Bridging Gaps Between Metrics and Social Outcomes in Multi-Stakeholder Machine Learning

Serena Lutong Wang

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2024-67
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2024/EECS-2024-67.html

May 9, 2024

Bridging Gaps Between Metrics and Social Outcomes in Multi-Stakeholder Machine Learning

by

Serena Wang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael Jordan, Chair
Assistant Professor Nika Haghtalab
Assistant Professor Jacob Steinhardt
Dr. Preston McAfee, Google Research

Spring 2024

Bridging Gaps Between Metrics and Social Outcomes in Multi-Stakeholder Machine Learning

Abstract

Bridging Gaps Between Metrics and Social Outcomes in Multi-Stakeholder Machine Learning

by

Serena Wang

Doctor of Philosophy in Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Michael Jordan, Chair

With the rise of machine learning (ML), society has become increasingly driven by metrics and algorithms. Unfortunately, even well-intended metrics often do not align with desired social outcomes. For example, in healthcare, mandated reporting of hospital mortality rate metrics actually led to worsened health outcomes for severely ill patients. Such misalignments present a fundamental challenge to understanding and improving the societal impacts of ML, where a growing literature relies on formulating broad notions such as fairness as *metrics* for either evaluation or optimization.

A core challenge driving the misalignments between metrics and social outcomes is the fact that ML systems are also *multi-stakeholder* systems. Some of the highest stakes deployments of ML also have many diverse stakeholders with asymmetric information, power, and values. For example, in healthcare, stakeholders include doctors, patients, hospitals, insurers, and many more. In these multi-stakeholder settings, misalignments between metrics and social outcomes challenge both policymakers seeking to audit ML systems, and engineers and researchers formulating ML problems.

To bridge the gaps between the technical formulations of ML and its societal impacts, this thesis addresses two complementary challenges. The first part concerns the implementation of socially relevant desiderata of **fairness, robustness, and interpretability** in ML, which become metrics in the form of objectives or constraints. Specifically, we will consider how to algorithmically build these notions into modern ML systems under noisy data and evolving large-scale training protocols. The second part will zoom out from these particular notions to consider the wider role of **metrics in multi-stakeholder systems**. This part brings in ideas from economics to achieve a better understanding of the interdependence between metrics and the surrounding ecosystem of stakeholders with asymmetric information, power, and values. In reimagining the ML development process with stakeholder involvement, we ask, *"Who has information on how to improve metrics, and when would they share it?"*

To my village of family, friends, and mentors.

Contents

Acknowledgments

This thesis reflects a winding journey, built on the shoulders of an entire village of giants who believed in me. The path was not a straight one, and if I'm honest, often felt like bushwhacking through many roads less traveled.

In this process, I am immensely grateful to my PhD advisor, Mike Jordan, for tirelessly believing in me and always seeing the best version of what I could be. Mike is my ultimate role model for blazing new trails in research: he is fearless, but not rash; optimistic, but not naive; visionary, yet always open-minded. Data during my PhD has shown that Mike has never been wrong – if it appears that he is wrong, time just hasn't proven him right yet.

I would also like to thank my thesis committee members, Jacob Steinhardt, Nika Haghtalab, and Preston McAfee. Jacob inspired me to work on topics in robustness early on. I thank Nika for her clear, level-headed advice in navigating an inherently uncertain job market. Finally, I am grateful to Preston, not least for teaching me how to think like an economist. I am continually humbled by Preston's unyielding faith in me, from our first chat about metrics and Goodhart's Law, to our weekly research meetings in which Preston patiently taught me how to look for *insights* in mathematical models.

Of course, it has taken a village to support me through the twists and turns, and I will take this opportunity to thank the many mentors whom I've adopted and who have adopted me over the years.

I've been incredibly lucky to have found so much support and mentorship at Google Research both before and throughout my PhD. I wouldn't be a machine learning researcher without Maya Gupta, whose mentorship brought me into the field at Google. Maya has an incredible combination of focus, pragmatism, kindness, and empathy, making her an effective leader and beloved manager. From her I learned how to be efficient and direct, and actually get things done. I'm also grateful for the doubly awesome advising sessions with Jim Muller. I also thank Ravi Kumar for his open-minded advising and support at Google. Through a seemingly bottomless source of astute questions, Ravi has taught me how to be a better theorist. I'm also grateful to Katrina Ligett not only for being a great research mentor, but for patiently coaching me to be a better speaker, helping me with a number of practice talks of which I've now lost count. As the most recent leg of my journey has brought me into the world of economics, causal inference, and political science, I thank P. M. Aronow for supporting me in a research vision that transcends disciplines. I could not ask for a more kind and generous mentor who is also an absolute force of nature.

I owe my start as a computer science researcher to Margo Seltzer and James Mickens, whose mentorship started in my undergraduate days at Harvard, and has continued for the many years since.

In addition to the above, I am incredibly fortunate to have had the opportunity to work with and learn from numerous brilliant collaborators: Rediet Abebe, Stephen Bates, Tolani Britton, Jill Burstein, Kevin Canini, Andy Cotter, Mihaela Curmei, Peng Ding, Badih Ghazi, Wenshuo Guo, Sara Hooker, Austin Jang, Ghassen Jerfel, Heinrich Jiang, Anmol Kabra, Mina Karzand, Lydia Liu, Tosca Lechner, Erez Louidor, Michal Lukasik, Yian Ma, Pasin

Manurangsi, Aditya K. Menon, Hari Narasimhan, Taman Narayan, Molly Offer-Westort, Ben Recht, James Sharpnack, Nati Srebro, Yixin Wang, Clara Wong-Fannjiang, Seungil You, Sen Zhao, Yichen Zhou, and Tao Zhu. I also thank Luca Zappella for his kind support as my Apple Research Scholars mentor. I am the researcher I am today because of what I've learned from all of you, and I truly can't thank you enough.

For making my PhD actually logistically happen, I am grateful to Naomi Yamasaki, Shirley Salanio, Jean Nguyen, and the rest of the administrative staff at Berkeley.

At the end of the day, I would not have made it through this without the support of my friends at Berkeley and beyond. Thank you to the cohort who started this journey with me: Akosua Busia, Wenshuo Guo, and Clara Wong-Fannjiang. From Boucheron to movin' on, I'm so grateful to have shared this time with you. Also with me from the beginning was Lydia Liu, whose deep friendship I will always cherish. I am grateful to the folks from SAIL and the wider Berkeley ML/theory groups for always being a source of inspiration, support, commiseration, celebration, and just overall making grad school be about so much more than just research. Beyond Berkeley, I'm grateful to all of my dear friends who have shaped who I am during my PhD years and before. While I won't be able to list everyone here, I'll particularly mention the Mage crew, who listened to way too many practice talks and also know way too much possible blackmail material: Brad, Jamie, Alex, Lily, Theo, and Corey. Also among those in the Bay Area who helped me through some of the most stressful times were Josh, Dan, Lily, Dennis, and Eric. Thank you to the BFF DinDins for being an unending source of inspiration: Rebecca, Kristina, Sierra, Saad, Emily, and Jessica. Shout out to the College Station-in-laws, Lisa, Leo, Amy, Amy, Kevin, Kevin, Lily, Zach, Austen, and Macy.

I owe everything I was, am, and will be to my family. Thank you Mom, Dad, Woody, and Rosetta. I will never be able to say enough to thank you. Thank you Pam, Brian, and Matt for being the most supportive in-laws I could ask for. Thank you Olive for listening to all my talks, and Momo for the warmest moral support. Finally, thank you Nick, for listening to all of the talks, meeting all of the people, supporting me through the hardest of times, and bouncing back all of the ideas. Thank you for being my loudest cheerleader, greatest inspiration, and dearest friend.

Chapter 1

Introduction

With the rise of machine learning (ML), society has become increasingly driven by metrics and algorithms. Unfortunately, for many common data-driven metrics used for optimization or evaluation, improvement in the metric does not always lead to better social outcomes. Furthermore, it is often the most vulnerable who ultimately bear the cost of this misalignment. Facebook's optimization of engagement metrics exacerbated mental health issues in teenage girls [Wells et al., 2021]. The focus of algorithms in healthcare on predicting costs led to racially biased predictions of health risk [Obermeyer et al., 2019]. An ML admissions algorithm dropped personal statements due to failure to improve their accuracy metric, and the algorithm was discontinued after critique over potentially exacerbating gender and racial disparities [Burke, 2020].

This mismatch between metrics and social outcomes is not unique to this era of ML – it has been historically documented across disciplines, notably identified in economic and monetary policy [Goodhart, 1984, Lucas Jr, 1976], social program evaluation [Campbell, 1979], and auditing in education [Strathern, 1997]. An adage commonly referred to as "Goodhart's Law" describes some of the challenges that arise [Strathern, 1997]:

> *"When a measure becomes a target, it ceases to be a good measure."*

Phenomena like Goodhart's Law continue to prevail in modern ML systems, and a gap remains between this long history of social scientific literature and modern ML research. Bridging this interdisciplinary gap offers rich avenues for improving our understanding of the societal impacts of ML in the modern day. Still, the computational power and scale of modern ML has brought a new gravity to the social consequences of these mismatches, along with new technical frontiers.

In fact, Goodhart's Law only scratches the surface of the types of complications that can arise in modern ML systems. More generally, machine learning systems are also *multi-stakeholder* systems: some of the highest stakes deployments of ML often involve many diverse stakeholders with asymmetric in information, power, and values. For example, in education, stakeholders include students, teachers, and schools. In content recommender systems, stakeholders include content creators, users, and advertisers. In healthcare, stakeholders include patients, doctors, hospitals, insurers, and many more. Thus, core to understanding the relationship between metrics and social outcomes is contextualizing the usage and evaluation of ML in its surrounding ecosystem of stakeholders.

To bridge the gaps between the technical formulations of ML and its societal impacts, this thesis is divided into two parts that address two complementary challenges. The first part concerns the implementation of socially relevant desiderata of **fairness, robustness, and interpretability** in ML, which are commonly metricized as objectives or constraints. Specifically, we will consider how to algorithmically build these notions into modern ML systems under noisy data and evolving large-scale training protocols. The second part will zoom out from these particular notions to consider the wider role of **metrics in multi-stakeholder systems**. To do this, we bring in economic modeling to understand the interdependence between metrics and the surrounding ecosystem of stakeholders with asymmetric information, power, and values.

1.1 Overview of Structure

We outline the two major parts of this thesis in more detail below. The content in these parts is based on previously published work co-authored with P. M. Aronow, Stephen Bates, Andrew Cotter, Wenshuo Guo, Maya R. Gupta, Sara Hooker, Michael I. Jordan, Katrina Ligett, Michal Lukasik, Preston McAfee, Aditya Krishna Menon, Harikrishna Narasimhan, and Yichen Zhou [Wang et al., 2020a, 2023b, Wang and Gupta, 2020, Wang et al., 2023a, 2024].

Part I: Fairness, Robustness, and Interpretability in ML

Growing public concern for the societal impacts of ML [House, 2023] has been matched with a growing literature on developing ML systems that satisfy socially relevant desiderata such as fairness, interpretability, robustness, safety, accountability, and many others [see, e.g., Barocas et al., 2019, Doshi-Velez and Kim, 2017, Arrieta et al., 2020]. These notions have been turned into metrics in various ways as objectives or constraints in ML problems. However, implementing these notions in practice in real ML systems continues to be a challenge as practitioners face issues with data and algorithms that deviate from theoretical assumptions in the literature. For example, data might be noisy or biased in unanticipated ways. ML training paradigms themselves are also evolving in the age of larger and larger models, where training is increasingly done in multiple stages via protocols like knowledge distillation, self-training, and fine-tuning [Hinton et al., 2015, Xie et al., 2020, Pan and Yang, 2009, De Lange et al., 2021]. Thus, it is no longer clear how a fairness notion encoded as a single objective should be applied across these multi-stage training pipelines with multiple optimization sub-problems.

Part I addresses these practical challenges to implementing fairness, robustness, and interpretability in ML in three chapters. Chapter 2 considers how noisy data can bias fairness metrics, and specifically gives algorithms to maintain validity of group-based fairness constraints when protected group data is noisy [Wang et al., 2020a]. Chapter 3 focuses on evolving training paradigms, and presents methodology to improve robustness of ML trained via a two-stage knowledge distillation protocol [Wang et al., 2023b]. Chapter 4 discusses ethical concepts overlooked by statistical group-based fairness metrics, and presents shape-constraints as a bridge between fairness and interpretability [Wang and Gupta, 2020].

Part II: Metrics in Multi-Stakeholder Systems

Moving beyond specific fairness metrics and ML algorithms, Part II more generally considers the complications that arise when metrics impact multi-stakeholder systems. Specifically, this part focuses on the gaps between metrics and social outcomes in environments of diverse stakeholders with asymmetric information, power, and values.

Stakeholder interactions and asymmetries in information, power, and values can lead to backfiring of even most well-intended metric designs. For example, after the New York Health Department mandated that hospitals report mortality rate metrics in 1990, severely ill patients ended up experiencing dramatically worsened health outcomes, in part due to providers selectively treating healthier patients [Dranove et al., 2003]. Since then, the Centers for Medicare and Medicaid Services (CMS) has continued to invest billions of dollars in the development of quality metrics [Wadhera et al., 2020, Casalino et al., 2016], which have also become increasingly deeply embedded in patient decision-making at scale via *online ranking platforms* like the US News and World Report, the LeapFrog Hospital Safety Score, and Cal Hospital Compare [Rosenberg, 2013, Health and Agency, 2022]. Concurrently, studies have continued to question the relationship between these metrics and patient outcomes [see, e.g., Glance et al., 2021, Ryan et al., 2009, Gonzalez and Ghaferi, 2014, Hwang et al., 2014, Jha et al., 2008, Smith et al., 2017]. Designing quality metrics that are better aligned with social welfare in the face of strategic selection behavior is the subject of Chapter 5 [Wang et al., 2023a].

A prevailing challenge with quality metrics in healthcare, and metric design more generally, is that stakeholders often have more information about the shortcomings of metrics than platforms and institutions with the power to set the metrics. For instance, doctors may observe more patient characteristics than health departments have on record [Dranove et al., 2003]. Generalizing beyond the specific structure of treatment policies, Chapter 6 more broadly models incentives for information sharing between agents in a system. Specifically, we devise a model motivated by the question,

Who has information on how to improve metrics, and when would they share it?

Our approach brings in ideas from information design and mechanism design [Bergemann and Morris, 2019] to re-imagine the ML development process. Future work of this kind is a gateway to answering many open questions surrounding asymmetries in information, power, and values in multi-stakeholder ML systems.

Part I

Fairness, Robustness, and Interpretability in Machine Learning

Chapter 2

Robust Optimization for Fairness with Noisy Protected Groups

2.1 Introduction

As machine learning becomes increasingly pervasive in real-world decision making, the question of ensuring *fairness* of ML models becomes increasingly important. The definition of what it means to be "fair" is highly context dependent. Much work has been done on developing mathematical fairness criteria according to various societal and ethical notions of fairness, as well as methods for building machine-learning models that satisfy those fairness criteria [see, e.g., Dwork et al., 2012, Hardt et al., 2016b, Russell et al., 2017, Kusner et al., 2017, Zafar et al., 2017, Cotter et al., 2019d, Friedler et al., 2019, Wang and Gupta, 2020].

Many of these mathematical fairness criteria are *group-based*, where a target metric is equalized or enforced over subpopulations in the data, also known as *protected groups*. For example, the *equality of opportunity* criterion introduced by Hardt et al. [2016b] specifies that the true positive rates for a binary classifier are equalized across protected groups. The *demographic parity* [Dwork et al., 2012] criterion requires that a classifier's positive prediction rates are equal for all protected groups.

One important practical question is whether or not these fairness notions can be reliably measured or enforced if the protected group information is noisy, missing, or unreliable. For example, survey participants may be incentivized to obfuscate their responses for fear of disclosure or discrimination, or may be subject to other forms of response bias. Social desirability response bias may affect participants' answers regarding religion, political affiliation, or sexual orientation [Krumpal, 2011]. The collected data may also be outdated: census data collected ten years ago may not an accurate representation for measuring fairness today.

Another source of noise arises from estimating the labels of the protected groups. For various image recognition tasks (e.g., face detection), one may want to measure fairness across protected groups such as gender or race. However, many large image corpora do not include protected group labels, and one might instead use a separately trained classifier to estimate group labels, which is likely to be noisy Buolamwini and Gebru [2018]. Similarly, zip codes can act as a noisy indicator for socioeconomic groups.

In this paper, we focus on the problem of training binary classifiers with fairness constraints when only noisy labels, $\hat{G} \in \{1, ..., \hat{m}\}$, are available for m true protected groups, $G \in \{1, ..., m\}$, of interest. We study two aspects: First, if one satisfies fairness constraints for noisy protected groups $\hat{G}$, what can one say with respect to those fairness constraints for the true groups G? Second, how can side information about the noise model between $\hat{G}$ and G be leveraged to better enforce fairness with respect to the true groups G?

Contributions

Our contributions can be summarized as follows:

1. We provide a bound on the fairness violations with respect to the true groups G when the fairness criteria are satisfied for the noisy groups $\hat{G}$.
2. We introduce two new robust-optimization methodologies that satisfy fairness criteria on the true protected groups G while minimizing a training objective. These methodologies differ in convergence properties, conservatism, and noise model specification.
3. We show empirically that unlike the naïve approach, our two proposed approaches are able to satisfy fairness criteria with respect to the true groups G on average.

The first approach we propose (Section 2.5) is based on distributionally robust optimization (DRO) [Duchi and Namkoong, 2018, Ben-Tal et al., 2013]. Let p denote the full distribution of the data, $X, Y \sim p$. Let p_j be the distribution of the data conditioned on the true groups being j, so $X, Y|G = j \sim p_j$; and $\hat{p}_j$ be the distribution of X, Y conditioned on the noisy groups, so $X, Y|\hat{G} = j \sim \hat{p}_j$. Given an upper bound on the total variation (TV) distance $\gamma_j \geq TV(p_j, \hat{p}_j)$ for each $j \in \{1, ..., m\}$, we define $\tilde{p}_j$ such that the conditional distributions $(X, Y|\tilde{G} = j \sim \tilde{p}_j)$ fall within the bound γ_j with respect to $\hat{p}_j$: $\gamma_j \geq TV(\tilde{p}_j, \hat{p}_j)$. Thus, the set of all such $\tilde{p}_j$ is guaranteed to include the unknown true group distribution p_j, for all j. Because it is based on the well-studied DRO setting, this approach has the advantage of being easy to analyze. However, the results may be overly conservative unless tight bounds $\{\gamma_j\}_{j=1}^m$ can be given.

Our second robust optimization strategy (Section 2.6) uses a robust re-weighting of the data from soft protected group assignments, inspired by criteria proposed by Kallus et al. [2020] for auditing the fairness of ML models given imperfect group information. Extending their work, we *optimize* a constrained problem to achieve their robust fairness criteria, and provide a theoretically ideal algorithm that is guaranteed to converge to an optimal feasible point, as well as an alternative practical version that is more computationally tractable. Compared to DRO, this second approach uses a more precise noise model, $P(\hat{G} = k|G = j)$, between $\hat{G}$ and G for all pairs of group labels j, k, that can be estimated from a small auxiliary dataset containing ground-truth labels for both G and $\hat{G}$. An advantage of this more detailed noise model is that a practitioner can incorporate knowledge of any bias in the relationship between G and $\hat{G}$ (for instance, survey respondents favoring one socially preferable response over others), which causes it to be less likely than DRO to result in an overly-conservative model. Notably, this approach does *not* require that $\hat{G}$ be a direct approximation of G—in fact, G and $\hat{G}$ can represent distinct (but related) groupings, or even groupings of different sizes, with the noise model tying them together. For example, if G represents "language spoken at home," then $\hat{G}$ could be a noisy estimate of "country of residence."

2.2 Related Work

Constrained optimization for group-based fairness metrics. The simplest techniques for enforcing group-based constraints apply a post-hoc correction of an existing classifier Hardt et al. [2016b], Woodworth et al. [2017]. For example, one can enforce *equality of opportunity* by choosing different decision thresholds for an existing binary classifier for each protected group [Hardt et al., 2016b]. However, the classifiers resulting from these post-processing techniques may not necessarily be optimal in terms of accuracy. Thus, constrained optimization techniques have emerged to train machine-learning models that can more optimally satisfy the fairness constraints while minimizing a training objective [Goh et al., 2016, Cotter et al., 2019b,d, Zafar et al., 2017, Agarwal et al., 2018, Donini et al., 2018, Narasimhan et al., 2019a].

Fairness with noisy protected groups. Group-based fairness notions rely on the knowledge of *protected group* labels. However, practitioners may only have access to noisy or unreliable protected group information. One may naïvely try to enforce fairness constraints with respect to these noisy protected groups using the above constrained optimization techniques, but there is no guarantee that the resulting classifier will satisfy the fairness criteria with respect to the true protected groups [Gupta et al., 2018].

Under the conservative assumption that a practitioner has no information about the protected groups, Hashimoto et al. [2018] applied DRO to enforce what Lahoti et al. [2020] refer to as *Rawlsian Max-Min fairness*. In contrast, here we assume some knowledge of a noise model for the noisy protected groups, and are thus able to provide tighter results with DRO: we provide a practically meaningful maximum total variation distance bound to enforce in the DRO procedure. We further extend Hashimoto et al. [2018]'s work by applying DRO to problems equalizing fairness metrics over groups, which may be desired in some practical applications [Kolodny, 2019].

Concurrently, Lahoti et al. [2020] proposed an adversarial reweighting approach to improve group fairness by assuming that non-protected features and task labels are correlated with unobserved groups. Like Hashimoto et al. [2018], Lahoti et al. [2020] also enforce *Rawlsian Max-Min fairness* with unknown protected groups, whereas our setup includes constraints for parity based fairness notions.

Kallus et al. [2020] considered the problem of *auditing* fairness criteria given noisy groups. They propose a "robust" fairness criteria using soft group assignments and show that if a given model satisfies those fairness criteria with respect to the noisy groups, then the model will satisfy the fairness criteria with respect to the true groups. Here, we build on that work by providing an algorithm for training a model that satisfies their robust fairness criteria while minimizing a training objective.

Lamy et al. [2019] showed that when there are only two protected groups, one need only tighten the "unfairness tolerance" when enforcing fairness with respect to the noisy groups. Mozannar et al. [2020] showed that if the predictor is independent of the protected attribute, then fairness with respect to the noisy groups is the same as fairness with respect to the true groups. When there are more than two groups, and when the noisy groups are included as an input to the classifier, other robust optimization approaches may be necessary. When using post-processing instead of constrained optimization, Awasthi et al. [2020] showed that under certain conditional independence assumptions, post-processing using the noisy groups will not be worse in terms of fairness violations than not post-processing at all. In our work, we consider the problem of training the model subject to fairness constraints, rather than taking a trained model as given and only allowing post-processing, and we do not rely on conditional independence assumptions. Indeed, the model may include the noisy protected attribute as a feature.

Robust optimization. We use a minimax set-up of a two-player game where the uncertainty is adversarial, and one minimizes a worst-case objective over a feasible set [Ben-Tal et al., 2009, Bertsimas et al., 2011]; e.g., the noise is contained in a unit-norm ball around the input data. As one such approach, we apply a recent line of work on DRO which assumes that the uncertain distributions of the data are constrained to belong to a certain set [Namkoong and Duchi, 2016, Duchi and Namkoong, 2018, Li et al., 2019].

2.3 Optimization Problem Setup

We begin with the training problem for incorporating group-based fairness criteria in a learning setting [Goh et al., 2016, Hardt et al., 2016b, Donini et al., 2018, Agarwal et al., 2018, Cotter et al., 2019d]. Let $X \in \mathcal{X} \subseteq \mathbb{R}^D$ be a random variable representing a feature vector, with a random binary label $Y \in \mathcal{Y} = \{0, 1\}$ and random protected group membership $G \in \mathcal{G} = \{1, ..., m\}$. In addition, let $\hat{G} \in \hat{\mathcal{G}} = \{1, ..., \hat{m}\}$ be a random variable representing the noisy protected group label for each (X, Y), which we assume we have access to during training. For simplicity, assume that $\hat{\mathcal{G}} = \mathcal{G}$ (and $\hat{m} = m$). Let $\phi(X; \theta)$ represent a binary classifier with parameters $\theta \in \Theta$ where $\phi(X; \theta) > 0$ indicates a positive classification.

Then, training with fairness constraints [Goh et al., 2016, Hardt et al., 2016b, Donini et al., 2018, Agarwal et al., 2018, Cotter et al., 2019d] is:

$$\min_{\theta} \quad f(\theta) \quad \text{s.t.} \quad g_j(\theta) \leq 0, \forall j \in \mathcal{G}, \tag{2.1}$$

The objective function $f(\theta) = \mathbb{E}[l(\theta, X, Y)]$, where $l(\theta, X, Y)$ is any standard binary classifier training loss. The constraint functions $g_j(\theta) = \mathbb{E}[h(\theta, X, Y)|G = j]$ for $j \in \mathcal{G}$, where $h(\theta, X, Y)$

is the target fairness metric, e.g. $h(\theta, X, Y) = \mathbb{1}\left(\phi(X;\theta) > 0\right) - \mathbb{E}[\mathbb{1}\left(\phi(X;\theta) > 0\right)]$ when equalizing positive rates for the *demographic parity* [Dwork et al., 2012] criterion (see Cotter et al. [2019d] for more examples). Algorithms have been studied for problem (2.1) when the true protected group labels G are given [see, e.g., Eban et al., 2017, Agarwal et al., 2018, Cotter et al., 2019d].

2.4 Bounds for the Naïve Approach

When only given the noisy groups $\hat{G}$, one naïve approach to solving problem (2.1) is to simply re-define the constraints using the noisy groups [Gupta et al., 2018]:

$$\min_{\theta} \quad f(\theta) \quad \text{s.t.} \quad \hat{g}_j(\theta) \leq 0, \;\; \forall j \in \mathcal{G}, \tag{2.2}$$

where $\hat{g}_j(\theta) = \mathbb{E}[h(\theta, X, Y)|\hat{G} = j], \;\; j \in \mathcal{G}$.

This introduces a practical question: if a model was constrained to satisfy fairness criteria on the noisy groups, how far would that model be from satisfying the constraints on the true groups? We show that the fairness violations on the true groups G can at least be bounded when the fairness criteria are satisfied on the noisy groups $\hat{G}$, provided that $\hat{G}$ does not deviate too much from G.

Bounding fairness constraints using TV distance

Recall that $X, Y|G = j \sim p_j$ and $X, Y|\hat{G} = j \sim \hat{p}_j$. We use the TV distance $TV(p_j, \hat{p}_j)$ to measure the distance between the probability distributions p_j and $\hat{p}_j$ (see Appendix A.1 and Villani [2009]). Given a bound on $TV(p_j, \hat{p}_j)$, we obtain a bound on fairness violations for the true groups when naïvely solving the optimization problem (2.2) using only the noisy groups:

Theorem 1. *(proof in Appendix A.1.) Suppose a model with parameters θ satisfies fairness criteria with respect to the noisy groups $\hat{G}$: $\hat{g}_j(\theta) \leq 0, \;\; \forall j \in \mathcal{G}$. Suppose $|h(\theta, x_1, y_1) - h(\theta, x_2, y_2)| \leq 1$ for any $(x_1, y_1) \neq (x_2, y_2)$. If $TV(p_j, \hat{p}_j) \leq \gamma_j$ for all $j \in \mathcal{G}$, then the fairness criteria with respect to the true groups G will be satisfied within slacks γ_j for each group: $g_j(\theta) \leq \gamma_j, \;\; \forall j \in \mathcal{G}$.*

Theorem 1 is tight for the family of functions h that satisfy $|h(\theta, x_1, y_1) - h(\theta, x_2, y_2)| \leq 1$ for any $(x_1, y_1) \neq (x_2, y_2)$. This condition holds for any fairness metrics based on rates such as demographic parity, where h is simply some scaled combination of indicator functions. Cotter et al. [2019d] list many such rate-based fairness metrics. Theorem 1 can be generalized to functions h whose differences are not bounded by 1 by looking beyond the TV distance to more general Wasserstein distances between p_j and $\hat{p}_j$. We show this in Appendix A.1, but for all fairness metrics referenced in this work, formulating Theorem 1 with the TV distance is sufficient.

Estimating the TV distance bound in practice

Theorem 1 bounds the fairness violations of the naïve approach in terms of the TV distance between the conditional distributions p_j and $\hat{p}_j$, which assumes knowledge of p_j and is not always possible to estimate. Instead, we can estimate an upper bound on $TV(p_j, \hat{p}_j)$ from metrics that are easier to obtain in practice. Specifically, the following lemma shows that shows that if the prior on class j is unaffected by the noise, $P(G \neq \hat{G} | G = j)$ directly translates into an upper bound on $TV(p_j, \hat{p}_j)$.

Lemma 1. (proof in Appendix A.1.) Suppose $P(G = j) = P(\hat{G} = j)$ for a given $j \in \mathcal{G}$. Then $TV(p_j, \hat{p}_j) \leq P(G \neq \hat{G} | G = j)$.

In practice, an estimate of $P(G \neq \hat{G} | G = j)$ may come from a variety of sources. As assumed by Kallus et al. [2020], a practitioner may have access to an *auxiliary* dataset containing G and $\hat{G}$, but not X or Y. Or, practitioners may have some prior estimate of $P(G \neq \hat{G} | G = j)$: if $\hat{G}$ is estimated by mapping zip codes to the most common socioeconomic group for that zip code, then census data provides a prior for how often $\hat{G}$ produces an incorrect socioeconomic group.

By relating Theorem 1 to realistic noise models, Lemma 1 allows us to bound the fairness violations of the naïve approach using quantities that can be estimated empirically. In the next section we show that Lemma 1 can also be used to produce a *robust* approach that will actually guarantee full satisfaction of the fairness violations on the true groups G.

2.5 Robust Approach 1: Distributionally Robust Optimization (DRO)

While Theorem 1 provides an upper bound on the performance of the naïve approach, it fails to provide a guarantee that the constraints on the true groups are satisfied, i.e. $g_j(\theta) \leq 0$. Thus, it is important to find other ways to do better than the naïve optimization problem (2.2) in terms of satisfying the constraints on the true groups. In particular, suppose in practice we are able to assert that $P(G \neq \hat{G} | G = j) \leq \gamma_j$ for all groups $j \in \mathcal{G}$. Then Lemma 1 implies a bound on TV distance between the conditional distributions on the true groups and the noisy groups: $TV(p_j, \hat{p}_j) \leq \gamma_j$. Therefore, any feasible solution to the following constrained optimization problem is guaranteed to satisfy the fairness constraints on the true groups:

$$\min_{\theta \in \Theta} \quad f(\theta) \quad \text{s.t.} \quad \max_{\substack{\tilde{p}_j : TV(\tilde{p}_j, \hat{p}_j) \leq \gamma_j \\ \tilde{p}_j \ll p}} \tilde{g}_j(\theta) \leq 0, \quad \forall j \in \mathcal{G}, \tag{2.3}$$

where $\tilde{g}_j(\theta) = \mathbb{E}_{X,Y \sim \tilde{p}_j}[h(\theta, X, Y)]$, and $\tilde{p}_j \ll p$ denotes absolute continuity.

General DRO formulation

A DRO problem is a minimax optimization [Duchi and Namkoong, 2018]:

$$\min_{\theta \in \Theta} \max_{q: D(q,p) \leq \gamma} \mathbb{E}_{X,Y \sim q}[l(\theta, X, Y)], \tag{2.4}$$

where D is some divergence metric between the distributions p and q, and $l : \Theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Much existing work on DRO focuses on how to solve the DRO problem for different divergence metrics D. Namkoong and Duchi [2016] provide methods for efficiently and optimally solving the DRO problem for f-divergences, and other work has provided methods for solving the DRO problem for Wasserstein distances [Li et al., 2019, Esfahani and Kuhn., 2018]. Duchi and Namkoong [2018] further provide finite-sample convergence rates for the empirical version of the DRO problem.

Solving the DRO problem

An important and often difficult aspect of using DRO is specifying a divergence D and bound γ that are meaningful. In this case, Lemma 1 gives us the key to formulating a DRO problem that is guaranteed to satisfy the fairness criteria with respect to the true groups G.

The optimization problem (2.3) can be written in the form of a DRO problem (2.4) with TV distance by using the Lagrangian formulation. Adapting a simplified version of a gradient-based algorithm provided by Namkoong and Duchi [2016], we are able to solve the empirical formulation of problem (2.4) efficiently. Details of our empirical Lagrangian formulation and pseudocode are in Appendix A.2.

2.6 Robust Approach 2: Soft Group Assignments

While any feasible solution to the distributionally robust constrained optimization problem (2.3) is guaranteed to satisfy the constraints on the true groups G, choosing each $\gamma_j = P(G \neq \hat{G} | G = j)$ as an upper bound on $TV(p_j, \hat{p}_j)$ may be rather conservative. Therefore, as an alternative to the DRO constraints in (2.3), in this section we show how to optimize using the robust fairness criteria proposed by Kallus et al. [2020].

Constraints with soft group assignments

Given a trained binary predictor, $\hat{Y}(\theta) = \mathbb{1}(\phi(\theta; X) > 0)$, Kallus et al. [2020] proposed a set of robust fairness criteria that can be used to audit the fairness of the given trained model with respect to the true groups $G \in \mathcal{G}$ using the noisy groups $\hat{G} \in \hat{\mathcal{G}}$, where $\mathcal{G} = \hat{\mathcal{G}}$ is not required in general. They assume access to a *main dataset* with the noisy groups $\hat{G}$, true labels Y, and features X, as well an *auxiliary dataset* containing both the noisy groups $\hat{G}$ and the true groups G. From the main dataset, one obtains estimates of the joint distributions

$(\hat{Y}(\theta), Y, \hat{G})$; from the auxiliary dataset, one obtains estimates of the joint distributions $(\hat{G}, G)$ and a noise model $P(G = j|\hat{G} = k)$ $\forall j \in \mathcal{G}, k \in \hat{\mathcal{G}}$.

These estimates are used to associate each example with a vector of weights, where each weight is an estimated probability that the example belongs to the true group j. Specifically, suppose that we have a function $w : \mathcal{G} \times \{0,1\} \times \{0,1\} \times \hat{\mathcal{G}} \to [0,1]$, where $w(j \mid \hat{y}, y, k)$ estimates $P(G = j|\hat{Y}(\theta) = \hat{y}, Y = y, \hat{G} = k)$. We rewrite the fairness constraint $E[h(\theta, X, Y)|G = j] = \frac{E[h(\theta,X,Y)P(G=j|\hat{Y}(\theta),Y,\hat{G})]}{P(G=j)}$ (derivation in Appendix A.3), and estimate this using w. We also show how h can be adapted to the *equality of opportunity* setting in Appendix A.3.

