hiring processes, we must grapple with a long tradition of explicit and implicit human biases that have rendered the labor market prone to discriminatory practices. We hope that this work can suggest ways that algorithmic fairness interventions can shift these hiring strategies towards contributing to a better, fairer future.

While this chapter has shown that imposing the TLM hiring constraint ultimately leads to a group-symmetric outcome, we do not claim that ours is the only intervention able to produce such an equilibrium. The labor market pipeline in reality is an elaborate sequence of agent choices and social stages that is much more complex and heterogeneous than our model's pre-TLM, TLM, and

PLM periods. The true space of possible policy interventions dwarfs those considered in this work. Interventions aimed at reducing the economic inequalities that exist between black and white communities have been implemented at a variety of junctures in the standard social pipeline, ranging from direct governmental subsidy programs for childhood education costs in high-poverty areas to private companies' attempts at diversifying hiring by partnering with historically black colleges. As such, there may exist a multiplicity of intervention-types that all ultimately lead to group-egalitarian outcomes. Further analysis of the costs and efficiencies associated with each of these regimes will produce a richer understanding of potential fairness interventions and their concomitant welfare effects. Insofar as work in labor market fairness ought to inspire action and policy in the real world, these open questions will require both theoretical and empirical attention.

4

Fair Classification and Social Welfare

4.1 Introduction

In his 1979 Tanner Lectures, Amartya Sen noted that since nearly all egalitarian theories are founded on an equality of *some* sort, the heart of the issue rests on clarifying the "equality of what?" problem.[83] The field of fair machine learning has not escaped this essential question. Does machine learning have an obligation to assure probabilistic equality of outcomes across various social groups?[35,45]

Or does it simply owe an equality of treatment?[30] Does fairness demand that individuals (or groups) be subject to equal mistreatment rates?[95,11] Or does being fair refer only to avoiding some intolerable level of algorithmic error?

Currently, the task of accounting for fair machine learning cashes out in the comparison of myriad metrics—probability distributions, error likelihoods, classification rates—sliced up every way possible to reveal the range of inequalities that may arise before, during, and after the learning process. But as shown in work by Chouldechova[20] and Kleinberg et al.,[61] fundamental statistical incompatibilities rule out any solution that can satisfy all parity metrics. Fairness-constrained loss minimization offers little guidance on its own for choosing among the fairness desiderata, which appear incommensurable and result in different impacts on different individuals and groups. We are thus left with the harsh but unavoidable task of adjudicating between these measures and methods. How ought we decide? For a given application, who actually benefits from the operationalization of a certain fairness constraint? This is a basic but critical question that must be answered if we are to understand the impact that fairness constraints have on classification outcomes. Much research in fairness has been motivated by the well-documented negative impacts that these systems can have on already structurally disadvantaged groups. But do fairness constraints as currently formulated in fact earn their reputation as serving to improve the welfares of marginalized social groups?

When algorithms are adopted in social environments—consider, for example, the use of predictive systems in the financial services industry—classification outcomes directly bear on individuals' material well-beings. We, thus, view predictions as *resource allocations* awarded to individuals and by extension, to various social groups. In this chapter, we build out a method of analysis that takes in generic fair learning regimes and analyzes them from a welfare perspective.

Our main contributions, presented in Section 3, are methodological as well as substantive in the field of algorithmic fairness. We show that how "fair" a classifier is—how well it accords with a group parity constraint such as "equality of opportunity" or "balance for false positives"—does not neatly translate into statements about different groups' welfares are affected. Drawing on techniques from parametric programming and finding a SVM's regularization path, our method of analysis finds the optimal ε-fair Soft-Margin SVM solution for all values of a fairness tolerance parameter $\varepsilon \in [0, 1]$. We track the welfares of individuals and groups as a function of ε and identify those ranges of ε values that support solutions that are Pareto-dominated by neighboring ε values. Further, the algorithmic implementation of our analyses is computationally efficient, with a complexity on the same order as current standard SVM solvers that fit a single SVM model, and is thus practical as a procedure that translates fairness constraints into welfare effects for all ε.

Our substantive results show that a classifier that abides by a stricter fairness standard does not necessarily issue improved outcomes for the disadvantaged group. In particular, we prove two results: first, starting at any nonzero ε-fair optimal SVM solution, we express the range of $\Delta\varepsilon < 0$ perturbations that tighten the fairness constraint and lead to classifier-output allocations that are weakly Pareto dominated by those issued by the "less fair" original classifier. Second, there are nonzero ε-fair optimal SVM solutions, such that there exist $\Delta\varepsilon < 0$ perturbations that yield classifications that are strongly Pareto dominated by those issued by the "less fair" original classifier. We demonstrate these findings on the Adult dataset. In general, our results show that when notions of fairness rest entirely on leading parity-based notions, always preferring more fair machine learning classifiers does not accord with the Pareto Principle, an axiom typically seen as fundamental in social choice theory and welfare economics.

The purposes of our work are twofold. The first is simply to encourage a welfare-centric understanding of algorithmic fairness. Whenever machine learning is deployed within important social and economic processes, concerns for fairness arise when societal ideals are in tension with a decision-maker's interests. Most leading methodologies have focused on optimization of utility or welfare to the vendor but have rarely awarded those individuals and groups who are subject to these systems the same kind of attention to welfare effects. Our work explicitly focuses its analysis on the latter.

We also seek to highlight the limits of conceptualizing fairness only in terms of group-based parity measures. Our results show that at current, making a system "more fair" as defined by popular metrics can harm the vulnerable social populations that were ostensibly meant to be served by the imposition of such constraints. Though the Pareto Principle is not without faults, the frequency with which "more fair" classification outcomes are welfare-wise dominated by "less fair" ones occurs is troublesome and should lead scholars to reevaluate popular methodologies by which we understand the impact of machine learning on different social populations.

4.1.1 Related Work

Research in fair machine learning has largely centered on computationally defining "fairness" as a property of a classifier and then showing that techniques can be invented to satisfy such a notion.[53,30,97,35,96,52,45,79,17,64,60,95,11,56,28,3] Since most methods are meant to apply to learning problems generally, many such notions of fairness center on parity-based metrics about a classifier's behavior on various legally protected social groups rather than on matters of welfare.

Most of the works that do look toward a welfare-based framework for interpreting appeals to fairness sit at the intersection of computing and economics. Mullainathan[75] also makes a comparison between policies as set by machine learning systems and policies as set by a social planner. He argues that algorithmic systems that make explicit their description of a global welfare function are less likely to perpetrate biased outcomes and are more successful at ameliorating social inequalities. Heidari et al.[48] propose using social welfare functions as fairness constraints on loss minimization programs. They suggest that a learner ought to optimize her classifier while in Rawls' original position. As a result, their approach to social welfare is closely tied with considerations of risk. Rather than integrate social welfare functions into the supervised learning pipeline, we claim that the result of an algorithmic classification system can itself be considered a welfare-impacting allocation. Thus, our work simply takes a generic ε-fair learning problem as-is, and then considers the welfare implications of its full path of outcomes for all $\varepsilon \in [0,1]$ on individuals as well as groups. Attention to the potential harms of machine learning systems, is not new, of course. Within the field of algorithmic fairness, Corbett-Davies and Goel[23] and Liu et al.[68] both devote most of their analyses to the person-impacting effects of classificatory systems.

The techniques that we use to translate fair learning outcomes into welfare paths are related to a number of existing works. The proxy fairness constraint in our instantiation of the ε-fair SVM problem original appeared in Zafar et al.'s work on restricting the disparate impact of machine classifiers.[96] Their research introduces this particular proxy fairness constrained program and shows that it can be efficiently solved and well approximates target fairness constraints. We use the constraint to demonstrate our overall findings about the effect of fairness criteria on individual and group welfares. We share some of the preliminary formulations of the fair SVM problem with Donini et al.[28] though they focus on the statistical and fairness guarantees of the generalized ERM program. Though this area seems far afield from questions of fairness and welfare, our analysis on the effect of $\Delta\varepsilon$ fairness perturbations on welfare makes use of these general methods.[26,46,90,89,54]

4.2 Problem Formalization

Our framework and results are motivated by those algorithmic use cases in which considerations of fairness and welfare stand alongside those of efficiency. Because our chapter connects machine classification and notions of algorithmic fairness with conceptions of social welfare, we first provide an overview of the notation and assumptions that feature throughout our work.

In the empirical loss minimization problem, a learner seeks a classifier h that issues the most accurate predictions when trained on set of n data points $\{\mathbf{x}_i, z_i, y_i\}_{i=1}^n$. Each triple gives an individual's feature vector $\mathbf{x}_i \in \mathcal{X}$, protected class attribute $z_i \in \{0, 1\}$, and true label $y_i \in \{-1, +1\}$.[*] A classifier that assigns an incorrect label $h(\mathbf{x}_i) \neq y_i$ incurs a penalty.

The empirical risk minimizing predictor is given by

$$h^* := \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{n} \ell(h(\mathbf{x_i}), y_i)$$

where hypothesis $h : \mathcal{X} \to \mathbb{R}$ gives a learner's model, the loss function $\ell : \mathbb{R} \times \{-1, +1\} \to \mathbb{R}$ gives the penalty incurred by a prediction, and $\mathcal{H}$ is the hypothesis class under the learner's consideration. Binary classification systems issue predictions $h(\mathbf{x}) \in \{-1, +1\}$.

Notions of fairness have been formalized in a variety of ways in the machine learning literature. Though Dwork et al.'s[30] initial conceptualization remains prominent and influential, much work has since defined fairness as a parity notion applied across different protected class groups.[45,20,61,95,28,3] The following definition gives the general form of these types of fairness criteria.

Definition 5. *A classifier h satisfies a general group-based notion of* ε-fairness *if*

$$|\mathbb{E}[g(\ell, h, \mathbf{x}_i, y_i)|\mathcal{E}_{\mathbf{z}_i=1}] - \mathbb{E}[g(\ell, h, \mathbf{x}_i, y_i)|\mathcal{E}_{\mathbf{z}_i=0}]| \leqslant \varepsilon \tag{4.1}$$

where g is some function of classifier h performance, and $\mathcal{E}_{\mathbf{z}_i=0}$ *and* $\mathcal{E}_{\mathbf{z}_i=1}$ *are events that occur with respect to groups* $z = 0$ *and* $z = 1$ *respectively.*

Further specifications of the function g and the events $\mathcal{E}$ instantiate particular group-based fairness notions. For example, when $g(\ell, h, \mathbf{x}_i, y_i) = h(\mathbf{x}_i)$ and $\mathcal{E}_{\mathbf{z}_i}$ refers to the events in which $y_i = +1$ for each group $\mathbf{z}_i \in \{0, 1\}$, Definition 5 gives an ε-approximation of *equality of opportunity*.[45] When $g(\ell, h, \mathbf{x}_i, y_i) = \ell(h(\mathbf{x}_i), y_i)$ and $\mathcal{E}_{\mathbf{z}_i}$ refers to all classification events for each group $\mathbf{z}_i$, Definition

[*] Though individuals in a dataset will typically be coded with many protected class attributes, in this chapter we will consider only a single sensitive attribute of focus.

5 gives the notion of ε-approximation of *overall error rate balance*.[20] Notice that as ε increases, the constraint loosens, and the solution is considered "less fair." As ε decreases, the fairness constraint becomes more strict, and the solution is considered "more fair."

Mapping classification outcomes to changes in individuals' welfares gives a useful method of analysis for many data-based algorithmic systems that are involved in resource distribution pipelines. In particular, we consider tools that issue outcomes uniformly ranked, or preferred, by those individuals who are the subjects of the system. That is, individuals agree on which outcome is preferred. Examples of such systems abound: applicants for credit generally want to be found eligible; candidates for jobs generally want to be hired, or at least ranked highly in their pool. These realms are precisely those in which fairness considerations are urgent and where fairness-adjusted learning methods are most likely to be adopted.

4.3 Welfare Impacts of Fairness Constraints

The central inquiry of our work asks how fairness constraints as popularized in the algorithmic fairness community relate to welfare-based analyses that are dominant in economics and policy-making circles. Do fairness-adjusted optimization problems actually make marginalized groups better-off in terms of welfare? In this section, we work from an empirical risk minimization (ERM) program with generic fairness constraints parametrized by a tolerance parameter $\varepsilon > 0$ and trace individuals' and groups' welfares as a function of ε. We assume that an individual benefits from receiving a positive classification, and thus we define group welfare as

$$W_k = \frac{1}{n_k} \sum_{i|z_i=k} \frac{h(\mathbf{x}_i)+1}{2}, \qquad k \in \{0,1\} \tag{4.2}$$

where n_k give the number of individuals in group $z = k$. We note that W_k can be defined in ways other than (4.2), which assumes that positive classification are always and only welfare-enhancing.

Other work has considered the possibility that positive classifications may in fact make individuals worse-off if they are false positives.[68] The definition of W_k can be generalized to account for these cases.

First, in Section 4.3.1, we present an instantiation of the ϵ-fair ERM problem with a fairness constraint proposed in prior work in algorithmic fairness. We work from the Soft-Margin SVM program and derive the various dual formulations that will be of use in the following analyses. In Section 4.3.2, we move on to show how $\Delta\epsilon$ perturbations to the fairness constraint in the ϵ-fair ERM problem yield changes in classification outcomes for individuals and by extension, how they impact a group's overall welfare. Our approach, which draws a connection between fairness perturbations and searches for an optimal SVM regularization parameter, tracks changes in an individual's classification by taking advantage of the codependence of variables in the dual of the SVM. By perturbing the fairness constraint, we observe changes in not its own corresponding dual variable but in the corresponding dual of the margin constraints, which relay the classification fates of data points.

Leveraging this technique, we plot the "solution paths" of the dual variable as a function of ϵ, which in turn allows us to compute group welfares as a function of ϵ and draw out substantive results on the dynamics of how classification outcomes change in response to ϵ-fair learning. We prove that stricter fairness standards do not necessarily support welfare-enhancing outcomes for the disadvantaged group. In many such cases, the learning goal of ensuring group-based fairness is incompatible with the Pareto Principle.

Definition 6 (Pareto Principle)**.** *Let x, y be two social alternatives. Let $\succeq_i$ be the preference ordering of individuals $i \in [n]$, and $\succeq_P$ be the preference ordering of a social planner. The planner abides by the Pareto Principle if $x \succeq_P y$ whenever $x \succeq_i y$ for all i.*

In welfare economics, the Pareto Principle is a standard requirement of social welfare functionals—it would appear that the selection of an allocation that is Pareto dominated by an available alternative would be undesirable and even irresponsible! Nevertheless, we show that applying fairness criteria to loss minimization tasks in some cases do just that. We perform our analysis on the Soft-Margin SVM optimization problem and, for concreteness, work with a well-known fairness formulation in the literature. However, we note that our methods and results apply to fairness-constrained convex loss minimization programs more generally.

We also show that this method of analysis can form practical tools. In Section 4.3.3, we present a computationally efficient algorithmic implementation of our analyses, fitting full welfare solution paths for all $\varepsilon \in [0, 1]$ values in a time complexity that is on the same order as that of a single SVM fit. We close this section by working from the shadow price of the fairness constraint to derive local and global sensitivities of the optimal solution to $\Delta\varepsilon$ perturbations.

4.3.1 Setting up the ε-fair ERM program

The general fairness-constrained empirical loss minimization program can be written as

$$\begin{aligned} \underset{h \in \mathcal{H}}{\text{minimize}} \quad & \ell(h(\mathbf{x}), y) \\ \text{subject to} \quad & f_h(\mathbf{x}, y) \leqslant \varepsilon \end{aligned} \tag{4.3}$$

where $\ell(h(\mathbf{x}), y)$ gives the empirical loss of a classifier $h \in \mathcal{H}$ on the dataset $\mathcal{X}$. To maximize accuracy, the learner ought to minimize 0-1 loss; however because the loss function ℓ_{0-1} is non-convex, a convex surrogate loss such as hinge loss (ℓ_h) or log loss ($\ell_{\log}$) is frequently substituted in its place to ensure that globally optimal solutions may be efficiently found. $f_h(\mathbf{x}, y) \leqslant \varepsilon$ gives a group-based fairness constraint of the type given in Definition 5, where $\varepsilon > 0$ is the unfairness "tolerance parameter"—a greater ε permits a greater group disparity on a metric of interest; a smaller ε more tightly restricts the level of permissible disparity.

We examine the behavior of fairness-constrained linear SVM classifiers, though we note that our techniques generalize to nonlinear kernels SVMs, since interpretations of the dual of the SVM and the full SVM regularization path are the same with kernels.[46] Our learner minimizes hinge loss with L_1 regularization; equivalently, she seeks a Soft-Margin SVM that is "ε-fair." Both SVM models and "fair training" approaches are in broad circulation. The fair empirical risk minimization program is thus given as

$$\begin{aligned} \underset{\boldsymbol{\theta},\, b}{\text{minimize}} \quad & \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & y_i(\boldsymbol{\theta}^\top \mathbf{x}_i + b) - 1 + \xi_i \geqslant 0, \\ & \xi_i \geqslant 0, \\ & f_{\boldsymbol{\theta},b}(\mathbf{x}, y) \leqslant \varepsilon \end{aligned} \tag{ε-fair Soft-SVM}$$

where the learner seeks SVM parameters $\boldsymbol{\theta}, b$; ξ_i are non-negative slack variables that violate the margin constraint in the Hard-Margin SVM problem $y_i(\boldsymbol{\theta}^\top \mathbf{x}_i + b) - 1 \geqslant 0$, and $C > 0$ is a hyperparameter tunable by the learner to express the trade-off between preferring a larger margin and penalizing violations of the margin. $f_{\boldsymbol{\theta},b}(\mathbf{x}, y)$ is the group parity-based fairness constraint.

The abundant literature on algorithmic fairness presents a long menu of options for the various forms that $f_{\boldsymbol{\theta},b}$ could take, but generally speaking, the constraints are non-convex. As such, much work has enlisted methods that depart from directly pursuing efficient constraint-based convex programming techniques in order to solve them.[53,11,3,95] Researchers have also devised convex proxy alternatives, which have been shown to approximate the intended outcomes of original fairness constraints well.[96,28,93] In particular, in this chapter, we work with the proxy constraint proposed by Zafar et al.,[96] which constrains disparities in covariance between group membership and the (signed) distance between individuals' feature vectors and the hyperplane decision boundary:

$$f_{\boldsymbol{\theta},b}(\mathbf{x},y) = |\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(\boldsymbol{\theta}^\intercal \mathbf{x}_i + b)| \leqslant \varepsilon \tag{4.4}$$

$\bar{z}$ reflects the bias in the demographic makeup of $\mathcal{X}$: $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$. Let ($\varepsilon$-fair-SVM1-P) be the Soft-Margin SVM program with this covariance constraint. The corresponding Lagrangian is

$$\begin{aligned}\mathcal{L}_P(\boldsymbol{\theta}, b, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \gamma_1, \gamma_2) = \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\lambda_i - \sum_{i=1}^{n}\mu_i(y_i(\boldsymbol{\theta}^\intercal \mathbf{x}_i + b) - 1 + \xi_i) \\ - \gamma_1\big(\varepsilon - \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(\boldsymbol{\theta}^\intercal \mathbf{x}_i + b)\big) \\ - \gamma_2\big(\varepsilon - \frac{1}{n}\sum_{i=1}^{n}(\bar{z} - z_i)(\boldsymbol{\theta}^\intercal \mathbf{x}_i + b)\big)\end{aligned} \tag{ε-fair-SVM1-L}$$

where $\boldsymbol{\theta} \in \mathbb{R}^d, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^n$ are primal variables. The (non-negative) Lagrange multipliers $\boldsymbol{\lambda}, \boldsymbol{\mu} \in \mathbb{R}^n$ correspond to the n non-negativity constraints $\xi_i \geqslant 0$ and the margin-slack constraints $y_i(\boldsymbol{\theta}^\intercal \mathbf{x}_i + b) - 1 + \xi_i \geqslant 0$ respectively. The multipliers $\gamma_1, \gamma_2 \in \mathbb{R}$ correspond to the two linearized forms of the absolute value fairness constraint. By complementary slackness, dual variables reveal information about the satisfaction or violation of their corresponding constraints. The analyses in the subsequent two subsections will focus on these interpretations.

By the Karush-Kuhn-Tucker (KKT) conditions, at the solution of the convex program, the gradients of $\mathcal{L}$ with respect to $\boldsymbol{\theta}$, b, and ξ_i are zero. Plugging in these conditions, the dual Lagrangian is

$$\mathcal{L}_D(\boldsymbol{\mu}, \gamma) = -\frac{1}{2}\|\sum_{i=1}^{n}\mu_i y_i \mathbf{x}_i - \frac{\gamma}{n}\sum_{i=1}^{n}(z_i - \bar{z})\mathbf{x}_i\|^2 + \sum_{i=1}^{n}\mu_i - |\gamma|\varepsilon \tag{4.5}$$

where $\gamma = \gamma_1 - \gamma_2$. The dual maximizes this objective subject to the constraints $\mu_i \in [0, C]$ for all

$i \in [n]$ and $\sum_{i=1} \mu_i y_i = 0$. We thus arrive at the Wolfe dual problem

$$\begin{aligned} \underset{\boldsymbol{\mu}, \gamma, V}{\text{maximize}} \quad & -\frac{1}{2} \|\sum_{i=1}^{n} \mu_i y_i \mathbf{x}_i - \frac{\gamma}{n} \sum_{i=1}^{n} (z_i - \bar{z}) \mathbf{x}_i\|^2 + \sum_{i=1}^{n} \mu_i - V\varepsilon \\ \text{subject to} \quad & \mu_i \in [0, C], \qquad i = 1, \ldots, n, \qquad (\varepsilon\text{-fair-SVM1-D}) \\ & \sum_{i=1}^{n} \mu_i y_i = 0, \\ & \gamma \in [-V, V] \end{aligned}$$

where we have introduced the variable V to eliminate the absolute value function $|\gamma|$ in the objective. Notice that when $\gamma = 0$ and neither of the constraints bind, we recover the standard dual SVM program. Since we are concerned with fair learning that does alter an optimal solution, we consider cases where V is strictly positive. We introduce additional dual variables β_- and β_+, corresponding to the $\gamma \in [-V, V]$ constraint and derive the Lagrangian

$$\begin{aligned} \mathcal{L}(\boldsymbol{\mu}, \gamma, V, \beta_-, \beta_+) = & -\frac{1}{2} \|\sum_{i=1}^{n} \mu_i y_i \mathbf{x}_i - \frac{\gamma}{n} \sum_{i=1}^{n} (z_i - \bar{z}) \mathbf{x}_i\|^2 + \sum_{i=1}^{n} \mu_i \\ & - V\varepsilon + \gamma(\beta_- - \beta_+) + V(\beta_- + \beta_+) \end{aligned}$$

Under KKT conditions, $\beta_- + \beta_+ = \varepsilon$ and

$$\gamma^* = \frac{n(n(\beta_- - \beta_+) + \sum_{i=1}^{n} \mu_i y_i \langle \mathbf{x}_i, \mathbf{u} \rangle)}{\|\mathbf{u}\|^2} \tag{4.6}$$

where $\mathbf{u} = \sum_{i=1}^{n} (z_i - \bar{z}) \mathbf{x}_i$ geometrically gives some group-sensitive "average" of $\mathbf{x} \in \mathcal{X}$. We can now rewrite ($\varepsilon$-fair-SVM1-D) as

$$
\begin{aligned}
\underset{\boldsymbol{\mu},\beta_-,\beta_+}{\text{maximize}} \quad & -\frac{1}{2}\|\sum_{i=1}^{n}\mu_i y_i(I-P_{\mathbf{u}})\mathbf{x}_i\|^2 + \sum_{i=1}^{n}\mu_i + \frac{2n\sum_i \mu_i y_i\langle \mathbf{x}_i, \mathbf{u}\rangle + n^2(\beta_- - \beta_+)}{2\|\mathbf{u}\|^2}(\beta_- - \beta_+) \\
\text{subject to} \quad & \mu_i \in [0, C], \qquad i = 1, \ldots, n, \\
& \sum_{i=1}^{n}\mu_i y_i = 0, \qquad\qquad (\varepsilon\text{-fair SVM2-D}) \\
& \beta_-, \beta_+ \geqslant 0, \\
& \beta_- + \beta_+ = \varepsilon
\end{aligned}
$$

where $I, P_{\mathbf{u}} \in \mathbb{R}^{d\times d}$. The former is the identity matrix, and the latter is the projection matrix onto the vector $\mathbf{u}$. As was also observed by Donini et al., the $\varepsilon = 0$ version of (ε-fair SVM2-D) is equivalent to the standard formulation of the dual SVM program with Kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \langle (I-P_{\mathbf{u}})\mathbf{x}_i, (I-P_{\mathbf{u}})\mathbf{x}_j\rangle$.[28]

Since we are interested in the welfare impacts of fair learning when fairness constraints *do* have an impact on optimal solutions, we will assume that the fairness constraint binds. For clarity of exposition, we assume that the positive covariance constraint binds, and thus that $\beta_- = 0$ and $\beta_+ = \varepsilon$ in (ε-fair SVM2-D). This is without loss of generalization—the same analyses apply when the negative covariance constraint binds. The dual ε-fair SVM program becomes

$$
\begin{aligned}
\underset{\boldsymbol{\mu}}{\text{minimize}} \quad & \frac{1}{2}\|\sum_{i=1}^{n}\mu_i y_i(I-P_{\mathbf{u}})\mathbf{x}_i\|^2 - \sum_{i=1}^{n}\mu_i + \frac{n\varepsilon(2\sum_i \mu_i y_i\langle \mathbf{x}_i, \mathbf{u}\rangle - n\varepsilon)}{2\|\mathbf{u}\|^2} \\
\text{subject to} \quad & \mu_i \in [0, C], \qquad i = 1, \ldots, n, \qquad\qquad (\varepsilon\text{-fair SVM-D}) \\
& \sum_{i=1}^{n}\mu_i y_i = 0
\end{aligned}
$$

We will work from this formulation of the constrained optimization problem for the remainder of the chapter.