Given the main dataset and auxiliary dataset, we limit the possible values of the function $w(j \mid \hat{y}, y, k)$ using the law of total probability (as in Kallus et al. [2020]). The set of possible functions w is given by:

$$\mathcal{W}(\theta) = \left\{ w : \substack{\sum_{\hat{y},y\in\{0,1\}} w(j|\hat{y},y,k)P(\hat{Y}(\theta)=\hat{y},Y=y|\hat{G}=k)=P(G=j|\hat{G}=k), \\ \sum_{j=1}^{m} w(j|\hat{y},y,k)=1, w(j|\hat{y},y,k)\geq 0 \quad \forall \hat{y},y\in\{0,1\}, j\in\mathcal{G}, k\in\hat{\mathcal{G}}} \right\}. \tag{2.5}$$

The robust fairness criteria can now be written in terms of $\mathcal{W}(\theta)$ as:

$$\max_{w\in\mathcal{W}(\theta)} g_j(\theta, w) \leq 0, \ \ \forall j \in \mathcal{G} \quad \text{where} \quad g_j(\theta, w) = \frac{\mathbb{E}[h(\theta, X, Y)w(j|\hat{Y}(\theta), Y, \hat{G})]}{P(G = j)}. \tag{2.6}$$

Robust optimization with soft group assignments

We extend Kallus et al. [2020]'s work by formulating a robust optimization problem using soft group assignments. Combining the robust fairness criteria above with the training objective, we propose:

$$\min_{\theta\in\Theta} \quad f(\theta) \quad \text{s.t.} \max_{w\in\mathcal{W}(\theta)} g_j(\theta, w) \leq 0, \ \ \forall j \in \mathcal{G}, \tag{2.7}$$

where Θ denotes the space of model parameters. Any feasible solution is guaranteed to satisfy the original fairness criteria with respect to the true groups. Using a Lagrangian, problem (2.7) can be rewritten as:

$$\min_{\theta\in\Theta} \max_{\lambda\in\Lambda} \mathcal{L}(\theta, \lambda) \tag{2.8}$$

where the Lagrangian $\mathcal{L}(\theta, \lambda) = f(\theta) + \sum_{j=1}^{m} \lambda_j \max_{w\in\mathcal{W}(\theta)} g_j(\theta, w)$, and $\Lambda \subseteq \mathbb{R}_{+}^{m}$.

When solving this optimization problem, we use the empirical finite-sample versions of each expectation. As described in Proposition 9 of Kallus et al. [2020], the inner maximization (2.6) over $w \in \mathcal{W}(\theta)$ can be solved as a linear program for a given fixed θ. However, the Lagrangian problem (2.8) is not as straightforward to optimize, since the feasible set $\mathcal{W}(\theta)$ depends on θ through $\hat{Y}$. While in general the pointwise maximum of convex functions is convex, the dependence of $\mathcal{W}(\theta)$ on θ means that even if $g_j(\theta, w)$ were convex, $\max_{w\in\mathcal{W}(\theta)} g_j(\theta, w)$ is not necessarily convex. We first introduce a theoretically ideal algorithm that we prove converges to an optimal, feasible solution. This ideal algorithm relies on a minimization oracle, which is not always computationally tractable. Therefore, we further provide a practical algorithm using gradient methods that mimics the ideal algorithm in structure and computationally tractable, but does not share the same convergence guarantees.

Ideal algorithm

The minimax problem in equation (2.8) can be interpreted as a zero-sum game between the θ-player and λ-player. In Algorithm 1, we provide an iterative procedure for solving equation (2.8), where at each step, the θ-player performs a full optimization, i.e., a *best response* over θ, and the λ-player responds with a gradient ascent update on λ.

For a fixed θ, the gradient of the Lagrangian $\mathcal{L}$ with respect to λ is given by $\partial\mathcal{L}(\theta, \lambda)/\partial\lambda_j = \max_{w\in\mathcal{W}(\theta)} g_j(\theta, w)$, which is a linear program in w. The challenging part, however, is the best response over θ; that is, finding a solution $\min_\theta \mathcal{L}(\theta, \lambda)$ for a given λ, as this involves a max over constraints $\mathcal{W}(\theta)$ which depend on θ. To implement this best response, we formulate a nested minimax problem that decouples this intricate dependence on θ, by introducing Lagrange multipliers for the constraints in $\mathcal{W}(\theta)$. We then solve this problem with an oracle that jointly minimizes over both θ and the newly introduced Lagrange multipliers (details in Algorithm 4 in Appendix A.4).

The output of the best-response step is a stochastic classifier with a distribution $\hat{\theta}^{(t)}$ over a finite set of θs. Algorithm 1 then returns the average of these distributions, $\overline{\theta} = \frac{1}{T}\sum_{t=1}^{T}\hat{\theta}^t$, over T iterations. By extending recent results on constrained optimization [Cotter et al., 2019b], we show in Appendix A.4 that the output $\overline{\theta}$ is near-optimal and near-feasible for the robust optimization problem in equation (2.7). That is, for a given $\varepsilon > 0$, by picking T to be large enough, we have that the objective $\mathbb{E}_{\theta\sim\overline{\theta}}\left[f(\theta)\right] \leq f(\theta^*) + \varepsilon$, for any θ^* that is feasible, and the expected violations in the robust constraints are also no more than ε.

Algorithm 1 *Ideal* Algorithm

```
Require: learning rate $\eta_\lambda > 0$, estimates of $P(G = j|\hat{G} = k)$ to specify $\mathcal{W}(\theta)$, $\rho$, $\rho'$
for $t = 1, \ldots, T$ do
    Best response on $\theta$: run the oracle-based Algorithm 4 to find a distribution $\hat{\theta}^{(t)}$ over $\Theta$
        s.t. $\mathbb{E}_{\theta\sim\hat{\theta}^{(t)}}\left[\mathcal{L}(\theta, \lambda^{(t)})\right] \leq \min_{\theta\in\Theta} \mathcal{L}(\theta, \lambda^{(t)}) + \rho$.
    Estimate gradient $\nabla_\lambda\mathcal{L}(\hat{\theta}^{(t)}, \lambda^{(t)})$: for each $j \in \mathcal{G}$, choose $\delta_j^{(t)}$ s.t.
        $\delta_j^{(t)} \leq \mathbb{E}_{\theta\sim\hat{\theta}^{(t)}}\left[\max_{w\in\mathcal{W}(\theta)} g_j(\theta, w)\right] \leq \delta_j^{(t)} + \rho'$
    Ascent step on $\lambda$: $\tilde{\lambda}_j^{(t+1)} \leftarrow \lambda_j^{(t)} + \eta_\lambda\,\delta_j^{(t)}, \;\; \forall j \in \mathcal{G}; \quad \lambda^{(t+1)} \leftarrow \Pi_\Lambda(\tilde{\lambda}^{(t+1)})$
end for
return $\overline{\theta} = \frac{1}{T}\sum_{t=1}^{T}\hat{\theta}^{(t)}$
```

Practical algorithm

Algorithm 1 is guaranteed to converge to a near-optimal, near-feasible solution, but may be computationally intractable and impractical for the following reasons. First, the algorithm needs a nonconvex minimization oracle to compute a best response over θ. Second, there are multiple levels of nesting, making it difficult to scale the algorithm with mini-batch or stochastic updates. Third, the output is a distribution over multiple models, which can be be difficult to use in practice Narasimhan et al. [2019b].

Therefore, we supplement Algorithm 1 with a practical algorithm, Algorithm 5 (see Appendix A.5) that is similar in structure, but approximates the inner best response routine with two simple steps: a maximization over $w \in \mathcal{W}(\theta^{(t)})$ using a linear program for the current iterate $\theta^{(t)}$, and a gradient step on θ at the maximizer $w^{(t)}$. Algorithm 5 leaves room for other practical modifications such as using stochastic gradients. We provide further discussion in Appendix A.5.

2.7 Experiments

We compare the performance of the naïve approach and the two robust optimization approaches (DRO and soft group assignments) empirically using two datasets from UCI Dua and Graff [2017] with different constraints. For both datasets, we stress-test the performance of the different algorithms under different amounts of noise between the true groups G and the noisy groups $\hat{G}$. We take l to be the hinge loss. The specific constraint violations measured and additional training details can be found in Appendix A.6. All experiment code is available on GitHub at https://github.com/wenshuoguo/robust-fairness-code.

Generating noisy protected groups: Given the true protected groups, we synthetically generate noisy protected groups by selecting a fraction γ of data uniformly at random. For each selected example, we perturb the group membership to a different group also selected uniformly at random from the remaining groups. This way, for a given γ, $P(\hat{G} \neq G) \approx P(\hat{G} \neq G | G = j) \approx \gamma$ for all groups $j, k \in \mathcal{G}$. We evaluate the performance of the different algorithms ranging from small to large amounts of noise: $\gamma \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$.

Case study 1 (Adult): equality of opportunity

We use the Adult dataset from UCI [Dua and Graff, 2017] collected from 1994 US Census, which has 48,842 examples and 14 features (details in Appendix A.6). The classification task is to determine whether an individual makes over $50K per year. For the true groups, we use $m = 3$ race groups of “white,” “black,” and “other.” As done by Cotter et al. [2019d], Friedler et al. [2019], Zafar et al. [2019], we enforce *equality of opportunity* by equalizing true positive rates (TPRs). Specifically, we enforce that the TPR conditioned on each group is greater than or equal to the overall TPR on the full dataset with some slack α, which produces m true group fairness criteria, $\{g_j^{\text{TPR}}(\theta) \leq 0\} \;\; \forall j \in \mathcal{G}$ (details on the constraint function h in Appendix A.2 and A.3).

Case study 2 (Credit): equalized odds

We consider another application of group-based fairness constraints to credit default prediction. Fourcade and Healy [2013] provide an in depth study of the effect of credit scoring techniques on the credit market, showing that this scoring system can perpetuate inequity. Enforcing group-based fairness with credit default predictions has been considered in a variety of prior works [Hardt et al., 2016b, Berk et al., 2017b, Wang and Gupta, 2020, Aghaei et al., 2019, Bera et al., 2019, Grari et al., 2020, Friedler et al., 2019, Barocas et al., 2019]. Following Hardt et al. [2016b] and Grari et al. [2020], we enforce *equalized odds* [Hardt et al., 2016b] by equalizing both true positive rates (TPRs) and false positive rates (FPRs) across groups.

We use the "default of credit card clients" dataset from UCI Dua and Graff [2017] collected by a company in Taiwan Yeh and hui Lien [2009], which contains 30,000 examples and 24 features (details in Appendix A.6). The classification task is to determine whether an individual defaulted on a loan. We use $m = 3$ groups based on education levels: "graduate school," "university," and "high school/other" (the use of education in credit lending has previously been studied in the algorithmic fairness and economics literature [Gillis, 2020, Bera et al., 2019, Lazar and Vijaykumar, 2020]). We constrain the TPR conditioned on each group to be greater than or equal to the overall TPR on the full dataset with a slack α, and the FPR conditioned on each group to be less than or equal to the overall FPR on the full dataset. This produces $2m$ true group-fairness criteria, $\{g_j^{\text{TPR}}(\theta) \leq 0, g_j^{\text{FPR}}(\theta) \leq 0\}$ $\forall j \in \mathcal{G}$ (details on constraint functions h in Appendix A.2 and A.3).

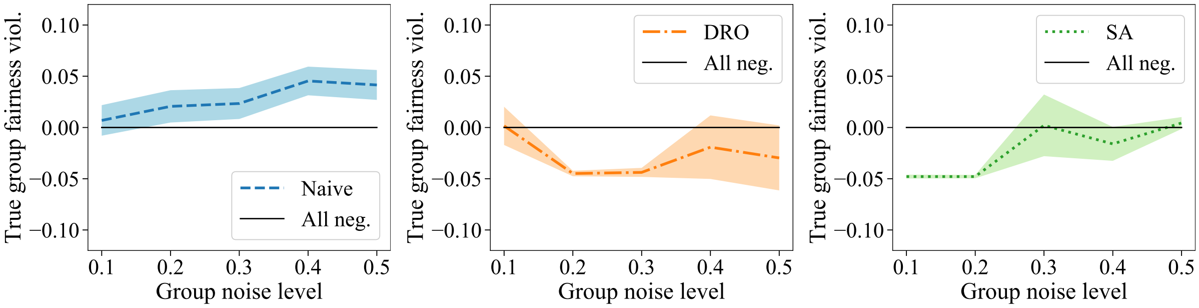


Figure 2.1: Case study 1 (Adult): maximum true group constraint violations on test set for the Naive, DRO, and soft assignments (SA) approaches for different group noise levels γ on the Adult dataset (mean and standard error over 10 train/val/test splits). The black solid line represents the performance of the trivial "all negatives" classifier, which has constraint violations of 0. A negative violation indicates satisfaction of the fairness constraints on the true groups.

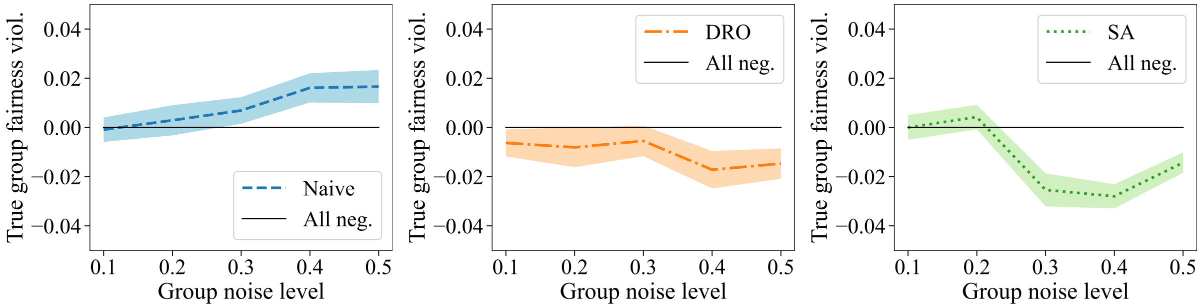


Figure 2.2: Case study 2 (Credit): maximum true group constraint violations on test set for the Naive, DRO, and soft assignments (SA) approaches for different group noise levels γ on the Credit dataset (mean and standard error over 10 train/val/test splits). This figure shows the max constraint violation over all TPR and FPR constraints, and Figure A.3 in Appendix A.6 shows the breakdown of these constraint violations into the max TPR and the max FPR constraint violations.

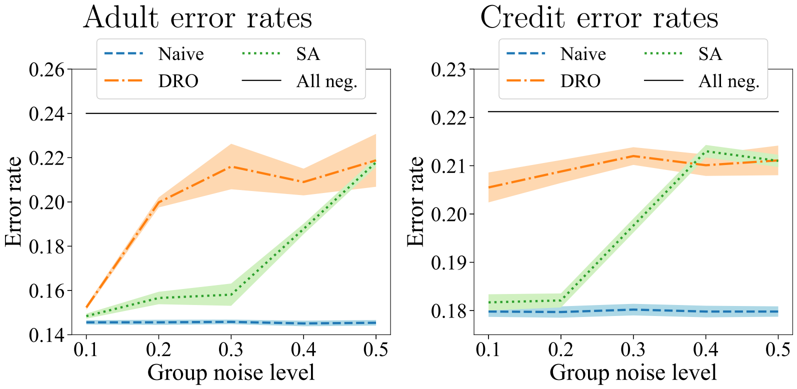


Figure 2.3: Error rates on test set for different group noise levels γ on the Adult dataset (*left*) and the Credit dataset (*right*) (mean and standard error over 10 train/val/test splits). The black solid line represents the performance of the trivial “all negatives” classifier. The soft assignments (SA) approach achieves lower error rates than DRO, and as the noise level increases, the gap in error rate between the naive approach and each robust approach increases.

Results

In case study 1 (Adult), the unconstrained model achieves an error rate of 0.1447 ± 0.0012 (mean and standard error over 10 splits) and a maximum constraint violation of 0.0234 ± 0.0164 on test set with respect to the true groups. The model that assumes knowledge of the true groups achieves an error rate of 0.1459 ± 0.0012 and a maximum constraint violation of -0.0469 ± 0.0068 on test set with respect to the true groups. As a sanity check, this demonstrates that when given access to the true groups, it is possible to satisfy the constraints on the test set with a reasonably low error rate.

In case study 2 (Credit), the unconstrained model achieves an error rate of 0.1797 ± 0.0013 (mean and standard error over 10 splits) and a maximum constraint violation of 0.0264 ± 0.0071 on the test set with respect to the true groups. The constrained model that assumes knowledge of the true groups achieves an error rate of 0.1796 ± 0.0011 and a maximum constraint violation of -0.0105 ± 0.0070 on the test set with respect to the true groups. For this dataset, it was possible to satisfy the constraints with approximately the same error rate on test as the unconstrained model. Note that the unconstrained model achieved a lower error rate on the train set than the constrained model (0.1792 ± 0.0015 unconstrained vs. 0.1798 ± 0.0024 constrained).

For both case studies, Figures 2.1 and 2.2 show that the robust approaches DRO (*center*) and soft group assignments (SA) (*right*) satisfy the constraints on average for all noise levels. As the noise level increases, the naïve approach (*left*) has increasingly higher true group constraint violations. The DRO and SA approaches come at a cost of a higher error rate than the naïve approach (Figure 2.3). The error rate of the naïve approach is close to the model optimized with constraints on the true groups G, regardless of the noise level γ. However, as the noise increases, the naïve approach no longer controls the fairness violations on the true groups G, even though it does satisfy the constraints on the noisy groups $\hat{G}$ (Figures A.1 and A.4 in Appendix A.6). DRO generally suffers from a higher error rate compared to SA (Figure 2.3), illustrating the conservatism of the DRO approach.

2.8 Conclusions

We explore the practical problem of enforcing group-based fairness for binary classification given noisy protected group information. In addition to providing new theoretical analysis of the naïve approach of only enforcing fairness on the noisy groups, we also propose two new robust approaches that guarantee satisfaction of the fairness criteria on the true groups. For the DRO approach, Lemma 1 gives a theoretical bound on the TV distance to use in the optimization problem. For the soft group assignments approach, we provide a theoretically ideal algorithm and a practical alternative algorithm for satisfying the robust fairness criteria proposed by Kallus et al. [2020] while minimizing a training objective. We empirically show that both of these approaches managed to satisfy the constraints with respect to the true groups, even under difficult noise models.

In follow-up work, Narasimhan et al. [2020] provide a general method for enforcing a large number of constraints at once, and enforce constraints concurrently on many possible realizations of noisy protected groups under a given noise model. This can be seen as an extension of the Soft Group Assignments approach that we propose in Section 2.6, which Narasimhan et al. [2020] describe in their Appendix.

One additional avenue of future work is to empirically compare the robust approaches when the noisy groups have different dimensionality from the true groups (Appendix A.2). Second, the looseness of the bound in Lemma 1 can lead to over-conservatism of the DRO approach, suggesting a need to better calibrate the DRO neighborhood. Finally, it would be valuable to study the impact of distribution mismatch between the main dataset and the auxiliary dataset.

2.9 Choices and Limitations of Fairness Metrics

As machine learning is increasingly employed in high stakes environments, any potential application has to be scrutinized to ensure that it will not perpetuate, exacerbate, or create new injustices. Aiming to make machine learning algorithms themselves intrinsically fairer, more inclusive, and more equitable plays an important role in achieving that goal. Group-based fairness [Hardt et al., 2016b, Friedler et al., 2019] is a popular approach that the machine learning community has used to define and evaluate fair machine learning algorithms. Until recently, such work has generally assumed access to clean, correct protected group labels in the data. Our work addresses the technical challenge of enforcing group-based fairness criteria under noisy, unreliable, or outdated group information. However, we emphasize that this technical improvement alone does not necessarily lead to an algorithm having positive societal impact, for reasons that we now delineate.

Choice of fairness criteria

First, our work does not address the choice of the group-based fairness criteria. Many different algorithmic fairness criteria have been proposed, with varying connections to prior sociopolitical framing Narayanan [2018], Hutchinson and Mitchell [2019]. From an algorithmic standpoint, these different choices of fairness criteria have been shown to lead to very different prediction outcomes and tradeoffs Friedler et al. [2019]. Furthermore, even if a mathematical criterion may seem reasonable (e.g., equalizing positive prediction rates with *demographic parity*), Liu et al. [2018] show that the long-term impacts may not always be desirable, and the choice of criteria should be heavily influenced by domain experts, along with awareness of tradeoffs.

Choice of protected groups

In addition to the specification of fairness criteria, our work also assumes that the true protected group labels have been pre-defined by the practitioner. However, in real applications, the selection of appropriate true protected group labels is itself a nontrivial issue.

First, the measurement and delineation of these protected groups should not be overlooked, as "the process of drawing boundaries around distinct social groups for fairness research is fraught; the construction of categories has a long history of political struggle and legal argumentation" Hanna et al. [2020]. Important considerations include the context in which the group labels were collected, who chose and collected them, and what implicit assumptions are being made by choosing these group labels. One example is the operationalization of race in the context of algorithmic fairness. Hanna et al. [2020] critiques "treating race as an attribute, rather than a structural, institutional, and relational phenomenon." The choice of categories surrounding gender identity and sexual orientation have strong implications and consequences as well Group [2014], with entire fields dedicated to critiquing these constructs. Jacobs and Wallach [2019] provide a general framework for understanding measurement issues for these sensitive attributes in the machine-learning setting, building on foundational work from the social sciences Bandalos [2017].

Another key consideration when defining protected groups is problems of *intersectionality* Crenshaw [1990], Hooks [1992]. Group-based fairness criteria inherently do not consider within-group inequality Kasy and Abebe [2020]. Even if we are able to enforce fairness criteria robustly for a given set of groups, the intersections of groups may still suffer Buolamwini and Gebru [2018].

Domain specific considerations

Finally, we emphasize that group-based fairness criteria simply may not be sufficient to mitigate problems of significant background injustice in certain domains. Abebe et al. [2020] argue that computational methods have mixed roles in addressing social problems, where they can serve as *diagnostics*, *formalizers*, and *rebuttals*, and also that "computing acts as synecdoche when it makes long-standing social problems newly salient in the public eye." Moreover, the use of the algorithm itself may perpetuate inequity, and in the case of credit scoring, create stratifying effects of economic classifications that shape life-chances Fourcade and Healy [2013]. We emphasize the importance of domain specific considerations ahead of time before applying any algorithmic solutions (even "fair" ones) in sensitive and impactful settings.

Chapter 3

Robust Distillation for Worst-class Performance

3.1 Introduction

Knowledge distillation, wherein one trains a *teacher* model and uses its predictions to train a *student* model of similar or smaller capacity, has proven to be a powerful tool that improves efficiency while achieving state-of-the-art classification accuracies [Hinton et al., 2015, Radosavovic et al., 2018, Anil et al., 2018, Pham et al., 2021]. Remarkably, the student accuracy under distillation is capable of even surpassing that of the teacher (e.g. Xie et al. [2020]).

However, recent work has shown that the gains in average accuracy may not be uniform across subgroups, and can hurt performance on subgroups that are rarer or more difficult to classify. This is particularly true of long-tailed classification settings, where the improved average accuracy often comes at the cost of poorer accuracies on the tail classes [Lukasik et al., 2022, Du et al., 2021], and model compression can further amplify these performance disparities [Hooker et al., 2020, Xu et al., 2021].

To mitigate the disparity between average and subgroup accuracy, a common remedy is to train a model to achieve low *worst-group* test error. Suitably modified robust optimization techniques have successfully achieved state-of-the-art worst-class performance with manageable computational overhead [Sagawa et al., 2020a, Sohoni et al., 2020]. However, the evaluation of these techniques has thus far primarily focused on the standard training setting involving a single model. In the increasingly popular distillation setting, which involves both a teacher and student model, there is limited understanding of how these approaches can be applied to achieve the best trade-offs between average and worst-class performance. In particular, it is unknown if the best results come from using a robust objective for the teacher, the student or *both*.

This work studies the interplay between robust training objectives for the teacher and student. We focus on a multi-class classification setting where we define worst-class accuracy as the lowest per-class recall. Empirically, we show that jointly modifying *both* the teacher and student objectives with robust objectives not only improves the worst-class accuracy of the student, but can provide Pareto improvements in the trade-off between average and worst-class performance. Theoretically, we analyze what makes a good teacher when training a robust student, and give to our knowledge the first concrete characterization of this by showing that the student's robustness depends on how *well-calibrated* the teacher's scores are for the individual classes.

Contributions

Our contributions proceed as follows:

1. We begin with the problem setup (Section 3.3), and adapt existing robust optimization objectives to a distillation setting, allowing for different combinations of modifications to *both* the teacher and student objectives (Section 3.4). We provide adapted algorithms to address practical training issues that arise when applying robust objectives to both the teacher and student (such as margin-based surrogate losses and shared validation set usage).
2. We demonstrate empirically on benchmark image datasets that the different combinations of student and teacher objectives not only improve the student's worst-class accuracy, but yield better trade-offs between average and worst-class performance than baselines (Section 3.6). Perhaps surprisingly, we find that the teacher's worst-class accuracy is not always predictive of the teacher's ability to yield robust students.
3. We show theoretically that the worst-class robustness of the student depends on the *per-class calibration* of the teacher, and additionally derive robustness guarantees for the student in terms of the teacher's errors (Section3.7).

3.2 Related Work

Worst-group robustness. The goal of achieving good worst-case performance across subgroups can be framed as a (group) distributionally robust optimization (DRO) problem, and can be solved by iteratively updating costs on the individual groups and minimizing the resulting cost-weighted loss [Chen et al., 2017]. Recent variants of this approach have sought to avoid over-fitting through group-specific regularization [Sagawa et al., 2020a,b] or margin-based losses [Narasimhan and Menon, 2021, Kini et al., 2021], and to handle unknown subgroups [Sohoni et al., 2020]. In the context of distillation, Lukasik et al. [2022] propose simple modifications to robustify the student's objective by controlling the strength of the teacher's labels for different groups. In contrast, we propose a more direct and theoretically-grounded procedure that seeks to explicitly optimize for the student's worst-case error.

Relationship to Narasimhan and Menon [2021]. This paper builds on the margin-based DRO framework of Narasimhan and Menon [2021], who also include preliminary distillation experiments on training the teacher with standard ERM and the student with a robust objective. However, this and other prior work [Lukasik et al., 2022] have only explored modifications to the student loss, while training the teacher using a standard procedure. Our robust distillation proposals build on this method, but carry out a more extensive analysis, exploring different combinations of teacher-student objectives and different trade-offs between average and worst-class performance. Additionally, we provide robustness guarantees for the student, equip the DRO algorithms to achieve different trade-offs between overall and worst-case error, and provide a rigorous analysis of different design choices, such as the use of teacher labels for the multiplier updates.

Long-tail learning. There has been much work on training classifiers from long-tail data, ranging from modifications to loss modifications [Cao et al., 2019a, Menon et al., 2021b, Cui et al., 2021] to architectural changes [Wang et al., 2020b, Cui et al., 2022]. All these methods focus on the standard single model training setup, and seek to maximize the balanced (and not the worst-class) accuracy. Recent attempts have sought to modify standard distillation for long-tail learning, by either re-balancing the student loss [Zhang et al., 2021], temperature-scaling the teacher predictions [He et al., 2021], employing multiple teachers [Xiang et al., 2020], and leveraging the teacher's intermediate embeddings [Iscen et al., 2021]. The common goal in most of these papers is to modify the student's objective to incorporate different forms of supervision from the teacher. In contrast, we seek to explore modifications to the teacher's training objective to improve the student's robustness.

Role of the teacher's objective. Few previous works have studied how the objective of the teacher affects the student performance. For example, multiple works have studied the effect label smoothing objectives of the teacher model, some finding it to harm the student performance [Müller et al., 2019], improve the student [Shen et al., 2021] or show varying impact depending on the temperature value [Chandrasegaran et al., 2022]. In another work, Lukasik et al. [2020] showed how applying noise correction objectives to the teacher often yield better result than only applying noise correction objectives in the student. We are not aware of a previous work studying the *interplay* between the student and the teacher objectives on the robustness of the student.

3.3 Problem Setup

We consider a multi-class classification problem with instance space $\mathcal{X}$ and output space $[m] = \{1, \ldots, m\}$. Let D denote the underlying data distribution over $\mathcal{X} \times [m]$, and $D_{\mathcal{X}}$ denote the marginal distribution over $\mathcal{X}$. Let Δ_m denote the $(m-1)$-dimensional probability simplex over m classes. We define the conditional-class probability as $\eta_y(x) = \mathbb{P}(Y = y|X = x)$ and the class priors $\pi_y = \mathbb{P}(Y = y)$. Note that $\pi_y = \mathbb{E}_{X \sim D_{\mathcal{X}}}[\eta_y(X)]$.

Learning objectives. Our goal is to learn a multiclass classifier $h : \mathcal{X} \to [m]$ that maps an instance $x \in \mathcal{X}$ to one of m classes. We will do so by first learning a scoring function $f : \mathcal{X} \to \mathbb{R}^m$ that assigns scores $[f_1(x), \ldots, f_m(x)] \in \mathbb{R}^m$ to a given instance x, and construct the classifier by predicting the class with the highest score: $h(x) = \arg\max j \in [m]\, f_j(x)$. We will denote a softmax transformation of f by $\text{softmax}_y(f(x)) = \frac{\exp(f_y(x))}{\sum_j \exp(f_j(x))}$, and use the notation $\text{softmax}_y(f(x)) \propto z_y$ to indicate that $\text{softmax}_y(f(x)) = \frac{z_y}{\sum_{j=1}^m z_j}$.

We measure the efficacy of the scoring function f using a loss function $\ell : [m] \times \mathbb{R}^m \to \mathbb{R}_+$ that assigns a penalty $\ell(y, z)$ for predicting score vector $z \in \mathbb{R}^m$ for true label y. Examples of loss functions include the 0-1 loss: $\ell^{0\text{-}1}(y, z) = \mathbb{1}\left(z \neq \arg\max j\, f_j(x)\right)$, and the softmax cross-entropy loss: $\ell^{\text{xent}}(y, z) = -f_y(x) + \log\left(\sum_{j\in[m]} \exp\left(f_j(x)\right)\right)$.

Standard objective: A standard machine learning goal entails minimizing the overall expected risk:

$$L^{\text{std}}(f) = \mathbb{E}\left[\ell(Y, f(X))\right]. \tag{3.1}$$

Balanced objective: In applications where the classes are severely imbalanced, i.e., the class priors π_y are non-uniform and significantly skewed, one may wish to instead optimize a *balanced* version of the above objective, where we average over the conditional loss for each class. Notice that the conditional loss for class y is weighted by the inverse of its prior:

$$\begin{aligned} L^{\text{bal}}(f) &= \frac{1}{m}\sum_{y\in[m]} \mathbb{E}\left[\ell(y, f(X)) \,|\, Y = y\right] \\ &= \frac{1}{m}\sum_{y\in[m]} \frac{1}{\pi_y}\mathbb{E}_X\left[\eta_y(X)\,\ell(y, f(X))\right]. \end{aligned} \tag{3.2}$$

Robust objective: A more stringent objective would be to focus on the worst-performing class, and minimize a *robust* version of equation (3.1) that computes the worst among the m conditional losses:

$$L^{\text{rob}}(f) = \max_{y\in[m]} \frac{1}{\pi_y}\mathbb{E}\left[\eta_y(X)\,\ell(y, f(X))\right]. \tag{3.3}$$

In practice, focusing solely on either the average or the worst-case performance may not be an acceptable solution, and therefore, in this paper, we will additionally seek to characterize the trade-off between the balanced and robust objectives. One way to achieve this trade-off is to minimize the robust objective, while constraining the balanced objective to be within an acceptable range. This constrained optimization can be equivalently formulated as optimizing a convex combination of the balanced and robust objectives, for trade-off $\alpha \in [0, 1]$:

$$L^{\text{tdf}}(f) = (1-\alpha)L^{\text{bal}}(f) + \alpha L^{\text{rob}}(f). \tag{3.4}$$

A similar trade-off can also be specified between the standard and robust objectives. To better understand the differences between the standard, balanced and robust objectives in equation (3.1)–equation (3.4), we look at the optimal scoring function for each given a cross-entropy loss:

Theorem 2 (**Bayes-optimal scorers**). *When ℓ is the cross-entropy loss ℓ^{xent}, the minimizers of equation (3.1)–equation (3.3) over all measurable functions $f : \mathcal{X} \to \mathbb{R}^m$ are given by:*

(i) $L^{\mathrm{std}}(f)$*:* $\mathrm{softmax}_y(f^*(x)) = \eta_y(x)$

(ii) $L^{\mathrm{bal}}(f)$*:* $\mathrm{softmax}_y(f^*(x)) \propto \frac{1}{\pi_y}\eta_y(x)$

(iii) $L^{\mathrm{rob}}(f)$*:* $\mathrm{softmax}_y(f^*(x)) \propto \frac{\lambda_y}{\pi_y}\eta_y(x)$

(iv) $L^{\mathrm{tdf}}(f)$*:* $\mathrm{softmax}_y(f^*(x)) \propto \frac{(1-\alpha)\frac{1}{m}+\alpha\lambda'_y}{\pi_y}\eta_y(x)$,

for class-specific constants $\lambda, \lambda' \in \mathbb{R}^m_+$ that depend on distribution D.

All proofs are provided in Appendix B.1. Interestingly, the optimal scorers for all four objectives involve a simple scaling of the conditional-class probabilities $\eta_y(x)$.

3.4 Distillation for Worst-Class Performance

We adopt the common practice of training both the teacher and student on the same dataset. Specifically, given a training sample $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn from D, we first train a teacher model $p^t : \mathcal{X} \to \Delta_m$, and use it to generate a student dataset $S' = \{(x_1, p^t(x_1)), \ldots, (x_n, p^t(x_n))\}$ by replacing the original labels with the teacher's predictions. We then train a student scorer $f^s : \mathcal{X} \to [m]$ using the re-labeled dataset, and use it to construct the final classifier.

Teacher and student objectives

In a typical setting, both the teacher and student are trained to optimize a version of the standard objective in equation (3.1), i.e., the teacher is trained to minimize the average loss against the original training labels, and the student is trained to minimize an average loss against the teacher's predictions:

$$\text{Teacher: } \hat{L}^{\mathrm{std}}(f^t) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i, f^t(x_i)\right); \tag{3.5}$$

$$\text{Student: } \hat{L}^{\text{std-d}}(f^s) = \frac{1}{n}\sum_{i=1}^{n}\sum_{y=1}^{m} p^t_y(x_i)\, \ell\left(y, f(x_i)\right),$$

where $p^t(x) = \mathrm{softmax}(f^t(x))$. It is also common to have the student use a mixture of the teacher and one-hot labels. For concreteness, we consider a simpler distillation setup without this mixture, though extensions with this mixture would be straightforward to add. This work takes a wider view and explores *what combinations of student and teacher objectives* facilitate better worst-group performance for the student. Our experiments evaluate all *nine* combinations of standard, balanced, and robust teacher objectives, paired with standard, balanced, and robust student objectives.

Given the choice of teacher objective, the student will either optimize a distilled version of the balanced objective in equation (3.2):

$$\hat{L}^{\text{bal-d}}(f^s) = \frac{1}{m} \sum_{y \in [m]} \frac{1}{\hat{\pi}^t_y} \frac{1}{n} \sum_{i=1}^{n} p^t_y(x_i)\, \ell\left(y, f^s(x_i)\right), \tag{3.6}$$

or a distilled version of the robust objective in equation (3.3):

$$\hat{L}^{\text{rob-d}}(f^s) = \max_{y \in [m]} \frac{1}{\hat{\pi}^t_y} \frac{1}{n} \sum_{i=1}^{n} p^t_y(x_i)\, \ell\left(y, f^s(x_i)\right). \tag{3.7}$$

In practice, the teacher's predictions may have a different marginal distribution from the underlying class priors, particularly when temperature scaling is applied to the teacher's logits to soften the predicted probabilities [Narasimhan and Menon, 2021]. To address this, in both equation (3.6) and equation (3.7) we have replaced the class priors π_y with the marginal distribution $\hat{\pi}^t_y = \frac{1}{n}\sum_{i=1}^{n} p^t_y(x_i)$ from the teacher's predictions.

In addition to exploring the combination of objectives that facilitates better worst-group performance for the student, we evaluate a more flexible approach – have both the teachers and the students trade-off between the balanced and robust objectives:

$$\begin{aligned} &\text{Teacher: } \hat{L}^{\text{tdf}}(f^t) = (1-\alpha^t)\hat{L}^{\text{bal}}(f^t) + \alpha^t \hat{L}^{\text{rob}}(f^t) \\ &\text{Student: } \hat{L}^{\text{tdf-d}}(f^s) = (1-\alpha^s)\hat{L}^{\text{bal-d}}(f^s) + \alpha^s \hat{L}^{\text{rob-d}}(f^s), \end{aligned} \tag{3.8}$$

where $\hat{L}^{\text{bal}}(f^t)$ and $\hat{L}^{\text{rob}}(f^t)$ are the respective empirical estimates of equation (3.2) and equation (3.3) from the training sample, and $\alpha^t, \alpha^s \in [0,1]$ are the respective tradeoff parameters for the teacher and student. We are thus able to evaluate the Pareto-frontier of balanced and worst-case accuracies, obtained from different combinations of the teachers and students, and trained with different trade-off parameters.

3.5 Robust Distillation Algorithms

The different objectives we consider – standard, balanced and robust – entail different loss objectives to ensure efficient optimization during training. For example, while training the standard teacher and student in equation (3.5), we take ℓ to be the softmax cross-entropy loss, and optimize it using SGD. For the balanced and robust models, we employ the margin-based surrogates that we detail below, which have shown to be more effective in training over-parameterized networks [Cao et al., 2019b, Menon et al., 2021b, Kini et al., 2021]. Across all objectives, at evaluation we take the loss ℓ in the student and teacher objectives to be the 0-1 loss.

Algorithm 2 Distilled Margin-based DRO

```
Inputs: Teacher $p^t$, Student hypothesis class $\mathcal{F}$, Training set $S$, Validation set $S^{\text{val}}$, Step-size $\gamma \in \mathbb{R}_+$, Number of iterations $K$, Loss $\ell$, Initial student $f^0 \in \mathcal{F}$, Initial multipliers $\lambda^0 \in \Delta_m$
Compute $\hat{\pi}_j^t = \frac{1}{n}\sum_{(x,y)\in S} p_j^t(x),\ \forall j \in [m]$
Compute $\hat{\pi}_j^{t,\text{val}} = \frac{1}{n^{\text{val}}}\sum_{(x,y)\in S^{\text{val}}} p_j^t(x),\ \forall j \in [m]$
For $k = 0$ to $K - 1$
    $\tilde{\lambda}_j^{k+1} = \lambda_j^k \exp\left(\gamma \hat{R}_j\right), \forall j \in [m]$    where $\hat{R}_j = \frac{1}{n^{\text{val}}}\frac{1}{\hat{\pi}_j^{t,\text{val}}}\sum_{(x,y)\in S^{\text{val}}} p_j^t(x)\,\ell(j, f^k(x))$
    $\lambda_y^{k+1} = \frac{\tilde{\lambda}_y^{k+1}}{\sum_{j=1}^m \tilde{\lambda}_j^{k+1}}, \forall y$
    $f^{k+1} \in \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \mathcal{L}^{\text{mar}}\left(p^t(x_i), f(x_i); \frac{\lambda^{k+1}}{\hat{\pi}^t}\right)$    // Replaced with a few steps of SGD
End For
Output: $\overline{f}^s : x \mapsto \frac{1}{K}\sum_{k=1}^K f^k(x)$
```

Margin-based surrogate for balanced objective

When the teacher or student model being trained is over-parameterized, i.e., has sufficient capacity to correctly classify all examples in the training set, the use of an outer weighting term in the objective (such as the inverse class marginals in equation (3.6)) can be ineffective. In other words, a model that yields zero training objective would do so irrespective of what outer weights we choose. To remedy this problem, we make use of the margin-based surrogate of Menon et al. [2021b], and incorporate the outer weights as margin terms within the loss. For the balanced student objective in equation (3.6), this would look like:

$$\widetilde{L}^{\text{bal-d}}(f^s) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}^{\text{mar}}\left(p^t(x_i), f^s(x_i); \mathbb{1}/\hat{\pi}^t\right), \tag{3.9}$$

$$\text{where} \quad \mathcal{L}^{\text{mar}}(\mathbf{p}, \mathbf{f}; \mathbf{c}) = \frac{1}{m}\sum_{y\in[m]} p_y \log\left(1 + \sum_{j\neq y} \exp\left(\log(c_y/c_j) - (f_y - f_j)\right)\right),$$

for teacher probabilities $\mathbf{p} \in \Delta_m$, student scores $\mathbf{f} \in \mathbb{R}^m$, and per-class costs $\mathbf{c} \in \mathbb{R}_+^m$. For the balanced teacher, the margin-based objective would take a similar form, but with one-hot labels.

We include a proof in Appendix B.1 showing that a scoring function that minimizes the surrogate objective in equation (3.9) also minimizes the the balanced objective in equation (3.6) (when ℓ is the cross-entropy loss, and the student is chosen from a sufficiently flexible function class). In practice, the margin term $\log(c_y/c_j)$ encourages a larger margin of separation for classes y for which the cost c_y is relatively higher.

Margin-based DRO for robust objective

Minimizing the robust objective with plain SGD can be difficult due to the presence of the outer "max" over m classes. The key difficulty is in computing reliable stochastic gradients for the max objective, especially given a small batch size. The standard approach is to instead use a (group) distributionally-robust optimization (DRO) procedure, which comes in multiple flavors Chen et al. [2017], Sagawa et al. [2020a], Kini et al. [2021]. We employ the margin-based variant of group DRO [Narasimhan and Menon, 2021] as it naturally extends the margin-based objective used in the balanced setting.

We illustrate below how this applies to the robust student objective in equation (3.7). The procedure for the robust teacher is similar, but involves one-hot labels. For a student hypothesis class $\mathcal{F}$, we first re-write the minimization in equation (3.7) over $f \in \mathcal{F}$ into an equivalent min-max optimization using per-class multipliers $\lambda \in \Delta_m$:

$$\min_{f\in\mathcal{F}} \max_{\lambda\in\Delta_m} \sum_{y\in[m]} \frac{\lambda_y}{\hat{\pi}^t_y} \frac{1}{n} \sum_{i=1}^{n} p^t_y(x_i)\, \ell\left(y, f(x_i)\right),$$

and then maximize over λ for fixed f, and minimize over f for fixed λ:

$$\lambda^{k+1}_y \propto \lambda^k_y \exp\left(\gamma \frac{1}{n\hat{\pi}^t_y} \sum_{i=1}^{n} p^t_y(x_i)\, \ell\left(y, f^k(x_i)\right)\right), \forall y$$
$$f^{k+1} \in \operatorname*{argmin}_{f\in\mathcal{F}} \sum_{y\in[m]} \frac{\lambda^{k+1}_y}{n\hat{\pi}^t_y} \sum_{i=1}^{n} p^t_y(x_i)\, \ell\left(y, f(x_i)\right),$$

where $\gamma > 0$ is a step-size parameter. The updates on λ implement exponentiated gradient (EG) ascent to maximize over the simplex [Shalev-Shwartz et al., 2011].

Following Narasimhan and Menon [2021], we make two modifications to the above updates when used to train over-parameterized networks that can fit the training set perfectly. First, we perform the updates on λ using a small held-out validation set $S^{\text{val}} = \{(x_1, y_1), \ldots, (x_{n^{\text{val}}}, y_{n^{\text{val}}})\}$, instead of the training set, so that the λs reflect how well the model generalizes out-of-sample. Second, in keeping with the balanced objective, we modify the weighted objective in the f-minimization step to include a margin-based surrogate. Algorithm 2 provides a summary of these steps and returns a scorer that averages over the K iterates: $\overline{f}^s(x) = \frac{1}{K}\sum_{k=1}^{K} f^k(x)$. While the averaging is needed for our theoretical analysis, in practice, we find it sufficient to return the last scorer f^K. In Appendix B.4, we describe how Algorithm 7 can be easily modified to trade-off between the balanced and robust objectives, as shown in equation (3.8).

3.6 Experiments

To empirically understand the interplay of teacher and student objectives, we explore the following questions: *what combination of teacher and student objectives yield the highest worst-*

class accuracy? Can some combinations improve worst-class accuracy without sacrificing average accuracy?

Datasets

We evaluate the proposed distillation protocols on benchmark image datasets: *(i)* CIFAR-10, *(ii)* CIFAR-100 [Krizhevsky, 2009], *(iii)* TinyImageNet (a subset of ImageNet with 200 classes) [Le and Yang, 2015], and *(iv)* ImageNet [Russakovsky et al., 2015]. We also include long-tailed versions of the first three datasets created by downsampling tail classes [Cui et al., 2019]. For both the original and long-tailed versions of the datasets, there are often biases in worst-class performance, possibly due to some classes being easier to learn [Lukasik et al., 2022, Hooker et al., 2020]. For all datasets, as done in prior work [Menon et al., 2021b, Narasimhan and Menon, 2021], we randomly split the original default test set in half to create a validation set and test set, and use the same validation and test sets for the long-tailed training sets as for the original versions.

Architectures

We evaluate our distillation protocols in both a self-distillation and compression setting. On all CIFAR datasets, all teachers were trained with the ResNet-56 architecture and students were trained with either ResNet-56 or ResNet-32. On TinyImageNet and ImageNet, teachers and students were trained with ResNet-18. More details on these architectures can be found in Lukasik et al. [2022] and He et al. [2016] (see, e.g., Table 7 in Lukasik et al. [2022]). Self-distillation results are reported in the main paper (teacher/student share the same architecture), and we include results with compressed students in Appendix B.6.

Hyperparameters

We apply temperature scaling to the teacher scores, i.e., compute $p^t(x) = \mathrm{softmax}(f^t(x)/\gamma)$, and vary the temperature parameter γ over a range of $\{1, 3, 5\}$. A higher temperature produces a softer probability distribution over classes [Hinton et al., 2015]. Unless otherwise specified, the temperature hyperparameters were chosen to achieve the highest worst-class accuracy on the validation set. We closely mimic the learning rate and regularization settings from prior work [Menon et al., 2021b, Narasimhan and Menon, 2021] (see Appendix B.5 for details).

Which objective combinations are most robust?

We begin by exploring the effect of the interaction between student and teacher objectives on worst-class accuracy. In Table 3.1, we search over combinations of the standard, balanced, and robust objectives for the teacher ($L^{\text{std}}, L^{\text{bal}}, L^{\text{rob}}$) and the student ($L^{\text{std-d}}, L^{\text{bal-d}}, L^{\text{rob-d}}$) (note that on the original datasets, L^{std} is equivalent to L^{bal}). For each combination, following prior conventions in long-tailed learning [Menon et al., 2021b, Lukasik et al., 2022], we report the *average accuracy* over all classes, and the *worst-class accuracy*, or minimum per-class recall over all classes (see equation (3.3)). For datasets with a long tail or high number of classes, we also report the *worst-k accuracy*, which is the average of the the worst k per-class recalls.

The first surprising finding in Table 3.1 is that *applying the robust objective twice isn't always best.* For all but one dataset, the $L^{\text{rob}}/L^{\text{rob-d}}$ teacher/student combination was outperformed by some other combination of either $L^{\text{std}}/L^{\text{rob-d}}$, $L^{\text{rob}}/L^{\text{std-d}}$, or $L^{\text{bal}}/L^{\text{rob-d}}$. Still, in the winning combination, at least one of the objectives was robust. This suggests that while the robust objective is effective for controlling worst-class accuracy, there may be some information loss in applying it twice to both the teacher and student.

To understand this information loss on the teacher's side, we highlight a second surprising finding that *the teacher with the best worst-class accuracy alone did not always produce the student with the best worst-class accuracy.* The robust teacher had the highest worst-class accuracy across all datasets, but for CIFAR-10 and all three long-tailed datasets, it was actually the L^{std} or L^{bal} teacher that produced the best robust student. This shows that there is more to a good teacher than just having good worst-class performance – in fact, we show theoretically in Section 3.7 that the property of the teacher that is most important for robust student performance is a form of *calibration* of per-class scores.

Trading off accuracy and robustness

Table 3.1 focuses on worst-class accuracy, but practitioners often must consider the trade-off between average accuracy and worst-class accuracy when deploying any model. To address this, we introduced the $L^{\text{tdf}}/L^{\text{tdf-d}}$ objectives for the teacher/student with trade-off parameters α^t, α^s. Figure 3.1 plots average and worst-class accuracies for a full spread of α^t, α^s parameters. First, we note that lower α^s usually leads to higher average accuracy (this is not always the case for α^t, which we show in more detail in Appendix B.6). Figure 3.1 also shows that combinations of α^t, α^s yield a roughly concave Pareto frontier of solutions with different average and worst-class accuracies to choose from. Selecting the best combination of trade-off parameters α^t, α^s in practice depends on domain-specific decisions regarding the importance of worst-class vs. average accuracy. Any selection criteria based on some trade-off of worst-class vs. average accuracy can be applied over the validation set to select α^t, α^s as hyperparameters. We demonstrate one such set of selection criteria here: in Tables 3.2 and 3.3, we select α^t, α^s to maximize worst-class accuracy on validation, subject to having at least as high average accuracy as standard distillation (within error margin) on the validation set. Other candidate criteria include weighted sums of worst-class accuracy and average accuracy, or constrained optimization criteria from Cotter et al. [2019d].

Table 3.1: Worst-class accuracy comparisons for different combinations of teacher/student objectives. Worst-1 test accuracy is reported (worst-10 for TinyImageNet-LT) (best in **bold**), and average test accuracy is shown in parentheses. Mean accuracies are reported over repeat trainings (see extended table in Appendix for standard errors). Note that on the original datasets, L^{std} and $L^{\text{std-d}}$ are equivalent to L^{bal} and $L^{\text{bal-d}}$.

Student Obj.	**CIFAR-10** Teacher Obj.		**CIFAR-100** Teacher Obj.	
	L^{std}	L^{rob}	L^{std}	L^{rob}
None	86.48 (93.74)	90.09 (92.67)	42.22 (72.42)	43.42 (68.81)
$L^{\text{std-d}}$	87.66 (94.34)	90.12 (94.07)	43.81 (74.61)	**45.33** (73.67)
$L^{\text{rob-d}}$	**90.94** (92.54)	85.14 (89.58)	42.96 (68.71)	27.59 (54.79)

Student Obj.	**TinyImageNet** Teacher Obj.	
	L^{std}	L^{rob}
None	8.42 (56.79)	11.87 (48.40)
$L^{\text{std-d}}$	6.32 (57.83)	10.53 (55.36)
$L^{\text{rob-d}}$	9.98 (49.84)	**16.58** (46.11)

Student Obj.	**CIFAR-10-LT** Teacher Obj.			**CIFAR-100-LT** Teacher Obj.		
	L^{std}	L^{bal}	L^{rob}	L^{std}	L^{bal}	L^{rob}
None	57.26 (76.27)	68.52 (79.85)	74.80 (80.29)	0.00 (43.33)	3.75 (47.55)	10.33 (44.27)
$L^{\text{std-d}}$	36.67 (69.50)	66.96 (79.25)	71.15 (80.95)	0.00 (43.86)	2.39 (48.95)	7.32 (47.93)
$L^{\text{bal-d}}$	71.23 (80.50)	70.52 (81.12)	72.96 (80.71)	4.39 (50.40)	7.08 (50.10)	7.19 (47.51)
$L^{\text{rob-d}}$	63.85 (76.81)	**75.56** (80.81)	69.21 (76.72)	9.05 (33.75)	**12.52** (34.05)	10.32 (36.83)

Student Obj.	**CIFAR-10-LT** Teacher Obj.		
	L^{std}	L^{bal}	L^{rob}
None	0.00 (33.15)	2.11 (35.96)	4.92 (27.23)
$L^{\text{std-d}}$	0.00 (26.05)	0.00 (27.21)	1.87 (25.34)
$L^{\text{bal-d}}$	0.20 (30.43)	2.82 (39.41)	4.77 (38.41)
$L^{\text{rob-d}}$	0.00 (22.66)	**4.93** (35.43)	3.32 (25.11)

Comparison to baselines

Finally, we contextualize the performance of the proposed $L^{\text{tdf}}/L^{\text{tdf-d}}$ objectives and the training protocol in Algorithm 2 by comparing to several state-of-the-art methods. In addition to *standard distillation* (training the teacher with L^{std} and the student with $L^{\text{std-d}}$), we compare the proposed objective combinations with two recent works focusing on robust distillation [Lukasik et al., 2022, Narasimhan and Menon, 2021], both of which use a standard objective for the teacher and modify only the student objective for worst-class performance. From Narasimhan and Menon [2021], we consider the following two methods: *(i) Post-shifting:*

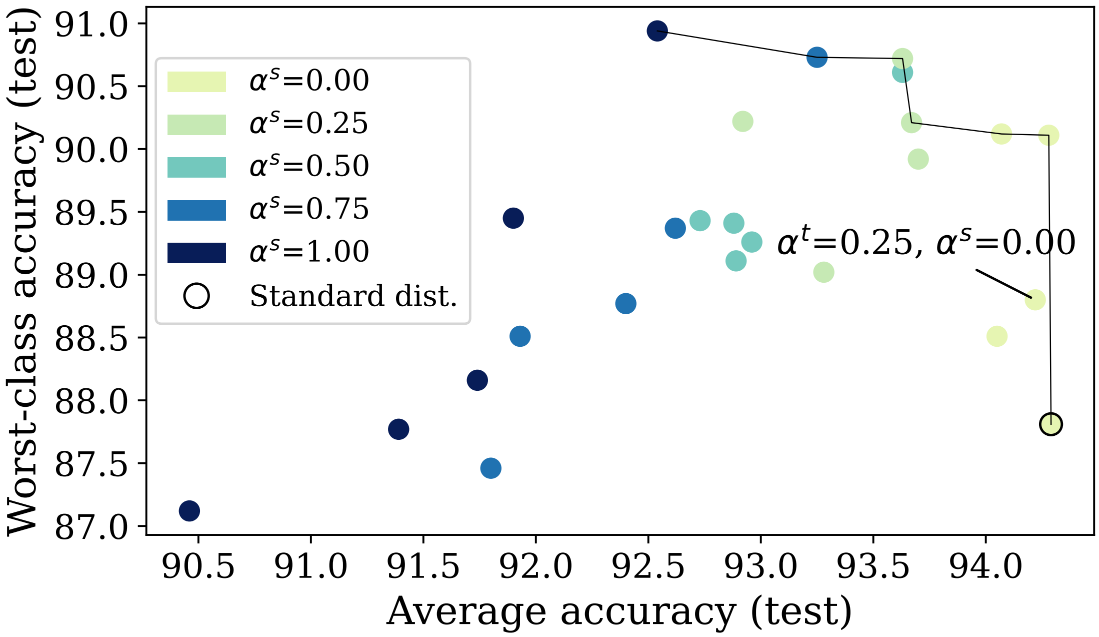


Figure 3.1: All α^t, α^s combinations for CIFAR-10 on test. The black line traces out the Pareto frontier. Average accuracy is roughly determined by α^s. The labeled point corresponds to the "best" combination selected in Table 3.2 based on validation criteria, but other domain-specific tradeoff criteria could yield any of these other points.

this non-distillation approach directly constructs a new scoring model by making post-hoc adjustments to the teacher, so as to maximize the robust accuracy on the validation sample. *(ii) Robust student:* this approach trains a student using $L^{\text{rob-d}}$ from a standard teacher. From Lukasik et al. [2022], we compare to their two proposed *AdaMargin* and *AdaAlpha* methods. Both methods are motivated by the observation that the margin defined for each class y by $\gamma_{\text{avg}}(y, p^{\text{t}}(x)) = p^{\text{t}}_y(x) - \frac{1}{m-1}\sum_{y' \neq y} p^{\text{t}}_{y'}(x)$ correlates with whether distillation improves over one-hot training [Lukasik et al., 2022]. AdaMargin uses that quantity as a margin in the distillation loss, whereas AdaAlpha uses it to adaptively mix between the one-hot and distillation losses. Additionally, for long-tailed datasets, we include a comparison to Menon et al. [2021b] which we refer to as *balanced student*, where the student is distilled with a balanced objective $L^{\text{bal-d}}$ from a standard teacher. Finally, we also include a comparison to the *Group DRO* method for subgroup robustness without distillation (Algorithm 1 in Sagawa et al. [2020a]). This method differs from our DRO procedure in that they do not apply a margin-based loss.

Tables 3.2 and 3.3 shows the average and worst-class accuracies on test for these baselines compared to the combination of α^t, α^s selected using the selection criteria previously described. The selection criteria for α^t, α^s are applied over the validation set, and thus do not directly

Table 3.2: Comparison to baselines on balanced datasets for the selected α^t, α^s combination on test data.

	CIFAR-10		**CIFAR-100**	
Method	Average acc.	Worst-1 acc.	Average acc.	Worst-1 acc.
Selected α^t, α^s combo	$\mathbf{94.28} \pm 0.06$	$\mathbf{90.11} \pm 0.23$	73.22 ± 0.26	$\mathbf{48.40} \pm 1.47$
Standard distillation	$\mathbf{94.34} \pm 0.07$	87.66 ± 0.40	$\mathbf{74.61} \pm 0.15$	43.81 ± 0.58
Post shift [NM'21]	92.16 ± 0.18	88.60 ± 0.35	61.22 ± 0.36	38.19 ± 0.40
Robust student [NM'21]	92.72 ± 0.05	89.90 ± 0.21	68.45 ± 0.13	43.62 ± 1.27
AdaMargin [LBMK'22]	93.69 ± 0.06	88.42 ± 0.36	73.58 ± 0.11	43.91 ± 1.11
AdaAlpha [LBMK'22]	$\mathbf{94.31} \pm 0.01$	88.33 ± 0.14	74.15 ± 0.08	45.46 ± 0.67
Group DRO [SKHL'20]	92.34 ± 0.07	89.32 ± 0.21	65.18 ± 0.08	43.89 ± 1.12

	TinyImageNet	
Method	Average acc.	Worst-1 acc.
Selected α^t, α^s combo	$\mathbf{58.09} \pm 0.13$	9.47 ± 1.76
Standard distillation	57.83 ± 0.13	6.32 ± 2.31
Post shift [NM'21]	43.02 ± 0.79	14.39 ± 1.13
Robust student [NM'21]	48.06 ± 0.24	$\mathbf{16.27} \pm 0.43$
AdaMargin [LBMK'22]	52.45 ± 0.08	$\mathbf{15.41} \pm 0.71$
AdaAlpha [LBMK'22]	57.22 ± 0.08	7.62 ± 2.17
Group DRO [SKHL'20]	48.78 ± 0.21	11.38 ± 1.79

translate to test performance: the selected α^t, α^s combination sometimes has lower average test accuracy than standard distillation. Still, overall, the selected α^t, α^s combination is Pareto efficient compared to all other baselines (dominant in at least one of average accuracy or worst-k accuracy). Among the rest of the different α^t, α^s candidates (as in Figure 3.1), there actually exist combinations that Pareto dominate all baselines in test performance (additional plots in Appendix B.6). While we only show results from our simple example selection criteria in Tables 3.2 and 3.3, this suggests that there is room for alternative selection criteria to yield even better results. The challenge, as with all hyperparameter selection, is that selection on the validation set comes with a generalization gap between validation and test.

3.7 Theoretical Analysis

Complementing our empirical findings, our theoretical analysis explores what constitutes a good teacher and how it aids a student in achieving robustness. To simplify our exposition, we present our theoretical analysis for a student trained using Algorithm 2 to yield good worst-class performance. Our results easily extend to the case where the student seeks to

Table 3.3: Comparison to baselines on long-tailed datasets for the selected α^t, α^s combination on test data.

	CIFAR-10-LT		CIFAR-100-LT	
Method	Average acc.	Worst-1 acc.	Average acc.	Worst-1 acc.
Selected α^t, α^s combo	79.02 ± 0.08	$\mathbf{75.43} \pm 0.39$	43.94 ± 0.16	$\mathbf{14.52} \pm 0.68$
Standard distillation	77.39 ± 0.10	60.12 ± 0.56	$\mathbf{46.01} \pm 0.16$	0.00 ± 0.00
Post shift [NM'21]	78.28 ± 0.05	74.33 ± 0.09	29.88 ± 0.61	10.01 ± 0.72
Robust student [NM'21]	80.05 ± 0.13	74.91 ± 0.24	30.79 ± 0.18	12.28 ± 0.46
Bal. student [MJRJVK'21]	$\mathbf{81.36} \pm 0.14$	71.60 ± 0.38	50.40 ± 0.12	4.39 ± 0.66
AdaMargin [LBMK'22]	72.69 ± 0.24	47.52 ± 0.95	31.26 ± 0.21	0.00 ± 0.00
AdaAlpha [LBMK'22]	70.83 ± 0.28	43.64 ± 1.09	42.52 ± 0.08	0.00 ± 0.00
Group DRO [SKHL'20]	74.39 ± 0.17	59.93 ± 0.59	40.47 ± 0.17	0.19 ± 0.17

	TinyImageNet-LT	
Method	Average acc.	Worst-10 acc.
Selected α^t, α^s combo	26.91 ± 0.16	$\mathbf{6.04} \pm 0.25$
Standard distillation	26.05 ± 0.18	0.00 ± 0.00
Post shift [NM'21]	21.32 ± 0.49	2.58 ± 0.42
Robust student [NM'21]	21.59 ± 0.19	1.55 ± 0.37
Bal. student [MJRJVK'21]	$\mathbf{30.43} \pm 0.06$	0.20 ± 0.18
AdaMargin [LBMK'22]	4.41 ± 0.09	0.00 ± 0.00
AdaAlpha [LBMK'22]	27.95 ± 0.14	0.00 ± 0.00
Group DRO [SKHL'20]	27.78 ± 0.13	0.00 ± 0.00

trade-off between average and worst-case performance.

What constitutes a good teacher?

We first characterize the properties of a good teacher when the student's goal is to minimize the robust population objective $L^{\text{rob}}(f^s)$ in equation (3.3). In particular, does the student's ability to perform well on this worst-case objective depend on the teacher also performing well on the same objective? Given scores from a teacher p^t, the student minimizes the robust distillation objective $\hat{L}^{\text{rob-d}}(f^s)$ in equation (3.7), and uses this as a proxy for the actual objective $L^{\text{rob}}(f^s)$ we care about. Intuitively, an *ideal* teacher would then be one that provides a good proxy for the student, and ensures that the difference $|\hat{L}^{\text{rob-d}}(f^s) - L^{\text{rob}}(f^s)|$ is as small as possible. Below, we provide a simple bound on this difference:

Theorem 3. *Suppose $\ell(y, z) \leq B, \forall x \in \mathcal{X}$ for some $B > 0$. Let $\pi^t_y = \mathbb{E}_x\left[p^t_y(x)\right]$, and let the following denote the per-class expected and empirical student losses respectively:*

$\phi_y(f^s) = \frac{1}{\pi^t_y}\mathbb{E}_x\left[p^t_y(x)\,\ell\left(y, f^s(x)\right)\right]$;
$\hat{\phi}_y(f^s) = \frac{1}{\hat{\pi}^t_y}\frac{1}{n}\sum_{i=1}^n p^t_y(x_i)\,\ell\left(y, f^s(x_i)\right)$.
Then for teacher p^t and student f^s:

$$\left|\hat{L}^{\text{rob-d}}(f^s) - L^{\text{rob}}(f^s)\right| \leq \underbrace{B\max_{y\in[m]}\mathbb{E}_x\left[\left|\frac{p^t_y(x)}{\pi^t_y} - \frac{\eta_y(x)}{\pi_y}\right|\right]}_{\textit{Calibration error}} + \underbrace{\max_{y\in[m]}\left|\phi_y(f^s) - \hat{\phi}_y(f^s)\right|}_{\textit{Estimation error}}.$$

The *calibration error* captures how well the teacher's predictions mimic the conditional-class distribution $\eta(x) \in \Delta_m$, up to per-class normalizations π. This suggests that even if p^t does not achieve good worst-class performance, as long as it is *well-calibrated* within each class (as measured by the calibration error), it will serve as a good teacher.

The *estimation error* captures how well the teacher aids in the student's out-of-sample generalization. The prior work by Menon et al. [2021a] study this question in detail for the standard student objective, and provide a bound that depends on the variance induced by the teacher's predictions on the student's objective: the lower the variance, the better the student's generalization. In Appendix B.2, we carry out a similar analysis with the estimation error in the theorem.

Calibration and worst-case error

We illustrate how, perhaps counterintuitively, a teacher with low worst-class accuracy might still have scores p^t that are well calibrated to match the true conditional-class distributions η. For this, we use a hypothetical "image classification" task with labels $y \in \{\text{cat}, \text{panda}, \text{other}\}$, and a single one-dimensional feature $x \in [0, 1]$ representing the fraction of black pixels in the image, uniformly distributed over the interval. Suppose the solid lines in Figure 3.2 below give the conditional-class distributions $\eta_y(x)$ for the cat and panda classes (pandas are rarer than cats in the dataset, with $\pi_{\text{cat}} = \frac{1}{2}$ and $\pi_{\text{panda}} = \frac{1}{4}$). Suppose the dashed lines in Figure 3.2 also give hypothetical teacher model scores $p^t_y(x)$, where $p^t_{\text{cat}}(x) = 2\eta_{\text{cat}}(x)$, and $p^t_{\text{panda}}(x) = \frac{1}{2}\eta_{\text{panda}}(x)$ (these arbitrary teacher scores do not necessarily correspond to softmax outputs from a neural network). This teacher model always outputs a higher score for the cat label than the panda label. However, the model still satisfies the necessary calibration property: $\frac{p^t_y(x)}{E_x[p^t_y(x)]} = \frac{\eta_y(x)}{\pi_y}$ for $y \in \{\text{cat}, \text{panda}\}$, despite the fact that the argmax predictions from this model has zero recall for the panda class. This illustrates that the important property of the teacher's scores is how well they mimic the *shape* of the conditional-class distributions, and not necessarily their worst-class predictive accuracy.

Relation to Bayes-optimal scorers

When the teacher outputs the conditional-class probabilities, i.e. $p^t(x) = \eta(x)$, the calibration error is trivially zero (recall that the normalization term $\pi^t_y = \pi_y$ in this case). Theorem

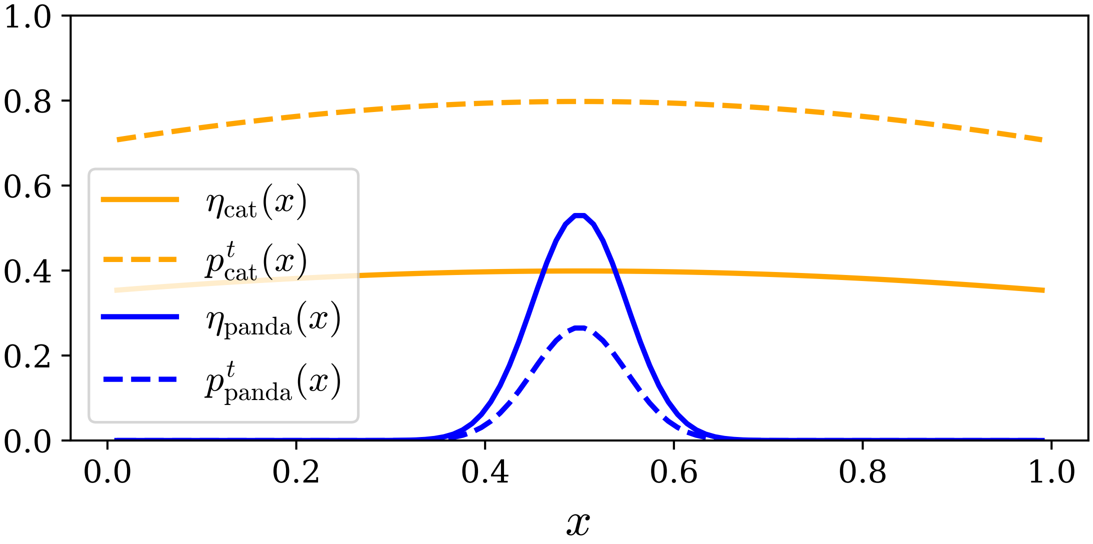


Figure 3.2: Hypothetical conditional-class distributions $\eta_y(x)$ and trained model scores $p^t_y(x)$ for $y \in \{\text{cat}, \text{panda}\}$.