4.3.2 Impact of Fair Learning on Individuals' Welfares

We now move on to investigate the effects of perturbing a fixed ε-fair SVM by some $\Delta\varepsilon$ on the classification outcomes that are issued. We ask, *"How are individuals' and groups' classifications, and thus their welfares, impacted when a learner tightens or loosens a fairness constraint?"* The key insight that drives our methods and results is that rather than perform sensitivity analysis directly on the dual variable corresponding to the fairness constraint—which, as we will see in Section 4.3.4, only gives information about the change in the learner's objective value—we track changes in the classifier's behavior by analyzing the effect of $\Delta\varepsilon$ perturbations on another set of dual variables: μ_i that correspond to the primal margin constraints. Each of these n dual variables indicate whether its corresponding vector $\mathbf{x}_i$ is correctly classified, lies in the margin, or incorrectly classified. Leveraging how these μ_i change as a function of ε thereby allows us to track the solution paths of individual points and by extension, compute group welfare paths.

Define a function $p(\varepsilon) : \mathbb{R} \rightarrow \mathbb{R}$ that gives the optimal value of the ε-fair loss minimizing program in (ε-fair SVM1-P), which by duality is also the optimal value of (ε-fair SVM-D). We begin at a solution $p(\varepsilon)$ and consider changes in classifications at the solution $p(\varepsilon + \Delta\varepsilon)$, where $\Delta\varepsilon$ are perturbations can be positive or negative, so long as $\varepsilon + \Delta\varepsilon > 0$. At an optimal solution, the classification fate of each data point $\mathbf{x}_i$ is encoded in the dual variable μ_i^*, which is a function of ε. $\mu_i(\varepsilon)$ is the ε-parameterized solution path of μ_i such that at any particular solution $p(\varepsilon)$, the optimal value of the dual variable $\mu_i^* = \mu_i(\varepsilon)$. As a slight abuse of notation, we reserve notation $\mu_i(\varepsilon)$ for the functional form of the solution path and write μ_i^ε; to refer to the value of the dual variable at a given ε.

Lemma 3. *The dual variable paths $\mu_i(\varepsilon)$ for all $i \in [n]$ are piecewise linear in ε.*

Though this lemma seems merely of technical interest, it is a workhorse result for both our methodological contributions—our analytical results and our computationally efficient algorithm, which converts fairness constraints to welfare paths—as well as our substantive fairness results about how fairness perturbations impact individual and learner welfares. The algorithm we present in Section 4.3.3, performs full welfare analysis for all values of ε in a computationally efficient manner by taking advantage of the piecewise linear form of individual and group welfares. Piecewise linearity also sets the stage for the later substantive results about the tension between fairness improvements and the Pareto Principle. We thus walk through the longer proof of this key result in the main text of the chapter as it provides important exposition, definitions, and derivations for subsequent results.

Proof. Let D^ε be the value of the objective function in (ε-fair SVM-D). By the dual formulation of the Soft-Margin SVM, we can use the value of $\frac{\partial D^\varepsilon}{\partial \mu_j}$ to partition the set of indices $j \in [n]$ in a way that corresponds to the classification fates of individual vectors $\mathbf{x}_j$ at the optimal solution:

$$\frac{\partial D^\varepsilon}{\partial \mu_j} > 0 \longrightarrow \mu_j^\varepsilon = 0, \text{ and } j \in \mathcal{F}^\varepsilon \tag{4.7}$$

$$\frac{\partial D^\varepsilon}{\partial \mu_j} = 0 \longrightarrow \mu_j^\varepsilon \in [0, C], \text{ and } j \in \mathcal{M}^\varepsilon \tag{4.8}$$

$$\frac{\partial D^\varepsilon}{\partial \mu_j} < 0 \longrightarrow \mu_j^\varepsilon = C, \text{ and } j \in \mathcal{E}^\varepsilon \tag{4.9}$$

Hence, $\mathbf{x}_j$ are either correctly classified free vectors (4.7), vectors in the margin (4.8), or error vectors (4.9). We track membership in these sets by letting $\{\mathcal{F}, \mathcal{M}, \mathcal{E}\}^\varepsilon$ be the index set partition at the ε-fair solution. To analyze the impact that applying a fairness constraint has on individuals' or groups' welfares, we track the behavior of $\frac{\partial D^\varepsilon}{\partial \mu_j}$ and observe how vector index membership in sets $\mathcal{F}^\varepsilon$, $\mathcal{M}^\varepsilon$, and $\mathcal{E}^\varepsilon$ change under a perturbation to ε. This information will in turn reveal how classifications change or remain stable upon tightening or loosening the fairness constraint.

Fairness perturbations do not always shuffle data points across the different membership sets $\mathcal{F}^\varepsilon$, $\mathcal{M}^\varepsilon$, and $\mathcal{E}^\varepsilon$. It is clear that for $j \in \{\mathcal{F}, \mathcal{E}\}^\varepsilon$, so long as a perturbation of $\Delta\varepsilon$ does not cause $\frac{\partial D^\varepsilon}{\partial \mu_j}$ to flip signs or to vanish to 0, j will belong to the same set and $h^\varepsilon(\mathbf{x}_j) = h^{\varepsilon+\Delta\varepsilon}(\mathbf{x}_j)$ where $h^\varepsilon(\mathbf{x}_j)$ gives the ε-fair classification outcome for $\mathbf{x}_j$. In these cases, an individual's welfare is unaffected by the change in the fairness tolerance level from ε to $\varepsilon + \Delta\varepsilon$.

In contrast, vectors $\mathbf{x}_j$ with $j \in \mathcal{M}^\varepsilon$ are subject to a different condition to ensure that they stay in the margin: $\frac{\partial D^\varepsilon}{\partial \mu_j} = \frac{\partial D^{\varepsilon+\Delta\varepsilon}}{\partial \mu_j} = 0$, i.e., perturbing by $\Delta\varepsilon$ does not lead to any changes in $\frac{\partial D^\varepsilon}{\partial \mu_j}$:

$$\frac{\partial D^\varepsilon}{\partial \mu_j} = \sum_{i=1}^{n} \mu_i y_i (I - P_\mathbf{u})\mathbf{x}_i y_j (I - P_\mathbf{u})\mathbf{x}_j + \frac{n\varepsilon y_j \langle \mathbf{x}_j, \mathbf{u} \rangle}{||\mathbf{u}||^2} + b y_j - 1 = 0 \tag{4.10}$$

for all $j \in \mathcal{M}^\varepsilon$. Let $r_j^{\varepsilon,\Delta\varepsilon}$ be the change in μ_j^ε upon perturbing ε by $\Delta\varepsilon$, then we have

$$\mu_j^{\varepsilon+\Delta\varepsilon} = \mu_j^\varepsilon + r_j^{\varepsilon,\Delta\varepsilon} \tag{4.11}$$

recalling that μ_j^ε is the value of μ_j at the optimal solution $p(\varepsilon)$. Let $\boldsymbol{r}^{\varepsilon,\Delta\varepsilon} \in \mathbb{R}^{n+1}$ be the vector of μ_i^ε sensitivities to perturbations $\Delta\varepsilon$ with $r_0^{\varepsilon,\Delta\varepsilon}$ as the change in the offset b. For all unshuffled $j \in \mathcal{M}^\varepsilon$, we can compute $r_j^{\varepsilon,\Delta\varepsilon}$ by taking the finite difference of (4.10) with respect to a $\Delta\varepsilon$ perturbation,

$$\sum_{i=1}^{n} r_i^{\varepsilon,\Delta\varepsilon} y_i y_j \langle (I - P_\mathbf{u})\mathbf{x}_i, (I - P_\mathbf{u})\mathbf{x}_j \rangle + r_0^{\varepsilon,\Delta\varepsilon} y_j = \frac{-n y_j \Delta\varepsilon}{||\mathbf{u}||^2} \langle \mathbf{u}, \mathbf{x}_j \rangle$$

It is clear that $r_i^{\varepsilon,\Delta\varepsilon} = 0$ for all i that are left unshuffled in the partition $\{\mathcal{F}, \mathcal{E}\}^\varepsilon$. For these "stable ranges" where no i changes its index set membership, we can simplify the previous expression by summing over only those $r_i^{\varepsilon,\Delta\varepsilon}$ where $i \in \mathcal{M}^\varepsilon$:

$$\sum_{i\in\mathcal{M}^\varepsilon} r_i^{\varepsilon,\Delta\varepsilon} y_i y_j \langle (I - P_{\mathbf{u}})\mathbf{x}_i, (I - P_{\mathbf{u}})\mathbf{x}_j \rangle + r_0^{\varepsilon,\Delta\varepsilon} y_j = \frac{-ny_j\Delta\varepsilon}{\|\mathbf{u}\|^2}\langle \mathbf{u}, \mathbf{x}_j \rangle$$

Thus we can compute $r_i^{\varepsilon,\Delta\varepsilon}$ by inverting the matrix

$$K^\varepsilon = \left(\begin{array}{c|c} 0 & y_1 \quad y_2 \quad \cdots \qquad y_{|\mathcal{M}^\varepsilon|} \\ \hline y_1 & \\ \vdots & y_i y_j \langle (I - P_{\mathbf{u}})\mathbf{x}_i, (I - P_{\mathbf{u}})\mathbf{x}_j \rangle \\ y_2 & \\ y_{|\mathcal{M}^\varepsilon|} & \end{array}\right) \in \mathbb{R}^{(|\mathcal{M}^\varepsilon|+1)\times(|\mathcal{M}^\varepsilon|+1)} \tag{4.12}$$

where indices are renumbered to only reflect $i, j \in \mathcal{M}^\varepsilon$. This matrix is invertible so long as the margin is not empty and the Kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \langle (I - P_{\mathbf{u}})\mathbf{x}_i, (I - P_{\mathbf{u}})\mathbf{x}_j \rangle$ forms a positive definite matrix. Since the objective function in (ε-fair SVM-D) is quadratic, a sufficient condition for K^ε to be invertible is that the objective is strictly convex—we assume this as a technical condition.[†] The

[†] We mention the case in which the margin is empty in Section 3.3, though we refer the interested reader to the Appendix for a full exposition of how μ_j^ε are updated when the margin is empty and as a result, we cannot compute how i move across index sets via the sensitivities $\boldsymbol{r}$.

sensitivities of μ_j^ε for $j \in \mathcal{M}^\varepsilon$ to $\Delta\varepsilon$ perturbations are given by

$$\boldsymbol{r}^{\varepsilon,\Delta\varepsilon} = \underbrace{(K^\varepsilon)^{-1}\Big(\frac{-n}{\|\mathbf{u}\|^2}\mathbf{v}\Big)}_{\boldsymbol{r}^\varepsilon}\Delta\varepsilon, \quad \text{where } \mathbf{v} = \begin{bmatrix} 0 \\ \vdots \\ y_j\langle \mathbf{u}, \mathbf{x}_j\rangle \\ \vdots \end{bmatrix} \in \mathbb{R}^{|\mathcal{M}^\varepsilon|+1} \tag{4.13}$$

Plugging this back into (4.11), we have

$$\mu_j^{\varepsilon+\Delta\varepsilon} = \mu_j^\varepsilon + \underbrace{\Big((K^\varepsilon)^{-1}\big(\frac{-n}{\|\mathbf{u}\|^2}\mathbf{v}\big)\Big)_j}_{r_j^\varepsilon}\Delta\varepsilon \tag{4.14}$$

Hence, for all $j \in \mathcal{M}^\varepsilon$ that stay in the margin, the solution path function $\mu_j(\varepsilon)$ is linear in ε. For $j \in \{\mathcal{F}, \mathcal{E}\}^\varepsilon$ that stay in their partition sets, $\mu_j(\varepsilon + \Delta\varepsilon) = \mu_j(\varepsilon)$, so the function is constant.

When $\Delta\varepsilon$ perturbations do result in changes in the partition, there are four ways that indices could be shuffled across sets:

1. $j \in \mathcal{E}^\varepsilon$ moves into $\mathcal{M}^{\varepsilon+\Delta\varepsilon}$
2. $j \in \mathcal{F}$ moves into $\mathcal{M}^{\varepsilon+\Delta\varepsilon}$
3. $j \in \mathcal{M}^\varepsilon$ moves into $\mathcal{F}^{\varepsilon+\Delta\varepsilon}$
4. $j \in \mathcal{M}^\varepsilon$ moves into $\mathcal{E}^{\varepsilon+\Delta\varepsilon}$

Since index transitions only occur by way of changes to the margin, we need now only confirm that each of these transitions maintains continuous $\mu_j(\varepsilon)$ paths for all $j \in [n]$ in order to conclude the proof that the paths are piecewise-linear. □

The linearity of paths $\mu_j(\varepsilon)$ for $j \in \mathcal{M}^\varepsilon$ gives conditions on the ranges of ε wherein individuals' classification outcomes do not change. As such, for any given tolerance parameter ε, we can compute the $\Delta\varepsilon$ perturbations that yield no changes to individuals' welfares. The following Proposition gives the analytical form of these stable regions, where although fairness appears to be "improving" or "worsening," the adjusted learning process has no material effects on the classificatory outcomes that individuals receive.

Proposition 7. *Denote the optimal μ_j^* values at an ε-fair SVM solution as μ_j^ε for $j \in [n]$. Let*

$$
\begin{aligned}
r_j &= \Big((K^\varepsilon)^{-1}\big(\frac{-n}{\|\mathbf{u}\|^2}\mathbf{v}\big)\Big)_j \quad \textit{with } K^\varepsilon \textit{ and } \mathbf{v} \textit{ as defined in (4.12) and (4.13)},\\
d_j &= \sum_{i \in \mathcal{M}^\varepsilon} r_i y_i y_j \langle (I - P_\mathbf{u})\mathbf{x}_i, (I - P_\mathbf{u})\mathbf{x}_j \rangle + r_0 y_j\\
g_j &= 1 - \Big(\sum_{i=1}^{n} \mu_i^\varepsilon y_i (I - P_\mathbf{u})\mathbf{x}_i y_j (I - P_\mathbf{u})\mathbf{x}_j + \frac{n\varepsilon y_j \langle \mathbf{x}_j, \mathbf{u}\rangle}{\|\mathbf{u}\|^2} + b y_j\Big)
\end{aligned}
\tag{4.15}
$$

All perturbations of ε in the range $\Delta\varepsilon \in \big(\max_j m_j, \min_j M_j\big)$ where

$$
m_j = \begin{cases} \begin{cases} \frac{g_j}{d_j}, & j \in \mathcal{F}^\varepsilon, d_j > 0 \\ -\infty, & j \in \mathcal{F}^\varepsilon, d_j < 0 \end{cases} \\ \min\{\frac{C - \mu_j^\varepsilon}{r_j}, \frac{-\mu_j^\varepsilon}{r_j}\}, \; j \in \mathcal{M}^\varepsilon \\ \begin{cases} -\infty, & j \in \mathcal{E}^\varepsilon, d_j > 0 \\ \frac{g_j}{d_j}, & j \in \mathcal{E}^\varepsilon, d_j < 0 \end{cases} \end{cases}
\quad
M_j = \begin{cases} \begin{cases} \infty, & j \in \mathcal{F}^\varepsilon, d_j > 0 \\ \frac{g_j}{d_j}, & j \in \mathcal{F}^\varepsilon, d_j < 0 \end{cases} \\ \min\{\frac{C - \mu_j^\varepsilon}{r_j}, \frac{-\mu_j^\varepsilon}{r_j}\}, \; j \in \mathcal{M}^\varepsilon \\ \begin{cases} \frac{g_j}{d_j}, & j \in \mathcal{E}^\varepsilon, d_j > 0 \\ \infty, & j \in \mathcal{E}^\varepsilon, d_j < 0 \end{cases} \end{cases}
\tag{4.16}
$$

yield no changes to index memberships in the partition $\{\mathcal{F}, \mathcal{M}, \mathcal{E}\}^\varepsilon$.

We defer the interested reader to the Appendix for the full proof of this Proposition, though we provide a sketch here. The result follows from observing that the sensitivities $r_i^\varepsilon \neq 0$ for $i \in \mathcal{M}^\varepsilon$ defined in (4.13) affect the values $\frac{\partial D^\varepsilon}{\partial \mu_j}$ for all $j \in [n]$, and additional conditions must hold to ensure that the vectors that are not on the margin are also unshuffled by the fairness perturbation. Define

$$g_j^\varepsilon = 1 - \Big(\sum_{i=1}^{n} \mu_i^\varepsilon y_i (I - P_{\mathbf{u}})\mathbf{x}_i y_j (I - P_{\mathbf{u}})\mathbf{x}_j + \frac{n\varepsilon y_j \langle \mathbf{x}_j, \mathbf{u}\rangle}{\|\mathbf{u}\|^2} + b y_j\Big) \tag{4.17}$$

$$d_j^\varepsilon = \frac{\partial D^\varepsilon}{\partial \mu_j \partial \varepsilon} = \sum_{i \in \mathcal{M}^\varepsilon} r_i^\varepsilon y_i y_j \langle (I - P_{\mathbf{u}})\mathbf{x}_i, (I - P_{\mathbf{u}})\mathbf{x}_j\rangle + r_0^\varepsilon y_j \tag{4.18}$$

The $\Delta\varepsilon$ condition for stability of vectors $\mathbf{x}_j$ for $j \notin \mathcal{M}^\varepsilon$ is given by

$$\frac{g_j^\varepsilon}{d_j^\varepsilon} \tag{4.19}$$

Recall the conditions of membership in sets $\mathcal{F}$ and $\mathcal{E}$ as given in (4.7) and (4.9) respectively. The following observations are critical to computing the bounds of the stable region:

For $j \in \mathcal{F}^\varepsilon$, perturbations $\Delta\varepsilon$ that increase g_j^ε do not threaten j's exiting the set; if $\Delta\varepsilon$ decreases g_j^ε, then j can enter $\mathcal{M}^{\varepsilon+\Delta\varepsilon}$.

Inversely, for $j \in \mathcal{E}^\varepsilon$, perturbations $\Delta\varepsilon$ that decrease g_j^ε ensure that j stays in the same partition, i.e., $j \in \mathcal{E}^{\varepsilon+\Delta\varepsilon}$. Perturbations that increase g_j^ε can cause j to shuffle into $\mathcal{M}^{\varepsilon+\Delta\varepsilon}$.

For $j \in \mathcal{M}^\varepsilon$ to stay in the margin, we need $\mu_j^{\varepsilon+\Delta\varepsilon} \in [0, C]$. Once μ_j^ε hits either endpoint of the interval, j risks shuffling across to $\mathcal{F}^{\varepsilon+\Delta\varepsilon}$ or $\mathcal{E}^{\varepsilon+\Delta\varepsilon}$.

Computing these transition inequalities results in a set of conditions that ensure that a partition is stable. Since $\Delta\varepsilon$ can be either positive or negative, we take the maximum of the lower bounds (m_j) and the minimum of the upper bounds (M_j) to arrive at the range of stable perturbations given in (4.16). We call the bounds of this interval the "breakpoints" of the solution paths.

This Proposition reveals a mismatch between the ostensible changes to the fairness level of an ε-fair Soft-Margin SVM learning process and the actual felt changes in outcomes by the individuals who are subject to the system. This results from the simple fact that the optimization problem captures changes in the learner's optimal solution but does not offer such fine-grained information on how individuals' outcomes vary as a result of $\Delta\varepsilon$ perturbations. So long as the fairness constraint is binding and its associated dual variable $\gamma > 0$, then tightening or loosening a fairness constraint *does* alter the loss of the optimal learner classifier—the actual SVM solution changes—yet analyzed from the perspective of the individual agents $\mathbf{x}_i$, so long as the $\Delta\varepsilon$ perturbation occurs within the range given by (4.16), classifications issued under this $\varepsilon + \Delta\varepsilon$-fair SVM solution are identical to those under the ε-fair solution. Thus despite the apparent more "fair" signal that a classifier abiding by $\varepsilon + \Delta\varepsilon < \varepsilon$ sends, agents are made no better off in terms of welfare. This result is summarized in the following Corollary.

Corollary 3. *Let $\{p(\varepsilon), W_0(\varepsilon), W_1(\varepsilon)\}$ be a triple expressing the welfares of the learner, group $z = 0$, and group $z = 1$ under the ε-fair SVM solution. Then for any $\Delta\varepsilon \in (\max_j m_j, 0)$ where m_j is defined in (4.16), $\{p(\varepsilon), W_0(\varepsilon), W_1(\varepsilon)\} \gtrsim \{p(\varepsilon + \Delta\varepsilon), W_0(\varepsilon + \Delta\varepsilon), W_1(\varepsilon + \Delta\varepsilon)\}$.*

Once we have demarcated the limits of $\Delta\varepsilon$ perturbations that yield no changes to the partition, i.e., $\{\mathcal{F}, \mathcal{M}, \mathcal{E}\}^{\varepsilon} = \{\mathcal{F}, \mathcal{M}, \mathcal{E}\}^{\varepsilon+\Delta\varepsilon}$, we can move on to consider the welfare effects of $\Delta\varepsilon$ perturbations that exceed the stable region outlined in Proposition 7. At each such breakpoint when $\Delta\varepsilon$ reaches $\max_j m_j$ or $\min_j M_j$ as defined in (4.16), the margin set changes: $\mathcal{M}^{\varepsilon} \neq \mathcal{M}^{\varepsilon+\Delta\varepsilon}$. As such, $r_j^{\varepsilon+\Delta\varepsilon}$ for $j \in \mathcal{M}^{\varepsilon+\Delta\varepsilon}$ must be recomputed via (4.13). These sensitivities hold until the next breakpoint when the set $\mathcal{M}$ updates again.

We can associate a group welfare with the classification scheme at each of the breakpoints. As already illustrated, index partitions are static in the stable regions around each breakpoint, so group welfares will also be unchanged in these regions. As such, we need only compute welfares at breakpoints to characterize the paths for $\varepsilon \in [0, 1]$. This method of analysis allows practitioners to straightforwardly determine whether the next ε breakpoint actually translates into better or worse outcomes for the group as a whole.

Of the four possible events that occur at a breakpoint, index transitions between the partitions

$\mathcal{M}$ and $\mathcal{E}$ correspond to changed classifications that affect group utilities. The following Proposition characterizes those breakpoint transitions that effect welfares triples $\{p(\varepsilon), W_0(\varepsilon), W_1(\varepsilon)\}$ for the learner, group $z = 0$, and group $z = 1$, that are *strictly Pareto dominated* by the welfare triple at a neighboring ε breakpoint. The full proof is left to the Appendix.

Proposition 8. *Consider the welfare triple at the optimal ε-fair SVM solution given by $\{p(\varepsilon), W_0(\varepsilon), W_1(\varepsilon)\}$. Let $b_L = \max_j m_j < 0$ be the neighboring lower breakpoint where index $\ell = \arg\max_j m_j$; let $b_U = \min_j M_j > 0$ be the neighboring upper breakpoint where index $u = \arg\min_j M_j$, assuming uniqueness in the* arg max *and* arg min. *If $\ell \in \mathcal{E}^\varepsilon$ and $y_\ell = -1$, or if $\ell \in \mathcal{M}^\varepsilon$ and $y_\ell = +1$, then*

$$\{p(\varepsilon + b_L), W_0(\varepsilon + b_L), W_1(\varepsilon + b_L)\}\} \prec \{p(\varepsilon), W_0(\varepsilon), W_1(\varepsilon)\}$$

If $u \in \mathcal{E}^\varepsilon$ and $y_u = +1$, or if $u \in \mathcal{M}^\varepsilon$ and $y_u = -1$, then

$$\{p(\varepsilon), W_0(\varepsilon), W_1(\varepsilon)\} \prec \{p(\varepsilon + b_U), W_0(\varepsilon + b_U), W_1(\varepsilon + b_U)\}$$

Thus minimizing loss in the presence of stricter fairness constraints need not correspond to monotonic gains or losses in the welfare levels of social groups. Fairness perturbations do not have a straightforward effect on classifications. Further, these results do not only arise as an unfortunate outcome of using the particular proxy fairness constraint suggested by Zafar et al.[96] So long as the ε parameter appears in the linear part of the dual Soft-Margin SVM objective function, the $\mu_j(\varepsilon)$ paths exhibit a piecewise linear form characterized by stable regions and breakpoints. Hence, these results apply to many proxy fairness criteria that have so far been proposed in the literature.[28,93,96] Even when the dual variable paths are not piecewise linear, so long as they are non-monotonic, fairer classification outcomes do not necessarily confer welfare benefits to the disadvantaged group. Monotonicity in welfare space is mathematically distinct from monotonicity in fairness space.

The preceding analyses show that although fairness constraints are often intended to improve classification outcomes for some disadvantaged group, they in general do not abide by the Pareto Principle, a common welfare economic axiom for deciding among social alternatives. That is, asking that an algorithmic procedure abide by a more stringent fairness criteria can lead to enacting classification schemes that actually make every stakeholder group worse-off. Here, the supposed "improved fairness" achieved by decreasing the unfairness tolerance parameter ε fails to translate into any meaningful improvements in the number of desirable outcomes issued to members of either group.

Theorem 5. *Consider two fairness-constrained ERM programs parameterized by ε_1 and ε_2 where $\varepsilon_1 < \varepsilon_2$. Then a decision-maker who always prefers the classification outcomes issued under the "more fair" ε_1-fair solution to those under the "less fair" ε_2-fair solution does not abide by the Pareto Principle.*

4.3.3 Algorithm and Complexity

We build upon the previous section of translating fairness constraints into individual welfare outcomes by considering the operationalization of our analysis and its practicality. The algorithmic procedure presented in this section computes ε breakpoints and tracks the solution paths of the $\mu_j(\varepsilon)$ for all individuals. Hence, the procedure enables the comparison of different social groups' welfares—where welfare is determined by the machine's allocative outcome—by aggregating the classification outcomes of all individuals j in a group z. Algorithm 1 outputs two useful fairness-relevant constructs that have as yet not been explored in the literature: 1) solution paths $\mu_j(\varepsilon)$ for $j \in [n]$ tracking individuals' welfares, and 2) full ε parameterized curves tracking groups' welfares.

The analysis of the previous section forms the backbone of the main update rules that construct the $\mu_j(\varepsilon)$ paths in Algorithm 1. In particular the values $r_j^\varepsilon, g_j^\varepsilon$, and d_j^ε as defined in (4.13), (4.17), and (4.18) respectively are key to computing the ε breakpoints, which in turn fully determine the piecewise linear form of $\mu_j(\varepsilon)$. There is, however, one corner case that the procedure must check that was not discussed in the preceding section. We had previously required that the matrix K^ε be invertible, which is the case whenever our objective function is strictly convex. But if the margin is empty, the standard update procedure, which computes sensitivities r_j^ε and K^ε, will not suffice. The KKT optimality condition $\sum_{i=1}^n \mu_i y_i = 0$ requires that the multiple indices moving in the margin at once must be positive and negative examples. For this reason we must refer to a different procedure to compute the ε breakpoint at which this transition occurs. For continuity of the main text of this chapter, the full exposition of this analysis is given in the Appendix.