2 shows that the Bayes-optimal scorer for the standard cross-entropy loss achieves this; however, in practice with finite data and model class limitations, a teacher trained with the cross-entropy loss is often far from approximating $\eta(x)$ exactly. In practice, it remains an open question what methodology might produce a teacher that most closely mimics these conditional-class distribution shapes for all classes in finite samples. For example, while the standard cross-entropy objective might lead to well calibrated model scores for a majority class, the scores may not match for rare classes. Our experiments explored training with different losses from Section 3.3 that encourage the teacher to approximate scaled versions of $\eta(x)$; however, future exploration of other practical training possibilities would be interesting to compare.

Robustness guarantee for the student

We next provide robustness guarantees for the student output by Algorithm 2 in terms of the calibration and estimation errors described above. We do so for a fixed teacher p^t, and a *self-distillation* setup where the student is chosen from the same function class $\mathcal{F}$ as the teacher, and can thus exactly mimic the teacher's predictions.

Proposition 1. *Suppose $p^t \in \mathcal{F}$ and $\mathcal{F}$ is closed under linear transformations. Let $\overline{\lambda}_y = (\prod_{k=1}^{K} \lambda^k_y / \pi^t_y)^{1/K}, \forall y$. Then the scoring function $\overline{f}^s(x) = \frac{1}{K}\sum_{k=1}^{K} f^k(x)$ output by Alg. 2 is of the form: $\text{softmax}_j(\overline{f}^s(x)) \propto \overline{\lambda}_j p^t_j(x),\ \forall j \in [m], \forall (x, y) \in S$.*

Theorem 4. *Suppose $p^t \in \mathcal{F}$ and $\mathcal{F}$ is closed under linear transformations. Suppose ℓ is the cross-entropy loss ℓ^{xent}, $\ell(y, z) \leq B$ and $\max_{y \in [m]} \frac{1}{\pi^t_y} \leq Z$, for some $B, Z > 0$. Furthermore, suppose for any $\delta \in (0, 1)$, the following bound holds on the estimation error in Theorem 3: with probability at least $1 - \delta$ (over draw of $S \sim D^n$), $\forall f \in \mathcal{F}$, $\max_{y \in [m]} \left|\phi_y(f) - \hat{\phi}_y(f)\right| \leq \Delta(n, \delta)$, for some $\Delta(n, \delta) \in \mathbb{R}_+$ that is increasing in $1/\delta$, and goes to 0 as $n \to \infty$. Then when the step size $\gamma = \frac{1}{2BZ}\sqrt{\frac{\log(m)}{K}}$ and $n^{\text{val}} \geq 8Z \log(2m/\delta)$, we have that with probability at least $1 - \delta$ (over draw of $S \sim D^n$ and $S^{\text{val}} \sim D^{n^{\text{val}}}$),*

$$L^{\text{rob}}(\bar{f}^s) \leq \min_{f \in \mathcal{F}} L^{\text{rob}}(f) + \underbrace{2\Delta(n^{\text{val}}, \delta/2) + 2\Delta(n, \delta/2)}_{\textit{Estimation error}}$$
$$+ \underbrace{2B \max_{y \in [m]} \mathbb{E}_x\left[\left|\frac{p^t_y(x)}{\pi^t_y} - \frac{\eta_y(x)}{\pi_y}\right|\right]}_{\textit{Calibration error}} + \underbrace{4BZ\sqrt{\frac{\log(m)}{K}}}_{\textit{EG convergence}}.$$

Proposition 1 shows the student not only learns to mimic the teacher on the training set, but improves upon it by making per-class adjustments to its predictions. Theorem 4 shows that these adjustments are chosen to close-in on the gap to the optimal robust scorer in $\mathcal{F}$. However, the student's convergence to the optimal scorer in $\mathcal{F}$ would still be limited by the teacher's calibration error: even when the sample sizes and number of iterations $n, n^{\text{val}}, K \to \infty$, the student's optimality gap may still be non-zero when the teacher is poorly calibrated.

3.8 Conclusions

We have demonstrated the value of applying different combinations of teacher/student objectives, not only for improving worst-class accuracy, but also to achieve efficient trade-offs between average and worst-class accuracy. Surprisingly, the teacher and students' objective functions can interact with each other in nontrivial ways: for example, applying a robust objective to both the teacher and the student does not always achieve the best worst-class accuracy (Table 1). Further exploring the trade-off between worst-class and average accuracy, we provided simple modifications to the teacher and student objectives that boosted worst-class accuracy with less degradation in average accuracy than prior methods that focus on worst-class accuracy. This confirms the key takeaway that the teacher's objective plays a crucial role in the student's robustness.

In a broader sense, our theory provides better understanding of the interplay between teacher and student objectives, and thus serves as a starting point for further development of methods to modify both the teacher and students' objectives jointly. An interesting future avenue for exploration would be to extend our distillation setup to incorporate other forms of teacher supervision such as intermediate embeddings or ensembled scores (e.g., Iscen et al. [2021]).

Training efficiency is another avenue for improvement, and future work in reducing the hyperparameter search space would be practically valuable. For settings where teacher retraining is particularly expensive, one could modify a given fixed teacher with some form of post-hoc logit adjustment [Narasimhan and Menon, 2021], or only fine-tune a subset of the teacher parameters with different values of α^t. These reductions in computational cost would improve the practicality of joint exploration of teacher and student objectives.

Chapter 4

Fairness Meets Interpretability with Monotonicity Shape Constraints

4.1 Introduction

As the use of machine-learned (ML) models broadens, strategies are sought to ensure machine-learned systems behave *responsibly*, and are *ethical* and *fair*. There may be many different reasonable and conflicting ethical stances for a given problem [Haidt, 2013, Singer, 2016, Gray and Graham, 2018, Corbett-Davies and Goel, 2018, Binns, 2018]. Thus, no single strategy for making machine learning fairer is likely to be sufficient.

In this paper, we show that nonlinear machine learning can easily produce trained models that violate social norms or ethics that can be described as: *certain inputs should not have a negative effect on an outcome*. For example, the toy example in Fig. 4.1 shows a nonlinear 1-d model trained to predict how highly a resume will be scored based on the candidate's years of experience. The best fit model (purple dashed line) sometimes penalizes candidates for having *more* job experience. For candidates with many years of experience, the dashed model picked up the age discrimination in the biased training samples. In addition, the dashed model also penalizes job experience in the lower-end of years of experience simply due to overfitting. Such unfair responses are very much a danger in sparse regions of a feature space with modern over-parameterized nonlinear models.

Fig. 4.1 also illustrates our proposed solution (blue solid line): train the model with monotonicity shape constraints to guarantee that the model can only use job experience as positive evidence. This is a *deontological* solution, and thus differs from many existing mathematical expressions of fairness that are *consequentialist* and *statistical* [Friedler et al., 2019, Hardt et al., 2016b, Binns, 2018, Sandvig et al., 2016]. Both types of goals may be of interest to a practitioner [Sandvig et al., 2016], and we explore how they relate theoretically and experimentally.

Contributions

The main contributions of this paper are: *(i)* identifying that violations of monotonicity in ML models may pose significant ethical issues and risk of societal harm, *(ii)* demonstrating experimentally that these violations can easily occur in practice with nonlinear machine learned models, *(iii)* showing that existing shape constrained ML can be used effectively to ameliorate such problems, and *(iv)* theoretically and empirically analyzing how the proposed monotonicity constraints relate to statistical fairness notions.

Model type	# shape	Train	Test
	constraints	Acc.	Acc.
GAM	0	94.93%	94.89%
GAM	2	94.90%	**94.97**%
DNN	0	94.97%	94.89%
GBT	0	**95.04**%	94.80%

Table 4.1: Law School Admissions Experiment Results. Two monotonicity constraints ensure that individuals aren't penalized for higher GPA or LSAT score.

4.2 Illustrating Unfair Penalization

To further illustrate the potential for machine-learning to produce objectionable models, consider the *Law School Admissions* dataset [Wightman, 1998]. Suppose this data is used to predict whether a person would pass the bar exam based on their LSAT score and undergraduate GPA, and that the classifier's score was used to guide law school admissions or scholarships. We trained a standard two-layer deep neural network (DNN) (more experimental details in Sec. 4.7) and show the model's output for each possible input in Fig. 4.2. The DNN sometimes penalizes people for having a higher GPA: for example, with an LSAT score

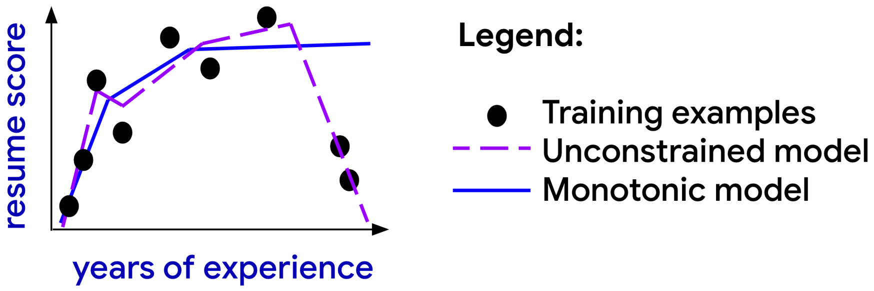


Figure 4.1: Toy example showing how a monotonicity constraint can protect against unfair penalization.

of 15, the DNN rewards students with a lower 2.7 GPA over students who earned a higher 3.5 GPA. Similarly, if a student has a GPA of 2.5, the DNN gives that student a higher score for scoring 10 on the LSAT than if they had scored 15. Thus this model violates merit-based social norms and the "best-qualified" ethical principle [Hunter and Schmidt, 1976] (we also acknowledge that the very use of standardized test scores or GPA for allocating goods may raise other ethical issues [Hunter and Schmidt, 1976]). Training with gradient boosted trees has the same problem: the model penalizes some people for raising their GPA or LSAT score.

Each of the two-dimensional models shown in Fig. 4.2 were trained on 19,064 training examples, which may sound like plenty of data to learn a model on two inputs, but the non-uniform distribution of the training data means that some regions of the feature space were sparse and the model may have overfit (for a density plot, see Fig. C.4 in the Appendix). These monotonicity violations can also occur from legitimate, clean data. To see this, note that how hard it is to get a good GPA can vary greatly between schools, and imagine as an extreme example that there was a large university that simply gave every student a 2.7 GPA, which could cause some of the problems seen in this model.

Our proposal is to train the model with monotonicity constraints. An example monotonic model is shown in *(c)* of Fig. 4.2. It is a generalized additive model (GAM) trained with the constraints that it never penalizes higher GPA's for any LSAT score, and that it never penalizes higher LSAT scores for any GPA. Training the same GAM *without* monotonicity constraints is shown in the lower right, and also produces an objectionable model. The test accuracies of these four models are given in Table 4.1 are similar (more experimental details in Sec. 4.7).

4.3 Monotonicity Fairness Constraints

The motivating examples in Figures 1 and 2 focused on the concern of *unfair penalization* – that there may be inputs that a responsible model may reward but should never penalize. A second ethical pattern we consider is to *favor the less fortunate* – there may be inputs that help us identify the less fortunate and favor them if there are no other relevant differences. Policies favoring the less fortunate have been well studied in economics [Coate and Loury, 1993]. As shown in Figures 1 and 2, we propose addressing these two principles by constraining the ML model to only respond positively to relevant inputs if all other inputs are fixed.

Throughout, we focus on ML models that produce a score that is used to determine some benefit, such as a better credit rating or a scholarship. Specifically, consider a learned model $f(x, z)$ where $x \in \mathbb{R}^D$ is a D-dimensional feature vector, and $z \in \mathbb{R}$ is another input that is of ethical interest (such as age or income). The model $f(x, z)$ satisfies a *positive monotonicity shape constraint* [Groeneboom and Jongbloed, 2014] with respect to z if for any $\delta > 0$ and any choice of x and z, $f(x, z + \delta) \geq f(x, z)$. If f is differentiable with respect to z, this constraint is equivalent to non-negative slope $\frac{\partial(f(x,z))}{\partial z} \geq 0$ for any x, z. *Strict monotonicity* replaces the $\geq$ sign above with $>$. Reverse definitions produce negative monotonicity constraints. To

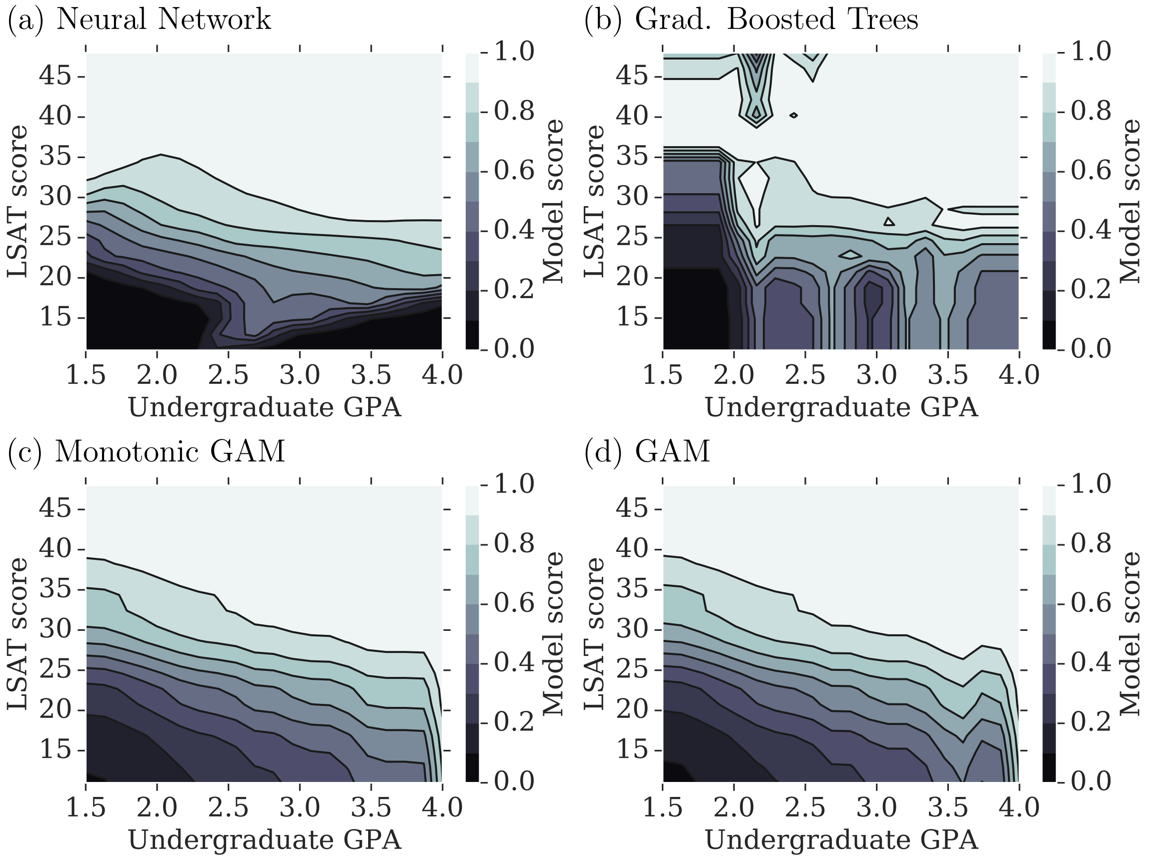


Figure 4.2: Real data example predicting if a student will bass the bar. Predictions by a neural network *(a)* and gradient boosted trees *(b)* penalize some students if they increase their GPA or LSAT score. We trained a generalized additive model (GAM) with constraints to be monotonically increasing in both GPA and LSAT score *(c)*. The same GAM model trained *without* monotonicity constraints also violates monotonicity in GPA *(d)*.

apply a monotonicity shape constraint to a categorical feature, one can express each category as a Boolean feature indicating membership in the category.

As defined above, and as is standard in the shape constraints literature [Groeneboom and Jongbloed, 2014], our proposed monotonicity shape restrictions are *ceterus paribus*: that is, $f(x)$ should behave monotonically with respect to increases in each protected feature z, but only *when all other features x are held fixed.* Here we take the input features x as given, but of course it is also important to have the best possible set of features x.

4.4 More Example Scenarios

We further illustrate the critical importance and breadth of situations in which we need ML models to behave consistently with a society's norms or ethical policies. Then in Section 4.7, we show experimentally on two more problems that nonlinear ML models do not naturally pick up such norms and policies, and that training with monotonicity constraints can be used to incorporate desired policies.

Crimes and misdemeanors. An ML system trained to determine fines for misdemeanors would likely be considered fairer if the fines were monotonically increasing with respect to the magnitude of the illegality, e.g. a larger fine for illegally parking one's car for longer, or for exceeding the speed limit by more [Allen et al., 2015]. Also, in many societies it is expected that a juvenile will not be penalized more heavily than an adult for the same crime, all else equal [Allen et al., 2015]. Many societies prefer not to give harsher penalties to first-time offenders than to repeat offenders [Allen et al., 2015].

Pay. People generally feel it is unfair to get paid less for doing more of the same work [Fehr and Schmidt, 1999]. Suppose a model is trained to advise parents on how much to pay their babysitter. It may be desirable to workers if the recommended pay were constrained to be a monotonically increasing function of the number of hours worked, all else equal. Similarly, consider an app that uses an ML model to advise people on how much to tip their waiter in America. Such a model would be better aligned with American societal norms if it recommended larger tips for more expensive meals, and if it recommended higher tip percentages for more upscale establishments, all else equal [Azar, 2004].

Medical triage. In some medical contexts, it is considered more ethical or more responsible to prioritize patients based on their risk or neediness [Iserson and Moksop, 2007]. For example, the United States Transplant Board has a policy of giving a sicker person a higher score to receive a transplant for ethical reasons [L. Bernstein, 2017]. Similarly an emergency room scoring patients for prioritization may require that patients that have waited longer are treated first, if all other relevant characteristics are equal [Iserson and Moksop, 2007]. More generally, *first-come first-served* is a common principle underpinning civil society [Beauchamp and Childress, 2001, Sugden, 1989].

4.5 Related Work

We outline two main categories of related work. First, we discuss literature in the intersection of ethics and ML. Then, we give context on training ML models with monotonicity shape constraints.

Ethics and ML

This work fits into the broader literature on *machine ethics* [Anderson and Anderson, 2007, Moor, 2006]. Ethics itself is a broad field of philosophy that encompasses many questions, and there may be many different reasonable ethical stances and criteria for a given problem [Binns, 2018, Thiroux and Krasemann, 2017, Haidt, 2013, Gray and Graham, 2018, Corbett-Davies and Goel, 2018, Singer, 2016, Gruetzemacher, 2018, Allen et al., 2015, Rawls, 1971].

Deontological ethics. Most recent work in machine learning fairness has been, in a broad sense, *consequentialist*: focusing on ensuring that machine-learned models deliver statistically-similar performance for different groups [Binns, 2018, Zafar et al., 2015, Goh et al., 2016, Hardt et al., 2016b, Zafar et al., 2017, Donini et al., 2018, Agarwal et al., 2018, Goel et al., 2018, Cotter et al., 2019a,c]. Those efforts solve a different important problem than the one we target here, and are complementary to this proposal. This proposal is instead *deontological*, in that it enables designers to impose rules on how the model can respond to inputs, producing an *implicit ethical agent* in the terminology of Moor [2006]. Sandvig et al. [2016] recently called for more research into deontological algorithms, noting that, "Applied ethics in real-world settings typically incorporates both rule-based and consequences-based reasoning."

Monotonicity and fairness. Concurrently, Cole and Williamson [2019] also recognized the importance of monotonicity in a fairness context. That work differs in framing: they use monotonicity to reduce unfair *resentment*. Our work further differs in its theoretical results and comparisons.

Individual fairness. Another fairness principle is that *similar individuals should receive similar treatment* [Dwork et al., 2012]. Individual fairness aims for equal outcomes for two examples, whereas this proposal allows for an asymmetric treatment whereby any unequal treatment is unequal in the appropriate direction, such as *favoring the less fortunate*.

Counterfactual fairness. Counterfactual fairness [Kusner et al., 2017, Pearl et al., 2016] says that changing a protected attribute A while holding things not causally dependent on A constant will not change the distribution of the model output. This is similar to the definition of monotonicity in section 4.3, but is focused on treating certain cases the *same* rather than prefering one case to another.

Continuous sensitive features. The proposed monotonicity constraints handles real-valued attributes natively. Other recent efforts have also been devised to handle continuous protected attributes, like age, for consequentialist fairness goals [Raff et al., 2018, Kearns et al., 2018, Komiyama et al., 2018].

Fair ranking and fair regression. Monotonicity constraints can be applied to ranking and regression models. Other notions of ethics have also been considered for ranking models [Berk et al., 2017a, Zehlike et al., 2017, Celis et al., 2018, Beutel et al., 2019, Singh and Joachims, 2018] and regression [Komiyama et al., 2018, Pérez-Suay et al., 2017, Berk et al., 2017a, Agarwal et al., 2019].

Training models with monotonicity constraints

Monotonicity shape constraints have long been used to capture prior knowledge and regularize estimation problems to improve a model's generalization to new test examples [Barlow et al., 1972, Chetverikov et al., 2018, Ben-David, 1992, Archer and Wang, 1993, Sill and Abu-Mostafa, 1997, Kotlowski and Slowinski, 2009, Groeneboom and Jongbloed, 2014, Gupta et al., 2016, Canini et al., 2016, You et al., 2017, Bonakdarpour et al., 2018]. In this paper, we point out that monotonicity shape constraints can and should *also* be used to ensure that machine-learned models behave consistently with societal norms and prima facie duties [Ross, 2002] that can be expressed as monotonic relationships.

Unlike its use for regularization, applying monotonicity constraints to impose ethical principles may actually hurt test accuracy if the training and test data is biased (as in the toy example of Fig. 1). However, in all three of our real-data experiments the test accuracy was little changed by adding these constraints (see Section 4.7 and Fig. 2 and Appendix C.4).

Constraining multi-dimensional models to obey monotonic shape constraints has been shown to work for a variety of function classes, including neural networks [Archer and Wang, 1993, Sill and Abu-Mostafa, 1997] and trees [Ben-David, 1992, Kotlowski and Slowinski, 2009, Bonakdarpour et al., 2018].

Our experiments use the open-source TensorFlow Lattice 2.0 package [Google AI Blog, 2020], which enables training GAMs and lattice models with monotonicity constraints [Gupta et al., 2016, Canini et al., 2016, You et al., 2017].

4.6 Relationship to Statistical Fairness

In this section we analyze how the proposed deontological monotonicity constraints interact with consequentialist statistical fairness goals that are based on aggregate outcomes. For example, suppose a national funding agency enforces that poorer schools are funded more often as richer schools on average (we will call this *one-sided statistical parity*). By only considering the national average funding rates, richer schools may still get funded more often than poorer schools in some states. In contrast, if the state were the only input X, then a deonotological monotonicity constraint would guarantee that for each state, poorer schools would get a higher funding rate than richer schools.

Hardt et al. [2016b] showed that any *oblivious* statistical fairness measure that doesn't depend on X or the function form $f(X, Z)$ can fail to identify forms of discrimination. Popular statistical notions like *statistical parity* [Dwork et al., 2012] and *equal opportunity*

[Hardt et al., 2016b] are oblivious, whereas monotonicity is not. In the next section we show specifically that satisfying *one-sided statistical parity* does not imply monotonicity, but that satisfying monotonicity can impact and bound one-sided statistical parity and equal opportunity (proofs in Appendix).

Bounds on one-sided statistical parity

Statistical parity is a well known measure of fairness for ML models [Dwork et al., 2012, Zafar et al., 2015, Cotter et al., 2019c]. Consider a model $f(X, Z)$ that takes as input a random feature vector $X \in \mathbb{R}^D$ and a random protected attribute Z, which may be categorical or real-valued. Suppose the model f outputs a real-valued score (we treat the special case of classifiers next). Then *statistical parity* requires $E[f(X, Z)|Z = j] = E[f(X, Z)|Z = k]$ for any j, k. Because we focus on asymmetric goals like favoring the less fortunate, we consider *one-sided* statistical parity: $E[f(X, Z)|Z = j] \leq E[f(X, Z)|Z = k]$ for any $j \leq k$.

We show in Appendix C.2 that due to Simpson's paradox [Bickel et al., 1975], that a monotonicity constraint on Z is not sufficient to guarantee one-sided statistical parity with respect to Z. However, monotonicity does imply a *bound* on the one-sided statistical parity violation: in Lemma 2, we show that if f is monotonic with respect to Z, then the one-sided statistical parity violation between $Z = j$ and $Z = k$ will be bounded by the maximum density ratio between the two groups.

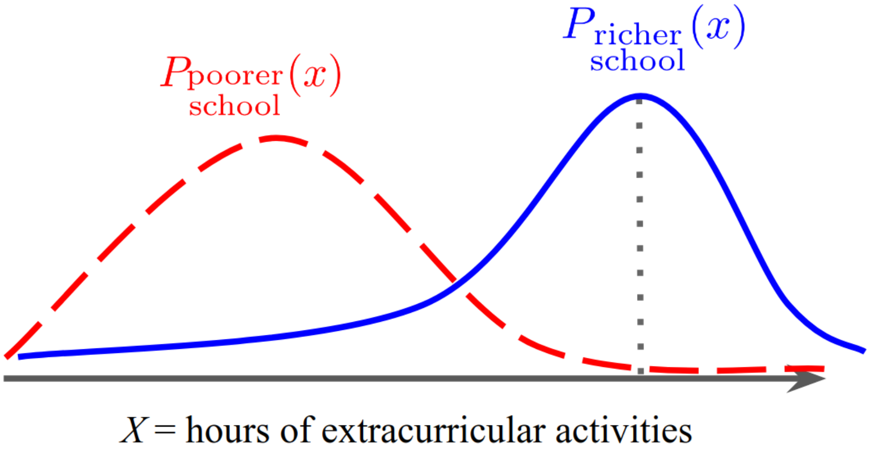


Figure 4.3: Illustration of Lemma 1. The dotted line shows the X value achieving the max density ratio C.

Lemma 2. Let $(\Omega, \mathcal{F})$ be a measurable space with a regular conditional probability property, and let $X : \Omega \to \mathbb{R}^D$, $Z : \Omega \to \mathbb{R}$ be $\mathcal{F}$-measurable random variables. Suppose P_j and P_k are σ-finite probability measures on $(\Omega, \mathcal{F})$, where P_j denotes the conditional probability measure of X given that $Z = j$, and P_k denote the same for $Z = k$, and P_j is absolutely continuous with respect to P_k. Let $f : \mathbb{R}^D \times \mathbb{R} \to \mathbb{R}$ be defined as in Section 4.3, and $f(x, z) \geq 0$ for all $x \in \mathbb{R}^D$, $z \in \mathbb{R}$. If the function f satisfies monotonicity in the second argument such that for $j \leq k$, $f(x, j) \leq f(x, k)$ for all $x \in \mathbb{R}^D$, and if the Radon-Nikodym derivative $\frac{dP_j}{dP_k}$ is bounded almost everywhere with respect to P_k by a finite constant $C > 0$, then

$$E[f(X, Z)|Z = j] \leq CE[f(X, Z)|Z = k]. \tag{4.1}$$

If the conditional probability distribution of X given $Z = j$ has density $p_{X|Z=j}(x)$, and X given $Z = k$ has density $p_{X|Z=k}(x)$, then Lemma 2 says that if the likelihood ratio $\frac{p_{X|Z=j}(x)}{p_{X|Z=k}(x)} \leq C$ almost everywhere with respect to P_k for finite $C > 0$, then the one-sided statistical parity bound (4.1) holds.

Fig. 4.3 illustrates Lemma 2 with an example: suppose Z is categorical and denotes either a poorer or richer high school, and suppose $D = 1$ with $X \in [0, 20]$ being the number of hours of extracurricular activities a student does each week. Let $f(X, Z)$ be a score used to determine a student's admission to some college. Then if f is monotonic in Z such that f never gives a lower admissions score to a poorer student if their X value is the same as a richer student, then Lemma 1 says that the average score for richer students will be no more than C times the average score for poorer students, where C is given by the maximum ratio of the two distributions over X.

While Lemma 1 shows that monotonicity implies a bound on the one-sided statistical parity violations, the converse does not hold: a model satisfying statistical parity can have arbitrarily high monotonicity violations (proof in Appendix C.2). This can be ethically problematic if overlooked by practitioners.

While Lemma 2 provides a *worst case bound* on the statistical parity violations of monotonic functions, we next ask, can imposing monotonicity ever make statistical parity violations worse? Lemma 3 shows that for any model f, the monotonic projection of f cannot have worse statistical parity violations *on average*.

Lemma 3. Let $f : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$, where $\mathcal{X} \subseteq \mathbb{R}^D$, $\mathcal{Z} \subseteq \mathbb{R}$. Assume that $\mathcal{X}, \mathcal{Z}$ are both finite, with $X \in \mathcal{X}$, $Z \in \mathcal{Z}$. Let $\tilde{f}$ be the projection of f onto the set of functions over $\mathcal{X} \times \mathcal{Z}$ that are monotonic with respect to Z such that for $j \leq k$, $f(x, j) \leq f(x, k)$. For $z_{(i)} \in \mathcal{Z}$, let $z_{(1)} \leq z_{(2)} \leq ... \leq z_{(|\mathcal{Z}|)}$. Define the average statistical parity violation:

$$R_f \triangleq \sum_{i=1}^{|\mathcal{Z}|} \frac{E[f(X, Z)|Z = z_{(i)}] - E[f(X, Z)|Z = z_{(i+1)}]}{|\mathcal{Z}|}$$

Then $R_{\tilde{f}} \leq R_f$.

When $|\mathcal{Z}| = 2$, Lemma 3 also bounds the worst case violation of $\tilde{f}$. However, for $|\mathcal{Z}| > 2$, there is no such worst case guarantee, and there may be pairs j, k where $\tilde{f}$ has a worse one-sided statistical parity violation than f (proof in Appendix C.2).

Bounds for binary classifiers

In binary classification, an example has an associated true label $Y \in \{0, 1\}$, and the model outputs a binary decision $\hat{Y} \in \{0, 1\}$. By definition, a monotonicity shape constraint on Z implies a one-sided bound on the conditional probabilities: for any $j \leq k$ and for all x, $P(\hat{Y} = 1|X = x, Z = j) \leq P(\hat{Y} = 1|X = x, Z = k)$.

Because the label Y can be modeled as a Bernoulli random variable, the goal of statistical parity is equivalent to *marginal independence*, that is, for any j, k, $P(\hat{Y} = 1|Z = j) = P(\hat{Y} = 1|Z = k)$. Correspondingly, the bound (4.1) on the statistical parity violation becomes a bound on the marginal probabilities: $P(\hat{Y} = 1|Z = j) \leq CP(\hat{Y} = 1|Z = k)$.

For binary classifiers, we can give a more explicit bound:

Lemma 4. Suppose X is a continuous (or with a straightforward extension, discrete) random variable, and let $\mathcal{S}$ be a nonempty set such that for all $x \in \mathcal{S}$, the joint probability density values $p_{X,\hat{Y}|Z=z}(x, 1) > 0$ for $z = j, k$. Suppose we have monotonicity where $f(x, j) \leq f(x, k)$ for $j \leq k$ for all $x \in \mathcal{S}$. For a binary classifier this implies $P(\hat{Y} = 1|X = x, Z = j) \leq P(\hat{Y} = 1|X = x, Z = k)$. Then we can bound one-sided statistical parity as follows:

$$\frac{P(\hat{Y} = 1|Z = j)}{P(\hat{Y} = 1|Z = k)} \leq \inf_{x \in \mathcal{S}} \frac{p_{X|Z=j}(x) p_{X|\hat{Y}=1,Z=k}(x)}{p_{X|Z=k}(x) p_{X|\hat{Y}=1,Z=j}(x)}$$

The bound in Lemma (4) contains two likelihood ratios: $\frac{p_{X|Z=j}(x)}{p_{X|Z=k}(x)}$ and $\frac{p_{X|\hat{Y}=1,Z=k}(x)}{p_{X|\hat{Y}=1,Z=j}(x)}$. The first is the same as in Lemma 2. The second is the inverse of that likelihood ratio, conditioned on $\hat{Y} = 1$. When the first likelihood ratio is low, the second inverse likelihood ratio may be high, producing a trade-off between these two ratios. We describe an example in Appendix C.3.

Similarly, for *equal opportunity* [Hardt et al., 2016b], we have Lemma 5:

Lemma 5. Let $Y \in \{0, 1\}$ be a random variable representing the target. Let $\mathcal{S}$ be a nonempty set such that for all $x \in \mathcal{S}$, the following joint probability density values are non-zero for $z = j, k$: $p_{X,Y,\hat{Y}|Z=z}(x, 1, 1) > 0$ and $p_{X,Y|\hat{Y}=1,Z=z}(x, 1) > 0$. Then,

$$\frac{P(\hat{Y} = 1|Y = 1, Z = j)}{P(\hat{Y} = 1|Y = 1, Z = k)} \leq \inf_{x \in \mathcal{S}} \frac{c_j(x)}{c_k(x)}$$

$$\text{where } c_z(x) = \frac{p_{X|Z=z}(x) P(Y = 1|\hat{Y} = 1, Z = z)}{p_{X|\hat{Y}=1,Z=z}(x) P(Y = 1|Z = z)}$$

We supplement these bounds with empirical results.

4.7 Experiments

We demonstrate experimentally with three public datasets that *(i)* nonlinear ML can violate common ethical policies or norms, *(ii)* training with monotonicity constraints can be used to impose such policies to the extent that the choice of model inputs x enables, *(iii)* test accuracy may not be hurt, *(iii)* statistical fairness violations may be reduced.

Model training details

Code for these experiments and further tutorials for using the open-source TensorFlow Lattice 2.0 library [Google AI Blog, 2020] are available at `https://github.com/tensorflow/lattice/blob/master/docs/tutorials/`. All experiments used a nonlinear generalized additive model (GAM) [Hastie and Tibshirani, 1990], also called a *calibrated linear* model in the TensorFlow Lattice, of the form $f(x) = \sum_{d=1}^{D} c_d(x[d]; \beta_d)$, where each $c_d(\cdot)$ is an one-dimensional piecewise-linear function parameterized by K interpolated (key, value) pairs where the keys are set to match the K quantiles of the training data and their corresponding values $\beta_d \in R^K$ are trained [Gupta et al., 2016]. We used a fixed $K = 20$ parameters for each of the D one-dimensional transforms, for all experiments. All DK parameters were jointly trained using projected stochastic gradient descent. For each input d, one can choose to constrain $f(x)$ to be monotonically increasing with respect to d by constraining its one-dimensional curve $c_d(x[d])$ to be monotonic [Gupta et al., 2016]. The TensorFlow Lattice package enables fitting more flexible monotonic models, but we felt it was most compelling to show that these monotonicity violations occur even with simple nonlinear models like GAMs, which are popular in the statistics literature. We randomly uniformly split each dataset into 70% training, 10% validation, and 20% test examples. We validated the learning rates using grid search by powers of 10 ensuring that the optimal learning rates did not fall on the extremes, and trained for 1000 epochs (more than sufficient for convergence).

Law School experiments

In Section 1 we partially described an experiment using the *Law School Admissions* dataset [Wightman, 1998]. The dataset has 27,234 total law students, and we use only two features in all models: GPA and LSAT scores. We plot the densities of these conditioned on whether the student passed the bar or not in Fig. C.4 in Appendix C.4. Models were trained as described above, where for the constrained model we constrained both LSAT score and undergraduate GPA to only be positive evidence, *ceterus paribus*. The two-layer neural network and gradient boosted trees models were trained with a similar number of model parameters to the GAMs and no monotonicity constraints. Resulting models were shown in Fig. 4.2, and train/test accuracies in Table 4.1.

Credit Default experiments

Next, we consider the Default of Credit Card Clients benchmark dataset from the UCI repository [Lichman, 2013, Yeh and hui Lien, 2009]. The data was collected from 30,000 Taiwanese credit card users and contains a binary label of whether or not a user defaulted on a payment in a time window. Features include marital status, gender, education, and how long a user is behind on payment of their existing bills, for each of the months of April-September 2005.

Repayment status is the integer number of months it has been since the user has repaid, and negative values mean the user has already repaid. Here we illustrate using monotonicity constraints to avoid *unfair penalization*: if the model were to be used to determine a user's credit score, it could feel unfair to many if they were penalized for paying their bills sooner, all else equal. Thus, we apply a monotonicity constraint that keeps the model from penalizing early payments.

Figure 4.4 *(top)* shows the average and standard deviation of the default rate of the training examples as a function of the months since the bills were paid. The data is very noisy for 3^+ months overdue payment (there are only 122 such training examples), and these noisy averages do not follow the reasonable principle of predicting a higher default rate if your bills are more over-due.

For ease of visualization, our first experiment uses only $D = 2$ features: marital status and April repayment status, with a monotonicity shape constraint on the repayment status to ensure that paying your bills on time doesn't hurt you. Our second experiment uses all $D = 24$ features, and we impose monotonicity shape constraints on all 6 repayment features to guarantee that the person is not penalized for paying early/on-time during any of the 6 months.

Fig. 4.5 *(top)* shows the predicted default rate for the unconstrained GAM. It mimics the average training labels, and thus unfairly rewards people who are 5-6 months overdue on their bills with a lower defaulting score than people who are only 2-3 months overdue. However, Figure 4.5 *(bottom)* shows the GAM trained with a monotonicity shape constraint: as requested, it does not penalize people for paying their bills early.

Table 4.2 shows that adding the monotoniciy constraints for the Credit Default problem had only a tiny effect on the train and test accuracy for the $D = 2$ experiments. For the $D = 24$ experiment, the test accuracy is slightly worse, but since the train accuracy was not hurt, we believe the lower test accuracy is simply due to the randomness of the sample.

Funding Proposals experiments

Next we demonstrate the use of monotonicity constraints to favor the less fortunate. We use the dataset from the KDD Cup 2014: Predicting Excitement at DonorsChoose.org [KDD Cup, 2014]. DonorsChoose.org is a platform through which teachers in K-12 schools can request funding and materials for proposed projects. Donors can search for and donate to projects. The dataset contains 619,327 examples of projects proposed by teachers. The label is a binary label that represents the outcome of the project, where the positive class means the project was deemed "exciting", a definition determined by DonorsChoose.org that includes whether the project was fully funded. Only 5.91% of examples are labeled "exciting". Figure 4.4 *(bottom)* shows the mean and std. dev. of the training examples for 2 of the 28 features: poverty level and number of students impacted. A machine learning model trained on this dataset could be used to rank projects to display to potential donors. Standard machine learning trained on just these two features would not favor schools at higher poverty levels, or projects with greater impact. We compare to training with monotonicity constraints for those two principles, both on just the shown two features, and on all 28 features (but only these two features constrained).

The results in Table 4.3 for the AUC metric (the same metric used in the KDD Cup 2014 competition) show that adding monotonicity constraints to prefer poorer schools and greater student impact did not hurt the test AUC. Figures comparing the standard and constrained $D = 2$ models are in Appendix C.5.

# features	# shape	Train	Test
	constraints	Acc.	Acc.
2	0	82.07%	81.55%
2	1	82.06%	81.60%
24	0	82.44%	82.02%
24	6	82.35%	80.86%

Table 4.2: Credit Default experiment results, where the monotonicity shape constraints ensure the model does not penalize people for paying their bills earlier.

# features	# shape	Train	Test
	constraints	AUC	AUC
2	0	0.520	0.517
2	2	0.514	0.518
28	0	0.752	0.746
28	2	0.751	0.746

Table 4.3: Funding Proposals experiment results, where the monotonicity shape constraints ensure that the model gives higher scores to projects for higher poverty schools or impacting more students, all else equal.

Comparison to statistical fairness

In Section 6 we gave *theoretical* results on how monotonicity constraints can bound and affect *one-sided statistical parity* and *one-sided equal opportunity* fairness violations. In Table 4.4 we show empirically what happens to the one-sided (1-s) statistical parity and equal opportunity violations (defined below) for the Funding Proposals experiment, where the protected groups $j, k \in \{0, 1, 2, 3\}$ are four different *poverty levels*, which we treat as four different ordinal protected groups for calculating these one-sided statistical metrics.

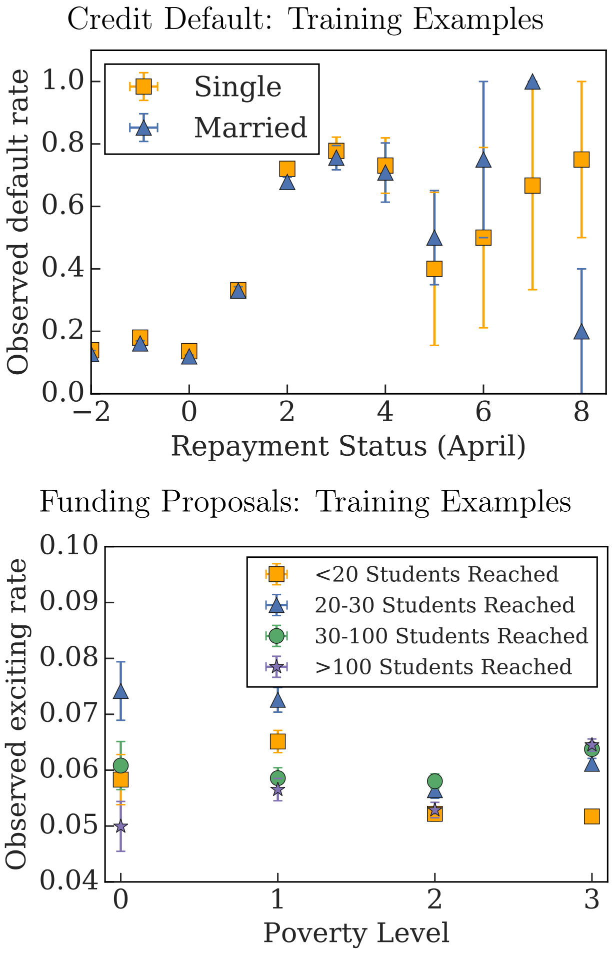


Figure 4.4: Mean and standard error of the label as a function of the inputs over the training dataset.

Max *One-sided Statistical Parity* Violation:

$$\max_{j<k}(0, P(\hat{Y}=1|Z=j) - P(\hat{Y}=1|Z=k)) \tag{4.2}$$

Max *One-sided Equal Opportunity* Violation:

$$\max_{j<k}(0,\ P(\hat{Y}=1|Z=j, Y=1) - P(\hat{Y}=1|Z=k, Y=1)) \tag{4.3}$$

Table 4.4 shows that the monotonically constrained models do lower violations of both fairness goals. The improvement is smaller when the model has $D = 28$ features than $D = 2$ features, which we believe is due to a weakening of the maximum likelihood ratios as described by Lemmas 4 and 5.

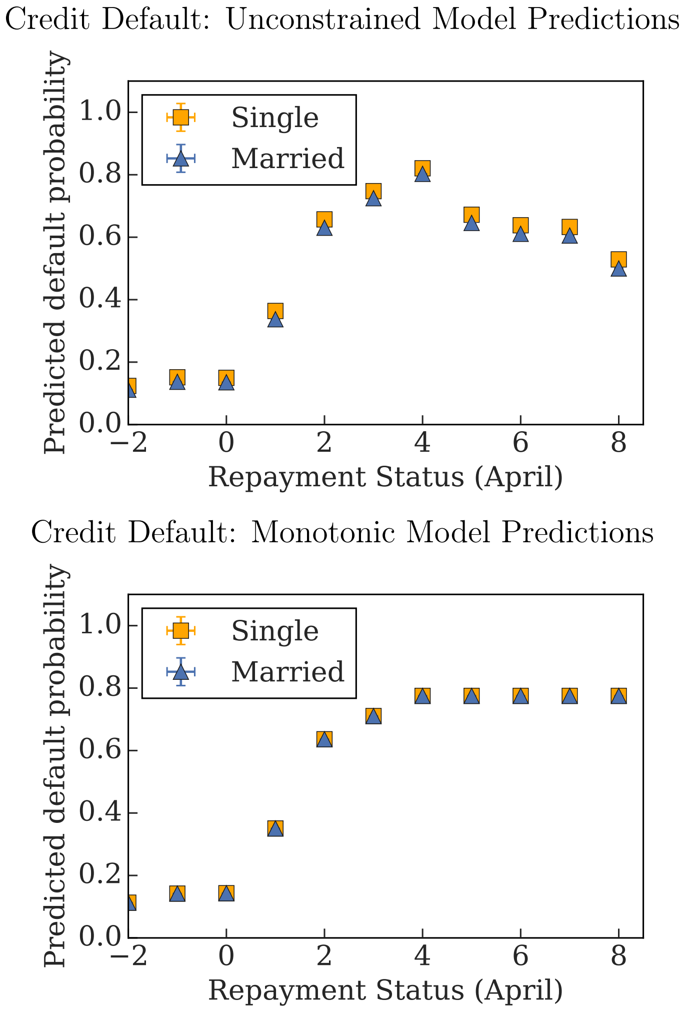


Figure 4.5: Credit Default: unconstrained model predictions *(top)* and constrained model predictions *(bottom)*.

D	# shape constraints	Max 1-s Stat. Par. Viol. (4.2)	Max 1-s Eq. Opp. Viol. (4.3)
2	0	0.00704	0.00707
2	2	0.00017	0.00037
28	0	0.00751	0.00397
28	2	0.00261	0.00331

Table 4.4: Statistical Fairness Violations for Funding Proposals Experiment.

4.8 Conclusions

We have demonstrated that nonlinear machine learned models can easily overfit noise or learn bias in a way that violates social norms or ethics about whether certain inputs should be allowed to negatively affect a score or decision. We have also shown that this problem can be ameliorated by training with monotonicity constraints to reflect the desired principle.

An advantage of monotonicity constraints is that their effect does not depend on the data distribution, so there are no questions of generalization to test data as there are with statistical fairness measures. While enforcing such constraints addresses deontological rather than consequentialist ethics, we have also shown theoretically and experimentally that monotonicity constraints can can improve or bound (consequentialist) one-sided statistical fairness violations.

Monotonicity constraints are fairly easy to explain and reason about for laypeople, as illustrated in the examples in this paper. The examples also illustrate the broad applicability of constraints to situations where the model should avoid unfair penalization of good attributes, and favor the less fortunate.

We conclude that monotonicity constraints are a necessary and useful tool for creating responsible AI, but certainly not sufficient or applicable to all situations. For this method to be completely effective, all relevant features must be identified and constrained. Non-sensitive information can be highly correlated with sensitive information, causing indirect discrimination [Hajian and Domino-Ferrer, 2013], which is also a problem for statistical fairness measures [Hardt et al., 2016b]. This method is also not directly applicable to unordered inputs like addresses, photos, or voice signals. Thus, this will be one of many tools and strategies needed to achieve responsible AI.

Appendix A

Deferred Proofs and Discussion for Chapter 2

A.1 Proofs for Section 2.4

This section provides proofs and definitions details for the theorems and lemmas presented in Section 2.4.

Proofs for TV distance

Definition 1. (TV distance) Let $c(x, y) = \mathbb{1}(x \neq y)$ be a metric, and let π be a coupling between probability distributions p and q. Define the total variation (TV) distance between two distributions p, q as

$$\begin{aligned} TV(p,q) = \inf_{\pi} \mathbb{E}_{X,Y\sim\pi}[c(X,Y)] \\ \text{s.t.} \int \pi(x,y)dy = p(x), \int \pi(x,y)dx = q(y). \end{aligned}$$

Theorem 1. *Suppose a model with parameters θ satisfies fairness criteria with respect to the noisy groups $\hat{G}$:*

$$\hat{g}_j(\theta) \leq 0 \ \ \forall j \in \mathcal{G}.$$

Suppose $|h(\theta, x_1, y_1) - h(\theta, x_2, y_2)| \leq 1$ for any $(x_1, y_1) \neq (x_2, y_2)$. If $TV(p_j, \hat{p}_j) \leq \gamma_j$ for all $j \in \mathcal{G}$, then the fairness criteria with respect to the true groups G will be satisfied within slacks γ_j for each group:

$$g_j(\theta) \leq \gamma_j \quad \forall j \in \mathcal{G}.$$

Proof. For any group label j,

$$g_j(\theta) = g_j(\theta) - \hat{g}_j(\theta) + \hat{g}_j(\theta) \leq |g_j(\theta) - \hat{g}_j(\theta)| + \hat{g}_j(\theta).$$

By Kantorovich-Rubenstein theorem (provided here as Theorem 5), we also have

$$|\hat{g}_j(\theta) - g_j(\theta)| = |\mathbb{E}_{X,Y\sim\hat{p}_j}[h(\theta, X, Y)] - \mathbb{E}_{X,Y\sim p_j}[h(\theta, X, Y)]| \leq TV(p_j, \hat{p}_j).$$

By assumption that θ satisifes fairness constraints with respect to the noisy groups $\hat{G}$, $\hat{g}_j(\theta) \leq 0$. Thus, we have the desired result that $g_j(\theta) \leq TV(p_j, \hat{p}_j) \leq \gamma_j$.

Note that if p_j and $\hat{p}_j$ are discrete, then the TV distance $TV(p_j, \hat{p}_j)$ could be very large. In that case, the bound would still hold, but would be loose. □

Theorem 5. *(Kantorovich-Rubinstein).*[1] *Call a function f Lipschitz in c if $|f(x) - f(y)| \leq c(x, y)$ for all x, y, and let $\mathcal{L}(c)$ denote the space of such functions. If c is a metric, then we have*

$$W_c(p, q) = \sup_{f \in \mathcal{L}(c)} \mathbb{E}_{X\sim p}[f(X)] - \mathbb{E}_{X\sim q}[f(X)].$$

As a special case, take $c(x, y) = \mathbb{I}(x \neq y)$ (corresponding to TV distance). Then $f \in \mathcal{L}(c)$ if and only if $|f(x) - f(y)| \leq 1$ for all $x \neq y$. By translating f, we can equivalently take the supremum over all f mapping to $[0, 1]$. This says that

$$TV(p, q) = \sup_{f:\mathcal{X}\to[0,1]} \mathbb{E}_{X\sim p}[f(X)] - \mathbb{E}_{X\sim q}[f(X)]$$

Lemma 1. *Suppose $P(G = i) = P(\hat{G} = i)$ for a given $i \in \{1, 2, ..., m\}$. Then $TV(p_i, \hat{p}_i) \leq P(G \neq \hat{G}|G = i)$.*

Proof. For probability measures p_i and $\hat{p}_i$, the TV distance is given by

$$TV(p_i, \hat{p}_i) = \sup\{|p_i(A) - \hat{p}_i(A)| : A \text{ is a measurable event}\}.$$

Fix A to be any measurable event for both p_i and $\hat{p}_i$. This means that A is also a measurable event for p, the distribution of the random variables X, Y. By definition of p_i, $p_i(A) = P(A|G = i)$. Then

$$\begin{aligned}
|p_i(A) - \hat{p}_i(A)| &= |P(A|G = i) - P(A|\hat{G} = i)| \\
&= |P(A|G = i, \hat{G} = i)P(\hat{G} = i|G = i) \\
&\quad + P(A|G = i, \hat{G} \neq i)P(\hat{G} \neq i|G = i) \\
&\quad - P(A|\hat{G} = i, G = i)P(G = i|\hat{G} = i) \\
&\quad - P(A|\hat{G} = i, G \neq i)P(G \neq i|\hat{G} = i)| \\
&= |P(A|G = i, \hat{G} = i)\left(P(\hat{G} = i|G = i) - P(G = i|\hat{G} = i)\right) \\
&\quad - P(\hat{G} \neq G|G = i)\left(P(A|G = i, \hat{G} \neq i) - P(A|\hat{G} = i, G \neq i)\right)|
\end{aligned}$$

[1] Edwards, D.A. On the Kantorovich–Rubinstein theorem. *Expositiones Mathematicae*, 20(4):387-398, 2011.

$$= |0 - P(\hat{G} \neq G|G = i)\Big(P(A|G = i, \hat{G} \neq i) - P(A|\hat{G} = i, G \neq i)\Big)\,|$$
$$\leq P(\hat{G} \neq G|G = i)$$

The second equality follows from the law of total probability. The third and the fourth equalities follow from the assumption that $P(G = i) = P(\hat{G} = i)$, which implies that $P(\hat{G} = G|G = i) = P(G = \hat{G}|\hat{G} = i)$ since

$$P(G = \hat{G}|G = i) = \frac{P(G = \hat{G}, G = i)}{P(G = i)} = \frac{P(G = \hat{G}, \hat{G} = i)}{P(\hat{G} = i)} = P(G = \hat{G}|\hat{G} = i).$$

This further implies that $P(\hat{G} \neq i|G = i) = P(G \neq i|\hat{G} = i)$.

Since $|p_i(A) - \hat{p}_i(A)| \leq P(\hat{G} \neq G|G = i)$ for any measurable event A, the supremum over all events A is also bounded by $P(\hat{G} \neq G|G = i)$. This gives the desired bound on the TV distance. □

Generalization to Wasserstein distances

Theorem 1 can be directly extended to loss functions that are Lipschitz in other metrics. To do so, we first provide a more general definition of Wasserstein distances:

Definition 2. (Wasserstein distance) Let $c(x, y)$ be a metric, and let π be a coupling between p and q. Define the Wasserstein distance between two distributions p, q as

$$W_c(p, q) = \inf_{\pi} \ \mathbb{E}_{X,Y\sim\pi}[c(X, Y)]$$
$$\text{s.t.} \int \pi(x, y)dy = p(x), \int \pi(x, y)dx = q(y).$$

As a familiar example, if $c(x, y) = ||x - y||_2$, then W_c is the earth-mover distance, and $\mathcal{L}(c)$ is the class of 1-Lipschitz functions. Using the Wasserstein distance W_c under different metrics c, we can bound the fairness violations for constraint functions h beyond those specified for the TV distance in Theorem 1.

Theorem 6. *Suppose a model with parameters θ satisfies fairness criteria with respect to the noisy groups $\hat{G}$:*

$$\hat{g}_j(\theta) \leq 0 \ \ \forall j \in \mathcal{G}.$$

Suppose the function h satisfies $|h(\theta, x_1, y_1) - h(\theta, x_2, y_2)| \leq c((x_1, y_1), (x_2, y_2))$ for any $(x_1, y_1) \neq (x_2, y_2)$ w.r.t a metric c. If $W_c(p_j, \hat{p}_j) \leq \gamma_j$ for all $j \in \mathcal{G}$, then the fairness criteria with respect to the true groups G will be satisfied within slacks γ_j for each group:

$$g_j(\theta) \leq \gamma_j \quad \forall j \in \mathcal{G}.$$

Proof. By the triangle inequality, for any group label j,

$$|g_j(\theta) - g(\theta)| \leq |g_j(\theta) - \hat{g}_j(\theta)| + \hat{g}_j(\theta)$$

By Kantorovich-Rubenstein theorem (provided here as Theorem 5), we also have

$$\begin{aligned}|\hat{g}_j(\theta) - g_j(\theta)| &= |\mathbb{E}_{X,Y\sim\hat{p}_j}[h(\theta, X, Y)] - \mathbb{E}_{X,Y\sim p_j}[h(\theta, X, Y)]| \\ &\leq W_c(p_j, \hat{p}_j).\end{aligned}$$

By the assumption that θ satisifes fairness constraints with respect to the noisy groups $\hat{G}$, $\hat{g}_j(\theta) \leq 0$. Therefore, combining these with the triangle inequality, we get the desired result. □

A.2 Additional details on DRO formulation for TV distance

Here we describe the details on solving the DRO problem (2.3) with TV distance using the empirical Lagrangian formulation. We also provide the pseudocode we used for the projected gradient-based algorithm to solve it.

Empirical Lagrangian Formulation

We rewrite the constrained optimization problem (2.3) as a minimax problem using the Lagrangian formulation. We also convert all expectations into expectations over empirical distributions given a dataset of n samples $(X_1, Y_1, G_1), ..., (X_n, Y_n, G_n)$.

Let n_j denote the number of samples that belong to a true group $G = j$. Let the empirical distribution $\hat{p}_j \in \mathbb{R}^n$ be a vector with i-th entry $\hat{p}_j^i = \frac{1}{n_j}$ if the i-th example has a noisy group membership $\hat{G}_i = j$, and 0 otherwise. Replacing all expectations with expectations over the appropriate empirical distributions, the empirical form of (2.3) can be written as:

$$\begin{aligned}\min_{\theta} \quad & \frac{1}{n}\sum_{i=1}^{n} l(\theta, X_i, Y_i) \\ \text{s.t.} \quad & \max_{\tilde{p}_j \in \mathbb{B}_{\gamma_j}(\hat{p}_j)} \sum_{i=1}^{n} \tilde{p}_j^i h(\theta, X_i, Y_i) \leq 0 \quad \forall j \in \mathcal{G}\end{aligned} \tag{A.1}$$

where $\mathbb{B}_{\gamma_j}(\hat{p}_j) = \{\tilde{p}_j \in \mathbb{R}^n : \frac{1}{2}\sum_{i=1}^{n} |\tilde{p}_j^i - \hat{p}_j^i| \leq \gamma_j, \sum_{i=1}^{n} \tilde{p}_j^i = 1, \tilde{p}_j^i \geq 0 \quad \forall i = 1, ..., n\}$.

For ease of notation, for $j \in \{1, 2, ..., m\}$, let

$$f(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(\theta, X_i, Y_i)$$

$$f_j(\theta, \tilde{p}_j) = \sum_{i=1}^{n} \tilde{p}_j^i h(\theta, X_i, Y_i).$$

Then the Lagrangian of the empirical formulation (A.1) is

$$\mathcal{L}(\theta, \lambda) = f(\theta) + \sum_{j=1}^{m} \lambda_j \max_{\tilde{p}_j \in \mathbb{B}_\gamma(\hat{p}_j)} f_j(\theta, \tilde{p}_j)$$

and problem (A.1) can be rewritten as

$$\min_{\theta} \max_{\lambda \geq 0} f(\theta) + \sum_{j=1}^{m} \lambda_j \max_{\tilde{p}_j \in \mathbb{B}_\gamma(\hat{p}_j)} f_j(\theta, \tilde{p}_j)$$

Moving the inner max out of the sum and rewriting the constraints as ℓ_1-norm constraints:

$$\begin{aligned} &\min_{\theta} \max_{\lambda \geq 0} \max_{\substack{\tilde{p}_j \in \mathbb{R}^n, \tilde{p}_j \geq 0, \\ j=1,\ldots,m}} f(\theta) + \sum_{j=1}^{m} \lambda_j f_j(\theta, \tilde{p}_j) \\ &\quad \text{s.t. } ||\tilde{p}_j - \hat{p}_j||_1 \leq 2\gamma_j, \quad ||\tilde{p}_j||_1 = 1 \quad \forall j \in \{1, ..., m\} \end{aligned} \tag{A.2}$$

Since projections onto the ℓ_1-ball can be done efficiently [Duchi et al., 2008], we can solve problem (A.2) using a projected gradient descent ascent (GDA) algorithm. This is a simplified version of the algorithm introduced by Namkoong and Duchi [2016] for solving general classes of DRO problems. We provide pseudocode in Algorithm 3, as well as an actual implementation in the attached code.

Projected GDA Algorithm for DRO

Algorithm 3 Project GDA Algorithm

```
Require: learning rates $\eta_\theta > 0$, $\eta_\lambda > 0$, $\eta_p > 0$, estimates of $P(G \neq \hat{G} | \hat{G} = j)$ to specify $\gamma_j$.
for $t = 1, \ldots, T$ do
    Descent step on $\theta$:
    $\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta_\theta \nabla_\theta f(\theta^{(t)}) - \eta_\theta \sum_{j=1}^{m} \lambda_j^{(t)} \nabla_\theta f_j(\theta^{(t)}, \tilde{p}_j^{(t)})$
    Ascent step on $\lambda$:
    $\lambda_j^{(t+1)} \leftarrow \lambda_j^{(t)} + \eta_\lambda f_j(\theta, \tilde{p}_j^{(t)})$
    for $j = 1, ..., m$ do
        Ascent step on $\tilde{p}_j$: $\tilde{p}_j^{(t+1)} \leftarrow \tilde{p}_j^{(t)} + \eta_p \lambda_j^{(t)} \nabla_{\tilde{p}_j} f_j(\theta^{(t)}, \tilde{p}_j^{(t)})$
        Project $\tilde{p}_j^{(t+1)}$ onto $\ell_1$-norm constraints: $||\tilde{p}_j^{(t+1)} - \hat{p}_j||_1 \leq 2\gamma_j, ||\tilde{p}_j^{(t+1)}||_1 = 1$
    end for
end for
return $\theta^{(t^*)}$ where $t^*$ denotes the best iterate that satisfies the constraints in (2.3) with the lowest objective.
```

Equalizing TPRs and FPRs using DRO

In the two case studies in Section 2.7, we enforce *equality of opportunity* and *equalized odds* Hardt et al. [2016b] by equalizing true positive rates (TPRs) and/or false positive rates (FPRs) within some slack α. In this section, we describe in detail the implementation of the constraints for equalizing TPRs and FPRs under the DRO approach.

To equalize TPRs with slack α under the DRO approach, we set

$$\tilde{g}_j^{\text{TPR}}(\theta) = \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1)\,\mathbb{1}(\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1)]} - \frac{\mathbb{E}_{X,Y\sim \tilde{p}_j}[\mathbb{1}(Y=1)\,\mathbb{1}(\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim \tilde{p}_j}[\mathbb{1}(Y=1)]} - \alpha. \tag{A.3}$$

The first term corresponds to the TPR for the full population. The second term estimates the TPR for group j. Setting $\alpha = 0$ exactly equalizes true positive rates.

To equalize FPRs with slack α under the DRO approach, we set

$$\tilde{g}_j^{\text{FPR}}(\theta) = \frac{\mathbb{E}_{X,Y\sim \tilde{p}_j}[\mathbb{1}(Y=0)\,\mathbb{1}(\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim \tilde{p}_j}[\mathbb{1}(Y=0)]} - \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0)\,\mathbb{1}(\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0)]} - \alpha. \tag{A.4}$$

The first term estimates the FPR for group j. The second term corresponds to the FPR for the full population. Setting $\alpha = 0$ exactly equalizes false positive rates.

To equalize TPRs for Case Study 1, we apply m constraints, $\left\{\max_{\tilde{p}_j : TV(\tilde{p}_j,\hat{p}_j)\le\gamma_j, \tilde{p}_j \ll p} \tilde{g}_j^{\text{TPR}}(\theta) \le 0\right\}\ \forall j \in \mathcal{G}$.

To equalize both TPRs and FPRs simultaneously for Case Study 2, we apply $2m$ constraints, $\left\{\max_{\tilde{p}_j : TV(\tilde{p}_j,\hat{p}_j)\le\gamma_j, \tilde{p}_j \ll p} \tilde{g}_j^{\text{TPR}}(\theta) \le 0, \max_{\tilde{p}_j : TV(\tilde{p}_j,\hat{p}_j)\le\gamma_j, \tilde{p}_j \ll p} \tilde{g}_j^{\text{FPR}}(\theta) \le 0\right\}\ \forall j \in \mathcal{G}$.

$h(\theta, X, Y)$ for equalizing TPRs and FPRs

Since the notation in Section 2.5 and in the rest of the paper uses generic functions h to express the group-specific constraints, we show in Lemma 6 that the constraint using $\tilde{g}_j^{\text{TPR}}(\theta)$ in Equation (A.3) can also be written as an equivalent constraint in the form of Equation (2.3), as

$$\tilde{g}_j^{\text{TPR}}(\theta) = \mathbb{E}_{X,Y\sim \tilde{p}_j}[h^{\text{TPR}}(\theta, X, Y)]$$

for some function $h^{\text{TPR}} : \Theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

Lemma 6. Denote $\hat{Y}$ as $\mathbb{1}(\phi(X;\theta) > 0)$. Let $h^{\text{TPR}}(\theta, X, Y)$ be given by

$$h^{\text{TPR}}(\theta, X, Y) = \frac{1}{2}\left(-\mathbb{1}(\hat{Y}=1, Y=1) - \mathbb{1}(Y=1)\left(\alpha - \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1,\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1)]}\right)\right).$$

Then

$$\frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1)\,\mathbb{1}(\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1)]} - \frac{\mathbb{E}_{X,Y\sim \tilde{p}_j}[\mathbb{1}(Y=1)\,\mathbb{1}(\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim \tilde{p}_j}[\mathbb{1}(Y=1)]} - \alpha \le 0$$
$$\iff \mathbb{E}_{X,Y\sim \tilde{p}_j}[h^{\text{TPR}}(\theta, X, Y)] \le 0.$$

Proof. Substituting the given function $h^{\text{TPR}}(\theta, X, Y)$, and using the fact that $\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(Y=1)] \geq 0$:

$$\begin{aligned}
&\mathbb{E}_{X,Y\sim\tilde{p}_j}[h^{\text{TPR}}(\theta, X, Y)] \leq 0 \\
&\iff \mathbb{E}_{X,Y\sim\tilde{p}_j}\left[\frac{1}{2}\left(-\mathbb{1}(\hat{Y}=1, Y=1) - \mathbb{1}(Y=1)\left(\alpha - \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1,\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1)]}\right)\right)\right] \leq 0 \\
&\iff -\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(\hat{Y}=1, Y=1)] - \mathbb{E}_{X,Y\sim\tilde{p}_j}\left[\mathbb{1}(Y=1)\left(\alpha - \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1,\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1)]}\right)\right] \leq 0 \\
&\iff -\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(\hat{Y}=1, Y=1)] - \alpha\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(Y=1)] \\
&\qquad + \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1,\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1)]}\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(Y=1)] \leq 0 \\
&\iff \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1,\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1)]} - \frac{\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(\hat{Y}=1, Y=1)]}{\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(Y=1)]} - \alpha \leq 0
\end{aligned}$$

□

By similar proof, we also show in Lemma 7 that the constraint using $\tilde{g}_j^{\text{FPR}}(\theta)$ in Equation (A.4) can also be written as an equivalent constraint in the form of Equation (2.3), as

$$\tilde{g}_j^{\text{FPR}}(\theta) = \mathbb{E}_{X,Y\sim\tilde{p}_j}[h^{\text{FPR}}(\theta, X, Y)]$$

for some function $h^{\text{FPR}} : \Theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

Lemma 7. Denote $\hat{Y}$ as $\mathbb{1}(\phi(X;\theta) > 0)$. Let $h^{\text{FPR}}(\theta, X, Y)$ be given by

$$h^{\text{FPR}}(\theta, X, Y) = \frac{1}{2}\left(\mathbb{1}(\hat{Y}=1, Y=0) - \mathbb{1}(Y=0)\left(\alpha + \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0,\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0)]}\right)\right).$$

Then

$$\begin{aligned}
\frac{\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(Y=0)\,\mathbb{1}(\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(Y=0)]} - \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0)\,\mathbb{1}(\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0)]} - \alpha \leq 0 \\
\iff \mathbb{E}_{X,Y\sim\tilde{p}_j}[h^{\text{FPR}}(\theta, X, Y)] \leq 0.
\end{aligned}$$

Proof. Substituting the given function $h^{\text{FPR}}(\theta, X, Y)$, and using the fact that $\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(Y=0)] \geq 0$:

$$\begin{aligned}
&\mathbb{E}_{X,Y\sim\tilde{p}_j}[h^{\text{FPR}}(\theta, X, Y)] \leq 0 \\
&\iff \mathbb{E}_{X,Y\sim\tilde{p}_j}\left[\frac{1}{2}\left(\mathbb{1}(\hat{Y}=1, Y=0) - \mathbb{1}(Y=0)\left(\alpha + \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0,\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0)]}\right)\right)\right] \leq 0
\end{aligned}$$

$$\begin{aligned}
\iff & \mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(\hat{Y}=1,Y=0)] - \mathbb{E}_{X,Y\sim\tilde{p}_j}\left[\mathbb{1}(Y=0)\left(\alpha + \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0,\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0)]}\right)\right] \leq 0 \\
\iff & \mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(\hat{Y}=1,Y=0)] - \alpha\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(Y=0)] \\
& - \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0,\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0)]}\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(Y=0)] \leq 0 \\
\iff & \frac{\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(\hat{Y}=1,Y=0)]}{\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(Y=0)]} - \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0,\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0)]} - \alpha \leq 0
\end{aligned}$$

□

DRO when $\hat{G}$ and G have different dimensionalities

The soft assignments approach is naturally formulated to be able to handle $G \in \mathcal{G} = \{1, ..., m\}$ and $\hat{G} \in \hat{\mathcal{G}} = \{1, ..., \hat{m}\}$ when $\hat{m} \neq m$. The DRO approach can be extended to handle this case by generalizing Lemma 1 to $TV(p_j, \hat{p}_i) \leq P(\hat{G} \neq i | G = j), j \in \mathcal{G}, i \in \hat{\mathcal{G}}$, and generalizing the DRO formulation to have the true group distribution p_j bounded in a TV distance ball centered at $\hat{p}_i$. Empirically comparing this generalized DRO approach to the soft group assignments approach when $\hat{m} \neq m$ is an interesting avenue of future work.