The following complexity result highlights the practicality of implementing the fairness-to-welfare mapping in Algorithm 1 to track the full solution paths of an ε-fair SVM program. We note that standard SVM algorithms such as `LibSVM` run in $O(n^3)$, and thus once the algorithm has been initialized with the unconstrained SVM solution, the complexity of computing both the full individual solution paths $\mu_j(\varepsilon)$ and the full group welfare curves $\{W_0(\varepsilon), W_1(\varepsilon)\}$ is on the same order as that of computing a single SVM solution.

Theorem 6. *Each iteration of Algorithm 1 runs in $O(n^2 + |\mathcal{M}|^2)$. For breakpoints on the order of n, the full run time complexity is $O(n^3 + n|\mathcal{M}|^2)$.*

Proof. Each iteration of the fairness-to-welfare algorithm requires the inversion of matrix $K^\varepsilon \in \mathbb{R}^{|\mathcal{M}^\varepsilon|+1}$ and the computations of $r_j^\varepsilon \in \mathbb{R}^{|\mathcal{M}^\varepsilon|}$ for $j \in \mathcal{M}^\varepsilon$, and g_j^ε and d_j^ε for $j \in \{\mathcal{F}, \mathcal{E}\}^\varepsilon$.

The standard Gauss-Jordan matrix inversion technique runs in $O(|\mathcal{M}|^3)$, but we take advantage of partition update rules to lower the number of computations: Since at each new breakpoint, the partition tends to change because of additions or eliminations of a single index j from the set $\mathcal{M}$, we can use the Cholesky decomposition rank-one update or downdate to ease the need to recompute the full matrix inverse at every iteration, thereby reducing the complexity of the operation to $O(|\mathcal{M}|^2)$. Computing the stability region conditions for $j \in \{\mathcal{F}, \mathcal{E}\}$ requires $O\big((n - |\mathcal{M}|)|\mathcal{M}|\big)$

steps. As such, at each breakpoint, the total computational cost is $O(|\mathcal{M}|^2 + n^2)$.

The number of breakpoints for each full run of the algorithm depends on the data distribution and how sensitive the solution is to the constraint. As a heuristic, datasets whose fairness constraints bind for smaller ε have fewer breakpoints. Previous empirical results on the full SVM path for L1 and L2 regularization have found that the number of breakpoints tends to be on the order of n.[46,90,89,54] Thus after initialization with 0-fair SVM solution, the final complexity for the algorithm is $O(n^3 + n|\mathcal{M}|^2)$. □

4.3.4 Impact of Fair Learning on Learner's Welfare

Having proven the main welfare-relevant sensitivity result for groups, we return to more standard analysis of the effect of $\Delta\varepsilon$ perturbations on the learner's loss. In this case, we directly solve for the dual variable of the fairness constraint. Recall γ^* from (A.13):

$$\gamma^* = \gamma_1^* - \gamma_2^* = \frac{n(n(\beta_- - \beta_+) + \sum_{i=1}^n \mu_i y_i \langle \mathbf{x}_i, \mathbf{u} \rangle)}{\|\mathbf{u}\|^2} \tag{4.20}$$

By complementary slackness, one of β_- and β_+ is zero, and the other is ε. In particular, if $\beta_- = 0$, then $\beta_+ = \varepsilon$, then we know that $\gamma > 0$. Thus the original fairness constraint that binds is the upper bound on covariance, suggesting that the optimal classifier must be constrained to limit its positive covariance with group $z = 1$. If $\beta_+ = 0$, then $\beta_- = \varepsilon$ and $\gamma < 0$, and the classifier must be constrained to limit its positive covariance with group $z = 0$.

We can interpret the value of the dual variable Lagrange multiplier as the shadow price of the fairness constraint. It gives the additional loss in the objective value that the learner would achieve if the fairness constraint were infinitesimally loosened. Whenever a fairness constraint binds, its shadow price is readily computable and is given by $|\gamma^*|$. It bears noting that because (ε-fair Soft-

SVM) is not a linear program, $|\gamma^*|$ can only be interpreted as a measure of *local* sensitivity, valid only in a small neighborhood around an optimal solution. But through an alternative lens of sensitivity analysis, we can derive a lower bound on global sensitivity due to changes in the fairness tolerance parameter ε. By writing ε as a perturbation variable, we can perform sensitivity analysis on the same ε-constrained problem. Returning to the perturbation function $p(\varepsilon)$, we have

$$p(\varepsilon) \geqslant \sup_{\boldsymbol{\mu},\gamma}\{\mathcal{L}(\boldsymbol{\mu}^*, \gamma^*) - \varepsilon|\gamma^*|\} \tag{4.21}$$

where $\mathcal{L}(\boldsymbol{\mu}^*, \gamma^*)$ gives the solution to the 0-fair SVM problem.

$$\mathcal{L}(\boldsymbol{\mu}^*, \gamma^*) = \max_{\boldsymbol{\mu}\in[0,C]^n,\gamma} -\frac{1}{2}\|\sum_{i=1}^{n}\mu_i y_i(I - P_u)\mathbf{x}_i\|^2 + \sum_{i=1}\mu_i \tag{4.22}$$

The perturbation formulation given in (4.21) is identical in form to the original program (ε-fair-SVM1-P) but gives a global bound on $p(\varepsilon)$ for all $\varepsilon \in [0, 1]$. Since (4.21) gives a lower bound, the global sensitivity bound yields an asymmetric interpretation.

Proposition 9. *If* $\Delta\varepsilon < 0$ *and* $|\gamma^*| \gg 0$*, then* $p(\varepsilon + \Delta\varepsilon) - p(\varepsilon) \gg 0$*. If* $\Delta\varepsilon > 0$ *and* $|\gamma^*| < \delta$ *for small* δ*, then* $p(\varepsilon + \Delta\varepsilon) - p(\varepsilon) \in [-\delta\Delta\varepsilon, 0]$*, and is thus also small in magnitude.*

Proposition 9 shows that tightening the fairness constraint when its shadow price is high leads to a great increase in learner loss, but loosening the fairness constraint when its shadow price is small leads only to a small decrease in loss.

4.4 Experiments

To demonstrate the efficacy of our approach, we track the impact of ε-fairness constrained SVM programs on the classification outcomes of individuals in the Adult dataset. The target variable in the dataset is a binary value indicating whether the individual has an annual income of more or less than \$50,000. If such a dataset were used to train a tool to be deployed in consequential resource allocation—say, for the purpose of determining access to credit—then classification decisions directly impact individuals' welfares.

Individual solution paths and relative group welfare changes are given in Figure 1. As ε increases from left to right, the fairness constraint is loosened, and outcomes become "less fair." In the case of the ε-fair SVM solution to the Adult dataset, the fairness constraint ceases to bind at the optimal solution when $\varepsilon \approx 0.175$. The top panel shows example individual piecewise linear paths of dual variables $\mu_i(\varepsilon)$, providing a visual depiction of how individual points can transition across index sets: from $\mu_i = 0, i \in \mathcal{F}$ and being correctly labeled, to $\mu_i \in (0, 1), i \in \mathcal{M}$, being correctly labeled but in margin; to $\mu_i = 1, i \in \mathcal{E}$ and being incorrectly labeled. Solid paths indicate individuals coded female; dashed paths indicate those coded males. As the top panel of Figure 1 shows, the actual "journey" of these paths are varied as ε changes.

As expected, tightening the fairness constraint in the ε-fair program does tend to lead to improved welfare outcomes for females as a group (more female individuals receive a positive classification), while males experience a relative decline in group welfare (receiving fewer positive classifications). However, as suggested by our results in Section 3.2, these welfare changes are not monotonic for either group. Tightening the fairness constraint could lead to declines in both groups' welfares, demonstrating that preferring more fair solutions in this predictive model does not abide by the Pareto Principle. We highlight an instance of this result in the bottom panel of Figure 1, where orange dashed lines to the left of black ones mark off solutions where "more fair" outcomes (orange) are Pareto-dominated by "less fair" (black) ones. A practitioner working in a domain in which welfare considerations might override parity-based fairness ones may prefer the outcomes of a fair learning procedure with $\varepsilon \approx 0.045$ to one with $\varepsilon \approx 0.015$. Additional plots showing absolute changes in group welfare and optimal learner value are given in the Appendix.

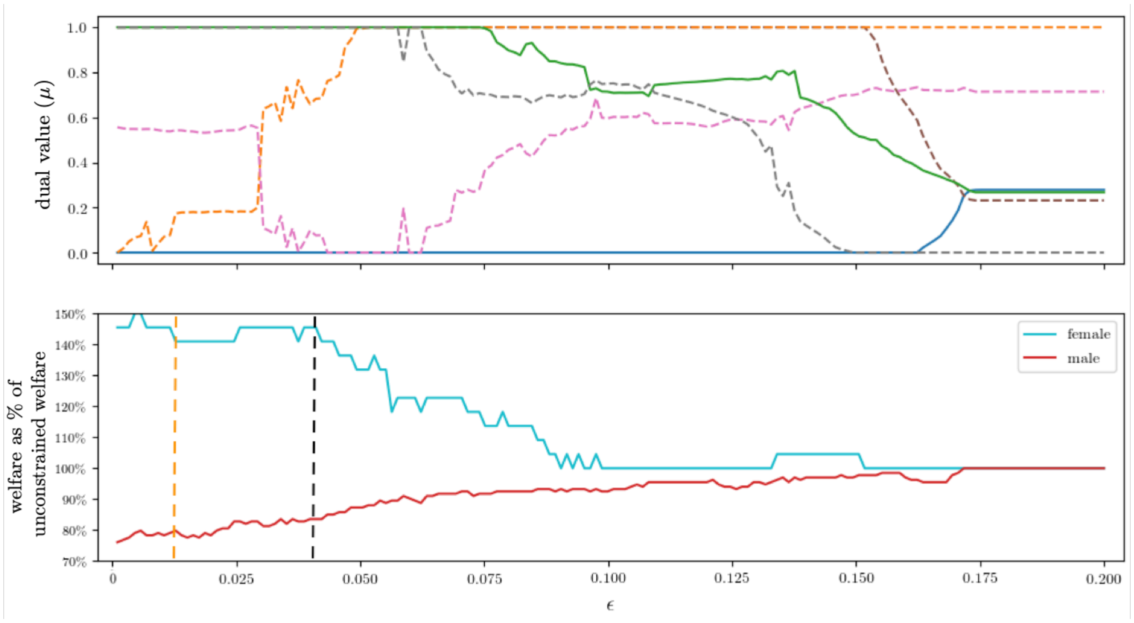


Figure 4.1: Fairness-to-welfare solution paths for individuals (top panel) and groups (bottom panel) on the Adult dataset.

4.5 Discussion

The question that leads off this chapter—*How do leading notions of fairness as defined by computer scientists map onto longer-standing notions of social welfare?*—sets an important agenda to come for the field of algorithmic fairness. It asks that the community look to disciplines that have long considered the problem of allocating goods in accordance with ideals of justice and fairness. For example, the notion of welfare in this chapter draws from work in welfare and public economics. The outcomes issued by an optimal classifier can, thus, be interpreted using welfare economic tools developed for considerations of social efficiency and equity. In an effort to situate computer scientists' notions of fairness within a broader understanding of distributive justice, we also show that loss minimization problems can indeed be mapped onto welfare maximization ones and vice versa. For reasons of continuity, analyses of this correspondence do not appear in the main text—we defer the interested reader to the Appendix—though we present an abbreviated overview here. We encourage readers to consider the main results of this chapter, which construct welfare paths out of fair learning algorithms, as a part of this larger project of bridging the two approaches.

4.5.1 Bridging Fair Machine Learning and Social Welfare Maximization

To highlight the correspondence between the machine learning and welfare economic approaches to allocation, we show that loss minimizing solutions can be understood as welfare maximizing ones under a particular social welfare function. In the Planner's Problem, a planner maximizes social welfare represented as the weighted sum of utility functions, where each individual's weight represents the value placed by society on her welfare. Inverting the Planner's Problem of social welfare maximization generates a question concerning social equity: *"Given a particular allocation, what is the presumptive social weight function that would yield it as optimal?"* We show that the set of predictions issued by the optimal classifier of any loss minimization task can be given as the set of allocations in the Planner's Problem over the same individuals endowed with a set of welfare weights. These weights lie at the heart of debates over fairness of distribution in economics. Analyzing the distribution of implied weights of individuals and groups offers a welfare economic way of considering the "fairness" of classifications.

4.5.2 Interpreting Welfare Alongside Fairness

Welfare economics can lend particular insights into formalizing notions of distributional fairness and general insights into building a technical field and methodology that grapples with normative questions. The field is concerned with what public policies ought to be, how to improve individuals' well-beings, and what distribution of outcomes are preferable. Answers to these questions appeal to values and judgments that refer to more than just descriptive or predictive facts about the world. The success of fair machine will largely hang on how well it can adapt to a similar ambitious task.

However, welfare economics is not the only—nor should it even serve as the main—academic resource for thinking through how goods ought to be provisioned in a just society. In this moment of broad appeal to the prowess of algorithmic systems, researchers in computing are called on to advise on matters beyond their specialized expertise and training. Many of these matters require explicit normative, political, and social-scientific reasoning. Insights and methods from across the arts, humanities, social sciences, and natural sciences bear fruit in answering these questions.

This chapter does not look to contribute a new fair learning algorithm or a new fairness definition. We take a popular classification algorithm, the Soft Margin SVM, append a parity-based fairness constraint, and analyze its implications on welfare. The constraint that we center in the chapter is just one concretization of a large menu of fairness notions that have been offered up to now. The method of analysis developed in the chapter applies generally to any convex formulations of these constraints, including versions of balance for false positives, balance for false negatives, and equality of opportunity that have circulated in the literature.[93,28,3] It is important future work to investigate the welfare implications of state-of-the-art fair classification algorithms that the community continues to develop, which can deal with a wider range of models and constraints, including non-convex ones.

This chapter asks that researchers in fair machine learning reevaluate not only their lodestars of optimality and efficiency but also their latest metrics of fairness. By viewing classification outcomes as allocations of a good, we incorporate considerations of individual and group utility in our analysis of classification regimes. The concept of "utility" in evaluations of social policy remains controversial, but in many cases of social distribution, utility considerations provide a partial but still important perspective on what is at stake within an allocative task. Utility-based notions of welfare can capture the relative benefit that a particular good can have on a particular individual. If machine learning systems are in effect serving as resource distribution mechanisms, then questions about fairness should align with questions of "Who benefits?" Our results show that many parity-based formulations of fairness in machine learning do not ensure that disadvantaged groups benefit. Preferring a classifier that better accords with a fairness measure can lead to selecting allocations that lower the welfare for every group. We note that nevertheless, there are several reasons in favor of limiting levels of inequality not reflected in utilitarian calculus. In some cases, the gap between groups is itself objectionable, and considerations of relational equality between groups overrides gains to the absolute utility level of disadvantaged groups. But without acknowledging and accounting for these reasons, well-intentioned optimization tasks that seek to be "fairer" can further disadvantage social groups for no reason but to satisfy a given fairness metric.

ALGORITHM 1: Fairness-to-welfare solution paths as a function of ε

```
Input: set $\mathcal{X}$ of $n$ data points $\{\mathbf{x}_i, z_i, y_i\}$
Output: solutions paths $\boldsymbol{\mu}(\varepsilon)$ and group welfare curves $\{W_0(\varepsilon), W_1(\varepsilon)\}$
$\boldsymbol{\mu}^0 = \arg\min_{\boldsymbol{\mu}} D(\boldsymbol{\mu})$ of (0-fair SVM-D);
$\varepsilon = 0, \Delta\varepsilon = 0$;
$|n_0| = \sum_{i=1}^{n} 1[z_i = 0], |n_1| = \sum_{i=1}^{n} 1[z_i = 1]$;
while $\varepsilon < 1$ do
    $W_0 = 0, W_1 = 0$;
    for each $\mu_i^\varepsilon$ do
        update $\{\mathcal{F}, \mathcal{M}, \mathcal{E}\}^\varepsilon$ according to (4.7), (4.8), (4.9);
        if $(\mu_i < C \,\&\, y_i = 1) \,||\, (\mu_i = C \,\&\, y_i = 0)$ then
            $W_{z_i} = W_{z_i} + 1$;
        end
    end
    $W_0(\varepsilon) = \frac{W_0}{n_0}; W_1(\varepsilon) = \frac{W_1}{n_1}$;
    if $|\mathcal{M}^\varepsilon| = 0$ then
        $\Delta\varepsilon = \min_i M_i$ as given in (A.16);
        update $\{\mathcal{F}, \mathcal{M}, \mathcal{E}\}^\varepsilon$ according to (A.18) and (A.19);
        $\varepsilon = \varepsilon + \Delta\varepsilon$;
    end
    compute $\boldsymbol{r}^\varepsilon, \boldsymbol{d}^\varepsilon$ according to (4.13), (4.18);
    $\Delta\varepsilon = \min_i M_i$ as given in (4.16);
    $\mu_i^{\varepsilon+\Delta\varepsilon} = \mu_i^\varepsilon + r_i^\varepsilon \Delta\varepsilon$ for $i \in \mathcal{M}^\varepsilon$, $\mu_i^\varepsilon = \mu_i^{\varepsilon+\Delta\varepsilon}$ for $i \in \{\mathcal{F}, \mathcal{E}\}^\varepsilon$;
    $\varepsilon = \varepsilon + \Delta\varepsilon$;
end
return $(\boldsymbol{\mu}(\varepsilon), W_0(\varepsilon), W_1(\varepsilon))$
```

5
Conclusion

The orientation that I adopt towards algorithmic fairness in this dissertation is one that focuses on the interaction between machine classifications and the broader societal contexts within which they are embedded. I consider changes to the incentive structures that data-based classification introduces, the strategic responses of agents who interact with such systems, and the welfare impacts of various fairness constraints that have been proffered in the field. Hence, the common theme that knits these works together is that of a perspective that foregrounds the *dynamics* of algorithmic fairness. Through the works contained in this dissertation, I hope to have shown the virtues of taking this approach. Models that re-embed algorithms in their social and economic environments provide insight into fairness that:

- inform how these tools actually operate in the real world to impact social outcomes,
- challenge standard wisdom in algorithmic fairness,
- better equip us to design systems that work against social inequality.

In the past several years, researchers in the field have increasingly come to adopt a broader sociotechnical framing of the problem of algorithmic fairness. One upshot of this gradual shift in perspective is that the dynamics-focused, wider lens analysis that I take in this dissertation is less distinctive than it was six years ago, when I first embarked on work in this area. I therefore want to conclude this dissertation by pivoting to propose new paths forward for theorizing about fair classification and thus, fair distribution, at a higher level of abstraction that also makes connections with some of my philosophical work on topics adjacent to algorithmic fairness.

My first suggestion is one that may be surprising given the field's shift in recent years away from putting forth new technical definitions of fairness and my preceding comments about these proposals. While I agree that all such formal accounts suffer from serious defects and fail as "definitions" of fairness, their introduction into the discourse has nevertheless been fruitful. They have forced us to articulate precisely what might be wrong with notions such as, say, *meritocratic fairness*[57] or *equalized odds*[45] or *counterfactual fairness*[64]—not just from a technical perspective about their effects on predictive accuracy or their difficulty in being translated into convex constrained optimization problems but from a distinctively *normative* perspective: why such notions are not adequate conceptions of fairness *on moral and political grounds*. Works such as Mitchell et al.'s "Algorithmic Fairness: Choices, Assumptions, and Definitions"[74] and Reubin Binns' "Fairness in Machine Learning: Lessons from Political Philosophy"[13] uncover and elucidate what value-laden modeling choices and assumptions are embedded in technical accounts of fairness and thus what their substantive ethical content is. Laying out formal definitions of fairness to be scrutinized in this manner thus yields generative cross-disciplinary dialogue, which is beneficial to the field as a whole. In the present world within which we live, machine learning algorithms play a crucial function in many important resource allocation pipelines. Computer scientists have thus been unwittingly cast as partial social planners. Given this state of affairs, it is paramount that the values that structure algorithm design are made explicit to broader society as well as to the engineers who build the systems themselves. Furthermore, technical formalization of fairness contributes also to our normative thinking about what constitutes fairness. For example, the field's so-called fairness "impossibility results" have spurred significant discussion about which if any of the three fairness criteria of calibration, balance for false positives, or balance for false negatives is necessary and/or sufficient for fairness[47,69]. The results have also prompted scholars to renovate accounts of discrimination in the law, consider which remedies towards fairness might be compatible with or at odds with what equal protection in the law requires, and evaluate which approaches to fairness are more likely to in fact make headway in addressing social inequalities[72,49,94]. Much of this progress was spurred by computer scientists' proposals of fairness definitions. Though they might have themselves been shown to be deficient, these formalizations generated fruitful critical discussion and exchange that both deepens our understanding of the normative matters at stake as well as strengthens our ability to build tools that may better meet our aims towards "fairness". More work that follows in this vein can thus be constructive for the field's development.

Second, the rapid rise of artificial intelligence and machine learning tools and the sheen of an exciting "newness" of these technologies can often make it seem as though the questions at the heart of the field of algorithmic fairness are also new and newly urgent. In truth, a field that is methodologically not so far off from our own, that of *welfare economics*, has for decades been concerned with the development of technical tools to probe social and fundamentally value-laden ethical questions of resource distribution. And so welfare economics, in my view, stands out as a natural point of connection for technical research in algorithmic fairness. The discipline can both provide specific insights into formalizing substantive notions of fairness in distribution and also general insights into how to build a technical field and methodology that more effectively grapples with normative questions. Welfare economics is carved out as the branch of economics that is explicitly concerned with what public policies *ought* to be, how to maximize individuals' well-beings, and what types of distributive outcomes are preferable. Answers to these questions appeal to values and judgments that do not refer only to descriptive or predictive facts about a state of affairs. It would appear that the success of fair machine will largely hang on how well it can adapt to a similar ambitious task. Since notions of fairness are invariably context-dependent and always informed by background normative views, it is unsurprising that there has been such wide disagreement within the community about which of the many fairness definitions is the "right" one. Insofar as developing a unified framework of analysis of these competing formal notions, their compatibilities with each other, and their impacts on other social values such as efficiency are key disciplinary aims for both algorithmic fairness and welfare economics, each community has great potential to grow from engagement with the other.

Finally, work on causal inference that studies the causal effects of social categories such as race and sex shares significant overlap with concerns about the discriminatory potential of machine learning systems. Much research in causal inference in the social sciences looks to identify race or sex causal estimands from observational data and thereby claims to quantify the extent of discrimination on the basis of race or sex. It is thus no surprise that causal methods have been imported into approaches towards fairness in data-based predictive systems.

Causal inference about race and sex has notoriously been the subject of decades-long methodological and conceptual disputes, which continues to this day. In recent works, prominent causal inference practitioners have debated whether one can even quantify the effect that race has on a decision that takes place downstream of other decisions that were themselves causally affected by race.[63,42,98] Cases of such multi-stage race-inflected decision-making processes abound. As an example, if race influences who police decide to stop such that administrative records of police encounters embody a selection effect, then certain race-causal estimands on outcomes downstream from stops will be biased absent strong untestable assumptions. This problem clearly bears directly on the use of any such data in machine learning-based classification systems. Work on the methodological challenges to producing unbiased estimates of such race-causal estimands and bounding such effect estimates will greatly inform the extent to which machine learning systems can take biased input data at "face-value". But furthermore, the problem of upstream racial bias presents also a *conceptual* problems for the prospect of quantifying the amount of "taint" that data have and thus the extent to which one can claim that an algorithm's outputs are not causally influenced by race. I argue elsewhere with legal scholar Issa Kohler-Hausmann that if causal inference practitioners start with the premise that there exists a race selection effect (i.e., racial discrimination or bias) on observational data from which their methods draw, then race-causal estimands quantified using such data are *mis*defined, because of a violation of the consistency axiom of causal inference.[50] Those who work in causal inference frequently debate which assumptions about the data generation process are necessary and/or sufficient in order for causal identification to be sound. Research in algorithmic fairness stands to greatly benefit from further contact with such work, as the methodological and conceptual problems at issue there are central too in our field. This is a line of research that I am pursuing in my own philosophical work, where I hope that bringing my technical and analytical tools to tackle foundational questions in causal reasoning in the (social) sciences can cross over to make an impact on how we think about what constitutes fairness and discrimination in algorithms.

A

Appendices

A.1 Appendix for Chapter 2

A.1.1 Proofs from Section 2.3.1

Proof of Proposition 1

We first construct the optimal learner classifier when facing only candidates of a single group. Suppose the learner encounters only group A candidates. Then using her knowledge that the true classifier h_A is based on a threshold $\tau_A \in [0,1]$, she can construct a classifier that admits those candidates with scores $x \geqslant \tau_A$ and rejects candidates $x < \tau_A$. Since the maximal manipulation cost that any candidate would be willing to undertake is 1, for all $x \in [0,1]$, $c_A(y) - c_A(x) \leqslant 1$ and therefore

$$y \leqslant c_A^{-1}(c_A(x) + 1)$$

Thus a candidate with feature $x = \tau_A$ would be able to move to any feature $y \leqslant \sigma_A$ where $\sigma_A = c_A^{-1}(c_A(\tau_A) + 1)$.

Repeating the same reasoning for group B, a candidate with feature $x = \tau_B$ would be willing to move to any feature $y \leqslant \sigma_B$ where $\sigma_B = c_B^{-1}(c_B(\tau_B) + 1)$.