A.3 Additional details for soft group assignments approach

Here we provide additional technical details regarding the soft group assignments approach introduced in Section 2.7.

Derivation for $\mathbb{E}[h(\theta, X, Y)|G = j]$

Here we show $\mathbb{E}[h(\theta,X,Y)|G=j] = \frac{\mathbb{E}[h(\theta,X,Y)P(G=j|\hat{Y},Y,\hat{G})]}{P(G=j)}$, assuming that $h(\theta, X, Y)$ depends on X through $\hat{Y}$, i.e. $\hat{Y} = \mathbb{1}(\phi(\theta, X) > 0)$. Using the tower property and the definition of conditional expectation:

$$\begin{aligned}
\mathbb{E}[h(\theta,X,Y)|G=j] &= \frac{\mathbb{E}[h(\theta,X,Y)\,\mathbb{1}(G=j)]}{P(G=j)} \\
&= \frac{\mathbb{E}[\mathbb{E}[h(\theta,X,Y)\,\mathbb{1}(G=j)|\hat{Y},Y,\hat{G}]]}{P(G=j)} \\
&= \frac{\mathbb{E}[h(\theta,X,Y)\mathbb{E}[\mathbb{1}(G=j)|\hat{Y},Y,\hat{G}]]}{P(G=j)} \\
&= \frac{\mathbb{E}[h(\theta,X,Y)P(G=j|\hat{Y},Y,\hat{G})]}{P(G=j)}
\end{aligned} \tag{A.5}$$

Equalizing TPRs and FPRs using soft group assignments

In the two case studies in Section 2.7, we enforce *equality of opportunity* and *equalized odds* Hardt et al. [2016b] by equalizing true positive rates (TPRs) and/or false positive rates (FPRs) within some slack α. In this section, we describe in detail the implementation of the constraints for equalizing TPRs and FPRs under the soft group assignments approach.

To equalize TPRs with slack α under the soft group assignments approach, we set

$$g_j^{\text{TPR}}(\theta, w) = \frac{\mathbb{E}[\mathbb{1}(Y=1)\,\mathbb{1}(\hat{Y}=1)]}{\mathbb{E}[\mathbb{1}(Y=1)]} - \frac{\mathbb{E}[\mathbb{1}(Y=1)\,\mathbb{1}(\hat{Y}=1)w(j|\hat{Y},Y,\hat{G})]}{\mathbb{E}[\mathbb{1}(Y=1)w(j|\hat{Y},Y,\hat{G})]} - \alpha. \tag{A.6}$$

The first term corresponds to the TPR for the full population. The second term estimates the TPR for group j as done by Kallus et al. [2020] in Equation (5) and Proposition 8. Setting $\alpha = 0$ exactly equalizes true positive rates.

To equalize FPRs with slack α under the soft group assignments approach, we set

$$g_j^{\text{FPR}}(\theta, w) = \frac{\mathbb{E}[\mathbb{1}(Y=0)\,\mathbb{1}(\hat{Y}=1)w(j|\hat{Y},Y,\hat{G})]}{\mathbb{E}[\mathbb{1}(Y=0)w(j|\hat{Y},Y,\hat{G})]} - \frac{\mathbb{E}[\mathbb{1}(Y=0)\,\mathbb{1}(\hat{Y}=1)]}{\mathbb{E}[\mathbb{1}(Y=0)]} - \alpha. \tag{A.7}$$

The first term estimates the FPR for group j as done previously for the TPR. The second term corresponds to the FPR for the full population. Setting $\alpha = 0$ exactly equalizes false positive rates.

To equalize TPRs for Case Study 1, we apply m constraints, $\left\{\max_{w\in\mathcal{W}(\theta)} g_j^{\text{TPR}}(\theta,w) \leq 0\right\}\ \forall j \in \mathcal{G}$. To equalize both TPRs and FPRs simultaneously for Case Study 2, we apply $2m$ constraints, $\left\{\max_{w\in\mathcal{W}(\theta)} g_j^{\text{TPR}}(\theta,w) \leq 0, \max_{w\in\mathcal{W}(\theta)} g_j^{\text{FPR}}(\theta,w) \leq 0\right\}\ \forall j \in \mathcal{G}$.

$h(\theta, X, Y)$ for equalizing TPRs and FPRs

Since the notation in Section 2.6 and in the rest of the paper uses generic functions h to express the group-specific constraints, we show in Lemma 8 that the constraint using $g_j^{\text{TPR}}(\theta,w)$ in Equation (A.6) can also be written as an equivalent constraint in the form of Equation (2.6), as

$$g_j^{\text{TPR}}(\theta,w) = \frac{\mathbb{E}[h^{\text{TPR}}(\theta,X,Y)w(j|\hat{Y},Y,\hat{G})]}{P(G=j)}$$

for some function $h^{\text{TPR}} : \Theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

Lemma 8. Denote $\hat{Y}$ as $\mathbb{1}(\phi(X;\theta) > 0)$. Let $h^{\text{TPR}}(\theta,X,Y)$ be given by

$$h^{\text{TPR}}(\theta,X,Y) = \frac{1}{2}\left(-\mathbb{1}(\hat{Y}=1, Y=1) - \mathbb{1}(Y=1)\left(\alpha - \frac{\mathbb{E}[\mathbb{1}(Y=1,\hat{Y}=1)]}{\mathbb{E}[\mathbb{1}(Y=1)]}\right)\right).$$

Then

$$\frac{\mathbb{E}[\mathbb{1}(Y=1)\,\mathbb{1}(\hat{Y}=1)]}{\mathbb{E}[\mathbb{1}(Y=1)]} - \frac{\mathbb{E}[\mathbb{1}(Y=1)\,\mathbb{1}(\hat{Y}=1)w(j|\hat{Y},Y,\hat{G})]}{\mathbb{E}[\mathbb{1}(Y=1)w(j|\hat{Y},Y,\hat{G})]} - \alpha \leq 0$$

$$\iff \frac{\mathbb{E}[h^{\mathrm{TPR}}(\theta, X, Y)w(j|\hat{Y}, Y, \hat{G})]}{P(G=j)} \leq 0.$$

for all $j \in \mathcal{G}, P(G=j) > 0$.

Proof. Substituting the given function $h^{\mathrm{TPR}}(\theta, X, Y)$, and using the fact that $P(G=j) > 0$ and $\mathbb{E}[\mathbb{1}(Y=1)w(j|\hat{Y}, Y, \hat{G})] \geq 0$:

$$\begin{aligned}
&\frac{\mathbb{E}[h^{\mathrm{TPR}}(\theta, X, Y)w(j|\hat{Y}, Y, \hat{G})]}{P(G=j)} \leq 0 \\
&\iff \mathbb{E}[h^{\mathrm{TPR}}(\theta, X, Y)w(j|\hat{Y}, Y, \hat{G})] \leq 0 \\
&\iff \mathbb{E}\left[\frac{1}{2}\left(-\mathbb{1}(\hat{Y}=1, Y=1) - \mathbb{1}(Y=1)\left(\alpha - \frac{\mathbb{E}[\mathbb{1}(Y=1, \hat{Y}=1)]}{\mathbb{E}[\mathbb{1}(Y=1)]}\right)\right)w(j|\hat{Y}, Y, \hat{G})\right] \leq 0 \\
&\iff -\mathbb{E}[\mathbb{1}(\hat{Y}=1, Y=1)w(j|\hat{Y}, Y, \hat{G})] \\
&\qquad - \mathbb{E}\left[\mathbb{1}(Y=1)\left(\alpha - \frac{\mathbb{E}[\mathbb{1}(Y=1, \hat{Y}=1)]}{\mathbb{E}[\mathbb{1}(Y=1)]}\right)w(j|\hat{Y}, Y, \hat{G})\right] \leq 0 \\
&\iff -\mathbb{E}[\mathbb{1}(\hat{Y}=1, Y=1)w(j|\hat{Y}, Y, \hat{G})] - \alpha\mathbb{E}[\mathbb{1}(Y=1)w(j|\hat{Y}, Y, \hat{G})] \\
&\qquad + \frac{\mathbb{E}[\mathbb{1}(Y=1, \hat{Y}=1)]}{\mathbb{E}[\mathbb{1}(Y=1)]}\mathbb{E}[\mathbb{1}(Y=1)w(j|\hat{Y}, Y, \hat{G})] \leq 0 \\
&\iff \frac{\mathbb{E}[\mathbb{1}(Y=1, \hat{Y}=1)]}{\mathbb{E}[\mathbb{1}(Y=1)]} - \frac{\mathbb{E}[\mathbb{1}(\hat{Y}=1, Y=1)w(j|\hat{Y}, Y, \hat{G})]}{\mathbb{E}[\mathbb{1}(Y=1)w(j|\hat{Y}, Y, \hat{G})]} - \alpha \leq 0
\end{aligned}$$

□

By similar proof, we also show in Lemma 9 that the constraint using $g_j^{\mathrm{FPR}}(\theta, w)$ in Equation (A.7) can also be written as an equivalent constraint in the form of Equation (2.6), as

$$g_j^{\mathrm{FPR}}(\theta, w) = \frac{\mathbb{E}[h^{\mathrm{FPR}}(\theta, X, Y)w(j|\hat{Y}, Y, \hat{G})]}{P(G=j)}$$

for some function $h^{\mathrm{FPR}} : \Theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

Lemma 9. Denote $\hat{Y}$ as $\mathbb{1}(\phi(X;\theta) > 0)$. Let $h^{\mathrm{FPR}}(\theta, X, Y)$ be given by

$$h^{\mathrm{FPR}}(\theta, X, Y) = \frac{1}{2}\left(\mathbb{1}(\hat{Y}=1, Y=0) - \mathbb{1}(Y=0)\left(\alpha + \frac{\mathbb{E}[\mathbb{1}(Y=0, \hat{Y}=1)]}{\mathbb{E}[\mathbb{1}(Y=0)]}\right)\right).$$

Then

$$\frac{\mathbb{E}[\mathbb{1}(Y=0)\,\mathbb{1}(\hat{Y}=1)w(j|\hat{Y}, Y, \hat{G})]}{\mathbb{E}[\mathbb{1}(Y=0)w(j|\hat{Y}, Y, \hat{G})]} - \frac{\mathbb{E}[\mathbb{1}(Y=0)\,\mathbb{1}(\hat{Y}=1)]}{\mathbb{E}[\mathbb{1}(Y=0)]} - \alpha \leq 0$$

$$\iff \frac{\mathbb{E}[h^{\text{FPR}}(\theta, X, Y)w(j|\hat{Y}, Y, \hat{G})]}{P(G = j)} \leq 0.$$

for all $j \in \mathcal{G}, P(G = j) > 0$.

Proof. Substituting the given function $h^{\text{FPR}}(\theta, X, Y)$, and using the fact that $P(G = j) > 0$ and $\mathbb{E}[\mathbb{1}(Y = 0)w(j|\hat{Y}, Y, \hat{G})] \geq 0$:

$$\begin{aligned}
&\frac{\mathbb{E}[h^{\text{FPR}}(\theta, X, Y)w(j|\hat{Y}, Y, \hat{G})]}{P(G = j)} \leq 0 \\
&\iff \mathbb{E}[h^{\text{FPR}}(\theta, X, Y)w(j|\hat{Y}, Y, \hat{G})] \leq 0 \\
&\iff \mathbb{E}\left[\frac{1}{2}\left(\mathbb{1}(\hat{Y} = 1, Y = 0) - \mathbb{1}(Y = 0)\left(\alpha + \frac{\mathbb{E}[\mathbb{1}(Y = 0, \hat{Y} = 1)]}{\mathbb{E}[\mathbb{1}(Y = 0)]}\right)\right) w(j|\hat{Y}, Y, \hat{G})\right] \leq 0 \\
&\iff \mathbb{E}[\mathbb{1}(\hat{Y} = 1, Y = 0)w(j|\hat{Y}, Y, \hat{G})] \\
&\qquad - \mathbb{E}\left[\mathbb{1}(Y = 0)\left(\alpha + \frac{\mathbb{E}[\mathbb{1}(Y = 0, \hat{Y} = 1)]}{\mathbb{E}[\mathbb{1}(Y = 0)]}\right) w(j|\hat{Y}, Y, \hat{G})\right] \leq 0 \\
&\iff \mathbb{E}[\mathbb{1}(\hat{Y} = 1, Y = 0)w(j|\hat{Y}, Y, \hat{G})] - \alpha\mathbb{E}[\mathbb{1}(Y = 0)w(j|\hat{Y}, Y, \hat{G})] \\
&\qquad - \frac{\mathbb{E}[\mathbb{1}(Y = 0, \hat{Y} = 1)]}{\mathbb{E}[\mathbb{1}(Y = 0)]}\mathbb{E}[\mathbb{1}(Y = 0)w(j|\hat{Y}, Y, \hat{G})] \leq 0 \\
&\iff \frac{\mathbb{E}[\mathbb{1}(\hat{Y} = 1, Y = 0)w(j|\hat{Y}, Y, \hat{G})]}{\mathbb{E}[\mathbb{1}(Y = 0)w(j|\hat{Y}, Y, \hat{G})]} - \frac{\mathbb{E}[\mathbb{1}(Y = 0, \hat{Y} = 1)]}{\mathbb{E}[\mathbb{1}(Y = 0)]} - \alpha \leq 0
\end{aligned}$$

□

A.4 Optimality and feasibility for the *Ideal* algorithm

We provide optimality and feasibility guarantees for Algorithm 1 and optimality guarantees for Algorithm 4.

Theorem 7 (**Optimality and Feasibility for Algorithm 1**). *Let $\theta^* \in \Theta$ be such that it satisfies the constraints $\max\limits_{w \in \mathcal{W}(\theta)} g_j(\theta^*, w) \leq 0,\ \forall j \in \mathcal{G}$ and $f_0(\theta^*) \leq f(\theta)$ for every $\theta \in \Theta$ that satisfies the same constraints. Let $0 \leq f_0(\theta) \leq B, \forall \theta \in \Theta$. Let the space of Lagrange multipliers be defined as $\Lambda = \{\lambda \in \mathbb{R}^m_+ \,|\, \|\lambda\|_1 \leq R\}$, for $R > 0$. Let $B_\lambda \geq \max_t \|\nabla_\lambda \mathcal{L}(\theta^{(t)}, \lambda^{(t)})\|_2$. Let $\bar{\theta}$ be the stochastic classifier returned by Algorithm 1 when run for T iterations, with the radius of the Lagrange multipliers $R = T^{1/4}$ and learning rate $\eta_\lambda = \frac{R}{B_\lambda\sqrt{T}}$ Then:*

$$\mathbf{E}_{\theta \sim \bar{\theta}}\left[f(\theta)\right] \,\leq\, f(\theta^*) \,+\, \mathcal{O}\left(\frac{1}{T^{1/4}}\right) \,+\, \rho$$

and

$$\mathbf{E}_{\theta\sim\bar{\theta}}\left[\max_{w\in\mathcal{W}(\theta)}\, g_j(\theta,w)\right] \le \mathcal{O}\left(\frac{1}{T^{1/4}}\right)+\rho'$$

Thus for any given $\varepsilon>0$, by solving Steps 2 and 4 of Algorithm 1 to sufficiently small errors ρ,ρ', and by running the algorithm for a sufficiently large number of steps T, we can guarantee that the returned stochastic model is ε-optimal and ε-feasible.

Proof. Let $\overline{\lambda}=\frac{1}{T}\sum_{t=1}^{T}\lambda^{(t)}$. We will interpret the minimax problem in equation (2.8) as a zero-sum between the θ-player who optimizes $\mathcal{L}$ over θ, and the λ-player who optimizes $\mathcal{L}$ over λ. We first bound the average regret incurred by the players over T steps. The best response computation in Step 2 of Algorithm 1 gives us:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\theta\sim\hat{\theta}^{(t)}}\left[\mathcal{L}(\theta,\lambda^{(t)})\right] &\le \frac{1}{T}\sum_{t=1}^{T}\min_{\theta\in\Theta}\mathcal{L}(\theta,\lambda^{(t)})\,+\,\varepsilon\\ &\le \min_{\theta\in\Theta}\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}(\theta,\lambda^{(t)})\,+\,\rho\\ &= \min_{\theta\in\Theta}\mathcal{L}(\theta,\overline{\lambda})\,+\,\rho\\ &\le \min_{\theta\in\Theta}\max_{\lambda\in\Lambda}\mathcal{L}(\theta,\lambda)\,+\,\rho\\ &\le f(\theta^*)+\rho. \end{aligned}\tag{A.8}$$

We then apply standard gradient ascent analysis for the projected gradient updates to λ in Step 4 of the algorithm, and get:

$$\max_{\lambda\in\Lambda}\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{m}\lambda_j\delta_j^{(t)} \ge \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{m}\lambda_j^{(t)}\delta_j^{(t)} - \mathcal{O}\left(\frac{R}{\sqrt{T}}\right).$$

We then plug the upper and lower bounds for the gradient estimates $\delta_j^{(t)}$'s from Step 3 of the Algorithm 1 into the above inequality:

$$\begin{aligned}&\max_{\lambda\in\Lambda}\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{m}\lambda_j\left(\mathbb{E}_{\theta\sim\hat{\theta}^{(t)}}\left[\max_{w\in\mathcal{W}(\theta)}g_j(\theta,w)\right]\,+\,\rho'\right)\\ &\ge\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{m}\lambda_j^{(t)}\mathbb{E}_{\theta\sim\hat{\theta}^{(t)}}\left[\max_{w\in\mathcal{W}(\theta)}g_j(\theta,w)\right]\,-\,\mathcal{O}\left(\frac{R}{\sqrt{T}}\right).\end{aligned}$$

which further gives us:

$$\max_{\lambda\in\Lambda}\left\{\sum_{j=1}^{m}\lambda_j\mathbb{E}_{\theta\sim\hat{\theta}^{(t)}}\left[\max_{w\in\mathcal{W}(\theta)}g_j(\theta,w)\right]\,+\,\|\lambda\|_1\rho'\right\}$$

$$\geq \sum_{j=1}^{m} \lambda_j^{(t)} \mathbb{E}_{\theta\sim\hat{\theta}^{(t)}}\left[\max_{w\in\mathcal{W}(\theta)} g_j(\theta, w)\right] - \mathcal{O}\left(\frac{R}{\sqrt{T}}\right).$$

Adding $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\hat{\theta}^{(t)}}\left[f(\theta)\right]$ to both sides of the above inequality, we finally get:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\hat{\theta}^{(t)}}\left[\mathcal{L}(\theta, \lambda^{(t)})\right] \geq \max_{\lambda\in\Lambda}\left\{\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\hat{\theta}^{(t)}}\left[\mathcal{L}(\theta, \lambda)\right] + \|\lambda\|_1\rho'\right\} - \mathcal{O}\left(\frac{R}{\sqrt{T}}\right). \quad (A.9)$$

Optimality. Now, substituting $\lambda = \mathbf{0}$ in equation (A.9) and combining with equation (A.8) completes the proof of the optimality guarantee:

$$\mathbb{E}_{\theta\sim\bar{\theta}}\left[f(\theta)\right] \leq f_0(\theta^*) + \mathcal{O}\left(\frac{R}{\sqrt{T}}\right) + \rho$$

Feasibility. To show feasibility, we fix a constraint index $j \in \mathcal{G}$. Now substituting $\lambda_j = R$ and $\lambda_{j'} = 0, \forall j' \neq j$ in equation (A.9) and combining with equation (A.8) gives us:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\hat{\theta}^{(t)}}\left[f(\theta) + R \max_{w\in\mathcal{W}(\theta)} g_j(\theta, w)\right] \leq f(\theta^*) + \mathcal{O}\left(\frac{R}{\sqrt{T}}\right) + \rho + R\rho'.$$

which can be re-written as:

$$\begin{aligned}\mathbb{E}_{\theta\sim\bar{\theta}}\left[\max_{w\in\mathcal{W}(\theta)} g_j(\theta, w)\right] &\leq \frac{f(\theta^*) - \mathbb{E}_{\theta\sim\bar{\theta}}\left[f(\theta)\right]}{R} + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \frac{\rho}{R} + \rho'. \\ &\leq \frac{B}{R} + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \frac{\rho}{R} + \rho',\end{aligned}$$

which is our feasibility guarantee. Setting $R = \mathcal{O}(T^{1/4})$ then completes the proof. □

Best Response over θ

We next describe our procedure for computing a best response over θ in Step 2 of Algorithm 1. We will consider a slightly relaxed version of the best response problem where the equality constraints in $\mathcal{W}(\theta)$ are replaced with closely-approximating inequality constraints.

Recall that the constraint set $\mathcal{W}(\theta)$ contains two sets of constraints equation (2.5), the total probability constraints that depend on θ, and the simplex constraints that do not depend on θ. So to decouple these constraint sets from θ, we introduce Lagrange multipliers μ for the total probability constraints to make them a part of the objective, and obtain a nested *minimax* problem over θ, μ, and w, where w is constrained to satisfy the simplex constraints alone. We then jointly minimize the inner Lagrangian over θ and μ, and perform gradient ascent updates on w with projections onto the simplex constraints. The joint-minimization over θ and μ is not necessarily convex and is solved using a minimization oracle.

Algorithm 4 Best response on θ of Algorithm 1

```
Require: $\lambda'$, learning rate $\eta_{\mathbf{w}} > 0$, estimates of $P(G = j|\hat{G} = k)$ to specify constraints $r_{g,\hat{g}}$'s, $\kappa$
for $q = 1, \ldots, Q$ do
    Best response on ($\theta$, $\boldsymbol{\mu}$): use an oracle to find find $\theta^{(q)} \in \Theta$ and $\boldsymbol{\mu}^{(q)} \in \mathcal{M}^m$ such that:
        $\ell(\theta^{(q)}, \boldsymbol{\mu}^{(q)}, \mathbf{w}^{(q)}; \lambda') \leq \min_{\theta \in \Theta, \boldsymbol{\mu} \in \mathcal{M}^m} \ell(\theta, \boldsymbol{\mu}, \mathbf{w}^{(q)}; \lambda') + \kappa,$
    for a small slack $\kappa > 0$.
    Ascent step on $\mathbf{w}$:
        $w_j^{(q+1)} \leftarrow \Pi_{\mathcal{W}_\Delta}\left(w_j^{(q)} + \eta_{\mathbf{w}} \nabla_{w_j} \ell(\theta^{(q)}, \boldsymbol{\mu}^{(q)}, \mathbf{w}^{(q)}; \lambda')\right),$
    where $\nabla_{w_j}\ell(\cdot)$ is a sub-gradient of $\ell$ w.r.t. $w_j$.
end for
return A uniform distribution $\hat{\theta}$ over $\theta^{(1)}, \ldots, \theta^{(Q)}$
```

We begin by writing out the best-response problem over θ for a fixed λ':

$$\min_{\theta \in \Theta} \mathcal{L}(\theta, \lambda') \;=\; \min_{\theta \in \Theta} f(\theta) + \sum_{j=1}^{m} \lambda'_j \max_{w_j \in \mathcal{W}(\theta)} g_j(\theta, w_j), \tag{A.10}$$

where we use w_j to denote the maximizer over $\mathcal{W}(\theta)$ for constraint g_j explicitly. We separate out the the simplex constraints in $\mathcal{W}(\theta)$ equation (2.5) and denote them by:

$$\mathcal{W}_\Delta = \left\{ w \in \mathbb{R}_+^{\mathcal{G} \times \{0,1\}^2 \times \hat{\mathcal{G}}} \;\middle|\; \sum_{j=1}^{m} w(j \mid \hat{y}, y, k) = 1,\ \forall k \in \hat{\mathcal{G}}, y, \hat{y} \in \{0,1\} \right\},$$

where we represent each w as a vector of values $w(i|\hat{y}, y, k)$ for each $j \in \mathcal{G}, \hat{y} \in \{0,1\}, y \in \{0,1\}$, and $k \in \hat{\mathcal{G}}$. We then relax the total probability constraints in $\mathcal{W}(\theta)$ into a set of inequality constraints:

$$\begin{aligned} P(G = j|\hat{G} = k) \;-\; \sum_{\hat{y}, y \in \{0,1\}} w(j \mid \hat{y}, y, k) P(\hat{Y}(\theta) = \hat{y}, Y = y|\hat{G} = k) \;-\; \tau \;&\leq\; 0 \\ \sum_{\hat{y}, y \in \{0,1\}} w(j \mid \hat{y}, y, k) P(\hat{Y}(\theta) = \hat{y}, Y = y|\hat{G} = k) \;-\; P(G = j|\hat{G} = k) \;-\; \tau \;&\leq\; 0 \end{aligned}$$

for some small $\tau > 0$. We have a total of $U = 2 \times m \times \hat{m}$ relaxed inequality constraints, and will denote each of them as $r_u(\theta, w) \leq 0$, with index u running from 1 to U. Note that each $r_u(\theta, w)$ is linear in w.

Introducing Lagrange multipliers μ for the relaxed total probability constraints, the optimization problem in equation (A.10) can be re-written equivalently as:

$$\min_{\theta\in\Theta} f(\theta) + \sum_{j=1}^{m} \lambda'_j \max_{w_j\in\mathcal{W}_\Delta} \min_{\mu_j\in\mathcal{M}} \left\{ g_j(\theta, w_j) - \sum_{u=1}^{U} \mu_{j,u}\, r_u(\theta, w_j) \right\},$$

where note that each w_j is maximized over only the simplex constraints $\mathcal{W}_\Delta$ which are independent of θ, and $\mathcal{M} = \{\mu_j \in \mathbb{R}_+^{m\times\hat{m}} \,|\, \|\mu_j\|_1 \leq R'\}$, for some constant $R' > 0$. Because each w_j and μ_j appears only in the j-th term in the summation, we can pull out the max and min, and equivalently rewrite the above problem as:

$$\min_{\theta\in\Theta} \max_{\mathbf{w}\in\mathcal{W}_\Delta^m} \min_{\boldsymbol{\mu}\in\mathcal{M}^m} \underbrace{f(\theta) + \sum_{j=1}^{m} \lambda'_j \Bigg(\underbrace{g_j(\theta, w_j) - \sum_{u=1}^{U} \mu_{j,u}\, r_u(\theta, w_j)}_{\omega(\theta,\mu_j,w_j)} \Bigg)}_{\ell(\theta,\boldsymbol{\mu},\mathbf{w};\lambda')}, \tag{A.11}$$

where $\mathbf{w} = (w_1, \ldots, w_m)$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$. We then solve this nested minimax problem in Algorithm 4 by using an minimization *oracle* to perform a full optimization of ℓ over (θ, μ), and carrying out gradient ascent updates on ℓ over w_j.

We now proceed to show an optimality guarantee for Algorithm 4.

Theorem 8 (**Optimality Guarantee for Algorithm 4**). *Suppose for every $\theta \in \Theta$, there exists a $\widetilde{w}_j \in \mathcal{W}_\Delta$ such that $r_u(\theta, \widetilde{w}_j) \leq -\gamma$, $\forall u \in [U]$, for some $\gamma > 0$. Let $0 \leq g_j(\theta, w_j) \leq B'$, $\forall \theta \in \Theta, w_j \in \mathcal{W}_\Delta$. Let $B_{\mathbf{w}} \geq \max_q \|\nabla_{\mathbf{w}}\, \ell(\theta^{(q)}, \boldsymbol{\mu}^{(q)}, \mathbf{w}^{(q)}; \lambda'))\|_2$. Let $\hat{\theta}$ be the stochastic classifier returned by Algorithm 4 when run for a given λ' for Q iterations, with the radius of the Lagrange multipliers $R' = B'/\gamma$ and learning rate $\eta_{\mathbf{w}} = \frac{R'}{B_{\mathbf{w}}\sqrt{T}}$. Then:*

$$\mathbb{E}_{\theta\sim\hat{\theta}}\left[\mathcal{L}(\theta, \lambda')\right] \leq \min_{\theta\in\Theta} \mathcal{L}(\theta, \lambda') + \mathcal{O}\left(\frac{1}{\sqrt{Q}}\right) + \kappa.$$

Before proving Theorem 8, we will find it useful to state the following lemma.

Lemma 10 (**Boundedness of Inner Lagrange Multipliers in equation (A.11)**). Suppose for every $\theta \in \Theta$, there exists a $\widetilde{w}_j \in \mathcal{W}$ such that $r_u(\theta, \widetilde{w}_j) \leq -\gamma$, $\forall u \in [U]$, for some $\gamma > 0$. Let $0 \leq g_j(\theta, w_j) \leq B'$, $\forall \theta \in \Theta, w_j \in \mathcal{W}_\Delta$. Let $\mathcal{M} = \{\mu_j \in \mathbb{R}_+^K \,|\, \|\mu_j\|_1 \leq R'\}$ with the radius of the Lagrange multipliers $R' = B'/\gamma$. Then we have for all $j \in \mathcal{G}$:

$$\max_{w_j\in\mathcal{W}_\Delta} \min_{\mu_j\in\mathcal{M}} \omega\big(\theta, \mu_j, w_j\big) = \max_{w_j\in\mathcal{W}_\Delta:\, r_u(\theta,w_j)\leq 0,\,\forall u} g_j(\theta, w_j).$$

Proof. For a given $j \in \mathcal{G}$, let $w_j^* \in \underset{w_j \in \mathcal{W}_\Delta :\, r_u(\theta, w_j) \le 0,\, \forall u}{\operatorname{argmax}} g_j(\theta, w_j)$. Then:

$$g_j(\theta, w_j^*) \;=\; \max_{w_j \in \mathcal{W}_\Delta} \min_{\mu_j \in \mathbb{R}_+^K} \omega\big(\theta, \mu_j, w_j\big), \tag{A.12}$$

where note that μ_j is minimized over all non-negative values. Since the ω is linear in both μ_j and w_j, we can interchange the min and max:

$$g_j(\theta, w_j^*) \;=\; \min_{\mu_j \in \mathbb{R}_+^K} \max_{w_j \in \mathcal{W}_\Delta} \omega\big(\theta, \mu_j, w_j\big).$$

We show below that the minimizer μ^* in the above problem is in fact bounded and present in $\mathcal{M}$.

$$\begin{aligned}
g_j(\theta, w_j^*) \;&=\; \max_{w_j \in \mathcal{W}} \omega\big(\theta, \mu_j^*, w_j\big)\\
&=\; \max_{w_j \in \mathcal{W}} \left\{ g_j(\theta, w_j) \,-\, \sum_{k=1}^{K} \mu_{j,k}^*\, r_k(\theta, w_j) \right\}\\
&\ge\; g_j(\theta, \widetilde{w}_j) \,-\, \|\mu_j^*\|_1 \max_{k \in [K]} r_k(\theta, \widetilde{w}_j)\\
&\ge\; g_j(\theta, w_j) \,+\, \|\mu_j^*\|_1 \gamma \;\ge\; \|\mu_j^*\|_1 \gamma.
\end{aligned}$$

We further have:

$$\|\mu_j^*\|_1 \;\le\; g_j(\theta, w_j)/\gamma \;\le\; B'/\gamma. \tag{A.13}$$

Thus the minimizer $\mu_j^* \in \mathcal{M}$. So the minimization in equation (A.12) can be performed over only $\mathcal{M}$, which completes the proof of the lemma. □

Equipped with the above result, we are now ready to prove Theorem 8.

Proof of Theorem 8. Let $\overline{w}_j = \frac{1}{Q}\sum_{q=1}^{Q} w_j^{(q)}$. The best response on θ and μ gives us:

$$\begin{aligned}
&\frac{1}{Q}\sum_{q=1}^{Q} \Big(f(\theta^{(q)}) \,+\, \sum_{j=1}^{m} \lambda'_j\, \omega\big(\theta^{(q)}, \mu_j^{(q)}, w_j^{(q)}\big) \Big)\\
&\le\; \frac{1}{Q}\sum_{q=1}^{Q} \min_{\theta \in \Theta,\, \boldsymbol{\mu} \in \mathcal{M}^m} \Big(f(\theta) \,+\, \sum_{j=1}^{m} \lambda'_j\, \omega\big(\theta, \mu_j, w_j^{(q)}\big) \Big) \,+\, \kappa\\
&=\; \frac{1}{Q}\sum_{q=1}^{Q} \Big(\min_{\theta \in \Theta} f(\theta) \,+\, \sum_{j=1}^{m} \lambda'_j \min_{\mu_j \in \mathcal{M}} \omega\big(\theta, \mu_j, w_j^{(q)}\big) \Big) \,+\, \kappa\\
&\qquad (j\text{-th summation term depends on } \mu_j \text{ alone})\\
&\le\; \min_{\theta \in \Theta} \frac{1}{Q}\sum_{q=1}^{Q} \Big(f(\theta) \,+\, \sum_{j=1}^{m} \lambda'_j \min_{\mu_j \in \mathcal{M}} \omega\big(\theta, \mu_j, w_j^{(q)}\big) \Big) \,+\, \kappa
\end{aligned}$$

$$\begin{aligned}
&\le \min_{\theta\in\Theta}\Big\{f(\theta) + \sum_{j=1}^{m}\lambda'_j \min_{\mu_j\in\mathcal{M}} \frac{1}{Q}\sum_{q=1}^{Q}\omega\big(\theta,\mu_j,w_j^{(q)}\big)\Big\} + \kappa \\
&= \min_{\theta\in\Theta}\Big\{f(\theta) + \sum_{j=1}^{m}\lambda'_j \min_{\mu_j\in\mathcal{M}} \omega\big(\theta,\mu_j,\overline{w}_j\big)\Big\} + \kappa \\
&\le \min_{\theta\in\Theta}\Big\{f(\theta) + \sum_{j=1}^{m}\lambda'_j \max_{w_j\in\mathcal{W}}\min_{\mu_j\in\mathcal{M}} \omega\big(\theta,\mu_j,w_j\big)\Big\} + \kappa \\
&\quad\text{(by linearity of } \omega \text{ in } w_j) \\
&= \min_{\theta\in\Theta}\Big\{f(\theta) + \sum_{j=1}^{m}\lambda'_j \max_{w_j:\, r_u(\theta,w_j)\le 0,\,\forall u} g_j(\theta,w_j)\Big\} + \kappa \\
&\quad\text{(from Lemma 10)} \\
&= \min_{\theta\in\Theta}\mathcal{L}(\theta,\lambda') + \kappa. \qquad\qquad (A.14)
\end{aligned}$$

Applying standard gradient ascent analysis to the gradient ascent steps on $\mathbf{w}$ (using the fact that ω is linear in $\mathbf{w}$)

$$\begin{aligned}
&\frac{1}{Q}\sum_{q=1}^{Q}\Big(f(\theta^{(q)}) + \sum_{j=1}^{m}\lambda'_j\,\omega\big(\theta^{(q)},\mu_j^{(q)},w_j^{(q)}\big)\Big) \\
&\ge \max_{\mathbf{w}\in\mathcal{W}_\Delta^m}\frac{1}{Q}\sum_{q=1}^{Q}\Big(f(\theta^{(q)}) + \sum_{j=1}^{m}\lambda'_j\,\omega\big(\theta^{(q)},\mu_j^{(q)},w_j\big)\Big) - \mathcal{O}\left(\frac{1}{\sqrt{Q}}\right) \\
&= \frac{1}{Q}\sum_{q=1}^{Q}\Big(f(\theta^{(q)}) + \sum_{j=1}^{m}\lambda'_j \max_{w_j\in\mathcal{W}_\Delta}\omega\big(\theta^{(q)},\mu_j^{(q)},w_j\big)\Big) - \mathcal{O}\left(\frac{1}{\sqrt{Q}}\right) \\
&\quad\text{(j-th summation term depends on } w_j \text{ alone)} \\
&\ge \frac{1}{Q}\sum_{q=1}^{Q}\Big(f(\theta^{(q)}) + \sum_{j=1}^{m}\lambda'_j \max_{w_j\in\mathcal{W}_\Delta}\min_{\mu_j\in\mathcal{M}}\omega\big(\theta^{(q)},\mu_j,w_j\big)\Big) - \mathcal{O}\left(\frac{1}{\sqrt{Q}}\right) \\
&\quad\text{(by linearity of } \omega \text{ in } w_j \text{ and } \mu_j) \\
&= \mathbb{E}_{\theta\sim\hat{\theta}}\left[f(\theta) + \sum_{j=1}^{m}\lambda'_j \max_{w_j\in\mathcal{W}_\Delta}\min_{\mu_j\in\mathcal{M}}\omega\big(\theta,\mu_j,w_j\big)\right] - \mathcal{O}\left(\frac{1}{\sqrt{Q}}\right) \\
&= \mathbb{E}_{\theta\sim\hat{\theta}}\left[f(\theta^{(q)}) + \sum_{j=1}^{m}\lambda'_j \max_{w_j\in\mathcal{W}_\Delta:\, r_u(\theta,w_j)\le 0,\,\forall u} g_j(\theta,w_j)\right] - \mathcal{O}\left(\frac{1}{\sqrt{Q}}\right) \\
&\quad\text{(from Lemma 10)} \\
&= \mathbb{E}_{\theta\sim\hat{\theta}}\left[\mathcal{L}(\theta,\lambda')\right] - \mathcal{O}\left(\frac{1}{\sqrt{Q}}\right). \qquad\qquad (A.15)
\end{aligned}$$

Combining equation (A.14) and equation (A.15) completes the proof. □

Algorithm 5 *Practical* Algorithm

```
Require: learning rates $\eta_\theta > 0$, $\eta_\lambda > 0$, estimates of
    $P(G = j|\hat{G} = k)$ to specify $\mathcal{W}(\theta)$
for $t = 1, \dots, T$ do
    Solve for $w$ given $\theta$ using linear programming or a gradient method:
    $w^{(t)} \leftarrow \max_{w \in \mathcal{W}(\theta^{(t)})} \sum_{j=1}^{m} \lambda_j^{(t)} g_j(\theta^{(t)}, w)$
    Descent step on $\theta$:
    $\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta_\theta \delta_\theta^{(t)}$, where
    $\delta_\theta^{(t)} = \nabla_\theta \left( f_0(\theta^{(t)}) + \sum_{j=1}^{m} \lambda_j^{(t)} g_j \left( \theta^{(t)}, w^{(t+1)} \right) \right)$
    Ascent step on $\lambda$:
    $\tilde{\lambda}_j^{(t+1)} \leftarrow \lambda_j^{(t)} + \eta_\lambda g_j \left( \theta^{(t+1)}, w^{(t+1)} \right) \quad \forall j \in \mathcal{G}$
    $\lambda^{(t+1)} \leftarrow \Pi_\Lambda(\tilde{\lambda}^{(t+1)})$,
end for
return $\theta^{(t^*)}$ where $t^*$ denotes the best iterate that satisfies the constraints in (2.7) with
    the lowest objective.
```

A.5 Discussion of the *Practical* algorithm

Here we provide the details of the *practical* Algorithm 5 to solve problem (2.8). We also further discuss how we arrive at Algorithm 5. Recall that in the minimax problem in equation (2.8), restated below, each of the m constraints contain a max over w:

$$\min_{\theta \in \Theta} \max_{\lambda \in \Lambda} f(\theta) + \sum_{j=1}^{m} \lambda_j \max_{w \in \mathcal{W}(\theta)} g_j(\theta, w).$$

We show below that this is equivalent to a minimax problem where the sum over j and max over w are swapped:

Lemma 11. The minimax problem in equation (2.8) is equivalent to:

$$\min_{\theta \in \Theta} \max_{\lambda \in \Lambda} \max_{w \in \mathcal{W}(\theta)} f(\theta) + \sum_{j=1}^{m} \lambda_j g_j(\theta, w). \tag{A.16}$$

Proof. Recall that the space of Lagrange multipliers $\Lambda = \{\lambda \in \mathbb{R}_+^m \mid \|\lambda\|_1 \leq R\}$, for $R > 0$. So the above maximization over Λ can be re-written in terms of a maximization over the m-dimensional simplex Δ_m and a scalar $\beta \in [0, R]$:

$$\min_{\theta \in \Theta} \max_{\beta \in [0,R],\, \nu \in \Delta_m} f(\theta) + \beta \sum_{j=1}^{m} \nu_j \max_{w \in \mathcal{W}(\theta)} g_j(\theta, w)$$

$$
\begin{aligned}
&= \min_{\theta\in\Theta} \max_{\beta\in[0,R]} f(\theta) + \beta \max_{\nu\in\Delta_m} \sum_{j=1}^{m} \nu_j \max_{w\in\mathcal{W}(\theta)} g_j(\theta, w) \\
&= \min_{\theta\in\Theta} \max_{\beta\in[0,R]} f(\theta) + \beta \max_{j\in\mathcal{G}} \max_{w\in\mathcal{W}(\theta)} g_j(\theta, w) \\
&= \min_{\theta\in\Theta} \max_{\beta\in[0,R]} f(\theta) + \beta \max_{w\in\mathcal{W}(\theta)} \max_{j\in\mathcal{G}} g_j(\theta, w) \\
&= \min_{\theta\in\Theta} \max_{\beta\in[0,R]} f(\theta) + \beta \max_{w\in\mathcal{W}(\theta)} \max_{\nu\in\Delta_m} \sum_{j=1}^{m} \nu_j g_j(\theta, w) \\
&= \min_{\theta\in\Theta} f(\theta) + \max_{\beta\in[0,R],\, \nu\in\Delta_m} \max_{w\in\mathcal{W}(\theta)} \sum_{j=1}^{m} \beta\nu_j g_j(\theta, w) \\
&= \min_{\theta\in\Theta} f(\theta) + \max_{\lambda\in\Lambda} \max_{w\in\mathcal{W}(\theta)} \sum_{j=1}^{m} \lambda_j g_j(\theta, w),
\end{aligned}
$$

which completes the proof. □

The practical algorithm outlined in Algorithm 5 seeks to solve the re-written minimax problem in equation (A.16), and is similar in structure to the ideal algorithm in Algorithm 1, in that it has two high-level steps: an approximate best response over θ and gradient ascent updates on λ. However, the algorithm works with deterministic classifiers $\theta^{(t)}$, and uses a simple heuristic to approximate the best response step. Specifically, for the best response step, the algorithm finds the maximizer of the Lagrangian over w for a fixed $\theta^{(t)}$ by e.g. using linear programming:

$$
w^{(t)} \leftarrow \max_{w\in\mathcal{W}(\theta^{(t)})} \sum_{j=1}^{m} \lambda_j^{(t)} g_j(\theta^{(t)}, w),
$$

uses the maximizer $w^{(t)}$ to approximate the gradient of the Lagrangian at $\theta^{(t)}$:

$$
\delta_\theta^{(t)} = \nabla_\theta\Big(f_0(\theta^{(t)}) + \sum_{j=1}^{m} \lambda_j^{(t)} f_j \left(\theta^{(t)}, w^{(t+1)}\right) \Big)
$$

and performs a single gradient update on θ:

$$
\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta_\theta \delta_\theta^{(t)}.
$$

The gradient ascent step on λ is the same as the ideal algorithm, except that it is simpler to implement as the iterates $\theta^{(t)}$ are deterministic:

$$
\tilde{\lambda}_j^{(t+1)} \leftarrow \lambda_j^{(t)} + \eta_\lambda f_j\Big(\theta^{(t+1)}, w^{(t+1)}\Big) \;\; \forall j \in \mathcal{G};
$$

$$
\lambda^{(t+1)} \leftarrow \Pi_\Lambda(\tilde{\lambda}^{(t+1)}).
$$

A.6 Additional experiment details and results

We provide more details on the experimental setup as well as further results.

Additional experimental setup details

This section contains further details on the experimental setup, including the datasets used and hyperparameters tuned. All categorical features in each dataset were binarized into one-hot vectors. All numerical features were bucketized into 4 quantiles, and further binarized into one-hot vectors. All code that we used for pre-processing the datasets from their publicly-downloadable versions can be found at https://github.com/wenshuoguo/robust-fairness-code.

For the naïve approach, we solve the constrained optimization problem (2.2) with respect to the noisy groups $\hat{G}$. For comparison, we also report the results of the unconstrained optimization problem and the constrained optimization problem (2.1) when the true groups G are known. For the DRO problem (2.3), we estimate the bound $\gamma_j = P(\hat{G} \neq G | G = j)$ in each case study. For the soft group assignments approach, we implement the *practical* algorithm (Algorithm 5).

In the experiments, we replace all expectations in the objective and constraints with finite-sample empirical versions. So that the constraints will be convex and differentiable, we replace all indicator functions with hinge upper bounds, as in Davenport et al. [2010] and Eban et al. [2017]. We use a linear model: $\phi(X;\theta) = \theta^T X$. The noisy protected groups $\hat{G}$ are included as a feature in the model, demonstrating that conditional independence between $\hat{G}$ and the model $\phi(X;\theta)$ is not required here, unlike some prior work [Awasthi et al., 2020]. Aside from being used to estimate the noise model $P(G = k|\hat{G} = j)$ for the soft group assignments approach[2], the true groups G are never used in the training or validation process.

Each dataset was split into train/validation/test sets with proportions 0.6/0.2/0.2. For each algorithm, we chose the *best* iterate $\theta^{(t^*)}$ out of T iterates on the train set, where we define *best* as the iterate that achieves the lowest objective value while satisfying all constraints. We select the hyperparameters that achieve the best performance on the validation set (details in Appendix A.6). We repeat this procedure for ten random train/validation/test splits and record the mean and standard errors for all metrics[3].

Adult dataset

For the first case study, we used the Adult dataset from UCI [Dua and Graff, 2017], which includes 48,842 examples. The features used were *age*, *workclass*, *fnlwgt*, *education*, *educa-*

[2] If $P(G = k|\hat{G} = j)$ is estimated from an auxiliary dataset with a different distribution than test, this could lead to generalization issues for satisfying the true group constraints on test. In our experiments, we lump those generalization issues in with any distributional differences between train and test.

[3] When we report the "maximum" constraint violation, we use the mean and standard error of the constraint violation for the group j with the maximum mean constraint violation.

tion_num, *marital_status*, *occupation*, *relationship*, *race*, *gender*, *capital_gain*, *capital_loss*, *hours_per_week*, and *native_country*. Detailed descriptions of what these features represent are provided by UCI [Dua and Graff, 2017]. The label was whether or not *income_bracket* was above $50,000. The true protected groups were given by the *race* feature, and we combined all examples with race other than "white" or "black" into a group of race "other." When training with the noisy group labels, we did *not* include the true *race* as a feature in the model, but included the noisy race labels as a feature in the model instead. We set $\alpha = 0.05$ as the constraint slack.

The constraint violation that we report in Figure 2.1 is taken over a test dataset with n examples $(X_1, Y_1, G_1), ..., (X_n, Y_n, G_n)$, and is given by:

$$\max_{j \in \mathcal{G}} \quad \frac{\sum_{i=1}^n \mathbb{1}(\hat{Y}(\theta)_i = 1, Y_i = 1)}{\sum_{i=1}^n \mathbb{1}(Y_i = 1)} - \frac{\sum_{i=1}^n \mathbb{1}(\hat{Y}(\theta)_i = 1, Y_i = 1, G_i = j)}{\sum_{i=1}^n \mathbb{1}(Y_i = 1, G_i = j)} - \alpha,$$

where $\hat{Y}(\theta)_i = \mathbb{1}(\phi(\theta; X_i) > 0)$.

Section A.3 shows how we specifically enforce equality of opportunity using the soft assignments approach, and Section A.2 shows how we enforce equality of opportunity using DRO.

Credit dataset

For the second case study, we used default of credit card clients dataset from UCI Dua and Graff [2017] collected by a company in Taiwan Yeh and hui Lien [2009], which contains 30000 examples and 24 features. The features used were *amount_of_the_given_credit*, *gender*, *education*, *education*, *marital_status*, *age*, *history_of_past_payment*, *amount_of_bill_statement*, *amount_of_previous_payment*. Detailed descriptions of what these features represent are provided by UCI [Dua and Graff, 2017]. The label was whether or not *default* was true. The true protected groups were given by the *education* feature, and we combined all examples with education level other than "graduate school" or "university" into a group of education level "high school and others". When training with the noisy group labels, we did *not* include the true *education* as a feature in the model, but included the noisy education level labels as a feature in the model instead. We set $\alpha = 0.03$ as the constraint slack.

The constraint violation that we report in Figure 2.1 is taken over a test dataset with n examples $(X_1, Y_1, G_1), ..., (X_n, Y_n, G_n)$, and is given by:

$$\max_{j \in \mathcal{G}} \quad \max(\Delta_j^{\text{TPR}}, \Delta_j^{\text{FPR}})$$

where

$$\Delta_j^{\text{TPR}} = \frac{\sum_{i=1}^n \mathbb{1}(\hat{Y}(\theta)_i = 1, Y_i = 1)}{\sum_{i=1}^n \mathbb{1}(Y_i = 1)} - \frac{\sum_{i=1}^n \mathbb{1}(\hat{Y}(\theta)_i = 1, Y_i = 1, G_i = j)}{\sum_{i=1}^n \mathbb{1}(Y_i = 1, G_i = j)} - \alpha$$

and

$$\Delta_j^{\text{FPR}} = \frac{\sum_{i=1}^n \mathbb{1}(\hat{Y}(\theta)_i = 1, Y_i = 0, G_i = j)}{\sum_{i=1}^n \mathbb{1}(Y_i = 0, G_i = j)} - \frac{\sum_{i=1}^n \mathbb{1}(\hat{Y}(\theta)_i = 1, Y_i = 0)}{\sum_{i=1}^n \mathbb{1}(Y_i = 0)} - \alpha$$

and $\hat{Y}(\theta)_i = \mathbb{1}(\phi(\theta; X_i) > 0)$.

Section A.3 shows how we specifically enforce equalized odds using the soft assignments approach, and Section A.2 shows how we enforce equalized odds using DRO.

Optimization code

For all case studies, we performed experiments comparing the naïve approach, the DRO approach (Section 2.5) and the soft group assignments approach (Section 2.6). We also compared these to the baselines of optimizing without constraints and optimizing with constraints with respect to the true groups. All optimization code was written in Python and TensorFlow [4]. All gradient steps were implemented using TensorFlow's Adam optimizer [5], though all experiments can also be reproduced using simple gradient descent without momentum. We computed full gradients over all datasets, but minibatching can also be used for very large datasets. Implementations for all approaches are included in the attached code. Training time was less than 10 minutes per model.

Table A.1: Hyperparameters tuned for each approach

Hparam	Values tried	Relevant approaches	Description
η_θ	{0.001,0.01,0.1}	all approaches	learning rate for θ
η_λ	{0.25,0.5,1.0,2.0}	all except unconstrained	learning rate for λ
$\eta_{\tilde{p}_j}$	{0.001, 0.01, 0.1}	DRO	learning rate for $\tilde{p}_j$
η_w	{0.001, 0.01, 0.1}	soft assignments	learning rate using gradient methods for w

Hyperparameters

The hyperparameters for each approach were chosen to achieve the best performance on the validation set on average over 10 random train/validation/test splits, where "best" is defined as the set of hyperparameters that achieved the lowest error rate while satisfying all constraints relevant to the approach. The final hyperparameter values selected for each method were neither the largest nor smallest of all values tried. A list of all hyperparameters tuned and the values tried is given in Table A.1.

For the naïve approach, the constraints used when selecting the hyperparameter values on the validation set were the constraints with respect to the noisy group labels given in Equation (2.2). For the DRO approach and the soft group assignments approach, the respective robust

[4] Abadi, M. et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. tensorflow.org.

[5] https://www.tensorflow.org/api_docs/python/tf/compat/v1/train/AdamOptimizer

constraints were used when selecting hyperparameter values on the validation set. Specifically, for the DRO approach, the constraints used were those defined in Equation (2.3), and for the soft group assignments approach, the constraints used were those defined in Equation (2.7). For the unconstrained baseline, no constraints were taken into account when selecting the best hyperparameter values. For the baseline constrained with access to the true group labels, the true group constraints were used when selecting the best hyperparameter values.

Hinge relaxations of all constraints were used during training to achieve convexity. Since the hinge relaxation is an upper bound on the real constraints, the hinge-relaxed constraints may require some additional slack to maintain feasibility. This positive slack β was added to the original slack α when training with the hinge-relaxed constraints, and the amount of slack β was chosen so that the relevant hinge-relaxed constraints were satisfied on the training set.

All approaches ran for 750 iterations over the full dataset.

Additional experiment results

This section provides additional experiment results. All results reported here and in the main paper are on the test set (averaged over 10 random train/validation/test splits).

Case study 1 (Adult)

This section provides additional experiment results for case study 1 on the Adult dataset.

Figure A.1 shows that the naïve approach, DRO approach, and soft assignments approaches all satisfied the fairness constraints for the noisy groups on the test set.

Figure A.2 confirms that the DRO approach and the soft assignments approaches both managed to satisfy their respective robust constraints on the test set on average. For the DRO approach, the constraints measured in Figure A.2 come from Equation (2.3), and for the soft assignments approach, the constraints measured in Figure A.2 come from Equation (2.7). We provide the exact error rate values and maximum violations on the true groups for the Adult dataset in Table A.2.

Case study 2 (Credit)

This section provides additional experiment results for case study 2 on the Credit dataset.

Figure A.3 shows the constraint violations with respect to the true groups on test separated into TPR violations and FPR violations. For all noise levels, there were higher TPR violations than FPR violations. However, this does not mean that the FPR constraint was meaningless – the FPR constraint still ensured that the TPR constraints weren't satisfied by simply adding false positives.

Figure A.4 confirms that the naïve approach, DRO approach, and soft assignments approaches all satisfied the fairness constraints for the noisy groups on the test set.

Figure A.5 confirms that the DRO approach and the soft assignments approaches both managed to satisfy their respective robust constraints on the test set on average. For the

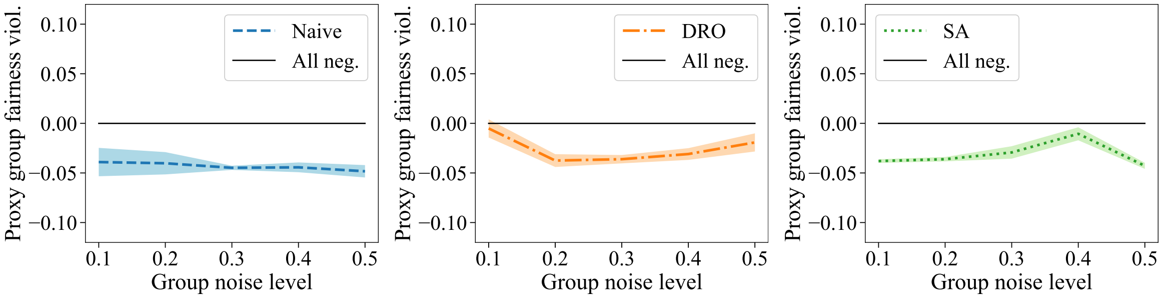


Figure A.1: Maximum fairness constraint violations with respect to the noisy groups $\hat{G}$ on the test set for different group noise levels γ on the Adult dataset. For each noise level, we plot the mean and standard error over 10 random train/val/test splits. The black solid line illustrates a maximum constraint violation of 0. While the naïve approach (*left*) has increasingly higher fairness constraints with respect to the true groups as the noise increases, it always manages to satisfy the constraints with respect to the noisy groups $\hat{G}$

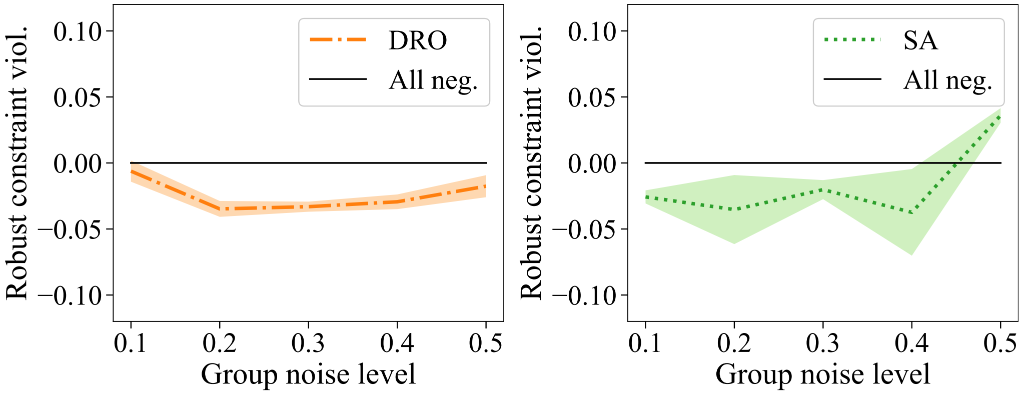


Figure A.2: Maximum robust constraint violations on the test set for different group noise levels $P(\hat{G} \neq G)$ on the Adult dataset. For each noise level, we plot the mean and standard error over 10 random train/val/test splits. The black dotted line illustrates a maximum constraint violation of 0. Both the DRO approach (*left*) and the soft group assignments approach (*right*) managed to satisfy their respective robust constraints on the test set on average for all noise levels.

DRO approach, the constraints measured in Figure A.5 come from Equation (2.3), and for the soft assignments approach, the constraints measured in Figure A.5 come from Equation (2.7).

We provide the exact error rate values and maximum violations on the true groups for

Table A.2: Error rate and fairness constraint violations on the true groups for the Adult dataset (mean and standard error over 10 train/test/splits).

	DRO		Soft Assignments	
Noise	Error rate	Max G Viol.	Error rate	Max G Viol.
0.1	0.152 ± 0.001	0.002 ± 0.019	0.148 ± 0.001	-0.048 ± 0.002
0.2	0.200 ± 0.002	-0.045 ± 0.003	0.157 ± 0.003	-0.048 ± 0.002
0.3	0.216 ± 0.010	-0.044 ± 0.004	0.158 ± 0.005	0.002 ± 0.030
0.4	0.209 ± 0.006	-0.019 ± 0.031	0.188 ± 0.003	-0.016 ± 0.016
0.5	0.219 ± 0.012	-0.030 ± 0.032	0.218 ± 0.002	0.004 ± 0.006

the Credit dataset in Table A.3.

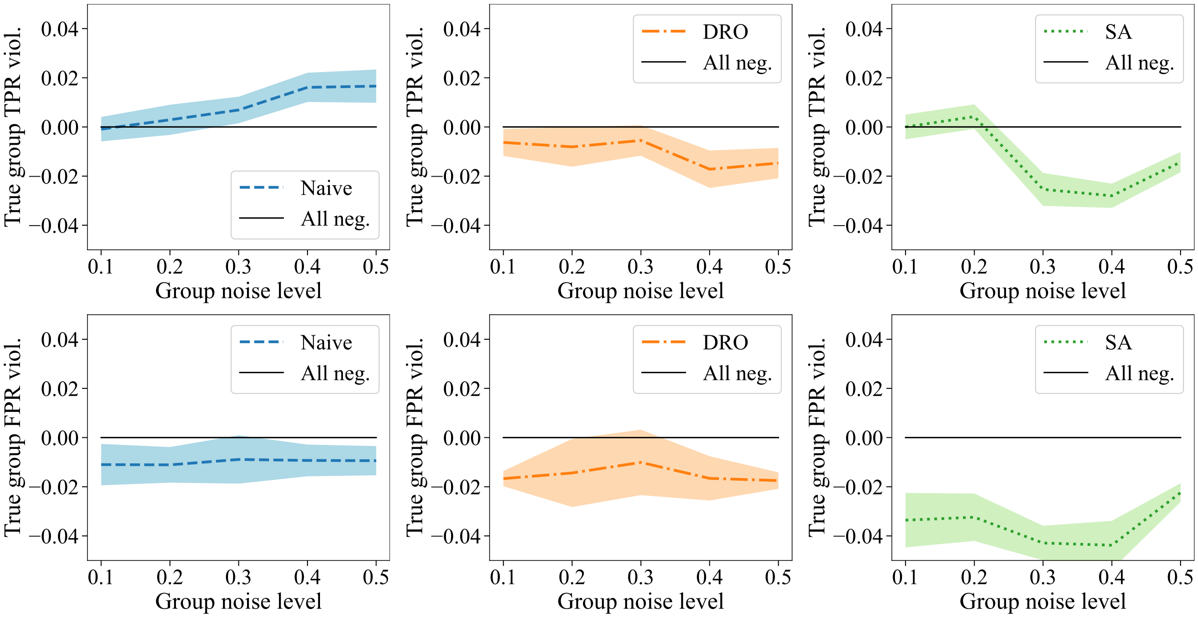


Figure A.3: Case study 2 (Credit): Maximum true group TPR (top) and FPR (bottom) constraint violations for the Naive, DRO, and soft assignments (SA) approaches on test set for different group noise levels γ on the Credit dataset (mean and standard error over 10 train/val/test splits). The black solid line represents the performance of the trivial "all negatives" classifier, which has constraint violations of 0. A negative violation indicates satisfaction of the fairness constraints on the true groups.

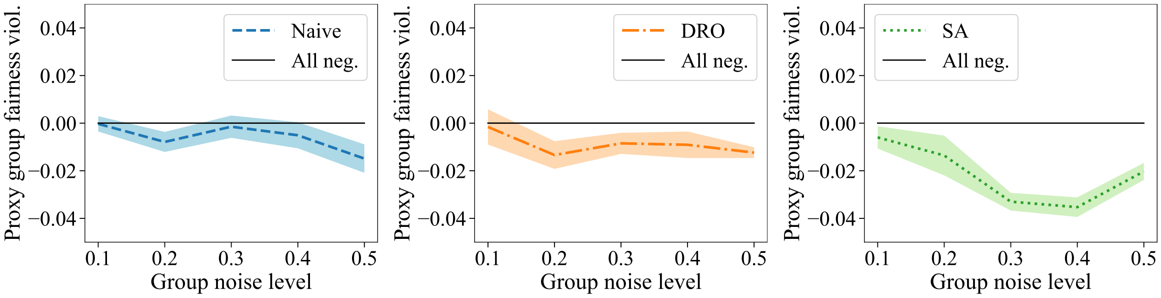


Figure A.4: Maximum fairness constraint violations with respect to the noisy groups $\hat{G}$ on the test set for different group noise levels γ on the Credit dataset. For each noise level, we plot the mean and standard error over 10 random train/val/test splits. The black solid line illustrates a maximum constraint violation of 0. While the naïve approach (*left*) has increasingly higher fairness constraints with respect to the true groups as the noise increases, it always manages to satisfy the constraints with respect to the noisy groups $\hat{G}$

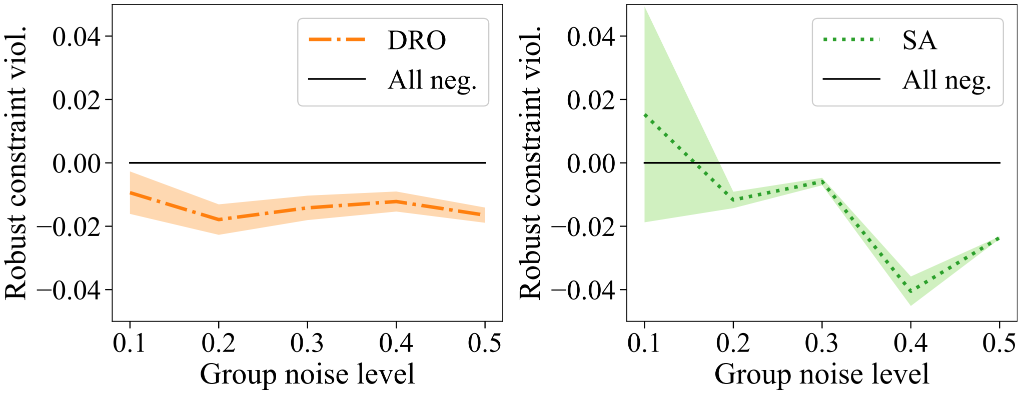


Figure A.5: Maximum robust constraint violations on the test set for different group noise levels $P(\hat{G} \neq G)$ on the Credit dataset. For each noise level, we plot the mean and standard error over 10 random train/val/test splits. The black dotted line illustrates a maximum constraint violation of 0. Both the DRO approach (*left*) and the soft group assignments approach (*right*) managed to satisfy their respective robust constraints on the test set on average for all noise levels.

Table A.3: Error rate and fairness constraint violations on the true groups for the Credit dataset (mean and standard error over 10 train/test/splits).

	DRO		Soft Assignments	
Noise	Error rate	Max G Viol.	Error rate	Max G Viol.
0.1	0.206 ± 0.003	-0.006 ± 0.006	0.182 ± 0.002	0.000 ± 0.005
0.2	0.209 ± 0.002	-0.008 ± 0.008	0.182 ± 0.001	0.004 ± 0.005
0.3	0.212 ± 0.002	-0.006 ± 0.006	0.198 ± 0.001	-0.025 ± 0.007
0.4	0.210 ± 0.002	-0.017 ± 0.008	0.213 ± 0.001	-0.028 ± 0.005
0.5	0.211 ± 0.003	-0.015 ± 0.006	0.211 ± 0.001	-0.014 ± 0.004

Appendix B

Deferred Proofs and Discussion for Chapter 3

B.1 Proofs

Proof of Theorem 2

(i) The first result follows from the fact that the cross-entropy loss is a proper composite loss [Williamson et al., 2016] with the softmax function as the associated (inverse) link function.

(ii) For a proof of the second result, please see Menon et al. [2021b].

(iii) Below, we provide a proof for the third result.

The minimization of the robust objective in equation (3.3) over f can be re-written as a min-max optimization problem:

$$\min_{f:\mathcal{X}\to\mathbb{R}^m} L^{\text{rob}}(f) = \min_{f:\mathcal{X}\to\mathbb{R}^m} \max_{\lambda\in\Delta_m} \underbrace{\sum_{y=1}^{m} \frac{\lambda_y}{\pi_y} \mathbb{E}\left[\eta_y(X)\,\ell(y, f(X))\right]}_{\omega(\lambda, f)}. \tag{B.1}$$

The min-max objective $\omega(\lambda, f)$ is clearly linear in λ (for fixed f) and with ℓ chosen to be the cross-entropy loss, is convex in f (for fixed λ), i.e., $\omega(\lambda, \kappa f_1 + (1-\kappa) f_2) \leq \kappa\omega(\lambda, f_1) + (1-\kappa)\omega(\lambda, f_2), \forall f_1, f_2 : \mathcal{X} \to \mathbb{R}^m, \kappa \in [0, 1]$. Furthermore, Δ_m is a convex compact set, while the domain of f is convex. It follows from Sion's minimax theorem [Sion, 1958] that:

$$\min_{f:\mathcal{X}\to\mathbb{R}^m} \max_{\lambda\in\Delta_m} \omega(\lambda, f) = \max_{\lambda\in\Delta_m} \min_{f:\mathcal{X}\to\mathbb{R}^m} \omega(\lambda, f). \tag{B.2}$$

Let (λ^*, f^*) be such that:

$$\lambda^* \in \operatorname*{argmax}_{\lambda\in\Delta_m} \min_{f:\mathcal{X}\to\mathbb{R}^m} \omega(\lambda, f); \qquad f^* \in \operatorname*{argmin}_{f:\mathcal{X}\to\mathbb{R}^m} \max_{\lambda\in\Delta_m} \omega(\lambda, f),$$

Such a λ^* exists for the following reason: for any fixed $\lambda \in \Delta_m$, owing to the use of the cross-entropy loss, a minimizer over always exists for $\omega(\lambda, f)$, and is given by $f_y(x) = \log\left(\frac{\lambda_y}{\pi_y}\eta_y(x)\right) + C$, for some $C \in \mathbb{R}$; therefore $\min_{f:\mathcal{X}\to\mathbb{R}^m} \omega(\lambda, f)$ is bounded above for any λ, and Δ_m being compact set gives us there exits a maximizer λ^* over this set. Similarly, such an f^* exists for the following reason: the objective $\max_{\lambda\in\Delta_m} \omega(\lambda, f)$ takes a bounded value when $f = \eta$, and any minimizer of $\max_{\lambda\in\Delta_m} \omega(\lambda, f)$ yields a value below that; because $\omega(\lambda, f) \geq 0$ and is convex in f, the minimizer f^* exits.