Now we want to show that $[\sigma_B, \sigma_A]$ marks an interval of undominated strategies. First we prove the ordering that $\sigma_B \leqslant \sigma_A$ for all cost functions c_B and c_A and all thresholds $\tau_B \leqslant \tau_A$. Recall that since $h_A(x) = 1 \implies h_B(x) = 1$, we have $\tau_B \leqslant \tau_A$. Although we cannot order $c_B(\tau_B)$ and $c_A(\tau_A)$, we have, by monotonicity of c_B

$$c_B(\tau_B) \leqslant c_B(\tau_A).$$

Let $\Delta = c_B(\tau_A) - c_B(\tau_B)$. Notice that if $\Delta \geqslant 1$, $c_B(\tau_B) + 1 \leqslant c_B(\tau_A)$, and so

$$\sigma_B = c_B^{-1}(c_B(\tau_B) + 1) \leqslant \tau_A < \sigma_A,$$

where the last inequality is due to monotonicity of c_A. ✓

Let us consider the $\Delta \in (0, 1)$ case. By the cost condition, we can write $c'_B(\tau_A) \geqslant c'_A(\tau_A)$. This implies that

$$c_B^{-1}(c_B(\tau_A) + 1) \leqslant c_A^{-1}(c_A(\tau_A) + 1)$$

Substituting in $c_B(\tau_A) = c_B(\tau_B) + \Delta$, we have

$$c_B^{-1}(c_B(\tau_B) + \Delta + 1) \leqslant c_A^{-1}(c_A(\tau_A) + 1) = \sigma_A.$$

By monotonicity of c_B, the left hand side is $\geqslant \sigma_B$, and we have that $\sigma_B \leqslant \sigma_A$ as desired. ✓

Notice that for all $\sigma < \sigma_B$, the learner commits false positive errors on candidates from group B, since σ_B is optimal for group B classification. She commits more false positives on group A candidates as well and does not commit any fewer false negatives because of the monotonicity of c_B and c_A. Thus for any error function with $C_{FP} > 0$, the threshold classifier σ_B dominates σ.

Similarly, for all $\sigma > \sigma_A$, the learner commits false negative errors on candidates from group A, since σ_A is optimal for group A classification. She also commits more false negatives on group B while committing no fewer false positives. Thus for any error function with $C_{FN} > 0$, the threshold classifier σ_A dominates σ.

For all $\sigma \in [\sigma_B, \sigma_A]$, the learner trades off false negatives on group B for false positives on group A, and we call this range of threshold strategies undominated. □

Proof of Proposition 2

We compute the cost of a learner's threshold strategy $\sigma \in [\sigma_B, \sigma_A]$ by first examining its performance on each group individually.

Recall from Proposition 1 that the optimal learner threshold that perfectly classifies all B candidates is σ_B. Thus for all threshold strategies based on $\sigma \in (\sigma_B, \sigma_A]$, the learner commits false negative errors on group B.

To compute which members of group B are subject to these errors, consider a learner classifier f based on a threshold σ. In order to manipulate to reach the feature threshold σ, a group B candidate must have an unmanipulated x such that

$$c_B(\sigma) - c_B(x) \leqslant 1,$$

$$x \geqslant c_B^{-1}(c_B(\sigma) + 1) = \ell_B(\sigma).$$

We know that $\tau_B \leqslant \ell_B(\sigma)$ by monotonicity of c_B, and thus for all group B candidates with feature $x \in [\tau_B, \ell_B(\sigma))$, the learner issues classification $f(x) = 0$, even though $h_B(x) = 1$. These are the false negative errors issued on group B for which the learner bears cost

$$C_{FN} p_B P_{x \sim \mathcal{D}_B}\big[x \in [\tau_B, \ell_B(\sigma))\big] \tag{A.1}$$

Following the same reasoning, notice that since σ_A is the optimal threshold policy for a learner facing only group A candidates, a classifier f based on any $\sigma \in [\sigma_B, \sigma_A)$ commits false positive errors on some group A candidates. Then repeating the steps that we carried out for group B, we see that for all group A candidates with x such that

$$x \geqslant c_A^{-1}(c_A(\sigma) + 1 = \ell_A(\sigma)$$

the classifier f issues a positive classification; $f(x) = 1$. Since $\ell_A(\sigma) \leqslant \tau_A$, candidates with features $x \in [\ell_A(\sigma), \tau_A)$, have true label $h_A(x) = 0$, and the learner commits false positive errors that bear cost

$$C_{FP}p_A P_{x \sim \mathcal{D}_A}\big[x \in [\ell_A(\sigma), \tau_A)\big] \tag{A.2}$$

Combining (A.1) and (A.2), the total cost of any classifier f based on a threshold $\sigma \in [\sigma_B, \sigma_A]$, we obtain our desired result. □

Proofs of Corollaries 1 and 2

These results follow by considering strategies σ_B, which commits no errors on group B and thus only bears the cost given in (A.2), and σ_A, which commits no errors on group A and thus only bears the cost given in (A.1). □

Proof of Proposition 3

Under the assumption of uniform feature distributions for both groups, minimizing a classifier's probability of error amounts to choosing the threshold σ as

$$\operatorname*{arg\,min}_{\sigma \in [\sigma_B, \sigma_A]} \ell_B(\sigma) - \ell_A(\sigma).$$

With proportional group costs $c_A(x) = qc_B(x)$ for $q \in (0,1)$, we have that

$$\begin{aligned}
\ell_B'(\sigma) &= \frac{(c_B)'(\sigma)}{\Big(c_B\Big)'\Big(\big(c_B\big)^{-1}\big(c_B(\sigma) - 1\big)\Big)} \\
&= \frac{(c_B)'(\sigma)}{\Big(c_B\Big)'\Big(\ell_B(\sigma)\Big)}
\end{aligned}$$

and

$$\begin{aligned}
\ell_A'(\sigma) &= \frac{(c_A)'(\sigma)}{\Big(c_A\Big)'\Big(\big(c_A\big)^{-1}\big(c_A(\sigma) - 1\big)\Big)} \\
&= \frac{(qc_B)'(\sigma)}{\Big(qc_B\Big)'\Big(\big(c_A\big)^{-1}\big(c_A(\sigma) - 1\big)\Big)} \\
&= \frac{(c_B)'(\sigma)}{\Big(c_B\Big)'\Big(\big(c_A\big)^{-1}\big(c_A(\sigma) - 1\big)\Big)} \\
&= \frac{(c_B)'(\sigma)}{\Big(c_B\Big)'\Big(\ell_A(\sigma)\Big)}.
\end{aligned}$$

When c_A and c_B are strictly concave, since $\ell_B(\sigma) > \ell_A(\sigma)$, $(c_B)'(\ell_A(\sigma)) > (c_B)'(\ell_B(\sigma))$ and therefore $\ell_A'(\sigma) < \ell_B'(\sigma)$ for all $\sigma \in [\sigma_B, \sigma_A]$, and the quantity $\ell_B(\sigma) - \ell_A(\sigma)$ is monotonically increasing in σ. Thus the optimal classifier threshold is $\sigma^* = \sigma_B$.

Similarly, when c_A and c_B are strictly convex, $\ell_A'(\sigma) > \ell_B'(\sigma)$ for all $\sigma \in [\sigma_B, \sigma_A]$, and the quantity $\ell_B(\sigma) - \ell_A(\sigma)$ is monotonically decreasing in σ. Thus the optimal classifier threshold is $\sigma^* = \sigma_A$. Thus the optimal classifier threshold is $\sigma^* = \sigma_A$.

Finally, when c_A and c_B are affine, $\ell_A'(\sigma) = \ell_B'(\sigma)$ for all $\sigma \in [\sigma_B, \sigma_A]$, and the quantity $\ell_B(\sigma) - \ell_A(\sigma)$ is constant for all $\sigma \in [\sigma_B, \sigma_A]$. Thus the learner is indifferent between all thresholds $\sigma \in [\sigma_B, \sigma_A]$. □

A.1.2 Proofs from Section 2.3.2

Proof of Lemma 1

Consider a candidate with unmanipulated feature $\mathbf{x} \in [0,1]^d$ and manipulation cost $\sum_{i=1}^d c_i x_i$ who faces a classifier $f(\mathbf{y})$ with linear decision boundary given by $\sum_{i=1}^d g_i y_i = g_0$. Recall that the utility a candidate receives for presenting feature $\mathbf{y} \geqslant \mathbf{x}$ is given by $f(\mathbf{y}) - c(\mathbf{x}, \mathbf{y})$. When $f(\mathbf{x}) = 1$, it is trivial that the candidate's best response to select $\mathbf{y} = \mathbf{x}$. ✓

Notice that if for all $i \in [d]$, $f(\mathbf{x} + \frac{1}{c_i}\mathbf{e}_i) = 0$, then we have that $\mathbf{g}^\intercal \mathbf{x} + \frac{g_k}{c_k} < g_0$, so

$$\frac{c_k(g_0 - \mathbf{g}^\intercal \mathbf{x})}{g_k} > 1$$

The manipulation from $\mathbf{x}$ to $\mathbf{y} = \mathbf{x} + \sum_{i\in K} \frac{t_i}{c_i}\mathbf{e}_i$ such that $\mathbf{g}^\intercal \mathbf{y} = g_0$ entails cost

$$c(\mathbf{y}) - c(\mathbf{x}) = \sum_{i\in K} t_i = \frac{c_k(g_0 - \mathbf{g}^\intercal \mathbf{x})}{g_k} > 1$$

and manipulating to achieve a positive classification using only components in K would require a cost > 1. By definition, keeping the sum $\sum_{i\in K} t_i$, but selecting different t_i such that some $i \notin K$, $t_i > 0$ would yield an even lower value $\mathbf{g}^\intercal \mathbf{x} + \sum_{i=1}^d \frac{g_i t_i}{c_i}$.

Thus manipulating from $\mathbf{x}$ to $\mathbf{y}$ such that $f(\mathbf{y}) = 1$ entails a cost $c(\mathbf{y}) - c(\mathbf{x}) > 1$, and the candidate would not move at all, since the utility for moving $1 - (c(\mathbf{y}) - c(\mathbf{x})) < 0$ makes her worse-off than being subject to a negative classification without expending any cost on feature manipulation. Thus she selects $\mathbf{y} = \mathbf{x}$. ✓

Now we consider the case where $f(\mathbf{x}) = 0$ and there exists $i \in [d]$ such that $f(\mathbf{x} + \frac{1}{c_i}\mathbf{e}_i) = 1$.

Let $k \in K = \arg\max_{i\in[d]} \frac{g_i}{c_i}$. We prove that the best-response manipulation for candidates with these $\mathbf{x}$ moves to

$$\mathbf{y} = \mathbf{x} + \sum_{i=1}^{d} \frac{t_i}{c_i}\mathbf{e}_i \tag{A.3}$$

where $t_i \geqslant 0, t_j = 0$ for all $j \notin K$, and $\mathbf{g}^\mathsf{T}(\mathbf{x} + \sum_{i \in K} \frac{t_i}{c_i}\mathbf{e}_i) = g_0$. Note that such a $\mathbf{y}$ may not be unique—there may be multiple best-response manipulated features that achieve the same candidate utility, since they all result in the same candidate cost, and thus regardless of choices $i \in K$, we have that

$$\sum_{i \in K} t_i = \frac{c_k(g_0 - \mathbf{g}^\mathsf{T}\mathbf{x})}{g_k} \tag{A.4}$$

The utility of any move to $\mathbf{y}$ satisfying (A.3) is given by

$$f(\mathbf{y}^*) - c(\mathbf{x}, \mathbf{y}^*) = 1 - \sum_{i=1} t_i$$

Let us pick any such $\mathbf{y}$ and call it $\mathbf{y}^*$ since we will show that all other manipulations that are not of the form given in (A.3) generate lower utility for the candidate than $\mathbf{y}^*$.

We now show that for any manipulation to $\mathbf{y}$, $\sum_{i=1}^{d} t_i \leqslant 1$. By assumption, for some i, we have

$$f(\mathbf{x} + \frac{1}{c_i}\mathbf{e}_i) = 1 \implies \mathbf{g}^\mathsf{T}\mathbf{x} + \frac{g_i}{c_i} \geqslant g_0$$

Thus by (A.4), we have that $\sum_{i \in K} t_i \leqslant \frac{c_k \frac{g_i}{c_i}}{g_k}$. By definition of k, this is at most one since $\frac{g_k}{c_k} \geqslant \frac{g_i}{c_i}$ for all $i \in [d]$. ✓

Suppose on the contrary that there exists another manipulated feature $\hat{\mathbf{y}} \neq \mathbf{y}^*$ that is optimal and is not of the form (A.3):

$$f(\hat{\mathbf{y}}) - (c(\hat{\mathbf{y}}) - c(\mathbf{x})) \geqslant 1 - \frac{c_k(g_0 - \mathbf{g}^\mathsf{T}\mathbf{x})}{g_k} \geqslant 0$$

Then it must be the case that moving to $\hat{\mathbf{y}}$ achieves a positive classification with a lower cost burden. We write

$$\hat{\mathbf{y}} = \mathbf{x} + \sum_{i=1} \hat{t}_i \mathbf{e}_i$$

where $\mathbf{e}_i$ is the i^{th} standard basis vector, and $\hat{t}_j = \hat{y}_j - x_j$ to highlight the components that have been manipulated from $\mathbf{x}$ to $\hat{\mathbf{y}}$.

First, we suppose that $\hat{\mathbf{y}}$ is such that there exists some component $\hat{\mathbf{y}}_j > 0$ where $j \notin K = \arg\max_{i \in [d]} \frac{g_i}{c_i}$. Now we construct a feature $\hat{\mathbf{y}}'$ by selecting this component, and decreasing $\hat{t}_j = 0$ and increasing a component $k \in K$ by $\frac{c_j \hat{t}_j}{c_k}$. That is

$$\hat{\mathbf{y}}' = \hat{\mathbf{y}} - \hat{t}_j \mathbf{e}_j + \frac{c_j \hat{t}_j}{c_k} \mathbf{e}_k$$

The cost of manipulation from $\mathbf{x}$ to $\hat{\mathbf{y}}'$ is the same as that for manipulation to $\hat{\mathbf{y}}$:

$$c(\hat{\mathbf{y}}') - c(\mathbf{x}) = \sum_{i=1}^{d} c_i \hat{y}_i - \hat{t}_j c_j + c_k \frac{c_j \hat{t}_j}{c_k} = \sum_{i=1}^{d} c_i \hat{y}_i$$

Notice that now we have

$$\sum_{i=1}^{d} g_i \hat{y}'_i = \sum_{i=1}^{d} g_i \hat{y}_i - g_j \hat{t}_j + \frac{g_k c_j \hat{t}_j}{c_k} > \sum_{i=1}^{d} g_i \hat{y}_i \geqslant g_0.$$

Thus the candidate can manipulate to $\hat{\mathbf{y}}'$ by expending the same cost with

$$\sum_{i=1}^{d} g_i \hat{y}'_i > g_0$$

Then by continuity of g, there must exist some $\bar{\mathbf{y}} \leqslant \hat{\mathbf{y}}'$ such that $\sum_{i=1}^{d} g_i \bar{y}_i \in [g_0, \sum_{i=1}^{d} g_i \hat{y}'_i)$. Thus since costs are monotonically increasing, $c(\mathbf{x}, \bar{\mathbf{y}}) < c(\mathbf{x}, \hat{\mathbf{y}})$ and since $\bar{\mathbf{y}}$ reaches the same classification, and we have shown that $\hat{\mathbf{y}}$ could not have been optimal, which is a contradiction. ✓

Now we consider the case where $\hat{\mathbf{y}} = \mathbf{x} + \sum_{i=1}^{d} \hat{t}_i \mathbf{e}_i$ is such that $\hat{t}_j = 0$ for all $j \notin K$, but $\mathbf{g}^\intercal \hat{\mathbf{y}} \neq g_0$. If $\mathbf{g}^\intercal \hat{\mathbf{y}} < g_0$, then $\hat{\mathbf{y}}$ is negatively classified and thus trivially receives a lower utility than manipulating to any feature $\mathbf{y}$ that is positively classified and associated with total cost $\sum_i t_i \leqslant 1$.

If $\mathbf{g}^\intercal \hat{\mathbf{y}} > g_0$, then there are two possibilities: If $c(\hat{\mathbf{y}}) - c(\mathbf{x}) \geqslant 1$, then once again, she receives at most a utility of 0, and thus manipulating to $\hat{\mathbf{y}}$ is a suboptimal move. If $c(\hat{\mathbf{y}}) - c(\mathbf{x}) < 1$, then we show the optimal manipulation is the one that moves from $\mathbf{x}$ to

$$\mathbf{y} = \mathbf{x} + \sum_{i=1} t_i \mathbf{e}_i$$

where $\mathbf{g}^\intercal \mathbf{y} = g_0$ and $t_j = 0, \forall j \notin K$—the move dictated by (A.3). This feature $\mathbf{y}$ also achieves a positive classification, but we argue that it does so at a lower cost than $\hat{\mathbf{y}}$. Since $\mathbf{g}^\intercal \hat{\mathbf{y}} > g_0$, we can define

$$\Delta = \mathbf{g}^\intercal \hat{\mathbf{y}} - g_0 > 0$$

The manipulation from $\mathbf{x}$ to $\hat{\mathbf{y}} - \frac{\Delta}{g_k} \mathbf{e}_k$ for any choice of k attains a higher utility since it receives the same classification since

$$\mathbf{g}^\intercal (\hat{\mathbf{y}} - \frac{\Delta}{c_k} \mathbf{e}_k) = g_0$$

but does so at a cost

$$c(\mathbf{y}) - c(\mathbf{x}) = c(\hat{\mathbf{y}}) - c(\mathbf{x}) - \Delta$$

Since we already showed that all manipulations to $\mathbf{y}$ of the form given in (A.3) bear the same cost, then we have shown that all such $\mathbf{y}$ are preferable to $\hat{\mathbf{y}}$. By monotonicity of $c(\mathbf{y}) - c(\mathbf{x})$ and $\sum_{i=1}^{d} g_i x_i$, all manipulations with lower cost entail a negative classification and thus a lower utility, and such only those manipulations to $\mathbf{y}$ are optimal. □

Proof of Theorem 1

We first prove that a learner who has access to the linear decision boundary for the true classifier can construct a classifier that commits no errors on any candidates from a single group; thus, in our setting, perfect classifiers exist for groups A and B. We then prove that all undominated classifiers commit no false positives on group B and no false negatives on group A.

Suppose true classifiers are given by h_A and h_B based on decision boundaries $\sum_{i=1}^{d} w_{A,i}x_i = \tau_A$ and $\sum_{i=1}^{d} w_{B,i}x_i = \tau_B$, costs are $c_A(\mathbf{x}) = \sum_{i=1}^{d} c_{A,i}x_i$ and $c_B(\mathbf{x}) = \sum_{i=1}^{d} c_{B,i}x_i$.

Claim 1: When facing candidates from a single group, a learner who has access to true decision boundary $\sum_{i=1}^{d} w_i x_i = \tau$ and manipulation costs $\sum_{i=1}^{d} c_i x_i$ can construct a perfect classifier.

Proof. Consider those features $\bar{\mathbf{x}} \in [0,1]^d$ that lie on the true decision boundary $\sum_{i=1}^{d} w_i x_i = \tau$ and thus have true labels 1. For each of these $\bar{\mathbf{x}}$, we construct $\Delta(\bar{\mathbf{x}})$ as defined in (2.12) to represent the candidate's space of potential manipulation to form the set $\{\Delta(\bar{\mathbf{x}})\}$ for all $\bar{\mathbf{x}}$ on the boundary. Notice that when all candidates face the same cost, the set of j^{th} vertices of each of the simplices $\Delta(\bar{\mathbf{x}})$, given by $\mathbf{v}_j(\bar{\mathbf{x}}) = \bar{\mathbf{x}} + \frac{1}{c_j}\mathbf{e}_j$, are coplanar. Each of these hyperplanes can be described as a set

$$\left\{\mathbf{y} : \sum_{i=1}^{d} w_i y_i = \tau + \frac{w_j}{c_j}\right\}.$$

Let $k \in \arg\max_j \frac{w_j}{c_j}$. We define g_1 to be a notational shortcut for the hyperplane corresponding to feature k, so

$$g_1 = \left\{\mathbf{y} : \sum_{j=1}^{d} g_{1,j} y_j = g_{1,0}\right\},$$

where $g_{1,0} = \tau + \frac{w_k}{c_k}$ and $g_{1,i} = w_i$ for all $i \in \{1, \ldots, d\}$. We define a classifier f_1 based on the hyperplane g_1:

$$f_1(\mathbf{y}) = \begin{cases} 1 & \sum_{j=1}^{d} g_{1,j} y_j \geqslant g_{1,0}, \\ 0 & \sum_{j=1}^{d} g_{1,j} y_j < g_{1,0}. \end{cases} \tag{A.5}$$

To show that f_1 is a perfect classifier of all candidates with these generic costs, we show that it commits no false positive errors and no false negative errors. Notice that since g_1 was constructed to be precisely the hyperplane that contains all vertices $\mathbf{v}_k(\bar{\mathbf{x}}) = \bar{\mathbf{x}} + \frac{1}{c_k}$ of the simplices $\Delta(\bar{\mathbf{x}})$ where $k \in \arg\max_{j \in [d]} \frac{w_j}{c_j}$, then all $\bar{\mathbf{x}}$ on the true decision boundary $\sum_{i=1}^{d} w_i x_i = \tau$ can indeed manipulate to $\mathbf{v}_k(\bar{\mathbf{x}})$ and reach g_1 to gain a positive classification.

Similarly, all candidates with features $\mathbf{x}$ such that $\sum_{i=1}^{d} w_i x_i > \tau$, can move to the k^{th} vertex of the simplex $\Delta(\mathbf{x})$ given by $\mathbf{v}_k(\mathbf{x}) = \mathbf{x} + \frac{1}{c_k}\mathbf{e}_k$ in order to be classified positively since

$$\sum_{i=1}^{d} w_i v_{k,i}(\mathbf{x}) > \tau + \frac{w_k}{c_k} \implies \sum_{i=1}^{d} g_{1,j} v_{k,i}(\mathbf{x}) > g_{1,0}.$$

Thus f_1 correctly classifies all these candidates positively and permits no false negatives. ✓

Consider the optimal manipulation for all true negative candidates $\mathbf{x}$. By Lemma 1, the optimal manipulation would be either to not move at all, guaranteeing a negative classification, or to move $\mathbf{x}$ to some point $\mathbf{y} = \mathbf{x} + \sum_{i=1}^{d} \frac{t_i}{c_i}\mathbf{e}_i$ where $t_j = 0$ for all $j \notin \arg\max_{j \in [d]} \frac{g_{1,j}}{c_j}$. But since $\sum_{i=1}^{d} w_i x_i < \tau$, then for all such $\mathbf{y}$,

$$\sum_{i=1}^{d} w_i y_i \leqslant \sum_{i=1}^{d} w_i x_i + \frac{w_k}{c_k} < \tau + \frac{w_k}{c_k} \implies \sum_{i=1}^{d} g_{1,j} y_j < g_{1,0}$$

and thus the classifier based on the hyperplane g_1 also issues a classification $f_1(\mathbf{x}) = 0$ and admits no false positives. ✓

Thus we have shown that the hyperplane g_1 supports a perfect classifier f_1 as defined in (A.5). □

Now we move on to group-specific claims, where groups have distinct costs and potentially distinct true decision boundaries, but we continue to use the constructions of f_1 and g_1 from Claim 1.

Claim 2: Let f_1^A be the classifier based on boundary g_1 for group A, and let f_1^B be the classifier based on boundary g_1 for group B, as in (A.5), but with group-specific costs and true decision boundary parameters. Then $\forall \mathbf{y} \in [0,1]^d$,

$$f_1^A(\mathbf{y}) = 1 \implies f_1^B(\mathbf{y}) = 1.$$

Proof. We first prove the claim for the case in which $h_A = h_B$ with decision bounday $\sum_{i=1}^d w_i x_i = \tau$. We then show that it also holds when the two are not equal.

By the cost condition that $c_A(\mathbf{y}) - c_A(\mathbf{x}) \leqslant c_B(\mathbf{y}) - c_B(\mathbf{x})$ for all $\mathbf{x} \in [0,1]^d$ and $\mathbf{y} \geqslant \mathbf{x}$, we know that for any given $\mathbf{x}$,

$$\Delta_B(\mathbf{x}) \subseteq \Delta_A(\mathbf{x}).$$

Let $k_A \in \arg\max_{j\in[d]} \frac{w_j}{c_{A,j}}$ and $k_B \in \arg\max_{j\in[d]} \frac{w_j}{c_{B,j}}$, so that g_1^A and g_1^B are defined as

$$\sum_{i=1}^d w_i y_i = \tau + \frac{w_{k_A}}{c_{A,k_A}} \iff g_1^A : \sum_{j=1}^d g_{1,j}^A y_j = g_{1,0}^A,$$

$$\sum_{i=1}^d w_i y_i = \tau + \frac{w_{k_B}}{c_{B,k_B}} \iff g_1^B : \sum_{j=1}^d g_{1,j}^B y_j = g_{1,0}^B.$$

Then since for all $i \in [d]$, $c_{A,i} \leqslant c_{B,i}$, we must have that

$$\tau + \frac{w_{k_A}}{c_{A,k_A}} \geqslant \tau + \frac{w_{k_B}}{c_{A,k_B}} \geqslant \tau + \frac{w_{k_B}}{c_{B,k_B}},$$

and thus $g_{1,0}^A \geqslant g_{1,0}^B$. Since f_1^A is the classifier based on g_1^A and f_1^B is based on g_1^B, we have that $\forall \mathbf{y} \in$

$[0,1]^d$,

$$f_1^A(\mathbf{y}) = 1 \implies f_1^B(\mathbf{y}) = 1.$$

Now consider the case in which h_A and h_B differ. Recall the assumption $h_A(\mathbf{x}) = 1 \implies h_B(\mathbf{x}) = 1$ for all $\mathbf{x} \in [0,1]^d$. Thus for all $\mathbf{x} \in [0,1]^d$,

$$\sum_{i=1} w_{A,i} x_i \geqslant \tau_A \implies \sum_{i=1} w_{B,i} x_i \geqslant \tau_B. \tag{A.6}$$

Recall that the hyperplanes g_1^A, g_1^B are constructed as shifts of $\sum_{i=1} w_{A,i} x_i \geqslant \tau_A$ and $\sum_{i=1} w_{B,i} x_i \geqslant \tau_B$ by the set of simplices $\{\Delta_A(\bar{\mathbf{x}}_A)\}$ and $\{\Delta_B(\bar{\mathbf{x}}_B)\}$ for $\bar{\mathbf{x}}_A$ such that $\sum_{i=1} w_{A,i} \bar{x}_{A,i} = \tau_A$ and $\bar{\mathbf{x}}_B$ such that $\sum_{i=1} w_{B,i} \bar{x}_{B,i} = \tau_B$. Since $\Delta_B(\mathbf{x}) \subseteq \Delta_A(\mathbf{x})$, g_1^A and g_1^B support classifiers f_1^A and f_1^B such that

$$f_1^A(\mathbf{y}) = 1 \implies f_1^B(\mathbf{y}) = 1.$$

□

Claim 3: All undominated classifiers commit no false negative errors on group A members and no false positive errors on group B members when candidates best respond.

Proof. Fix a classifier f and consider a group A candidate with true feature vector $\bar{\mathbf{x}}$ who manipulates to best response $\bar{\mathbf{y}}$ such that $h_A(\bar{\mathbf{x}}) = 1$ but $f(\bar{\mathbf{y}}) = 0$. Thus the classifier f makes a false negative error on this candidate. We show that we can construct another classifier $\hat{f}$ that correctly classifies $\bar{\mathbf{x}}$ under its optimal manipulation with respect to $\hat{f}$.

We prove that $\hat{f}$ commits no more errors than does f and commits strictly fewer errors since it commits no false negatives on group A candidates.

Construct the classifier $\hat{f}$ such that

$$\hat{f}(\mathbf{y}) = \begin{cases} 1 & f(\mathbf{y}) = 1 \text{ or } f_1^A(\mathbf{y}) = 1, \\ 0 & \text{otherwise,} \end{cases} \tag{A.7}$$

where $f_1^A(\mathbf{y})$ is based on the boundary $\sum_{j=1} g_{1,j}^A y_j = g_{1,0}^A$.

We first argue that f and $\hat{f}$ make exactly the same set of false positive errors.

Consider a potential false positive error that $\hat{f}$ issues on a candidate with feature $\mathbf{x}$ from group A. Such a candidate cannot manipulate to a feature $\mathbf{y}$ to "trick" classifier f_1^A, since we have shown in Claim 1 that f_1^A perfectly classifies all group A candidates, and thus does not admit false positives. Thus any potential false positive error must be due to $f(\mathbf{y}) = 1$, in which case $\hat{f}$ and f issue the same false positive error.

Now we consider a potential false positive error that $\hat{f}$ issues on a candidate with feature $\mathbf{x}$ from group B. By Claim 2, $f_1^A(\mathbf{y}) = 1 \implies f_1^B(\mathbf{y}) = 1$, and thus we would have that the candidate with feature $\mathbf{x}$ was able to manipulate to some feature $\mathbf{y}$ such that $f_1^B(\mathbf{y}) = 1$. But this is a contradiction, since we know that f_1^B commits no false positives on group B members, and thus $f_1^A(\mathbf{y})$ does not commit false positives on group B. Thus if $\hat{f}$ commits a false positive, then it must be the case that f committed the same false positive.

Consider a potential false negative error that $\hat{f}$ issues on a candidate with feature $\mathbf{x}$ from group B. Then it must be the case that $\mathbf{x}$ can manipulate to some $\mathbf{y}$ such that *both* $f(\mathbf{y}) = 0$ and $f_1^A(\mathbf{y}) = 0$, and thus it be the case that f commits the same false negative.

Lastly, consider a potential false negative error on a candidate from group A. By claim 1, this candidate must have been able to manipulate to some feature vector $\mathbf{y}$ such that $f_1^A(\mathbf{y}) = 1$, since f_1^A commits no errors on group A members. Thus when a candidate with unmanipulated feature $\mathbf{x}$ can manipulate to some $\mathbf{y}$ such that $f_1^A(\mathbf{y}) = 1$ yet can only present a (possibly different) feature $\mathbf{y}$

such that $f(\mathbf{y}) = 0$, then $\hat{f}$ correctly classifies this candidate positively, even when f does not. Thus $\hat{f}$ makes no false negative errors on group B.

Thus $\hat{f}$ commits strictly fewer errors than f—none of which are false negatives on group A members—and f is dominated by $\hat{f}$. ✓

The second half of the claim can be proved through an analogous argument. □

Combining Claims 1 and 3, we conclude that we can construct perfect classifiers for group A that commit only false negative errors on group B and perfect classifiers for group B that commit only false positive errors on group A. f_1^A and f_1^B are examples of such classifiers, though they are not unique.

□

Proof of Lemma 2

$\Longrightarrow$ direction: Assume a group m candidate with feature $\mathbf{x}$ can move to $\mathbf{y}$ such that $f(\mathbf{y}) = 1$ and $c_m(\mathbf{y}) - c_m(\mathbf{x}) \leqslant 1$, we show that necessarily $\mathbf{x} \geqslant \ell$ for some $\ell \in \mathcal{L}_m(g)$.

If $\mathbf{x}$ can move to $\mathbf{y}$, then $\mathbf{x} \in \Delta^{-1}(\mathbf{y})$. By the definition of $\ell_m(\mathbf{y})$, $\mathbf{x} \geqslant \bar{\mathbf{x}}$ for some $\bar{\mathbf{x}} \in \ell_m(\mathbf{y})$. Then by monotonicity of g, we have that

$$\sum_{i=1}^{d} g_i x_i \geqslant \sum_{i=1}^{d} g_i \bar{x}_i \geqslant \min_{x \in \ell_m(\mathbf{y})} \sum_{i=1}^{d} g_i x_i$$

Thus $\mathbf{x} \geqslant \ell$ for some $\ell \in \mathcal{L}_m(g)$. ✓

$\Longleftarrow$ direction: Assume some group m candidate has feature $\mathbf{x} \geqslant \ell$ for some $\ell \in \mathcal{L}_m(g)$. Then she can move to some $\mathbf{y}$ such that $f(\mathbf{y}) = 1$ and $c_m(\mathbf{y}) - c_m(\mathbf{x}) \leqslant 1$.

If $\mathbf{x} \geqslant \ell$ for some $\ell \in \mathcal{L}_m(g)$, then

$$\sum_{i=1}^{d} g_i x_i \geqslant \sum_{i=1}^{d} g_i \ell_i,$$

where $\ell \in \Delta_m^{-1}(\mathbf{y})$ for some $\mathbf{y}$ such that $\sum_{i=1} g_i y_i = g_0$ and $f(\mathbf{y}) = 1$. Since ℓ is defined as $\arg\min_{x \in \ell_m(\mathbf{y})} \sum_{i=1} g_i x_i$, then we have

$$\sum_{i=1}^{d} g_i \ell_i = \sum_{i=1}^{d} \big(g_i y_i - \max_{t_i} \sum_{i=1}^{d} \frac{g_i t_i}{c_{m,i}}\big),$$

where $t_i \geqslant 0$ and $\sum_{i=1} t_i = 1$ as shown before. Then substituting $\sum_{i=1}^{d} g_i y_i = g_0$, we have that

$$\sum_{i=1} g_i \ell_i + \frac{g_{k_m}}{c_{m,k_m}} = g_0,$$

where $k_m \in \arg\max_{i=[d]} \frac{g_i}{c_i}$. Since $\mathbf{x} \geqslant \ell$, $\mathbf{x}$ can also manipulate to some $\mathbf{y}$ with $f(\mathbf{y}) = 1$, bearing a cost $\leqslant 1$. □

Proof of Proposition 4

If a learner publishes an undominated classifier f, then by Theorem 1, the hyperplane $g : \mathbf{g}^\intercal \mathbf{x} = g_0$ that supports this classifier can only commit inequality-reinforcing errors: only false positives on group A members and only false negatives on group B members.

As proved in Lemma 2, the set $\mathcal{L}_m(g)$ determines the effective threshold on unmanipulated features $\mathbf{x}$ for a candidate of group m. We have already shown that for any two $\ell_1, \ell_2 \in \mathcal{L}_m(g)$,

$$\sum_{i=1} g_i \ell_{1,i} = \sum_{i=1} g_i \ell_{2,i} = g_0 - \frac{g_{k_m}}{c_{k_m}}$$

where $k_m \in \arg\max_{i=[d]} \frac{g_i}{c_{m,i}}$. For any $\ell \in \mathcal{L}_B(g)$, we have

$$\sum_{i=1}^{d} g_i \ell_i + \frac{g_{k_B}}{c_{B,k_B}} = g_0$$

Thus combining these results, those group B candidates with features $\mathbf{x} \in [0,1]^d$ in the intersection

$$\mathbf{g}^\intercal \mathbf{x} < g_0 - \frac{g_{k_B}}{c_{B,k_B}} \bigcap \mathbf{w}_B^\intercal \mathbf{x} \geqslant \tau_B$$

are classified as false negatives. For group A, we consider $\ell \in \mathcal{L}_A(g)$:

$$\sum_{i=1}^{d} g_i \ell_i + \frac{g_{k_A}}{c_{A,k_A}} = g_0$$

and thus group A candidates with features $\mathbf{x} \in [0,1]^d$ in the intersection

$$\mathbf{w}_A^\intercal \mathbf{x} < \tau_A \bigcap \mathbf{g}^\intercal \mathbf{x} \geqslant g_0 - \frac{g_{k_A}}{c_{A,k_A}}$$

are classified as false positives. Thus the cost publishing g is

$$C_{FN} P_{x \sim \mathcal{D}_B}\big[\mathbf{x} \in \big(\mathbf{g}^\intercal \mathbf{x} < g_0 - \frac{g_{k_B}}{c_{k_B}} \bigcap \mathbf{w}_B^\intercal \mathbf{x} \geqslant \tau_B\,)\big]$$
$$+\, C_{FP} P_{x \sim \mathcal{D}_A}\big[\mathbf{x} \in \big(\mathbf{w}_A^\intercal \mathbf{x} < \tau_A \bigcap \mathbf{g}^\intercal \mathbf{x} \geqslant g_0 - \frac{g_{k_A}}{c_{k_A}})\big]$$

□

A.1.3 Proofs from Section 2.4

Reduction from the d-dimensional setting to the one-dimensional setting

We first show that under certain conditions of a learner's equilibrium classifier strategy, a d-dimensional subsidy analysis is equivalent to a one-dimensional subsidy analysis.

In general d-dimensions, those features $\mathbf{y}$ attainable from an unmanipulated feature $\mathbf{x} \in [0,1]^d$, where $f(\mathbf{x}) = 0$, is given by

$$\mathbf{y} \leqslant \mathbf{x} + \sum_{i=1}^{d} \frac{t_i}{c_i} \mathbf{e}_i \text{ where } \sum_{i=1}^{d} t_i = 1$$

where the right hand side gives the simplex $\Delta(\mathbf{x})$ of potential manipulation. By Lemma 1, if a candidate moves from $\mathbf{x}$ to $\mathbf{y} \neq \mathbf{x}$, then she selects $\mathbf{t}$ such that $t_j = 0$ for all $j \notin K = \arg\max_{i=[d]} \frac{g_i}{c_i}$. Staying within the simplex implies $\sum_{i=1}^{d} t_i \leqslant 1$.

Increasing the candidate's available cost to expend from 1 to n increases her range of motion such that now she can move to any

$$\mathbf{y} \leqslant \mathbf{x} + \sum_{i=1}^{d} \frac{t_i}{c_i} \mathbf{e}_i \text{ where } \sum_{i=1}^{d} t_i = n$$

She continues to manipulate in the spirit of Lemma 1—optimal moves entail choices of $\mathbf{t}$ such that $t_j = 0$ for all $j \notin K$—however now, she is willing to manipulate if $\exists i \in [d]$ such that

$$f(\mathbf{x} + \frac{n}{c_i} \mathbf{e}_i) = 1$$

and thus chooses $\mathbf{t}$ such that $\sum_{i=1} t_i \leqslant n$.

Since offering a subsidy does not change the form of the group B cost function, a candidate from group B will pursue the same manipulation strategy given by the vector $\mathbf{t}$ under subsidy regimes as long as the classifier's decision boundaries stay the same. By definition, all such choices of $\mathbf{y}$ resulting from a manipulation via $\mathbf{t}$ have equivalent values $\mathbf{g}^\intercal \mathbf{y}$.

When costs are subsidized through a flat α or a proportional β subsidy, a candidate with feature $\mathbf{x}$ can manipulate to any $\mathbf{y}_\alpha, \mathbf{y}_\beta \geqslant \mathbf{x}$ that satisfies

$$\mathbf{y}_\alpha \in [\mathbf{x}, \mathbf{x} + \sum_{i=1}^{d} \frac{t_i}{c_i} \mathbf{e}_i] \text{ where } \sum_{i=1}^{d} t_i = 1 + \alpha \tag{A.8}$$

$$\mathbf{y}_\beta \in [\mathbf{x}, \mathbf{x} + \sum_{i=1}^{d} \frac{t_i}{c_i} \mathbf{e}_i] \text{ where } \sum_{i=1}^{d} t_i = \frac{1}{\beta} \tag{A.9}$$

We can pursue a dimensionality reduction by mapping each feature $\mathbf{x} \in [0,1]^d$ to $\mathbf{g}^\intercal \mathbf{x} \in \mathbb{R}_+$. Rather than considering an optimal manipulation in d-dimensions from $\mathbf{x}$ to $\mathbf{y}$, we instead consider the relationship between the cost of the manipulation and the change from $\mathbf{g}^\intercal \mathbf{x}$ to $\mathbf{g}^\intercal \mathbf{y}$:

$$\sum_{i=1}^{d} c_i (y_i - x_i) \iff \sum_{i=1}^{d} g_i (y_i - x_i)$$

where g_i gives the coefficients of the linear decision boundary that supports f, and $\mathbf{x}$ optimally manipulates to $\mathbf{y}$. We want to show that such a relationship is linear.

Consider optimal manipulations: If a candidate chooses not to manipulate at all, she will incur a cost of 0 and will also move from $\sum_{i=1}^{d} g_i(y_i - x_i) = 0$. Since optimal manipulations (under any "budget" constraint) only are along k^{th} components, a move from $\mathbf{x}$ to $\mathbf{y}$ always entails a total cost of

$$\sum_{i \in K}^{d} c_i (y_i - x_i)$$

accompanied with

$$\sum_{i \in K}^{d} g_i (y_i - x_i) = \mathbf{g}^\intercal (\mathbf{y} - \mathbf{x})$$

Thus we can write her total cost c for a move from $\mathbf{x}$ to $\mathbf{y}$ as

$$\frac{c_k}{g_k}(\mathbf{g}^\mathsf{T}\mathbf{y} - \mathbf{g}^\mathsf{T}\mathbf{x}) \tag{A.10}$$

for any $k \in K$. Recall that by Lemma 1, optimal non-stationary manipulations move from $\mathbf{x}$ to $\mathbf{y} > \mathbf{x}$ such that $\sum_{i=1}^{d} g_i y_i = g_0$, so in these cases, we can also write the above as

$$\frac{c_k}{g_k}(g_0 - \mathbf{g}^\mathsf{T}\mathbf{x})$$

Thus we can consider candidates' unmanipulated d-dimensional features $\mathbf{x}$ as one-dimensional features $\mathbf{g}^\mathsf{T}\mathbf{x}$ and classifiers f based on d-dimensional hyperplanes $g \ : \ \sum_{i=1}^{d} g_i x_i \ = \ g_0$ as imposing one-dimensional thresholds g_0.

However a learner may also choose a different optimal subsidy strategy, thus publishing a classifier that now admits candidates differently. Formally, suppose a learner first publishes a classifier f_1 based on a decision boundary $g_1 \ : \ \sum_{i=1}^{d} g_{1,i} x_i \ = \ g_{1,0}$ to which a candidate's optimal response follows the form given in Lemma 1 with $k_1 \in \arg\max_{i \in [d]} \frac{g_{1,i}}{c_i}$. If a learner then chooses to change her strategy when implementing a subsidy, thus publishing a different classifier f_2 based on decision boundary $g_2 \ : \ \sum_{i=1}^{d} g_{2,i} x_i \ = \ g_{2,0}$, a candidate's optimal manipulation strategy will continue to adhere to Lemma 1, however, now, $k_2 \in \arg\max_{i \in [d]} \frac{g_{2,i}}{c_i}$. Whereas the corresponding one-dimensional cost function $c(\mathbf{y}) - c(\mathbf{x})$ for best-response manipulations when facing classifier f_1 was given by

$$\frac{c_{k_1}}{g_{1,k_1}}(\mathbf{g}_1^\mathsf{T}(\mathbf{y} - \mathbf{x}))$$

Her corresponding cost function when facing classifier f_2 is

$$\frac{c_{k_2}}{g_{2,k_2}}(\mathbf{g}_2^\mathsf{T}(\mathbf{y} - \mathbf{x}))$$

When these cost functions are the same, as when the coefficients $g_{1,i} = g_{2,i}$ for all i, the agent's strategies when facing f_1 and f_2 are identical when reduced to one-dimension. This case arises, for example, when the learner continues to perfectly classify a single group in both the non-subsidy regime and the subsidy regime. In these cases, we can transition to considering just one-dimensional manipulations from $\mathbf{g}^\intercal \mathbf{y}$ to $\mathbf{g}^\intercal \mathbf{x}$, where candidates bear linear costs of manipulation given in (A.10).

Proof of Proposition 5

Working from the subsidy and no-subsidy comparisons given in Proposition 2, we show that all three parties would have preferred the outcomes of a non-manipulation world to those in both of the manipulation cases.

To facilitate comparisons of welfare across classification regimes, we formalize group-wide utilities in the following definition.

Definition 7 (Group welfare under a proportional subsidy). *The average welfare of group B under classifier f_{prop} and a proportional subsidy with parameter β is given by*

$$W_B(f_{prop}, \beta) = \int_{R_1} P_{x \sim \mathcal{D}_B}(x) dx + \int_{R_2} \big(1 - \beta(c_B(y(x)) - c_B(x))\big) P_{x \sim \mathcal{D}_B}(x) dx,$$

$$W_A(f_{prop}, 1) = \int_{R_1} P_{x \sim \mathcal{D}_A}(x) dx + \int_{R_2} \big(1 - (c_A(y(x)) - c_A(x))\big) P_{x \sim \mathcal{D}_A}(x) dx,$$

where $y(x)$ is the best response of a candidate with unmanipulated feature x, R_1 sums over those candidates who are positively classified by f_{prop} without expending any cost, and R_2 sums over those candidates who are positively classified after manipulating their features. Since group A members do not receive subsidy benefits, their welfare form is the same across no-subsidy and subsidy regimes.

We use $W_A(f_{prop})$ to denote $W_A(f_{prop}, 1)$, the average welfare for group A under classifier f_{prop} with *no subsidy.*

Definition 8 (Group welfare in a non-manipulation setting). *The average welfare of group m under classifier f_0 in a non-manipulation setting is given by*

$$W_m(f_0) = \int_R P_{x\sim\mathcal{D}_m}(x)dx$$

where R sums over candidates who are positively classified by f_0.

Proposition 10. *There exist cost functions c_A and c_B satisfying the cost conditions, learner distributions $\mathcal{D}_A$ and $\mathcal{D}_B$, true classifiers with threshold τ_A and τ_B, population proportions p_A and p_B, and learner penalty parameters C_{FN}, C_{FP}, and λ, such that*

$$W_A(f^*_{prop}) < W_A(f^*_0), \qquad W_B(f^*_{prop}, \beta^*) < W_B(f^*_0),$$

$$W_A(f^*_1) < W_A(f^*_0), \qquad W_B(f^*_1) < W_B(f^*_0),$$

$$C(f^*_{prop}, \beta^*) > C(f^*_0), \qquad C(f^*_1) > C(f^*_0)$$

where f_0 is the equilibrium classifier in the non-manipulation regime, f^*_1 is the equilibrium classifier in the manipulation regime, and (f^*_{prop}, β^*) is the equilibrium classifier in the subsidy regime. The average welfare of each group, $W_m(\cdot)$, as well as the learner, $1 - C(\cdot)$, is higher at the equilibrium of the non-manipulation game compared with the equilibria of the Strategic Classification Game with proportional subsidies and compared with the equilibrium of the Strategic Classification Game with no subsidies.

Example 2. *Now we consider a case in which candidates have linear cost functions $c_A(x) = 3x$ and $c_B(x) = 4x$. To show that diminished welfare for both candidate groups can occur without requiring distortions of probability distributions or cost functions, we consider a learner who seeks to avoid errors* *on group B in both the subsidy and the non-subsidy regimes by penalizing false negatives twice as much as false positives, with* $C_{FN} = \frac{2}{3}$, $C_{FP} = \frac{1}{3}$, *and* $\lambda = \frac{3}{4}$. *As in the previous example, we assume that the underlying unmanipulated features for both groups are uniformly distributed with* $p_A = p_B = \frac{1}{2}$, *and that* $\tau_A = 0.4$ *and* $\tau_B = 0.3$.

Now the equilibrium learner classifier without subsidies is based on threshold $\sigma_1^* = \sigma_B = 0.55$, *which perfectly classifies all candidates from group B, while permitting false positives on candidates from group A with features* $x \in [0.217, 0.4)$. *Under a proportional subsidy intervention, the learner's equilibrium action is to choose threshold* $\sigma_{prop}^* = \sigma_B^\beta \approx 0.552$ *and* $\beta^* = 0.994$, *which again perfectly classifies B candidates. Notice that now her optimal threshold commits fewer false positive errors on group A members, while still committing false positives on those members with features* $x \in [0.219, 0.4)$.

Here, even when the learner has a cost penalty that is explicitly concerned with mistakenly excluding group B candidates and then seeks to offer a subsidy benefit to further alleviate their costs, group B members are still no better off. They receive the same classifications as before and it can be shown that all candidates who manipulate must spend more to reach the higher threshold, even while accounting for the subsidy benefit! Some group A candidates are also worse off since the threshold has increased, and they receive no subsidy benefits. As before, only the learner gains from the intervention.

Example 3. *This example is based on Example 2. Now we consider the case a learner seeks* $\sigma_1^* \in [\sigma_B, \sigma_A]$ *where* $\sigma_A = 0.733$ *and* $\sigma_B = 0.55$. *Suppose she seeks to equalize the number of false positives she commits on group A and the number of false negatives for group B and thus chooses* $\sigma_1^* = 0.64$ *such that*

$$\ell_A(\sigma_1^*) = 0.31$$

$$\ell_B(\sigma_1^*) = 0.39$$

Thus group B candidates with features $x \in [0.3, 0.39)$ *are mistakenly excluded, and group A candi-*

dates with features $x \in [0.31, 0.4)$ *are mistakenly admitted.*

Upon implementing a subsidy and minimizing the same error penalty as in Example 1, the learner selects an optimal proportional β subsidy such that

$$\sigma^*_{prop} = \sigma_A = 0.733; \beta = 0.806$$

Under this regime, group B members are worse-off because many more candidates now receive false negative classifications

$$x \in [0.3, 0.423)$$

Others who do secure positive classifications must pay more to do so. Candidates in group A are now perfectly classified, though this actually entails a welfare decline, since some candidates lose their false positive benefits. The learner is also strictly better off with a total penalty decline

$$C(\sigma^*_0) = 0.183 \to C(\sigma^*_{prop}, \beta^*) = 0.128$$

Recall that the learner's utility is given by $1 - C(\cdot)$*. Thus we have that*

$$W_A(\sigma^*_{prop}, \beta^*) < W_A(\sigma^*_1)$$

$$W_B(\sigma^*_{prop}, \beta^*) < W_B(\sigma^*_1)$$

$$C(\sigma^*_1) > C(\sigma^*_{prop}, \beta^*)$$

Now consider a non-manipulation regime, in which the learner selects to equalize the number of false negatives for group B and the number of false positives for group A, she now chooses a threshold on un-

manipulated features

$$\sigma_0^* = 0.35$$

Some group A candidates lose false positive benefits in the manipulation regime, though on the whole, the group fares better off because all those candidates with features

$$x \in [0.39, 0.64)$$

need not expend any costs in order to receive a positive classification. Group B candidates are strictly better off since they both receive fewer false negatives and need not pay to manipulate. The learner is also better off here because she reduces her error down to $C(\tau^*) = 0.1$. *Thus comparing the non-manipulation regime, the no-subsidy manipulation regime, and the subsidy regime, we have that utility comparisons for all three parties is given by*

$$W_A(\sigma_0^*) > W_A(\sigma_1^*) > W_A(\sigma_{prop}^*, \beta^*)$$

$$W_B(\sigma_0^*) > W_B(\sigma_1^*) > W_B(\sigma_{prop}^*, \beta^*)$$

$$1 - C(\sigma_0^*) > 1 - C(\sigma_{prop}^*, \beta^*) > 1 - C(\sigma_1^*)$$

A.1.4 Flat Subsidies

Here we give analogous definitions and results for flat subsidies in which the learner absorbs up to a flat α amount from each group B candidate's costs and show that qualitatively similar results hold.

Definition 9 (Flat subsidy). *Under a flat subsidy plan, the learner pays an* $\alpha > 0$ *benefit to all members of group B. As such, a group B candidate who manipulates from an initial score* $\mathbf{x}$ *to a final score* $\mathbf{y} \geqslant \mathbf{x}$ *bears a cost of* $\max\{0, c_B(\mathbf{y}) - c_B(\mathbf{x}) - \alpha\}$.

A learner's strategy now consists of both a choice of α and a choice of classifier f to issue. The learner's goal is to minimize her penalty

$$C_{FP} \sum_{m \in \{A,B\}} p_m P_{\mathbf{x} \sim \mathcal{D}_m}[h_m(\mathbf{x}) = 0, f(\mathbf{y}) = 1] + C_{FN} \sum_{m \in \{A,B\}} p_m P_{\mathbf{x} \sim \mathcal{D}_m}[h_m(\mathbf{x}) = 1, f(\mathbf{y}) = 0] + \lambda cost(f, \alpha),$$

We can define

$$\ell_B^\alpha(y) = c_B^{-1}\Big(c_B(y) - (1+\alpha)\Big).$$

Under the α subsidy, for an observed feature y, the group B candidate must have unmanipulated feature $x \geqslant \ell_B^\alpha(y)$.

From these functions, we define σ_B^α and σ_B^β such that $\ell_B^\alpha(\sigma_B^\alpha) = \tau_B$, and $\ell_B^\beta(\sigma_B^\beta) = \tau_B$. Under a flat α subsidy, setting a threshold at σ_B^α correctly classifies all group B members; under a proportional β subsidy, a threshold at σ_B^β correctly classifies all group B members.

From this, we define σ_B^α such that $\ell_B^\alpha(\sigma_B^\alpha) = \tau_B$. Under a flat α subsidy, setting a threshold at σ_B^α correctly classifies all group B members.

In order to compute the cost of a subsidy plan, we must determine the number of group B candidates who will take advantage of a given subsidy benefit. Since manipulation brings no benefit in itself, candidates will still only choose to manipulate and use the subsidy if it will lead to a positive classification. For the flat α subsidy, $cost(f, \alpha)$ is given by

$$\int_{c_B^{-1}(c_B(\sigma)-\alpha)}^{\sigma} [c_B(\sigma) - c_B(x)] P_{D_B}(x) dx + \alpha \int_{\ell_B^\alpha(\sigma)}^{c_B^{-1}(c_B(\sigma)-\alpha)} P_{D_B}(x) dx,$$

where σ is the threshold for classifier f. The first integral refers to the benefits paid out to candidates with manipulation costs less than the α amount offered. The latter refers to the total sum of full α payments offered to those with costs greater than α.

Definition 10 (Group welfare under a flat subsidy). *The average welfare of group B under classifier f* *and a flat subsidy with parameter α is given by*

$$W_B(f, \alpha) = \int_{R_1} P_{x \sim \mathcal{D}_B}(x)dx + \int_{R_2} (1 - c_B(y(x) - c_B(x)))P_{x \sim \mathcal{D}_B}(x)dx$$

where $y(x)$ is the best response of a candidate with unmanipulated feature x, R_1 sums over those candidates who are positively classified without expending any cost, and R_2 sums over those candidates who are positively classified after manipulating their features. Note that under the flat subsidy, group B costs have the form $\max\{0, c_B(y) - c_B(x) - \alpha\}$ *The formulation of average group A welfare is the same in this setting and follows the same form given in Definition 5.*

Theorem 7 (Subsidies can harm both groups). *There exist cost functions c_A and c_B satisfying the cost conditions, learner distributions $\mathcal{D}_A$ and $\mathcal{D}_B$, true classifiers with threshold τ_A and τ_B, population proportions p_A and p_B, and learner penalty parameters C_{FN}, C_{FP}, and λ, such that*

$$W_A(f^*_{prop}) < W_A(f^*_0), \qquad W_B(f^*_{prop}, \alpha^*) < W_B(f^*_0),$$

where f^*_{prop} and α^* are the learner's equilibrium classifier and subsidy choice in the Strategic Classification Game with flat subsidies and f^*_0 is the learner's equilibrium classifier in the Strategic Classification Game with no subsidies.

A.2 Appendix for Chapter 3

Proof of Proposition 1

We want to show that the firm-set reputation threshold $\hat{\Pi}^{t'} = p_H - \Delta_{t'}$, where t' is the time since the last wage update, enforces a worker strategy of effort exertion akin to that of the one-shot game, in which a worker exerts high effort if she can afford to do so and low effort otherwise. The firm, by setting its reputation threshold $\approx p_H$, is correctly restricting its membership to workers who appear to be consistently exerting high effort. By the Law of Large Numbers, a worker's recent time t' individual reputation $\Pi_i^{t'} \to p_H$ almost surely as $t' \to \infty$ as long as she continuously exerts high effort at each time step. Moreover, since the relationship between effort exertion and G or B outcomes can considered Bernoulli trials with $p = p_H$, we use the law of the iterated logarithm to bound individual good workers' reputational deviations away from the theoretical mean p_H as t increases and have that for all $t = \tau$,

$$|\Pi^{\tau} - \hat{\Pi}^{\tau}| \leqslant \sqrt{\tau^{-1}(2 * 0.25 \log \log \tau)} \tag{A.11}$$

Rubinstein and Yaari[81] have shown that, for a similar setup of imperfect observability and moral hazard in repeated interactions between insurers and clients, the enforceability of the insurers' strategies is dependent on the choice of the forgiveness buffer sequence. In our case, as long as $\Delta_\tau > \sqrt{\tau^{-1}(0.5 \log \log \tau)}$ and the sequence $\Delta_{t'} \to 0$ monotonically, the Rubinstein-Yaari result carries over into employment relationships, and workers will always exert high effort when they can afford to do so. Importantly, our scenario does differ from theirs in two ways: 1) Workers do not stay in the labor market for an infinite number of rounds, 2) A firm must pay the labor-market-wide wage upon hiring a worker and cannot unilaterally deviate from the set price. Since workers exit the market according to a Poisson parameter λ and the wage premium $w_t = w(g_{t'}) > 0$ is set to always provide a higher payoff for a worker than failing to be hired at all (due to the normalization with respect to the unskilled job wage), the memoryless death process ensures that a worker i with qualifications ρ will always find it within her interest to pursue the skilled job as long as it is individually rational for her to do so, i.e. $e_\rho(\theta_i) \leqslant w_t(p_H - p_\rho)$.

Proof of Theorem 3

The TLM hiring constraint effects two guarantees: 1) It retains the fundamental equality of groups' ability level distributions $F(\theta)$ within the labor market; 2) It results in statistical parity in the proportion of workers offered skilled jobs in the TLM. Since the instantaneous time t contributions to groups' full population societal reputations π^μ are equivalent to g^μ up to the same constant factor (ℓ proportion who enter the TLM), showing that the g^μ values converge is sufficient to show that group reputations π^μ do as well.

Consider $g_{t+1} = \xi(g_t)$ as a self-mapping $\xi : X \to X$ where X is the unit interval $[0, 1]$. Groups μ and ν have the same functional form of ξ differing only in a few particular parameters, which will be addressed in the decomposition of ξ into two separate functions. Assuming the two groups begin with unequal societal reputations, we suppose that (without loss of generality) $\pi_\nu < \pi_\mu$. We want to show that regardless of initial values $\pi_0^\nu < \pi_0^\mu$, hiring outcomes will converge to achieve equal group outcomes system-wide under labor market dynamics with the TLM fairness constraint.

Due to effect 1) of the TLM hiring constraint and the fact that both groups experience the same labor-market-wide wage $w(g_t)$, the PLM ability thresholds $\widehat{\theta_Q}$ and $\widehat{\theta_U}$ are also equivalent across groups. Thus the difference between the g_t^μ and g_t^ν arises due to the different corresponding proportions of qualified workers $\gamma_t^\nu < \gamma_t^\mu$ at time t. As such, we construct the function φ as a mapping of $\gamma_t \in [0, 1]$ to $g_{t+1} \in [0, 1]$, such that $g_{t+1} = \varphi(\gamma_t)$. The function φ is generic across the two groups, and group differences are entirely encoded in the distinct inputs γ_t^μ and γ_t^ν.

Let's call $g_{t+1}^\mu = \varphi(\gamma_t^\mu)$ and $g_{t+1}^\nu = \varphi(\gamma_t^\nu)$, where we treat γ^μ and γ^ν as distinct points of the mapping φ. Then, we have

$$\begin{aligned} g^{\mu}_{t+1} =& p_H[1 - F(\widehat{\theta}_Q)\gamma^{\mu}_t - F(\widehat{\theta}_U)(1 - \gamma^{\mu}_t)] + p_Q F(\widehat{\theta}_Q)\gamma^{\mu}_t \\ &+ p_U F(\widehat{\theta}_U)(1 - \gamma^{\mu}_t) \end{aligned} \tag{A.12}$$

The difference $|g^{\mu}_{t+1} - g^{\nu}_{t+1}|$ is thus equivalent to the following

$$\begin{aligned} |\varphi(\gamma^{\mu}_t) - \varphi(\gamma^{\nu}_t)| &= | - p_H F(\widehat{\theta}_Q)(\gamma^{\mu}_t - \gamma^{\nu}_t) + p_H F(\widehat{\theta}_U)(\gamma^{\mu}_t - \gamma^{\nu}_t) \\ &+ p_Q F(\widehat{\theta}_Q)(\gamma^{\mu}_t - \gamma^{\nu}_t) - p_U F(\widehat{\theta}_U)(\gamma^{\mu}_t - \gamma^{\nu}_t)| \\ &= (\gamma^{\mu}_t - \gamma^{\nu}_t)|p_H[F(\widehat{\theta}_U) - F(\widehat{\theta}_Q)] + p_Q F(\widehat{\theta}_Q) - p_U F(\widehat{\theta}_U)| \end{aligned}$$

We rewrite the quantity inside the absolute value:

$$|\underbrace{F(\widehat{\theta}_U)[p_H - p_U]}_{\in(0,1)} + \underbrace{F(\widehat{\theta}_Q)[p_Q - p_H]}_{\in(-1,0)}| = |\varepsilon_t| < 1$$

Together, $|g^{\mu}_{t+1} - g^{\nu}_{t+1}| = |\varphi(\gamma^{\mu}_t) - \varphi(\gamma^{\nu}_t)| \leqslant |\varepsilon_t|(\gamma^{\mu}_t - \gamma^{\nu}_t), \forall \gamma^{\mu}_t, \gamma^{\nu}_t \in [0, 1]$, and with the bound on ε, φ is a contraction mapping.

Since group reputation considers the proportion of *all* members in a group who are producing good outcomes, statistical parity also has the upshot that a particular instantaneous time t group reputation π^{μ} exactly scales with g^{μ} as each group is proportionally represented within the labor market according to its population-wide demographic share, so we need only consider g^{μ}_t values to determine the feedback loop property of collective reputation π^{μ}_t and group cost functions c_{μ} and c_{ν}. Thus, the mapping $\psi : X \to X$, which maps normalized $g_{t+1} \in X = [0, 1]$ to $\gamma_{t+1} \in X = [0, 1]$ such that $\gamma_{t+1} = \psi(g_{t+1})$, is a weakly contracting map.

We can now rewrite the recursive system $g_{t+1} = \xi(g_t)$ as a composition: $g_{t+1} = \xi(g_t) =$

$\varphi(\psi(g_t))$, where we have shown that φ is a contraction and ψ is a short map. Then their composition ξ, which represents the recursive self-map determining the evolution of group-wide employment outcomes, is also a contraction map.

Then by the Banach Fixed Point Theorem, there is a unique fixed-point $\tilde{g} = \xi(\tilde{g})$ such that all initial points $g_i \in [0,1]$ converge to $\tilde{g}$ via a sequence of applications of the recursive relation ξ as in (A.12): For any two group reputations π^μ and π^ν corresponding to initial points g_0^μ and g_0^ν, there exists a T such that $\forall t > T$, $\pi_t^\mu = \pi_t^\nu = \tilde{\pi}$ (similarly with g^μ). At equilibrium, there is a unique wage $\tilde{w}$ corresponding to $\tilde{g}$, and the system admits group fairness. □

Proof of Theorem 4

To show that the contraction and convergence assured by statistical parity hiring is not guaranteed under group-blind hiring, note that when $\pi_B < \pi_W$, necessarily $1 - F(\widetilde{\theta_B}) < 1 - F(\widetilde{\theta_W})$, and the composition of workers granted entry into the TLM does not satisfy statistical parity. We call the proportion of workers in the TLM belonging to groups B and W, k^B and $1 - k^B$ respectively. Similarly to the proof of Theorem 3, we decompose g_t^μ into the feed-forward labor market flow effect and the feedback natural reputational effect. However, since γ^μ for the two groups are the same, and $F(\widetilde{\theta_\mu})$ values differ, we instead write labor market flow as a function of $F(\widetilde{\theta_\mu})$, call it φ^*.

Then $g_{t+1}^W = \varphi(F(\widetilde{\theta_Q}))$ and $g_{t+1}^B = \varphi(F(\widetilde{\theta_B}))$, and

$$|\varphi(F(\widetilde{\theta_Q})) - \varphi(F(\widetilde{\theta_B}))| = (F(\widetilde{\theta_B}) - F(\widetilde{\theta_Q}))\gamma(p_H - p_Q)$$

Since $\gamma(p_H - p_Q) < 1$, φ thus also contracts in the feed-forward mechanism, however the function only captures the proportional g_t^μ dynamics from the TLM into the PLM, which does not scale with group reputation π^μ since statistical parity is not guaranteed. Instead, under group-

[*] Note that in this proof, we also assume that $\widetilde{\theta_B} \quad < \quad \widetilde{\theta_U}$, but the proof carries through in the exact same manner when this is not true.

blind hiring, group reputation, which captures the proportion of *all* workers in the group who are producing good outcomes in the skilled labor, is a function of k^B, or the bottleneck of group proportionality created by the group-blind investment threshold. Thus the particular time t normalized group societal reputation $\pi_t^B \propto \frac{k^B g_t^B}{\sigma_B} < g_t^B$ and $\pi_t^W \propto \frac{(1-k^B) g_t^W}{1-\sigma_B} > g_t^W$, and as a result, $|\pi_t^W - \pi_t^B| > |g_{t+1}^W - g_{t+1}^B|$. Since the mapping from $g_t^\mu \rightarrow \pi_t^\mu$ is not a contraction, the reputation feedback is not guaranteed to contract either. The system may thus reach an asymmetric equilibrium in which groups B and W maintain distinct investment costs and equal group reputations are never recovered.

We now show that this asymmetric outcome is Pareto-dominated by the hiring constraint-produced symmetric steady-state when PLM firms' demand for workers is not saturated and $w(\widetilde{g_t}) = \bar{w}$. For the two groups, B and W, group-blind hiring imposes a single investment threshold $\tilde{\eta}$ such that hired workers in both labor markets have the same probability of being qualified regardless of group membership: $\gamma^\mu = \gamma^\nu = \gamma$. Suppose group reputations are not equal as in the case of the group-blind asymmetric equilibrium just proven, then group-blind hiring results in effective ability thresholds that may be ranked with respect to the threshold $\bar{\theta}$ under statistical parity hiring. If $\pi_B < \pi_W$, then $\widetilde{\theta_W} < \bar{\theta} < \widetilde{\theta_B}$. Note that throughout the chapter, it is assumed that not all workers in the TLM are able to be hired in the PLM; therefore the ability threshold for exerting on-the-job effort is greater than the ability threshold resulting from the investment threshold under statistical parity-constrained hiring: $\widehat{\theta_Q} > \bar{\theta}$.

When $\widehat{\theta_Q} < \widetilde{\theta_B}$, then TLM group-blind hiring leaves behind high ability workers in group B who would have otherwise been hired in the PLM. In particular, all qualified workers in group B with ability level $\theta \in [\widehat{\theta_Q}, \widetilde{\theta_B})$ are only hired in the fairness constrained equilibrium; under group-blind hiring, they are barred from entering the TLM. This result accords with the vicious circle of the asymmetric equilibrium, since the reputation gap $|\pi_t^B - \pi_t^W|$ and consequently, differences in group investment costs are maintained.

Further, since $1 - F_g(\widehat{\theta_\rho}) < 1 - F_f(\widehat{\theta_\rho})$ where F_g and F_f are the ability CDFs under the group-blind and fair regime respectively, in a labor market that demands more workers yet cannot sustain a higher wage ($\tilde{w} = w$)[†], firms strictly prefer the steady-state equilibrium under the fairness constraint. This is because the effective higher ability threshold for group B under the group-blind TLM strategy is inefficient, leaving behind an untapped resource of skilled and qualified individuals in group B who would have otherwise been hired in the PLM. Even those workers in group W with ability level $\theta \in [\widetilde{\theta_W}, \widehat{\theta_Q})$ who are only allowed to enter the TLM in the group-blind regime do not fare better, since all such workers have ability level lower than the PLM reputation threshold and are not hired at equilibrium anyway. Thus since some workers in group B are strictly better off and workers in group W no worse off, the asymmetric equilibria under group-blind hiring is Pareto-dominated by the symmetric one of the fair case.

The proof of this result for the statistical discriminatory hiring regime follows similarly. If $\xi_W > \xi_B$, then $P(Q|W, \eta) > P(Q|B, \eta)$, and the groups face different incentive compatibility constraints. Self-confirming asymmetric equilibria also exist under this regime,[22] and using the same argument about lost efficiency due to inequitable ability thresholds in the TLM for group B, these equilibria are also Pareto-dominated by hiring that abides by statistical parity.

[†] There are a variety of reasons why an association of firms that demand more workers would be unable or unwilling to raise its wage higher $\tilde{w} = w$: A higher wage may encourage lower ability workers to apply and exert effort, and in reality, probabilities of success p_H may be variable according to ability; thus the firm may want to *a priori* exclude such workers. Wage caps may also result from firm-firm collusion on price.

A.3 Appendix for Chapter 4

A.3.1 Dual derivations of the ε-fair SVM program

In this Appendix section, we walk through the preliminary setup of the ε-fair SVM program given in Section 5.1 and present intermediate derivations omitted from the main text.

Recall that the fair empirical risk minimization program of central focus is

$$\begin{aligned} \underset{\boldsymbol{\theta}, b}{\text{minimize}} \quad & \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & y_i(\boldsymbol{\theta}^\intercal \mathbf{x}_i + b) - 1 + \xi_i \geqslant 0, \\ & \xi_i \geqslant 0, \\ & f_{\boldsymbol{\theta},b}(\mathbf{x}, y) \leqslant \varepsilon \end{aligned} \tag{ε-fair Soft-SVM}$$

The hyperplane parameters are $\boldsymbol{\theta} \in \mathbb{R}^d$ and $b \in \mathbb{R}$. The non-negative ξ_i allow the margin constraints to have some slack—this is why these variables are commonly called "slack variables." In the Soft-Margin (as opposed to the Hard-Margin) SVM, the margin is permitted to be less than 1. A slack variable $\xi_i > 0$ corresponds to a point $\mathbf{x}_i$ having a functional margin of less than 1. There is a cost associated with this margin violation, even though it need not correspond to a classification error. $C > 0$ is a hyperparameter tunable by the learner to optimize this trade-off between preferring a larger margin and penalizing violations of the margin.

When we combine the general Soft-Margin SVM with the covariance parity constraint in (4.4) proposed by Zafar et al.[96], we have the program

$$\begin{aligned}
&\underset{\boldsymbol{\theta}, b}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^{n}\xi_i \\
&\text{subject to} \qquad y_i(\boldsymbol{\theta}^\intercal \mathbf{x}_i + b) - 1 + \xi \geqslant 0, \\
&\qquad\qquad\quad |\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(\boldsymbol{\theta}^\intercal \mathbf{x}_i + b)| \leqslant \varepsilon
\end{aligned} \tag{ε-fair-SVM1-P}$$

where $\bar{z}$ reflects the bias in the demographic makeup of $\mathcal{X}$: $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$. The corresponding Lagrangian is

$$\begin{aligned}
&\mathcal{L}_P(\boldsymbol{\theta}, b, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \gamma_1, \gamma_2) \\
&\quad = \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\lambda_i - \sum_{i=1}^{n}\mu_i(y_i(\boldsymbol{\theta}^\intercal \mathbf{x}_i + b) - 1 + \xi_i) \\
&\quad - \gamma_1\big(\varepsilon - \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(\boldsymbol{\theta}^\intercal \mathbf{x}_i + b)\big) \\
&\quad - \gamma_2\big(\varepsilon - \frac{1}{n}\sum_{i=1}^{n}(\bar{z} - z_i)(\boldsymbol{\theta}^\intercal \mathbf{x}_i + b)\big)
\end{aligned} \tag{ε-fair-SVM1-L}$$

where $\boldsymbol{\theta} \in \mathbb{R}^d, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^n$ are Primal variables. The (non-negative) Lagrange multipliers $\boldsymbol{\lambda}, \boldsymbol{\mu} \in \mathbb{R}^n$ correspond to the n non-negativity constraints $\xi_i \geqslant 0$ and the margin-slack constraints $y_i(\boldsymbol{\theta}^\intercal \mathbf{x}_i + b) - 1 + \xi_i \geqslant 0$ respectively. The multiplier μ_i relays information about the functional margin of its corresponding point $\mathbf{x}_i$. If the margin is greater than 1 in the Primal, *i.e.*, there is slack in the constraint), then by complementary slackness, $\mu_i = 0$. Otherwise, if the constraint holds with equality, $\mu_i \in (0, C]$. When the classifier commits an error on $\mathbf{x}_i$, $y_i(\boldsymbol{\theta}^\intercal \mathbf{x}_i + b) < 1$, and then by the KKT conditions, $\mu_i = C$.

The multipliers $\gamma_1, \gamma_2 \in \mathbb{R}$ correspond to the two linearized forms of the absolute value fairness constraint. Notice that these two constraints cannot simultaneously hold with equality for $\varepsilon > 0$.

Thus, by complementary slackness again, we know that at least one of γ_1, γ_2 is zero, and the other is strictly positive.

By the Karush-Kuhn-Tucker conditions, at the solution of the convex program, the gradients of $\mathcal{L}$ with respect to $\boldsymbol{\theta}$, b, and ξ_i are zero:

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} &:= 0 \Rightarrow \boldsymbol{\theta} = \sum_{i=1}^{n} \mu_i y_i \mathbf{x}_i - \frac{\gamma}{n}\Big(\sum_{i=1}^{n} (z_i - \bar{z})\mathbf{x}_i\Big) \\
\frac{\partial \mathcal{L}}{\partial b} &:= 0 \Rightarrow \sum_{i=1}^{n} \mu_i y_i = \frac{\gamma}{n}\sum_{i=1}^{n} (z_i - \bar{z}) = 0 \\
\frac{\partial \mathcal{L}}{\partial \xi_i} &:= 0 \Rightarrow \lambda_i + \mu_i = C, \qquad i = 1, \dots, n
\end{aligned}
$$

Plugging in these optimality conditions, the dual Lagrangian is

$$
\mathcal{L}_D(\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \gamma_1, \gamma_2) = -\frac{1}{2}\|\sum_{i=1}^{n} \mu_i y_i \mathbf{x}_i - \frac{\gamma}{n}\sum_{i=1}^{n} (z_i - \bar{z})\mathbf{x}_i\|^2 + \sum_{i=1}^{n} \mu_i - |\gamma|\varepsilon
$$

where we have $\gamma = \gamma_1 - \gamma_2$, since at most one side of the fairness constraint binds, thereby ensuring that at least one of γ_1 or γ_2 is 0. The dual maximizes this objective subject to the constraints $\mu_i \in [0, C]$ for all i and $\sum_{i=1} \mu_i y_i = 0$. Hence, we derive the full dual problem

$$
\begin{array}{lll}
\underset{\boldsymbol{\mu}, \gamma, V}{\text{maximize}} & -\frac{1}{2}\|\sum_{i=1}^{n} \mu_i y_i \mathbf{x}_i - \frac{\gamma}{n}\sum_{i=1}^{n} (z_i - \bar{z})\mathbf{x}_i\|^2 + \sum_{i=1}^{n} \mu_i - V\varepsilon & \\
\text{subject to} & \mu_i \in [0, C], \qquad i = 1, \dots, n, & (\varepsilon\text{-fair-SVM1-D}) \\
& \sum_{i=1}^{n} \mu_i y_i = 0, & \\
& \gamma \in [-V, V] &
\end{array}
$$

where we have introduced the variable V to eliminate the absolute value function $|\gamma|$ in the objective. Notice that when $\gamma = 0$ and neither of the constraints bind, we recover the standard dual

SVM program. Since we are concerned with fair learning that does in fact alter an optimal solution, we consider cases in which V is strictly positive. From this program, we introduce additional dual variables β_- and β_+, corresponding to the $\gamma \in [-V, V]$ constraint and derive the Lagrangian

$$\begin{aligned}\mathcal{L}(\boldsymbol{\mu}, \gamma, V, \beta_-, \beta_+) = &-\frac{1}{2}\|\sum_{i=1}^{n}\mu_i y_i \mathbf{x}_i - \frac{\gamma}{n}\sum_{i=1}^{n}(z_i - \bar{z})\mathbf{x}_i\|^2 + \sum_{i=1}^{n}\mu_i \\ &- V\varepsilon + \gamma(\beta_- - \beta_+) + V(\beta_- + \beta_+)\end{aligned}$$

Under KKT conditions, $\beta_- + \beta_+ = \varepsilon$ and

$$\gamma^* = \frac{n(n(\beta_- - \beta_+) + \sum_{i=1}^{n}\mu_i y_i \langle \mathbf{x}_i, \mathbf{u}\rangle)}{\|\mathbf{u}\|^2} \tag{A.13}$$

where $\mathbf{u} = \sum_{i=1}^{n}(z_i - \bar{z})\mathbf{x}_i$ gives some group-sensitive geometric "average" of $\mathbf{x} \in \mathcal{X}$. We can subsequently rewrite (ε-fair-SVM1-D) as

$$\begin{aligned}\underset{\boldsymbol{\mu}, \beta_-, \beta_+}{\text{maximize}} \quad & -\frac{1}{2}\|\sum_{i=1}^{n}\mu_i y_i (I - P_{\mathbf{u}})\mathbf{x}_i\|^2 + \sum_{i=1}^{n}\mu_i \\ & + \frac{2n\sum_i \mu_i y_i \langle \mathbf{x}_i, \mathbf{u}\rangle + n^2(\beta_- - \beta_+)}{2\|\mathbf{u}\|^2}(\beta_- - \beta_+) \\ \text{subject to} \quad & \mu_i \in [0, C], \qquad i = 1, \ldots, n, \\ & \sum_{i=1}^{n}\mu_i y_i = 0, \\ & \beta_-, \beta_+ \geqslant 0, \\ & \beta_- + \beta_+ = \varepsilon\end{aligned} \tag{ε-fair SVM2-D}$$

where $I, P_{\mathbf{u}} \in \mathbb{R}^{d \times d}$. The former is the identity matrix, and the latter is the projection matrix onto the vector $\mathbf{u}$. As was also observed by Donini et al., the $\varepsilon = 0$ version of (ε-fair SVM2-D) is equivalent to the standard formulation of the dual SVM program with Kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \langle (I - P_{\mathbf{u}})\mathbf{x}_i, (I - P_{\mathbf{u}})\mathbf{x}_j \rangle$.[28]

Since we are interested in the welfare impacts of fair learning when fairness constraints *do* have an impact on optimal solutions, we will assume that the fairness constraint binds. For clarity of exposition, we assume that the positive covariance constraint binds, and thus that $\beta_- = 0$ and $\beta_+ = \varepsilon$ in (ε-fair SVM2-D). This is without loss of generalization—the same analyses apply when the negative covariance constraint binds. The dual ε-fair SVM program becomes

$$\begin{aligned}
\underset{\boldsymbol{\mu}}{\text{minimize}} \quad & \frac{1}{2}\|\sum_{i=1}^{n} \mu_i y_i (I - P_{\mathbf{u}})\mathbf{x}_i\|^2 - \sum_{i=1}^{n} \mu_i + \frac{n\varepsilon(2\sum_i \mu_i y_i \langle \mathbf{x}_i, \mathbf{u}\rangle - n\varepsilon)}{2\|\mathbf{u}\|^2} \\
\text{subject to} \quad & \mu_i \in [0, C], \qquad i = 1, \ldots, n, \qquad\qquad (\varepsilon\text{-fair SVM-D}) \\
& \sum_{i=1}^{n} \mu_i y_i = 0
\end{aligned}$$

A.3.2 Algorithms

Finding the next breakpoint when $|\mathcal{M}^\varepsilon| = 0$

When $|\mathcal{M}^\varepsilon| = 0$, the standard procedure that finds the next breakpoint by computing sensitivities to μ_i in the margin ($i \in \mathcal{M}^\varepsilon$) by inverting the matrix K in (4.12) fails. Without r_i^ε, we also cannot compute changes to d_i for i not in the margin ($i \in \{\mathcal{F}, \mathcal{E}\}^\varepsilon$) as defined in (4.18) to track when points enter the margin. As a result, we need a special procedure to find the next breakpoint when the margin becomes empty.

If the solution is to remain optimal, it must continue to abide by KKT conditions; in particular $\sum_{i=1}^{n} \mu_i y_i = 0$. Notice then that if the margin is empty, we have that $\sum_{i \in \mathcal{E}^\varepsilon} \mu_i y_i = 0 = C \sum_{i \in \mathcal{E}^\varepsilon} y_i$, which means that there are equal numbers of $+1$ and -1 vectors that are misclassified. Thus at the next breakpoint, both $+1$ and -1 vectors will enter the margin at the same time, offsetting each other exactly to retain the optimality of the solution.

Tracking how vectors enter the margin at the solution $p(\varepsilon)$ requires tracking sign changes of $\frac{\partial D^\varepsilon}{\partial \mu}$:

$$\sum_{i=1}^{n} \mu_i y_i (I - P_{\mathbf{u}})\mathbf{x}_i y_j (I - P_{\mathbf{u}})\mathbf{x}_j + \frac{n\varepsilon y_j \langle \mathbf{x}_j, \mathbf{u} \rangle}{\|\mathbf{u}\|^2} + b y_j - 1 \underset{\mathcal{E}^\varepsilon}{\overset{\mathcal{F}^\varepsilon}{\gtrless}} 0$$

We can perturb ε by $\Delta\varepsilon$ and narrow the range of eligible optimal b. Consider how the SVM boundary splits the dataset. On the positive side of the boundary, we have

$$b > y_i \Big(1 - \sum_{i=1}^{n} \mu_i y_i (I - P_{\mathbf{u}})\mathbf{x}_i y_j (I - P_{\mathbf{u}})\mathbf{x}_j - \frac{n\varepsilon y_j \langle \mathbf{x}_j, \mathbf{u} \rangle}{\|\mathbf{u}\|^2} \Big)$$

for i with $y_i = +1$ and $y_i \in \mathcal{F}^\varepsilon$, as well as $y_i = -1$ and $y_i \in \mathcal{E}^\varepsilon$. Call this set of indices R. On the other hand,

$$b < y_i \Big(1 - \sum_{i=1}^{n} \mu_i y_i (I - P_{\mathbf{u}})\mathbf{x}_i y_j (I - P_{\mathbf{u}})\mathbf{x}_j - \frac{n\varepsilon y_j \langle \mathbf{x}_j, \mathbf{u} \rangle}{\|\mathbf{u}\|^2} \Big)$$

for i with $y_i = -1$ and $y_i \in \mathcal{F}^\varepsilon$, as well as $y_i = +1$ and $y_i \in \mathcal{E}^\varepsilon$. Call this set of indices L. Let

$$s(\varepsilon) = 1 - \sum_{i=1}^{n} \mu_i y_i (I - P_{\mathbf{u}})\mathbf{x}_i y_j (I - P_{\mathbf{u}})\mathbf{x}_j - \frac{n\varepsilon y_j \langle \mathbf{x}_j, \mathbf{u} \rangle}{\|\mathbf{u}\|^2}$$

.Then we have the range

$$b \in \mathcal{B}_\varepsilon = [\max_{i \in R} y_i s(\varepsilon), \min_{i \in L} y_i s(\varepsilon)] \tag{A.14}$$

Perturbations of $\Delta\varepsilon$ result in changes of

$$t(\Delta\varepsilon) = -y_i \frac{n\Delta\varepsilon \langle \mathbf{x}_j, \mathbf{u} \rangle}{\|\mathbf{u}\|^2}$$

so we can write

$$\mathcal{B}_\varepsilon(\Delta\varepsilon) = [\max_{i \in R} y_i s(\varepsilon) - t(\Delta\varepsilon), \min_{i \in L} y_i s(\varepsilon) - t(\Delta\varepsilon)] \tag{A.15}$$

In increasing the magnitude of $\Delta\varepsilon$, the interval $\mathcal{B}_\varepsilon(\Delta\varepsilon)$ shrinks until it collapses onto a single value of b. The $\Delta\varepsilon$ be the perturbation when

$$\max_{i \in R} y_i s(\varepsilon) - t(\Delta\varepsilon) = \min_{i \in L} y_i s(\varepsilon) - t(\Delta\varepsilon) \tag{A.16}$$

determines the next breakpoint. The indices

$$k = \arg\max_{i \in R} y_i s(\varepsilon) - t(\Delta\varepsilon), \qquad \ell = \arg\min_{i \in L} y_i s(\varepsilon) - t(\Delta\varepsilon) \tag{A.17}$$

leave their respective sets and enter the margin. The partition is updated as:

$$\mathcal{M}^{\varepsilon+\Delta\varepsilon} = \{k, \ell\} \tag{A.18}$$

$$\{\mathcal{F}, \mathcal{E}\}^{\varepsilon+\Delta\varepsilon} = \{\mathcal{F}, \mathcal{E}\}^{\varepsilon} - \{k, \ell\} \tag{A.19}$$

A.3.3 Additional Figures

Figure 2 gives more information on the welfare impacts of ε-fair SVM-solutions on the Adult dataset. Increasing ε from left to right loosens fairness constraint, and classification outcomes become "less fair." Paths level off at $\varepsilon \approx 0.175$ when constraint ceases to bind at the optimal solution. The top panel shows that the learner objective value monotonically decreases as the fairness constraint loosens. The bottom panel gives the group-specific welfare change at an ε-fair SVM solution given as an absolute change in the number of positively labeled examples compared to the unconstrained solution baseline.

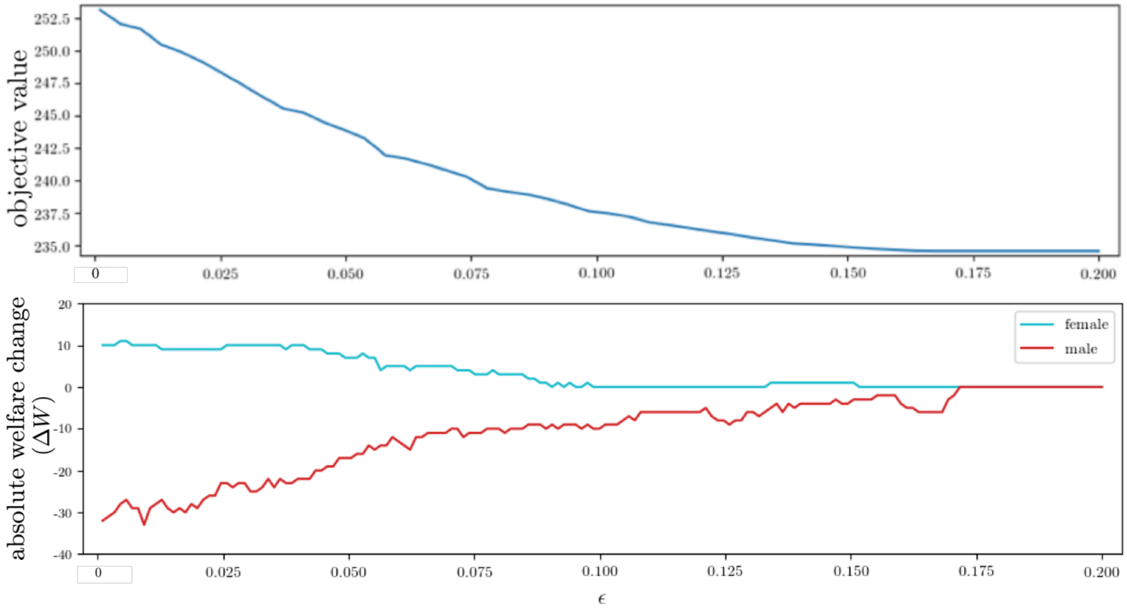


Figure A.1: Impact of fair SVM learning on learner objective value (top panel) and group welfare given as absolute welfare changes for female and male groups (bottom panel) on the Adult dataset.

A.3.4 Results on the Correspondence between Loss Minimization and Social Welfare Maximization

In the Planner's Problem, a planner maximizes a social welfare functional (SWF) given as a weighted sum of individual utilities, $W = \sum_{i=1}^{n} w_i u_i$. An individual i's contribution to society's total welfare is a product of her utility u_i and her social weight $w_i \in [0, 1]$ normalized so that $\sum_{i}^{n} w_i = 1$. Utility functions $u_i : \mathcal{X} \rightarrow \mathbb{R}_+$ assign positive utilities to a set of attributes or goods $\mathbf{x}_i$. We suppose a utility function is everywhere continuous and differentiable with respect to its inputs.

Since a planner who allocates a resource h impacts her recipients' utilities, she solves $h^{SWF}(\mathbf{x}; \boldsymbol{w}) := \arg\max_{\boldsymbol{h}} \sum_{i=1}^{n} w_i u(\mathbf{x}_i, h_i)$ under a budget constraint: $\sum_{i=1}^{n} h_i \leqslant B$. Since we consider cases of social planning in which a desirable good is being allocated, it is natural to suppose that u is strictly monotone with respect to h. As is common in welfare economics, we take u to be concave in h, so that receiving the good exhibits diminishing marginal returns. Further, we require that the social welfare functional W be symmetric: $W(\boldsymbol{h}; \mathbf{x}, \boldsymbol{w}) = W(\sigma(\boldsymbol{h}); \sigma(\mathbf{x}), \sigma(\boldsymbol{w}))$ for all possible permutations of $\sigma(\cdot)$. This property implies that the utility functions in the Planner's Problem are not individualized. In the case of binary classification, the planner decides whether to allocate the discrete good to individual i or not ($h_i \in \{0, 1\}$).

To highlight the correspondence between the machine learning and welfare economic approaches to social allocation, we first show that we can understand loss minimizing solutions to also be welfare maximizing ones, albeit under a particular instantiation of the social welfare function. Since social welfare is given as the weighted sum of individuals' utilities, it is clear that manipulating weights $\boldsymbol{w}$ significantly alters the planner's solution. Thus just as we can compute optimal allocations under a fixed set of welfare weights, we can also begin with an optimal allocation and find welfare weights that would support them. In welfare economics, the form of $\boldsymbol{w}$ corresponds to societal preferences about what constitutes a fair distribution. For example, the commonly-called "Rawlsian" social welfare function named after political philosopher John Rawls, can be written as $W_{Rawls} = \min_i u_i$ where u_i gives the utility of individual i. This function is equivalent to the general form $\sum_{i=1}^{n} w_i u_i$ where the individual i with the lowest utility u_i has welfare weight $w_i = 1$ and all individuals $k \neq i$ have weight $w_k = 0$. On the other hand, the commonly-called "Benthamite" social welfare function named after the founder of utilitarianism Jeremy Bentham, aggregates social welfare such that an extra unit of utility contributes equally to the social welfare regardless of who receives it. Benthamite weights are equal across all individuals: $w_i = \frac{1}{n}$ for all $i \in [n]$.

Thus associating an optimal (possibly fairness constrained) loss minimizing allocation with a set of welfare weights that would make it socially optimal lends insight into how socially "fair" a classification is from a welfare economic perspective. The following Proposition formally states this correspondence between loss minimization and social welfare maximization.

Proposition 11. *For any vector of classifications $h^{ML}(\mathbf{x}_i)$ that solves a loss minimization task, there* *exists a set of welfare weights $\boldsymbol{w}$ with $\sum_{i=1}^{n} w_i = 1$ such that the planner who maximizes social welfare W with a budget B selects an optimal allocation $h^{SWF}(\mathbf{x}_i) = h^{ML}(\mathbf{x}_i)$ for all $i \in [n]$.*

Proof. First, we know that since $W(\mathbf{x}, \boldsymbol{w})$ is a weighted sum of functions u, which are concave in h, the planner can indeed find a social welfare maximizing allocation $\boldsymbol{h}^{SWF}$. Let $h^{ML}(\mathbf{x})$ be the empirical loss-minimizing classifier for $\{\mathbf{x}_i, z_i, y_i\}_{i=1}^{n}$. With these allocations given, we can invert the social welfare maximization problem to find the weights that $\boldsymbol{w}$ support them.

For a given utility function u, we evaluate $\frac{\partial u(\mathbf{x},h)}{\partial h}\Big|_{\{\mathbf{x}_i, h^{ML}(\mathbf{x}_i)\}} = m_i \; \forall i \in [n]$, which gives the marginal gain in utility for individual i from having received an infinitesimal additional allocation of h. Notice that at a welfare maximizing allocation $\boldsymbol{h}$, we must have that

$$w_i \frac{\partial u(\mathbf{x}, h)}{\partial h}\Big|_{\{\mathbf{x}_i, h_i\}} = w_j \frac{\partial u(\mathbf{x}, h)}{\partial h}\Big|_{\{\mathbf{x}_j, h_j\}} \text{ for all } i, j \in [n] \tag{A.20}$$

When the allocation $h^{ML}(\mathbf{x})$ has been fixed, we must have that $w_i m_i = w_j m_j = k$, where the constant k is set by the planner's budget B, for all i, j along with $\sum_{i=1}^{n} w_i = 1$. Since u is strictly monotone with respect to h, $m_i > 0$ for all i. We thus have a non-degenerate system of n equations with n variables, and there exists a unique solution of welfare weights $\boldsymbol{w}$ that support the allocation.

□

Note that in the case of binary classification $h^{ML}(\mathbf{x}) \in \{-1, +1, \}$, so allocations are not awarded at a fractional level. Thus rather than the partial $\frac{\partial u(\mathbf{x},h)}{\partial h}$, the planner must consider the margin gain of receiving a positive classification. Nevertheless, Proposition 1 still holds, and the proof carries through with $\Delta u(\mathbf{x}, h(\mathbf{x})) = u(\mathbf{x}, 1) - u(\mathbf{x}, 0)$ in place of partial derivatives $\frac{\partial u(\mathbf{x},h)}{\partial h}$.

The equations given in (A.20) set an optimality condition for the planner. Its structure, though simple, reveals that welfare weights must be inversely proportional to an individuals' marginal utility gain from receiving an allocation. This result is formalized in the Proposition below.

Proposition 12. *For any set of optimal allocations* $\boldsymbol{h} = \arg\max_{\boldsymbol{h}} \sum_{i=1}^{n} \bar{w}_i u(\mathbf{x}_i, h_i)$ *with strictly monotonic utility function u concave in h, the supporting welfare weights have the form* $\bar{w}_i = \frac{k}{m_i}$ *where* $m_i = \frac{\partial u(\mathbf{x}_i)}{\partial h}|_{\{\mathbf{x}_i, h_i\}}$ *and* $k > 0$ *is a constant set by the planner's budget* $B = \sum_{i=1}^{n} h_i$*.*

By associating a set of classification outcomes with a set of implied welfare weights, one can inquire about the social fairness of the allocation scheme by investigating the distribution of welfare weights across individuals or across groups. While there may not be a single distribution of welfare weights that can be said to be "most fair," theoretical and empirical work in economics has been conducted on the range of fair distributions of societal weights. [37,82] This research has considered weights as implied by current social policies, [1,99,21] philosophical notions of justice, [2,38] and individuals' preferences in surveys and experiments. [1,65,82] They thus offer substantive notions of fairness currently uncaptured by many current algorithmic fairness approaches.

An Algorithm that Records All Possible Labelings

In the previous section, we showed that for any vector of classifications, one can compute the implied societal welfare weights of the generic SWF that would yield the same allocations in the Planner's Problem. In this section, we work in the converse direction: Beginning with a planner's social welfare maximization problem, does there exist a classifier $h^{ML} \in \mathcal{H}$ that generates the same classification as the planner's optimal allocation such that for all $i \in [n]$, $h^{ML}(\mathbf{x}_i) = h^{SWF}(\mathbf{x}_i)$?

We answer this question for the hypothesis class of linear decision boundary-based classifiers by providing an algorithm that accomplishes a much more general task: Given a set $\mathcal{X}$, containing n d-dimensional nondegenerate data points $\mathbf{x} \in \mathbb{R}^d$, our algorithm enumerates all linearly separable labelings and can output a hyperplane parameterized by $\boldsymbol{\theta} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ that achieves that set of labels. In order to build intuition for its construction, we first consider a hyperplane separation technique that applies to a very specific case: a case in which a hyperplane separates sets A and B, intersecting A at a single point and intersecting B at $d-1$ points.

Lemma 4. *Consider linearly separable sets A and B of points $\mathbf{x} \in \mathbb{R}^d$. For any $d-1$-dimensional hyperplane h_V with $h_V \cap A = \mathbf{v}$ and $h_V \cap B = P$ where $|P| = d-1$ that separates A and B into closed halfspaces $\bar{h}_V^+$ and $\bar{h}_V^-$, one can construct a $d-1$-dimensional hyperplane h that separates A and B into open halfspaces h^+ and h^-.*

Because its techniques are not of primary relevance for this Section, we defer the full proof of this Lemma to the Appendix but provide a brief exposition. The construction on which the Lemma relies is a "pivot-and-translate" maneuver. A hyperplane as described can separate points in open halfspaces by first pivoting (infinitesimally) on a $d-2$-dimensional facet P of a convex hull $C(B)$ away from $\mathbf{v} \in C(A)$ and then translating (infinitesimally) back toward $\mathbf{v}$ and away from $C(B)$. We show that all separable convex sets can be separated by such a hyperplane and procedure.

Note that since we seek enumerations of all labelings achievable by a linear separator on a given dataset, we are not *a priori* given convex hulls to separate. That is, we want to know which points *can* be made into distinct convex hulls and which cannot. Thus we take the preceding procedure and invert it—the central idea is to begin with the separators and from there, search for all possible convex hulls: Beginning with an arbitrary $d-1$-dimensional hyperplane h defined by d data points, we construct convex hulls out of the points in each halfspace created by h. Then we can use the pivot-and-translate procedure to construct a separation of the two sets into two open halfspaces. We must show that such a procedure is indeed exhaustive.

Theorem 8. *Given a dataset $\mathcal{X}$ consisting of n nondegenerate points $\mathbf{x} \in \mathbb{R}^d$, Algorithm 2 enumerates all possible labelings achievable by a $d-1$-dimensional hyperplane in $O(n^d d)$ time and outputs hyperplane parameters $(\boldsymbol{\theta}, b)$ that achieve each one.*

ALGORITHM 2: Record all possible labelings on a dataset $\mathcal{X}$ by linear separators

```
Input: Set $\mathcal{X}$ of $n$ data points $\mathbf{x} \in \mathbb{R}^d$
Output: All possible partitions $A, B$ attainable via linear separators; supporting
        hyperplane $h$
for all $V \subset \mathcal{X}$ with $|V| = d$ do
    Construct $d - 1$-dimensional hyperplane $h_V$ defined by $\mathbf{v} \in V$;
    for each point $\mathbf{v} \in V$ do
        $P = V \backslash \mathbf{v}$;
        $h = pivot(h_V, P, \mathbf{v})$ ; // $h_V$ pivots around the $d - 2$-dimensional plane
          $P$ away from $\mathbf{v}$
        $h = translate(h, \mathbf{v})$ ;                 // $h$ translates toward $\mathbf{v}$
        Record $A = \{\mathbf{x}|\mathbf{x} \in h^+\}, B = \{\mathbf{x}|\mathbf{x} \in h^-\}, h$;
    end
end
```

Proof. We have already shown that the pivot-and-translate construction is sufficient to linearly separate two sets A and B in the very specific case given in the preceding Lemma. But we must prove that all linearly separable sets can be constructed via Algorithm 2. We prove it is exhaustive by contradiction.

Suppose there exists a separation of $\mathcal{X}$ that is not captured by Algorithm 2. Then there exists disjoint sets A and B such that their convex hulls $C(A)$ and $C(B)$ do not intersect. By the hyperplane separating theorem, there exists a $d - 1$-dimensional hyperplane h_{V_1} that separates A and B, defined by a set V_1 of d vertices $\mathbf{v}$, at least one of which is on the boundary of each convex hull. Without loss of generality, we assume that for all $\mathbf{x} \in A, \mathbf{x} \in h_{V_1}^+$ and for all $\mathbf{x} \in B, \mathbf{x} \in h_{V_1}^-$. Notice that this hyperplane is indeed "checked" by the Algorithm, and this hyperplane h_{V_1} correctly separates $\mathbf{x} \in \mathcal{X} \backslash V_1$ into the two sets A and B. Thus if the separation is not disclosed via the procedure, the omission must occur due to the pivot-and-translate procedure's being incomplete.

In Algorithm 2, the set V_1 is partitioned so that $V_1 = \mathbf{v}_{f,1} \cup P_1$ where $\mathbf{v}_{f,1}$ is the "free vertex" and P_1 is the pivot set consisting of $d - 1$ vertices. This partition occurs d times so that each vertex $\mathbf{v} \in V_1$

has its turn as the "free vertex." Thus we can view the pivot-and-translate procedure as constituting a second partition—a partition of the d vertices that define the initial separating hyperplane. By contradiction, we claim that there exists a partition $D_1, E_1 \subset V_1$ such that $D_1 \coprod E_1 = V_1$ where $D_1 \subset A$ and $E_1 \subset B$ that is unaccounted for in the d pivot-and-translate operations applied to h_{V_1}. Thus $|D_1|, |E_1| \geqslant 2$. We use a "gift-wrapping" argument, a technique common in algorithms that construct convex hulls, to show that the partition A and B is indeed covered by Algorithm 2.

Select $\mathbf{v} \in D_1$ to be the free vertex $\mathbf{v}_{f,1}$, and let the pivot set $P_1 = V_1 \backslash \mathbf{v}_{f,1}$. We pivot around P_1 and away from $\mathbf{v}_{f,1}$ so that $\mathbf{v}_{f,1} \in h^+_{V_1}$. Rotations in d-dimensions are precisely defined as being around $d-2$-dimensional planes. Thus pivoting around the ridge P_1 away from $\mathbf{v}_{f,1}$ is a well-defined rotation in $\mathbb{R}^d$. Since h_{V_1} is a supporting hyperplane to $C(B)$, E_1 constitutes a $|E_1| - 1$-dimensional facet of $C(B)$. There exists a vertex $\mathbf{v}_E \in C(B)$ such that $E_1 \cup \mathbf{v}_E$ gives a $|E_1|$-dimensional facet of $C(B)$. Let h_{V_2} be defined by the set $V_2 = P_1 \cup \mathbf{v}_E$. h_{V_2} continues to correctly separate all $\mathbf{x} \in \mathcal{X} \backslash V_2$.

We once again partition V_2 into sets D_2 and E_2 whose members must be ultimately classified in sets A and B respectively. Notice that $|D_2| = |D_1| - 1$, since h_{V_2} correctly classifies $\mathbf{v}_{f,1}$ as belonging to set A. Thus with each iteration of the pivot procedure, the separating classifier unhinges from a vertex in $C(A)$ and "wraps" around $C(B)$ just as in the gift wrapping algorithm to attach onto another vertex in $C(B)$. At each step, the hyperplane defined by d vertices continues to support and separate $C(A)$ and $C(B)$. Thus process iterates until in the $|D_1| - 1$-th round, the hyperplane $h_{V_{|D_1|-1}}$ has partition $D_{|D_1|-1}$ and $E_{|D_1|-1}$ with $|D_{|D_1|-1}| = 1$. Applying the full pivot-and-translate procedure ensures the desired separation of sets A and B into open halfspaces.

Thus starting from a separable hyperplane defined by d vertices on the convex hulls $C(A)$ and $C(B)$, which must exist in virtue of the separability of sets A and B, we were able to use the pivot procedure in order to "gift-wrap" around one convex hull until we arrived at a d-dimensional separating hyperplane with only one vertex $\mathbf{v}_f \in C(A)$. This hyperplane is obviously checked by the first for-loop of Algorithm 2. The subsequent for-loop that performs the second partition of the d

vertices into the free vector $\mathbf{v}_f$ and the pivot set P then directly applies and performs the pivot-and-translate procedure given in Algorithm 2 to achieve the desired separation. $\square$

Degeneracies in the dataset can be handled by combining Algorithm 2 with standard solutions to degeneracy problems in geometric algorithms, which perform slight perturbations to degenerate data points to transform them into nondegenerate ones.[31] In concert with these solutions, Algorithm 2 automatically reveals which social welfare maximization solutions are attainable on a given dataset $\mathcal{X}$ via hyperplane-based classification and the $0-1$ accuracy loss each entails.

A.3.5 Proofs

Proof of Proposition 7

Proof. For all $j \in \mathcal{F}^{\varepsilon}$, remaining in $\mathcal{F}^{\varepsilon+\Delta\varepsilon}$ after the perturbation requires that $\frac{\partial D}{\partial \mu_j} > 0$ after the perturbation. Let μ_i^{ε} be the optimal μ_i solution at $p(\varepsilon)$. Then following (4.10), we rewrite the quantity $\frac{\partial D}{\partial \mu_j}$ as

$$g_j = 1 - \Big(\sum_{i=1}^{n} \mu_i^{\varepsilon} y_i (I - P_{\mathbf{u}})\mathbf{x}_i y_j (I - P_{\mathbf{u}})\mathbf{x}_j + \frac{n\varepsilon y_j \langle \mathbf{x}_j, \mathbf{u}\rangle}{\|\mathbf{u}\|^2} + b y_j\Big) < 0$$

If $d_j \Delta\varepsilon > 0$, then $j \in \mathcal{F}^{\varepsilon+\Delta\varepsilon}$. Otherwise, for $d_j \Delta\varepsilon < 0$, if $\Delta\varepsilon < \frac{g_j}{d_j}$, then $\frac{\partial D}{\partial \mu_j^{\varepsilon+\Delta\varepsilon}} > 0$, and $j \in \mathcal{F}^{\varepsilon+\Delta\varepsilon}$ after the perturbation. ✓

The same reasoning follows for $j \in \mathcal{E}^{\varepsilon}$, except we have that $g_j > 0$. Thus if $d_j \Delta\varepsilon < 0$, then $j \in \mathcal{E}^{\varepsilon+\Delta\varepsilon}$. Otherwise, for $d_j \Delta\varepsilon > 0$, if $\Delta\varepsilon < \frac{g_j}{d_j}$, then $\frac{\partial D}{\partial \mu_j^{\varepsilon+\Delta\varepsilon}} > 0$, and $j \in \mathcal{E}^{\varepsilon+\Delta\varepsilon}$ after the perturbation. ✓

To ensure that margin vectors do not escape the margin, we can directly look to $r_j = \frac{\partial \mu_j}{\partial \varepsilon}$. Since for all $j \in \mathcal{M}^{\varepsilon}$, $\mu_j^{\varepsilon} \in [0, C]$, then staying in the margin and set $\mathcal{M}^{\varepsilon+\Delta\varepsilon}$ depends on the sign of r_j and requires that

$$r_j < 0 \longrightarrow \frac{C - \mu_j^\varepsilon}{r_j} < \Delta\varepsilon < \frac{-\mu_j^\varepsilon}{r_j} \tag{A.21}$$

$$r_j > 0 \longrightarrow \frac{-\mu_j^\varepsilon}{r_j} < \Delta\varepsilon < \frac{C - \mu_j^\varepsilon}{r_j} \tag{A.22}$$

Thus taking the minimum of the positive quantities gives an upper bound, while taking the maximum of the negative quantities gives a lower bound on $\Delta\varepsilon$ perturbations, such that $\{\mathcal{F}, \mathcal{M}, \mathcal{E}\}^\varepsilon = \{\mathcal{F}, \mathcal{M}, \mathcal{E}\}^{\varepsilon+\Delta\varepsilon}$. Let

$$m_j = \begin{cases} \begin{cases} \frac{g_j}{d_j}, & j \in \mathcal{F}, d_j > 0 \\ -\infty, & j \in \mathcal{F}, d_j < 0 \end{cases} \\ \min\{\frac{C-\mu_j^\varepsilon}{r_j}, \frac{-\mu_j^\varepsilon}{r_j}\}, j \in \mathcal{M} \\ \begin{cases} -\infty, & j \in \mathcal{E}, d_j > 0 \\ \frac{g_j}{d_j}, & j \in \mathcal{E}, d_j < 0 \end{cases} \end{cases} \quad , \quad M_j = \begin{cases} \begin{cases} \infty, & j \in \mathcal{F}, d_j > 0 \\ \frac{g_j}{d_j}, & j \in \mathcal{F}, d_j < 0 \end{cases} \\ \min\{\frac{C-\mu_j^\varepsilon}{r_j}, \frac{-\mu_j^\varepsilon}{r_j}\}, j \in \mathcal{M} \\ \begin{cases} \frac{g_j}{d_j}, & j \in \mathcal{E}, d_j > 0 \\ \infty, & j \in \mathcal{E}, d_j < 0 \end{cases} \end{cases}$$

Thus all perturbations of ε within the range

$$\Delta\varepsilon \in \big(\max_j m_j, \min_j M_j\big)$$

satisfy the necessary conditions to ensure stable sets $\{\mathcal{F}, \mathcal{M}, \mathcal{E}\}$. Stable classifications $\hat{y}_i$ follow. □

Proof of Corollary 3

Proof. For all $\Delta\varepsilon$ in the stable region given in (4.16), $W_i(\varepsilon) = W_i(\varepsilon + \Delta\varepsilon)$ where i gives group membership $z = i$. Thus the groups are welfare-wise indifferent between classifications at ε and $\Delta\varepsilon$. For all $\Delta\varepsilon < 0$, where the fairness constraint is tightened,$p(\varepsilon) \leqslant p(\varepsilon + \Delta\varepsilon)$. Since the learner prefers lower loss, we have that $p(\varepsilon) \geq p(\varepsilon + \Delta\varepsilon)$. Comparing the triples at each ε value, we thus have

$$\{p(\varepsilon), W_0(\varepsilon), W_1(\varepsilon)\} \geq \{p(\varepsilon + \Delta\varepsilon), W_0(\varepsilon + \Delta\varepsilon), W_1(\varepsilon + \Delta\varepsilon)\}$$

as desired. □

Proof of Proposition 9

Proof. Following much of the exposition in the main text, recall we have that the perturbation function in (4.21) is given as

$$p(\varepsilon) \geqslant \sup_{\boldsymbol{\mu},\gamma}\{\mathcal{L}(\boldsymbol{\mu}^*, \gamma^*) - \varepsilon|\gamma^*|\}$$

which gives a global lower bound. Thus when a perturbation $\Delta\varepsilon < 0$ causes $\mathcal{L}(\boldsymbol{\mu}^*, \gamma^*) - \varepsilon|\gamma^*|$ to increase, then $p(\varepsilon + \Delta\varepsilon)$ is guaranteed to increase by at least $\Delta\varepsilon|\gamma^*|$. Thus when $|\gamma^*| \gg 0, p(\varepsilon + \Delta\varepsilon) - p(\varepsilon) \gg 0$. The learner experience a significant increase in her optimal value $p(\varepsilon)$ (which she wishes to minimize).

On the other hand, when $\Delta\varepsilon > 0$, then $\mathcal{L}(\boldsymbol{\mu}^*, \gamma^*) - \varepsilon|\gamma^*|$ decreases. But the decrease gives only the lower bound, and thus when $|\gamma^*|$ is small, her optimal value $p(\varepsilon)$ decreases but it is guaranteed not to decrease by much. □

Proof of Proposition 8

Proof. Fix $\varepsilon \in (0, 1)$ and consider the stable region of $\Delta\varepsilon$ perturbations given by (b_L, b_U). Suppose $b_L = \frac{g_j}{d_j}$ with $j \in \mathcal{E}$, then if $y_j = -1, \hat{y}_j = +1$. Thus at the breakpoint $\Delta\varepsilon = b_L, j$ moves into $\mathcal{M}^{\varepsilon + b_L}$ and $\hat{y}_j = +1$ and $u_{z_j}(\varepsilon + b_L) < u_{z_j}(\varepsilon)$ where z_j gives the group membership of $\mathbf{x}_j$. Since no other points transition, $u_{\bar{z}}(\varepsilon + b_L) = u_{\bar{z}}(\varepsilon)$ for all $\bar{z} \neq z_j$. Since $b_L < 0$, the fairness constraint is tightened and associated with a shadow price given by $\gamma > 0$ such that $p(\varepsilon + b_L) < p(\varepsilon)$. ✓

Suppose $b_L = \frac{C-\mu_j^\varepsilon}{r_j}$ and $j \in \mathcal{M}^\varepsilon$ with $y_j = +1$, then j moves into $j \in \mathcal{E}^{\varepsilon+b_L}$ such that $\hat{y}_j = -1$. Thus $u_{z_j}(\varepsilon + b_L) < u_{z_j}(\varepsilon)$ and $u_{\bar{z}}(\varepsilon + b_L) = u_{\bar{z}}(\varepsilon)$ where z_j is the group membership of $\mathbf{x}_j$ and $\bar{z} \neq z_j$, and $p(\varepsilon + b_L) \leqslant p(\varepsilon)$. ✓

Suppose $b_U = \frac{g_j}{d_j} > 0$ where $j \in \mathcal{E}^\varepsilon, y_j = +1$, and $\hat{y}_j = -1$. At the breakpoint, j moves into $\mathcal{M}^{\varepsilon+b_U}$ such that $y_j = -1$. Then $u_{z_j}(\varepsilon + b_U) > u_{z_j}(\varepsilon)$ where z_j is the group membership of $\mathbf{x}_j$. For $\bar{z} \neq z_j$, $u_{\bar{z}}(\varepsilon + b_U) = u_{\bar{z}}(\varepsilon)$, and since $b_U > 0$, the fairness constraint is loosened and $p(\varepsilon + b_U) > p(\varepsilon)$.

Suppose $b_U = \frac{C-\mu_j^\varepsilon}{r_j} > 0$ where $j \in \mathcal{M}^\varepsilon$ and $y_j = -1$. At the breakpoint, j moves into $\mathcal{E}^{\varepsilon+b_U}$ such that $\hat{y}_j = +1$. Then $u_{z_j}(\varepsilon + b_U) > u_{z_j}(\varepsilon)$ where z_j gives the group membership of $\mathbf{x}_j$. For $\bar{z} \neq z_j$, $u_{\bar{z}}(\varepsilon + b_U) = u_{\bar{z}}(\varepsilon)$, and since $b_U > 0$, the fairness constraint is loosened and $p(\varepsilon + b_U) \geqslant p(\varepsilon)$. ✓ □

Proof of Theorem 5

Proof. Theorem 5 follows from Lemma 3.2, Proposition 7, Corollary 3, and Proposition 8. □

Proof of Lemma 4 from Appendix Section 6.4

Proof. Let A and B be a pair of disjoint non-empty convex sets that partition $\mathcal{X} \subset \mathbb{R}^d$: $A \coprod B = \mathcal{X}$. Then by the hyperplane separation theorem, there exists a pair $(\boldsymbol{\theta}, b)$ such that for all $\mathbf{x} \in A$, $\boldsymbol{\theta}^\intercal \mathbf{x} \geqslant b$—call this closed halfspace $\bar{h}^+$—and for all $\mathbf{x} \in B$, $\boldsymbol{\theta}^\intercal \mathbf{x} \leqslant b$—call this closed halfspace $\bar{h}^-$. One such hyperplane can be constructed to separate the convex hulls of A and B

$$C(A) = \Big\{ \sum_{i=1}^{|A|} \alpha_i \mathbf{x}_i | \mathbf{x}_i \in A, \alpha_i \geqslant 0, \sum_{i=1}^{|A|} \alpha_i = 1 \Big\}$$

$$C(B) = \Big\{ \sum_{i=1}^{|B|} \alpha_i \mathbf{x}_i | \mathbf{x}_i \in B, \alpha_i \geqslant 0, \sum_{i=1}^{|B|} \alpha_i = 1 \Big\}$$

Let h_V be the $d-1$-dimensional hyperplane defined by the set V with $|V| = d$ such that $V \cap C(A) \neq \varnothing$ and $V \cap C(B) \neq \varnothing$. In order for the hyperplane to separate $C(A)$ and $C(B)$, h_V must also support each hull—we know that such a hyperplane always exists. In order to separate $C(A)$ and $C(B)$ so they are contained within open halfspaces h_V^+ and h_V^-, we wiggle the hyperplane so that it no longer passes through vertices $\mathbf{v} \in V$ but still maintains convex hull separation. This "wiggle" step is the final step of separating A and B.

Suppose V can be partitioned into a single vertex $\mathbf{v}_A$ in $C(A)$ and a set $P = \{\mathbf{v} | \mathbf{v} \in C(B)\}$ with $|P| = d-1$. The set P defines a ridge on $C(B)$, since it is a $d-2$-dimensional facet of $C(B)$. Rotations in d-dimensions are precisely defined as being around $d-2$-dimensional planes. Thus pivoting h_V around the ridge P away from $\mathbf{v}_A$ is a well-defined rotation in $\mathbb{R}^d$. Selecting any infinitesimally small rotation angle ρ will be enough to have $C(A) \in h_V^+$. After the pivot, we translate h_V away from the ridge P back toward $\mathbf{v}_A$. An infinitesimal translation is sufficient, since we simply wish to dislodge h_V from the ridge P, so that $C(B) \in h_V^-$. □

References

[1] Ackert, L. F., Martinez-Vazquez, J., & Rider, M. (2007). Social preferences and tax policy design: some experimental evidence. *Economic Inquiry*, 45(3), 487–501.

[2] Adler, M. (2012). *Well-being and fair distribution: beyond cost-benefit analysis*. Oxford University Press.

[3] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*.

[4] Akyol, E., Langbort, C., & Basar, T. (2016). Price of transparency in strategic machine learning. CoRR arXiv:1610.08210.

[5] Altonji, J. G. & Blank, R. M. (1999). Race and gender in the labor market. *Handbook of Labor Economics*, 3, 3143–3259.

[6] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica, May*, 23.

[7] Antonovics, K. (2006). Statistical discrimination and intergenerational income mobility. *Unpublished manuscript, University of California at San Diego*.

[8] Arrow, K. et al. (1973). The theory of discrimination. *Discrimination in Labor Markets*, 3(10), 3–33.

[9] Auer, P. & Cesa-Bianchi, N. (1998). On-line learning with malicious noise and the closure algorithm. *Annals of mathematics and artificial intelligence*, 23(1-2), 83–99.

[10] Ballwieser, W., Bamberg, G., Beckmann, M., Bester, H., Blickle, M., Ewert, R., Feichtinger, G., Firchau, V., Fricke, F., Funke, H., et al. (2012). *Agency theory, information, and incentives*. Springer Science & Business Media.

[11] Bechavod, Y. & Ligett, K. (2017). Learning fair classifiers: A regularization-inspired approach. *arXiv preprint arXiv:1707.00044*.

[12] Becker, G. S. (1971). *The economics of discrimination*. University of Chicago Press.

[13] Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency* (pp. 149–159).: PMLR.

[14] Bowles, S., Loury, G. C., & Sethi, R. (2014). Group inequality. *Journal of the European Economic Association*, 12(1), 129–152.

[15] Brückner, M. & Scheffer, T. (2011). Stackelberg games for adversarial prediction problems. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[16] Cain, G. G. (1986). The economic analysis of labor market discrimination: A survey. *Handbook of Labor Economics*, 1, 693–785.

[17] Calmon, F. P., Wei, D., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized data pre-processing for discrimination prevention. *arXiv preprint arXiv:1704.03354*.

[18] Card, D. & Rothstein, J. (2007). Racial segregation and the black–white test score gap. *Journal of Public Economics*, 91(11–12), 2158–2184.

[19] Chaudhuri, S. & Sethi, R. (2008). Statistical discrimination with peer effects: can integration eliminate negative stereotypes? *The Review of Economic Studies*, 75(2), 579–596.

[20] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2), 153–163.

[21] Christiansen, V. & Jansen, E. S. (1978). Implicit social preferences in the norwegian system of indirect taxation. *Journal of Public Economics*, 10(2), 217–245.

[22] Coate, S. & Loury, G. C. (1993). Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, (pp. 1220–1240).

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

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

[25] Datta, A., Fredrikson, M., Ko, G., Mardziel, P., & Sen, S. (2017). Proxy non-discrimination in data-driven systems. CoRR arXiv:1707.08120.

[26] Diehl, C. P. & Cauwenberghs, G. (2003). Svm incremental learning, adaptation and optimization. In *Neural Networks, 2003. Proceedings of the International Joint Conference on*, volume 4 (pp. 2685–2690).: IEEE.

[27] Dong, J., Roth, A., Schutzman, Z., Waggoner, B., & Wu, Z. S. (2018). Strategic classification from revealed preferences. In *Proceedings of the ACM Conference on Economics and Computation*.

[28] Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., & Pontil, M. (2018). Empirical risk minimization under fairness constraints. *arXiv preprint arXiv:1802.08626*.

[29] Dutta, S., Wei, D., Yueksel, H., Chen, P.-Y., Liu, S., & Varshney, K. (2020). Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International Conference on Machine Learning* (pp. 2803–2813).: PMLR.

[30] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214–226).: ACM.

[31] Edelsbrunner, H. & Mücke, E. P. (1990). Simulation of simplicity: a technique to cope with degenerate cases in geometric algorithms. *ACM Transactions on Graphics (tog)*, 9(1), 66–104.

[32] Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018). Runaway feedback loops in predictive policing. In *Proceedings of the Conference on Fairness, Accountability and Transparency*.

[33] Esteban, J. & Ray, D. (2006). Inequality, lobbying, and resource allocation. *American Economic Review*, 96(1), 257–279.

[34] Eubanks, V. (2018). *Automating inequality: How High-tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

[35] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 259–268).: ACM.

[36] Fish, B., Kun, J., & Lelkes, Á. D. (2016). A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM international conference on data mining* (pp. 144–152).: SIAM.

[37] Fleurbaey, M. & Maniquet, F. (2011). *A theory of fairness and social welfare*, volume 48. Cambridge University Press.

[38] Fleurbaey, M., Maniquet, F., et al. (2015). *Optimal taxation theory and principles of fairness*. Technical report, Université catholique de Louvain, Center for Operations Research and ?

[39] Frankel, A. & Kartik, N. (Forthcoming, 2018). Muddled information. *Journal of Political Economy*.

[40] Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im)possibility of fairness. *arXiv preprint arXiv:1609.07236*.

[41] Fryer, R. G., Pager, D., & Spenkuch, J. L. (2013). Racial disparities in job finding and offered wages. *The Journal of Law and Economics*, 56(3), 633–689.

[42] Gaebler, J., Cai, W., Basse, G., Shroff, R., Goel, S., & Hill, J. (2022). A causal framework for observational studies of discrimination. *Statistics and Public Policy*, (just-accepted), 1–61.

[43] Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2018). Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[44] Hardt, M., Megiddo, N., Papadimitriou, C., & Wootters, M. (2016a). Strategic classification. In *Proceedings of the ACM Conference on Innovations in Theoretical Computer Science*.

[45] Hardt, M., Price, E., Srebro, N., et al. (2016b). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* (pp. 3315–3323).

[46] Hastie, T., Rosset, S., Tibshirani, R., & Zhu, J. (2004). The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5(Oct), 1391–1415.

[47] Hedden, B. (2021). On statistical criteria of algorithmic fairness. *Philosophy and Public Affairs*, 49(2).

[48] Heidari, H., Ferrari, C., Gummadi, K. P., & Krause, A. (2018). Fairness behind a veil of ignorance: A welfare analysis for automated decision making. *arXiv preprint arXiv:1806.04959*.

[49] Hellman, D. (2020). Measuring algorithmic fairness. *Virginia Law Review*, 106(4), 811–866.

[50] Hu, L. & Kohler-Hausmann, I. (2022). Of misdefined causal questions: The case of race and multi-stage outcomes. *Unpublished manuscript*.

[51] Johnson, K. D., Foster, D. P., & Stine, R. A. (2016). Impartial predictive modeling: Ensuring fairness in arbitrary models. CoRR arXiv:1608.00528.

[52] Joseph, M., Kearns, M., Morgenstern, J. H., & Roth, A. (2016). Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems* (pp. 325–333).

[53] Kamishima, T., Akaho, S., & Sakuma, J. (2011). Fairness-aware learning through regularization approach. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on* (pp. 643–650).: IEEE.

[54] Karasuyama, M., Harada, N., Sugiyama, M., & Takeuchi, I. (2012). Multi-parametric solution-path algorithm for instance-weighted support vector machines. *Machine learning*, 88(3), 297–330.

[55] Kearns, M. & Li, M. (1993). Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4), 807–837.

[56] Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning* (pp. 2564–2572).: PMLR.

[57] Kearns, M., Roth, A., & Wu, Z. S. (2017). Meritocratic fairness for cross-population selection. In *International Conference on Machine Learning* (pp. 1828–1836).

[58] Kephart, A. & Conitzer, V. (2015). Complexity of mechanism design with signaling costs. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*.

[59] Kephart, A. & Conitzer, V. (2016). The revelation principle for mechanism design with reporting costs. In *Proceedings of the ACM Conference on Economics and Computation*.

[60] Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*.

[61] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference* (pp. 43:1–43:23).: ACM.

[62] Kleinberg, J. & Raghavan, M. (2018). How do classifiers induce agents to invest effort strategically? CoRR arXiv:1807.05307.

[63] Knox, D., Lowe, W., & Mummolo, J. (2020). Administrative records mask racially biased policing. *American Political Science Review*, 114(3), 619–637.

[64] Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems* (pp. 4066–4076).

[65] Kuziemko, I., Norton, M. I., Saez, E., & Stantcheva, S. (2015). How elastic are preferences for redistribution? evidence from randomized survey experiments. *American Economic Review*, 105(4), 1478–1508.

[66] Laffont, J.-J. & Martimort, D. (2009). *The Theory of Incentives: The Principal-Agent Model*. Princeton University Press.

[67] Levin, J. et al. (2009). The dynamics of collective reputation. *The BE Journal of Theoretical Economics*, 9(1), 1–25.

[68] Liu, L., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2018). Delayed impact of fair machine learning. In *International Conference on Machine Learning* (pp. 3156–3164).

[69] Long, R. (2021). Fairness in machine learning: against false positive rate equality as a measure of fairness. *Journal of Moral Philosophy*, 1(aop), 1–30.

[70] Loury, G. C. & Kim, Y.-C. (2014). Collective reputation and the dynamics of statistical discrimination.

[71] Loury, G. C. & Loury, G. C. (2009). *The anatomy of racial inequality*. Harvard University Press.

[72] Mayson, S. G. (2018). Bias in, bias out. *Yale lJ*, 128, 2218.

[73] Milli, S., Miller, J., Dragan, A. D., & Hardt, M. (Forthcoming, 2019). The social cost of strategic classification.

[74] Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8, 141–163.

[75] Mullainathan, S. (2018). Algorithmic fairness and the social welfare function. In *Proceedings of the 2018 ACM Conference on Economics and Computation* (pp. 1–1).: ACM.

[76] Neal, D. A. & Johnson, W. R. (1996). The role of premarket factors in black-white wage differences. *Journal of Political Economy*, 104(5), 869–895.

[77] O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books.

[78] Phelps, E. S. (1972). The statistical theory of racism and sexism. *The American Economic Review*, 62(4), 659–661.

[79] Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. In *Advances in Neural Information Processing Systems* (pp. 5680–5689).

[80] Qureshi, B., Kamiran, F., Karim, A., & Ruggieri, S. (2016). Causal discrimination discovery through propensity score analysis. CoRR arXiv:1608.03735.

[81] Rubinstein, A. & Yaari, M. E. (1983). Repeated insurance contracts and moral hazard. *Journal of Economic Theory*, 30(1), 74–97.

[82] Saez, E. & Stantcheva, S. (2016). Generalized social marginal welfare weights for optimal tax theory. *American Economic Review*, 106(1), 24–45.

[83] Sen, A. (1980). *Equality of What?* Cambridge University Press: Cambridge. Reprinted in John Rawls et al., Liberty, Equality and Law (Cambridge: Cambridge University Press, 1987).

[84] Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, 87(3), 355–374.

[85] Spence, M. (1978). Job market signaling. In *Uncertainty in Economics* (pp. 281–306).

[86] Spremann, K. (1987). Agent and principal. In *Agency theory, information, and incentives* (pp. 3–37). Springer.

[87] Sweeney, L. (2013). Discrimination in online ad delivery. *Queue*, 11(3), 10.

[88] Tirole, J. (1996). A theory of collective reputations (with applications to the persistence of corruption and to firm quality). *The Review of Economic Studies*, 63(1), 1–22.

[89] Wang, G., Yeung, D.-Y., & Lochovsky, F. H. (2007). A kernel path algorithm for support vector machines. In *Proceedings of the 24th international conference on Machine learning* (pp. 951–958).: ACM.

[90] Wang, L., Zhu, J., & Zou, H. (2006). The doubly regularized support vector machine. *Statistica Sinica*, 16(2), 589.

[91] Wick, M., Tristan, J.-B., et al. (2019). Unlocking fairness: a trade-off revisited. *Advances in neural information processing systems*, 32.

[92] Winfree, J. A. & McCluskey, J. J. (2005). Collective reputation and quality. *American Journal of Agricultural Economics*, 87(1), 206–213.

[93] Woodworth, B., Gunasekar, S., Ohannessian, M. I., & Srebro, N. (2017). Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*.

[94] Yang, C. S. & Dobbie, W. (2020). Equal protection under algorithms: A new statistical and legal framework. *Michigan Law Review*, 119(2), 291–395.

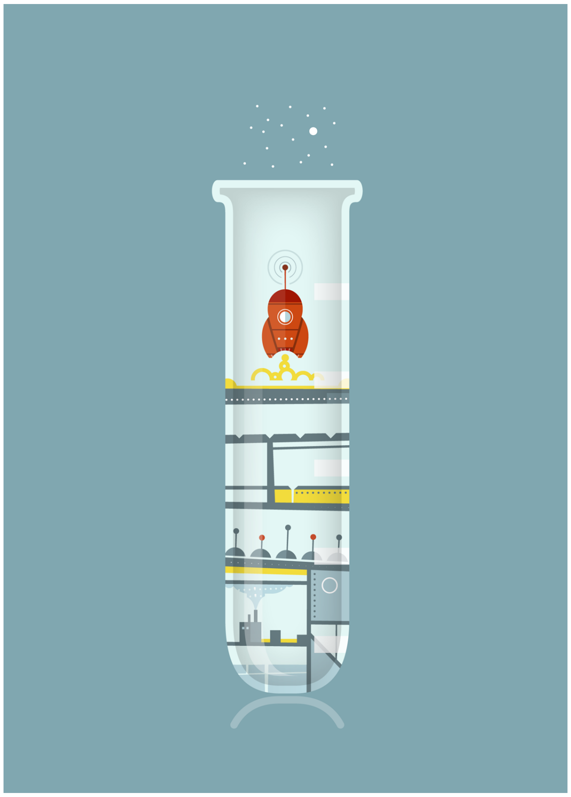
[95] Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 1171–1180).: International World Wide Web Conferences Steering Committee.

[96] Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2015). Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*.

[97] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)* (pp. 325–333).

[98] Zhao, Q., Keele, L. J., Small, D. S., & Joffe, M. M. (2022). A note on posttreatment selection in studying racial discrimination in policing. *American Political Science Review*, 116(1), 337–350.

[99] Zoutman, F. T., Jacobs, B., & Jongen, E. L. (2013). Optimal redistributive taxes and redistributive preferences in the netherlands. *Erasmus University Rotterdam*.



This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration, "Science Experiment 02", was created by Ben Schlitter and released under CC BY-NC-ND 3.0. A template that can be used to format a PhD thesis with this look and feel has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/Dissertate or from its author, Jordan Suchow, at suchow@post.harvard.edu.