We then have from equation (B.2):

$$\begin{aligned}\omega(\lambda^*, f^*) &\leq \max_{\lambda\in\Delta_m} \omega(\lambda, f^*)\\ &= \min_{f:\mathcal{X}\to\mathbb{R}^m} \max_{\lambda\in\Delta_m} \omega(\lambda, f) \;=\; \max_{\lambda\in\Delta_m} \min_{f:\mathcal{X}\to\mathbb{R}^m} \omega(\lambda, f)\\ &= \min_{f:\mathcal{X}\to\mathbb{R}^m} \omega(\lambda^*, f) \;\leq\; \omega(\lambda^*, f^*),\end{aligned}$$

which tells us that there exists (λ^*, f^*) is a saddle-point for equation (B.1), i.e.,

$$\omega(\lambda^*, f^*) = \max_{\lambda\in\Delta_m} \omega(\lambda, f^*) \;=\; \min_{f:\mathcal{X}\to\mathbb{R}^m} \omega(\lambda^*, f).$$

Consequently, we have:

$$L^{\text{rob}}(f^*) = \max_{\lambda\in\Delta_m} \omega(\lambda, f^*) \;=\; \min_{f:\mathcal{X}\to\mathbb{R}^m} \max_{\lambda\in\Delta_m} \omega(\lambda, f) \;=\; \min_{f:\mathcal{X}\to\mathbb{R}^m} L^{\text{rob}}(f).$$

We thus have that f^* is a minimizer of $L^{\text{rob}}(f)$. Furthermore, because f^* is also a minimizer of $\omega(\lambda^*, f)$ over f, i.e.,

$$f^* \in \operatorname*{argmin}_{f:\mathcal{X}\to\mathbb{R}^m} \sum_{y=1}^{m} \frac{\lambda^*_y}{\pi_y}\mathbb{E}\left[\eta_y(X)\, \ell(y, f(X))\right],$$

it follows that:

$$\text{softmax}_y(f^*(x)) \;\propto\; \frac{\lambda^*_y}{\pi_y}\eta_y(x).$$

(iv) For the fourth result, we expand the traded-off objective, and re-write it as:

$$\begin{aligned}L^{\text{tdf}}(f) &= (1-\alpha)L^{\text{bal}}(f) + \alpha L^{\text{rob}}(f)\\ &= (1-\alpha)\frac{1}{m}\sum_{y=1}^{m}\frac{1}{\pi_y}\mathbb{E}\left[\eta_y(X)\, \ell(y, f(X))\right] \;+\; \alpha \max_{\lambda\in\Delta_m}\sum_{y=1}^{m}\frac{\lambda_y}{\pi_y}\mathbb{E}\left[\eta_y(X)\, \ell(y, f(X))\right]\\ &= \max_{\lambda\in\Delta_m} \underbrace{\sum_{y=1}^{m}\left((1-\alpha)\frac{1}{m} + \alpha\lambda_y\right)\frac{1}{\pi_y}\mathbb{E}\left[\eta_y(X)\, \ell(y, f(X))\right].}_{\omega(\lambda, f)}\end{aligned}$$

For a fixed λ, $\omega(\lambda, f)$ is convex in f (as the loss ℓ is the cross-entropy loss), and for a fixed f, $\omega(\lambda, f)$ is linear in λ. Following the same steps as the proof of (iii), we have that there exists (λ^*, f^*) such that

$$L^{\mathrm{tdf}}(f^*) \;=\; \max_{\lambda\in\Delta_m} \,\omega(\lambda, f^*) \;=\; \min_{f:\mathcal{X}\to\mathbb{R}^m} \,L^{\mathrm{tdf}}(f),$$

and

$$f^* \in \operatorname*{argmin}_{f:\mathcal{X}\to\mathbb{R}^m} \sum_{y=1}^{m} \left((1-\alpha)\frac{1}{m} + \alpha\lambda^*_y\right)\frac{1}{\pi_y}\mathbb{E}\left[\eta_y(X)\,\ell(y, f(X))\right],$$

which, owing to the properties of the cross-entropy loss, then gives us the desired form for f^*.

Proof of Theorem 3

Proof. Expanding the left-hand side, we have:

$$\begin{aligned}
|\hat{L}^{\text{rob-d}}(f) - L^{\text{rob}}(f)| &\leq |\hat{L}^{\text{rob-d}}(f) - L^{\text{rob-d}}(f) + L^{\text{rob-d}}(f) - L^{\text{rob}}(f)|\\
&\leq |\hat{L}^{\text{rob-d}}(f) - L^{\text{rob-d}}(f)| + |L^{\text{rob-d}}(f) - L^{\text{rob}}(f)|\\
&= |\hat{L}^{\text{rob-d}}(f) - L^{\text{rob-d}}(f)| + \left|\max_{y\in[m]} \frac{\mathbb{E}_x\left[p^t_y(x)\,\ell(y, f(x))\right]}{\mathbb{E}_x\left[p^t_y(x)\right]} - \max_{y\in[m]} \frac{\mathbb{E}_x\left[\eta_y(x)\,\ell(y, f(x))\right]}{\pi_y}\right|\\
&\leq |\hat{L}^{\text{rob-d}}(f) - L^{\text{rob-d}}(f)| + \max_{y\in[m]}\left|\frac{\mathbb{E}_x\left[p^t_y(x)\,\ell(y, f(x))\right]}{\mathbb{E}_x\left[p^t_y(x)\right]} - \frac{\mathbb{E}_x\left[\eta_y(x)\,\ell(y, f(x))\right]}{\pi_y}\right|\\
&\leq |\hat{L}^{\text{rob-d}}(f) - L^{\text{rob-d}}(f)| + B\max_{y\in[m]}\mathbb{E}_x\left[\left|\frac{p^t_y(x)}{\mathbb{E}_x\left[p^t_y(x)\right]} - \frac{\eta_y(x)}{\pi_y}\right|\ell(y, f(x))\right]\\
&\leq |\hat{L}^{\text{rob-d}}(f) - L^{\text{rob-d}}(f)| + B\max_{y\in[m]}\mathbb{E}_x\left[\left|\frac{p^t_y(x)}{\mathbb{E}_x\left[p^t_y(x)\right]} - \frac{\eta_y(x)}{\pi_y}\right|\right],
\end{aligned}$$

where the second-last step uses Jensen's inequality and the fact that $\ell(y, f(x)) \geq 0$, and the last step uses the fact that $\ell(y, f(x)) \leq B$.

Further expanding the first term,

$$\begin{aligned}
|\hat{L}^{\text{rob-d}}(f) - L^{\text{rob}}(f)| &\leq \left|\max_{y\in[m]}\phi_y(f) \;-\; \max_{y\in[m]}\hat{\phi}_y(f)\right| + B\max_{y\in[m]}\mathbb{E}_x\left[\left|\frac{p^t_y(x)}{\mathbb{E}_x\left[p^t_y(x)\right]} - \frac{\eta_y(x)}{\pi_y}\right|\right]\\
&\leq \max_{y\in[m]}\left|\phi_y(f) \;-\; \hat{\phi}_y(f)\right| + B\max_{y\in[m]}\mathbb{E}_x\left[\left|\frac{p^t_y(x)}{\mathbb{E}_x\left[p^t_y(x)\right]} - \frac{\eta_y(x)}{\pi_y}\right|\right],
\end{aligned}$$

as desired. □

Calibration of Margin-based Loss $\mathcal{L}^{\text{mar}}$

To show that minimizer of the margin-based objective in equation (3.9) also minimizes the balanced objective in equation (3.6), we state the following general result:

Lemma 12. Suppose $p^t \in \mathcal{F}$ and $\mathcal{F}$ is closed under linear transformations. Let

$$\hat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}^{\text{mar}}\left(p^t(x_i), f(x_i); \mathbf{c}\right) \tag{B.3}$$

for some cost vector $\mathbf{c} \in \mathbb{R}_+^m$. Then:

$$\hat{f}_y(x_i) = \log\left(c_y p_y^t(x_i)\right) + C_i, \quad \forall i \in [n],$$

for some example-specific constant constants $C_i \in \mathbb{R}, \forall i \in [n]$. Furthermore, for any assignment of example weights of $w \in \mathbb{R}_+^n$, $\hat{f}$ is also the minimizer of the weighted objective:

$$\hat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} w_i \sum_{y=1}^{m} c_y \, p_y^t(x_i) \, \ell\left(y, f(x_i)\right). \tag{B.4}$$

Proof. Following Menon et al. [2021b] (e.g. proof of Theorem 1), we have that for class probabilities $\mathbf{p} \in \Delta_m$ and costs $\mathbf{c} \in \mathbb{R}_+^m$, the margin-based loss in equation (3.9)

$$\mathcal{L}^{\text{mar}}\left(\mathbf{p}, \mathbf{f}; \mathbf{c}\right) = \frac{1}{m} \sum_{y \in [m]} p_y \log\left(1 + \sum_{j \neq y} \exp\left(\log(c_y/c_j) \; - \; (f_y - f_j)\right)\right).$$

is minimized by:

$$f_y^* = \log\left(c_y p_y\right) + C,$$

for any $C > 0$. To see why this is true, note that the above loss can be equivalently written as:

$$\mathcal{L}^{\text{mar}}\left(\mathbf{p}, \mathbf{f}; \mathbf{c}\right) = -\frac{1}{m} \sum_{y \in [m]} p_y \log\left(\frac{\exp\left(f_y - \log(c_y)\right)}{\sum_{j=1}^{m} \exp\left(f_j - \log(c_j)\right)}\right).$$

This the same as the softmax cross-entropy loss with adjustments made to the logits, the minimizer for which is of the form:

$$f_y^* - \log(c_y) = \log\left(p_y\right) + C \quad \text{ or } \quad f_y^* = \log\left(c_y p_y\right) + C.$$

It follows that any minimizer $\hat{f}$ of the average margin-based loss in equation (B.3) over sample S, would do so point-wise, and therefore

$$\hat{f}_y(x_i) = \log\left(c_y p_y^t(x_i)\right) + C_i, \quad \forall i \in [n],$$

for some example-specific constant constants $C_i \in \mathbb{R}, \forall i \in [n]$.

To prove the second part, we note that for the minimizer $\hat{f}$ to also minimize the weighted objective:

$$\frac{1}{n}\sum_{i=1}^{n} w_i \sum_{y=1}^{m} c_y \, p_y^t(x_i) \, \ell\left(y, f(x_i)\right),$$

it would also have to do so point-wise for each $i \in [m]$, and so as long the weights w_i are non-negative, it suffices that

$$\hat{f}(x_i) \in \underset{\mathbf{f} \in \mathbb{R}^m}{\operatorname{argmin}} \sum_{y=1}^{m} c_y \, p_y^t(x_i) \, \ell\left(y, f(x_i)\right).$$

This is indeed the case when ℓ is the softmax cross-entropy loss, where the point-wise minimizer for each $i \in [m]$ would be of the form $\text{softmax}_y(f(x)) = c_y p_y^t(x)$, which is satisfied by $\hat{f}$. □

A similar result also holds in the population limit, when equation (B.3) and equation (B.4) are computed in expectation, and the per-example weighting in equation (B.4) is replaced by an arbitrary weighting function $w(x) \in \mathbb{R}_+$. Any scorer of the following form would then minimize both objectives:

$$\hat{f}_y(x) = \log\left(c_y p_y^t(x)\right) + C(x), \quad \forall x \in \mathcal{X},$$

where $C(x)$ is some example-specific constant.

Proof of Proposition 1

Proposition 1. *Suppose $p^t \in \mathcal{F}$ and $\mathcal{F}$ is closed under linear transformations. Then the final scoring function $\overline{f}^s(x) = \frac{1}{K}\sum_{k=1}^{K} f^k(x)$ output by Algorithm 2 is of the form:*

$$\text{softmax}_j(\overline{f}^s(x)) \propto \overline{\lambda}_j p_j^t(x), \qquad \forall j \in [m], \ \forall (x, y) \in S,$$

where $\overline{\lambda}_y = \left(\prod_{k=1}^{K} \lambda_y^k / \pi_y^t\right)^{1/K}$.

Proof. The proof follows from Lemma 12 with the costs $\mathbf{c}$ set to λ^k/π^t for each iteration k. The lemma tells us that each f^k is of the form:

$$f^k(x') = \log\left(\frac{\lambda_y^k}{\pi_y^t} p_y^t(x')\right) + C(x'), \quad \forall (x', y') \in S,$$

for some example-specific constant $C(x') \in \mathbb{R}$. Consequently, we have that:

$$\overline{f}_y^s(x') = \log(\overline{\lambda}_y p_y^t(x')) + \overline{C}(x'), \quad \forall (x', y') \in S,$$

where $\overline{\lambda}_y = \left(\prod_{k=1}^{K} \lambda_y^k / \pi_y^t\right)^{1/K}$ and $\overline{C}(x') \in \mathbb{R}$. Applying a softmax to $\overline{f}^s$ results in the desired form. □

Proof of Theorem 4

Theorem 4. *Suppose $p^t \in \mathcal{F}$ and $\mathcal{F}$ is closed under linear transformations. Suppose ℓ is the softmax cross-entropy loss ℓ^{xent}, $\ell(y, z) \leq B$ and $\max_{y \in [m]} \frac{1}{\pi^t_y} \leq Z$, for some $B, Z > 0$. Furthermore, suppose for any $\delta \in (0, 1)$, the following bound holds on the estimation error in Theorem 3: with probability at least $1 - \delta$ (over draw of $S \sim D^n$), for all $f \in \mathcal{F}$,*

$$\max_{y \in [m]} \left| \phi_y(f) - \hat{\phi}_y(f) \right| \leq \Delta(n, \delta),$$

for some $\Delta(n, \delta) \in \mathbb{R}_+$ that is increasing in $1/\delta$, and goes to 0 as $n \to \infty$. Fix $\delta \in (0, 1)$. Then when the step size $\gamma = \frac{1}{2BZ}\sqrt{\frac{\log(m)}{K}}$ and $n^{\text{val}} \geq 8Z \log(2m/\delta)$, with probability at least $1 - \delta$ (over draw of $S \sim D^n$ and $S^{\text{val}} \sim D^{n^{\text{val}}}$)

$$L^{\text{rob}}(\overline{f}^s) \leq \min_{f \in \mathcal{F}} L^{\text{rob}}(f) \,+\, \underbrace{2B \max_{y \in [m]} \mathbb{E}_x\left[\left|\frac{p^t_y(x)}{\pi^t_y} \,-\, \frac{\eta_y(x)}{\pi_y}\right|\right]}_{\textit{Approximation error}}$$
$$+\, \underbrace{2\Delta(n^{\text{val}}, \delta/2) \,+\, 2\Delta(n, \delta/2)}_{\textit{Estimation error}} \,+\, \underbrace{4BZ\sqrt{\frac{\log(m)}{K}}}_{\textit{EG convergence}}.$$

Before proceeding to the proof, we will find it useful to define:

$$\hat{\phi}^{\text{val}}_y(f^s) = \frac{1}{\hat{\pi}^{t,\text{val}}_y}\frac{1}{n^{\text{val}}} \sum_{(x', y') \in S^{\text{val}}} p^t_y(x')\, \ell\left(y, f^s(x')\right).$$

We then state a useful lemma.

Lemma 13. Suppose the conditions in Theorem 4 hold. Then with probability $\leq 1 - \delta$ (over draw of $S \sim D^n$ and $S^{\text{val}} \sim D^{n^{\text{val}}}$), at each iteration k,

$$\sum_{y=1}^m \lambda^{k+1}_y \phi_y(f^{k+1}) \,-\, \min_{f \in \mathcal{F}} \sum_{y=1}^m \lambda^{k+1}_y \phi_y(f) \leq 2\Delta(n, \delta);$$

and for any $\lambda \in \Delta_m$:

$$\left|\sum_{y=1}^m \lambda_y \hat{\phi}^{\text{val}}_y(f^{k+1}) \,-\, \sum_{y=1}^m \lambda_y \phi_y(f^{k+1})\right| \,\leq\, \Delta(n^{\text{val}}, \delta).$$

Proof. We first note that by applying Lemma 12 with $w_i = 1, \forall i$, we have that f^{k+1} is the minimizer of $\sum_{y=1}^m \lambda^{k+1}_y \hat{\phi}_y(f)$ over all $f \in \mathcal{F}$, and therefore:

$$\sum_{y=1}^m \lambda^{k+1}_y \hat{\phi}_y(f^{k+1}) \leq \sum_{y=1}^m \lambda^{k+1}_y \hat{\phi}_y(f),\ \forall f \in \mathcal{F}. \tag{B.5}$$

Further, for a fixed iteration k, let us denote $\tilde{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{y=1}^{m} \lambda_y^{k+1} \phi_y(f)$. Then for the first part, we have:

$$\begin{aligned}
&\sum_{y=1}^{m} \lambda_y^{k+1} \phi_y(f^{k+1}) \,-\, \sum_{y=1}^{m} \lambda_y^{k+1} \phi_y(\tilde{f}) \\
&\leq \sum_{y=1}^{m} \lambda_y^{k+1} \phi_y(f^{k+1}) \,-\, \sum_{y=1}^{m} \lambda_y^{k+1} \hat{\phi}_y(f^{k+1}) \,+\, \sum_{y=1}^{m} \lambda_y^{k+1} \hat{\phi}_y(f^{k+1}) \,-\, \sum_{y=1}^{m} \lambda_y^{k+1} \phi_y(\tilde{f}) \\
&\leq \sum_{y=1}^{m} \lambda_y^{k+1} \phi_y(f^{k+1}) \,-\, \sum_{y=1}^{m} \lambda_y^{k+1} \hat{\phi}_y(f^{k+1}) \,+\, \sum_{y=1}^{m} \lambda_y^{k+1} \hat{\phi}_y(\tilde{f}) \,-\, \sum_{y=1}^{m} \lambda_y^{k+1} \phi_y(\tilde{f}) \\
&\leq 2 \sup_{f \in \mathcal{F}} \left| \sum_{y=1}^{m} \lambda_y^{k+1} \hat{\phi}_y(f) \,-\, \sum_{y=1}^{m} \lambda_y^{k+1} \phi_y(f) \right| \\
&\leq 2 \sup_{f \in \mathcal{F}} \max_{\lambda \in \Delta_m} \left| \sum_{y=1}^{m} \lambda_y \hat{\phi}_y(f) \,-\, \sum_{y=1}^{m} \lambda_y \phi_y(f) \right| \\
&\leq 2 \sup_{f \in \mathcal{F}} \max_{\lambda \in \Delta_m} \sum_{y=1}^{m} \lambda_y \left| \hat{\phi}_y(f) \,-\, \phi_y(f) \right| \\
&= 2 \sup_{f \in \mathcal{F}} \max_{y \in [m]} \left| \hat{\phi}_y(f) - \phi_y(f) \right|.
\end{aligned}$$

where for the second inequality, we use equation (B.5). Applying the generalization bound assumed in Theorem 4, we have with probability $\leq 1 - \delta$ (over draw of $S \sim D^n$), for all iterations $k \in [K]$,

$$\sum_{y=1}^{m} \lambda_y^{k+1} \phi_y(f^{k+1}) \,-\, \sum_{y=1}^{m} \lambda_y^{k+1} \phi_y(\tilde{f}) \,\leq\, 2\Delta(n, \delta),$$

For the second part, note that for any $\lambda \in \Delta_m$,

$$\begin{aligned}
\left| \sum_{y=1}^{m} \lambda_y \hat{\phi}_y^{\text{val}}(f^{k+1}) \,-\, \sum_{y=1}^{m} \lambda_y \phi_y(f^{k+1}) \right| &\leq \sum_{y=1}^{m} \lambda_y \left| \hat{\phi}_y^{\text{val}}(f^{k+1}) \,-\, \phi_y(f^{k+1}) \right| \\
&\leq \max_{y \in [m]} \left| \hat{\phi}_y^{\text{val}}(f^{k+1}) \,-\, \phi_y(f^{k+1}) \right| \\
&\leq \sup_{f \in \mathcal{F}} \max_{y \in [m]} \left| \hat{\phi}_y^{\text{val}}(f) \,-\, \phi_y(f) \right|.
\end{aligned}$$

An application of the generalization bound assumed in Theorem 4 to empirical estimates from the validation sample completes the proof. □

We are now ready to prove Theorem 4.

Proof of Theorem 4. Note that because $\min_{y\in[m]} \pi^t_y \geq \frac{1}{Z}$ and $n^{\text{val}} \geq 8Z\log(2m/\delta)$, we have by a direct application of Chernoff's bound (along with a union bound over all m classes) that with probability at least $1-\delta/2$:

$$\min_{y\in[m]} \hat{\pi}^{t,\text{val}}_y \geq \frac{1}{2Z}, \forall y \in [m]$$

and consequently, $\hat{\phi}^{\text{val}}_y(f) \leq 2BZ, \forall f \in \mathcal{F}$. The boundedness of $\hat{\phi}^{\text{val}}_y$ will then allow us to apply standard convergence guarantees for exponentiated gradient ascent [Shalev-Shwartz et al., 2011]. For $\gamma = \frac{1}{2BZ}\sqrt{\frac{\log(m)}{K}}$, the updates on λ will give us with probability at least $1-\delta/2$:

$$\max_{\lambda\in\Delta_m} \frac{1}{K}\sum_{k=1}^{K}\sum_{y=1}^{m}\lambda_y\hat{\phi}^{\text{val}}_y(f^k) \;\leq\; \frac{1}{K}\sum_{k=1}^{K}\sum_{y=1}^{m}\lambda^k_y\hat{\phi}^{\text{val}}_y(f^k) \;+\; 4BZ\sqrt{\frac{\log(m)}{K}} \tag{B.6}$$

Applying the second part of Lemma 13 to each iteration k, we have with probability at least $1-\delta$:

$$\max_{\lambda\in\Delta_m} \frac{1}{K}\sum_{k=1}^{K}\sum_{y=1}^{m}\lambda_y\phi_y(f^k) \;\leq\; \frac{1}{K}\sum_{k=1}^{K}\sum_{y=1}^{m}\lambda^k_y\phi_y(f^k) \;+\; 4BZ\sqrt{\frac{\log(m)}{K}} \;+\; 2\Delta(n^{\text{val}},\delta/2),$$

and applying the first part of Lemma 13 to the RHS, we have with the same probability:

$$\begin{aligned}
&\max_{\lambda\in\Delta_m} \frac{1}{K}\sum_{k=1}^{K}\sum_{y=1}^{m}\lambda_y\phi_y(f^k)\\
&\leq \frac{1}{K}\sum_{k=1}^{K}\min_{f\in\mathcal{F}}\sum_{y=1}^{m}\lambda^k_y\phi_y(f) \;+\; 4BZ\sqrt{\frac{\log(m)}{K}} \;+\; 2\Delta(n^{\text{val}},\delta/2) \;+\; 2\Delta(n,\delta/2)\\
&\leq \min_{f\in\mathcal{F}} \frac{1}{K}\sum_{k=1}^{K}\sum_{y=1}^{m}\lambda^k_y\phi_y(f) \;+\; 4BZ\sqrt{\frac{\log(m)}{K}} \;+\; 2\Delta(n^{\text{val}},\delta/2) \;+\; 2\Delta(n,\delta/2).
\end{aligned}$$

Note that we have taken a union bound over the high probability statement in equation (B.6) and that in Lemma 13. Using the convexity of $\phi(\cdot)$ in $f(x)$ and Jensen's inequality, we have that $\sum_{y=1}^{m}\lambda_y\phi_y(\overline{f}^s) \leq \frac{1}{K}\sum_{k=1}^{K}\sum_{y=1}^{m}\lambda_y\phi_y(f^k)$. We use this to further lower bound the LHS in terms of the averaged scoring function $\overline{f}^s(x) = \frac{1}{K}\sum_{k=1}^{K} f^k(x)$:

$$\begin{aligned}
&\max_{\lambda\in\Delta_m} \sum_{y=1}^{m}\lambda_y\phi_y(\overline{f}^s)\\
&\leq \min_{f\in\mathcal{F}} \frac{1}{K}\sum_{k=1}^{K}\sum_{y=1}^{m}\lambda^k_y\phi_y(f) \;+\; 4BZ\sqrt{\frac{\log(m)}{K}} \;+\; 2\Delta(n^{\text{val}},\delta/2) \;+\; 2\Delta(n,\delta/2)
\end{aligned}$$

$$\begin{aligned}
&= \min_{f\in\mathcal{F}} \sum_{y=1}^{m} \tilde{\lambda}_y \phi_y(f) \,+\, 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta(n^{\text{val}}, \delta/2) \,+\, 2\Delta(n, \delta/2)\\
&\leq \max_{\lambda\in\Delta_m} \min_{f\in\mathcal{F}} \sum_{y=1}^{m} \lambda_y \phi_y(f) \,+\, 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta(n^{\text{val}}, \delta/2) \,+\, 2\Delta(n, \delta/2)\\
&= \min_{f\in\mathcal{F}} \max_{\lambda\in\Delta_m} \sum_{y=1}^{m} \lambda_y \phi_y(f) \,+\, 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta(n^{\text{val}}, \delta/2) \,+\, 2\Delta(n, \delta/2)\\
&= \min_{f\in\mathcal{F}} \max_{y\in[m]} \phi_y(f) \,+\, 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta(n^{\text{val}}, \delta/2) \,+\, 2\Delta(n, \delta/2), \qquad \text{(B.7)}
\end{aligned}$$

where in the second step $\tilde{\lambda}_y = \frac{1}{K}\sum_{k=1}^{K} \lambda_y^k$; in the fourth step, we swap the 'min' and 'max' using Sion's minimax theorem [Sion, 1958]. We further have from equation (B.7),

$$\max_{y\in[m]} \phi_y(\overline{f}^s) \leq \min_{f\in\mathcal{F}} \max_{y\in[m]} \phi_y(f) \,+\, 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta(n^{\text{val}}, \delta/2) \,+\, 2\Delta(n, \delta/2).$$

In other words,

$$L^{\text{rob-d}}(\overline{f}^s) \leq \min_{f\in\mathcal{F}} L^{\text{rob-d}}(f) \,+\, 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta(n^{\text{val}}, \delta/2) \,+\, 2\Delta(n, \delta/2).$$

To complete the proof, we need to turn this into a guarantee on the original robust objective L^{rob} in equation (3.3):

$$\begin{aligned}
L^{\text{rob}}(\overline{f}^s) &\leq \min_{f\in\mathcal{F}} L^{\text{rob}}(f) \,+\, 2\max_{f\in\mathcal{F}} \left|L^{\text{rob}}(f) - L^{\text{rob-d}}(f)\right|\\
&\qquad + 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta(n^{\text{val}}, \delta/2) \,+\, 2\Delta(n, \delta/2)\\
&\leq \min_{f\in\mathcal{F}} L^{\text{rob}}(f) \,+\, 2B\max_{y\in[m]} \mathbb{E}_x\left[\left|\frac{p_y^t(x)}{\pi_y^t} - \frac{\eta_y(x)}{\pi_y}\right|\right]\\
&\qquad + 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta(n^{\text{val}}, \delta/2) \,+\, 2\Delta(n, \delta/2),
\end{aligned}$$

where we have used the bound on the approximation error in the proof of Theorem 3. This completes the proof. □

B.2 Student Estimation Error

We now provide a bound on the estimation error in Theorem 4 using a generalization bound from Menon et al. [2021a].

Lemma 14. Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ be a given class of scoring functions. Let $\mathcal{V} \subseteq \mathbb{R}^{\mathcal{X}}$ denote the class of loss functions $v(x, y) = \ell(y, f(x))$ induced by scorers $f \in \mathcal{F}$. Let $\mathcal{M}_n = \mathcal{N}_\infty(\frac{1}{n}, \mathcal{V}, 2n)$ denote the uniform L_∞ covering number for $\mathcal{V}$. Fix $\delta \in (0, 1)$. Suppose $\ell(y, z) \leq B$, $\pi^t_y \leq \frac{1}{Z}, \forall y \in [m]$, and the number of samples $n \geq 8Z \log(4m/\delta)$. Then with probability $\geq 1 - \delta$ over draw of $S \sim D^n$, for any $f \in \mathcal{F}$ and $y \in [m]$:

$$\left|\phi_y(f) - \hat{\phi}_y(f)\right| \leq CZ\left(\sqrt{\mathbb{V}_{n,y}(f)\frac{\log(m\mathcal{M}_n/\delta)}{n}} + \frac{\log(m\mathcal{M}_n/\delta)}{n} + B\sqrt{\frac{\log(m/\delta)}{n}}\right),$$

where $\mathbb{V}_{n,y}(f)$ denotes the empirical variance of the loss values $\{p^t_y(x_i) \cdot \ell(y, f(x_i))\}_{i=1}^n$ for class y, and $C > 0$ is a distribution-independent constant.

Notice the dependence on the *variance* that the teacher's predictions induce on the loss. This suggests that the lower the variance in the teacher's predictions, the better is the student's generalization. Similar to Menon et al. [2021a], one can further show that when the teacher closely approximates the Bayes-probabilities $\eta(x)$, the distilled loss $p^t_y(x_i) \cdot \ell(y, f(x_i))$ has a lower empirical variance that the loss $\ell(y_i, f(x_i))$ computed from one-hot labels.

Proof of Lemma 14. We begin by defining the following intermediate term:

$$\tilde{\phi}_y(f) = \frac{1}{\pi^t_y}\frac{1}{n}\sum_{i=1}^n p^t_y(x_i)\,\ell\left(y, f(x_i)\right).$$

Then for any $y \in [m]$,

$$\left|\phi_y(f) - \hat{\phi}_y(f)\right| \leq \left|\phi_y(f) - \tilde{\phi}_y(f)\right| + \left|\tilde{\phi}_y(f) - \hat{\phi}_y(f)\right|. \tag{B.8}$$

We next bound each of the terms in equation (B.8), starting with the first term:

$$\begin{aligned}\left|\phi_y(f) - \tilde{\phi}_y(f)\right| &= \frac{1}{\pi^t_y}\left|\mathbb{E}_x\left[p^t_y(x)\,\ell\left(y, f(x)\right)\right] - \frac{1}{n}\sum_{i=1}^n p^t_y(x_i)\,\ell\left(y, f(x_i)\right)\right| \\ &\leq Z\left|\mathbb{E}_x\left[p^t_y(x)\,\ell\left(y, f(x)\right)\right] - \frac{1}{n}\sum_{i=1}^n p^t_y(x_i)\,\ell\left(y, f(x_i)\right)\right|,\end{aligned}$$

where we use the fact that $\pi^t_y \leq \frac{1}{Z}, \forall y$. Applying the generalization bound from Menon et al. [2021a, Proposition 2], along with a union bound over all m classes, we have with probability at least $1 - \delta/2$ over the draw of $S \sim D^n$, for all $y \in [m]$:

$$\left|\phi_y(f) - \tilde{\phi}_y(f)\right| \leq C'Z\left(\sqrt{\mathbb{V}_{n,y}(f)\frac{\log(m\mathcal{M}_n/\delta)}{n}} + \frac{\log(m\mathcal{M}_n/\delta)}{n}\right), \tag{B.9}$$

for a distribution-independent constant $C' > 0$.

We next bound the second term in equation (B.8):

$$\begin{aligned}\left|\tilde{\phi}_y(f) - \hat{\phi}_y(f)\right| &= \left|\frac{1}{\pi^t_y} - \frac{1}{\hat{\pi}^t_y}\right| \frac{1}{n}\sum_{i=1}^n p^t_y(x_i)\cdot \ell\left(y, f(x_i)\right)\\ &\leq B\left|\frac{1}{\pi^t_y} - \frac{1}{\hat{\pi}^t_y}\right|\\ &= \frac{B}{\pi^t_y\hat{\pi}^t_y}\left|\pi^t_y - \hat{\pi}^t_y\right|,\end{aligned}$$

where in the second step we use the fact that $\ell(y, f(x)) \leq B$ and $p^t_y(x) \leq 1$.

Further note that because $\min_{y\in[m]} \pi^t_y \geq \frac{1}{Z}$ and $n \geq 8Z\log(4m/\delta)$, we have by a direct application of Chernoff's bound (and a union bound over m classes) that with probability at least $1-\delta/4$:

$$\min_{y\in[m]} \hat{\pi}^t_y \geq \frac{1}{2Z}, \forall y \in [m]. \tag{B.10}$$

Therefore for any $y \in [m]$:

$$\left|\tilde{\phi}_y(f) - \hat{\phi}_y(f)\right| \leq 2BZ^2\left|\pi^t_y - \hat{\pi}^t_y\right|.$$

Conditioned on the above statement, a simple application of Hoeffding's inequality and a union bound over all $y \in [m]$ gives us that with probability at least $1-\delta/4$ over the draw of $S \sim D^n$, for all $y \in [m]$:

$$\left|\tilde{\phi}_y(f) - \hat{\phi}_y(f)\right| \leq 2BZ^2\left(\frac{1}{Z}\sqrt{\frac{\log(8m/\delta)}{2n}}\right) = 2BZ\sqrt{\frac{\log(8m/\delta)}{2n}}. \tag{B.11}$$

A union bound over the high probability statements in (B.9–B.11) completes the proof. To see this, note that, for any $\varepsilon > 0$ and $y \in [m]$,

$$\begin{aligned}&\mathbb{P}\left(\left|\phi_y(f) - \hat{\phi}_y(f)\right| \geq \varepsilon\right)\\ &\leq \mathbb{P}\left(\left(\left|\phi_y(f) - \tilde{\phi}_y(f)\right| \geq \varepsilon\right) \vee \left(\left|\tilde{\phi}_y(f) - \hat{\phi}_y(f)\right| \geq \varepsilon\right)\right)\\ &\leq \mathbb{P}\left(\left|\phi_y(f) - \tilde{\phi}_y(f)\right| \geq \varepsilon\right) + \mathbb{P}\left(\left|\tilde{\phi}_y(f) - \hat{\phi}_y(f)\right| \geq \varepsilon\right)\\ &\leq \mathbb{P}\left(\left|\phi_y(f) - \tilde{\phi}_y(f)\right| \geq \varepsilon\right) + \mathbb{P}\left(\hat{\pi}^t_y \leq \frac{1}{Z}\right)\cdot\mathbb{P}\left(\left|\tilde{\phi}_y(f) - \hat{\phi}_y(f)\right| \geq \varepsilon \;\middle|\; \hat{\pi}^t_y \leq \frac{1}{Z}\right)\\ &\qquad + \mathbb{P}\left(\hat{\pi}^t_y \geq \frac{1}{Z}\right)\cdot\mathbb{P}\left(\left|\tilde{\phi}_y(f) - \hat{\phi}_y(f)\right| \geq \varepsilon \;\middle|\; \hat{\pi}^t_y \geq \frac{1}{Z}\right)\\ &\leq \mathbb{P}\left(\left|\phi_y(f) - \tilde{\phi}_y(f)\right| \geq \varepsilon\right) + \mathbb{P}\left(\hat{\pi}^t_y \leq \frac{1}{Z}\right) + \mathbb{P}\left(\left|\tilde{\phi}_y(f) - \hat{\phi}_y(f)\right| \geq \varepsilon \;\middle|\; \hat{\pi}^t_y \geq \frac{1}{Z}\right),\end{aligned}$$

which implies that a union bound over (B.9–B.11) would give us the desired result in Lemma 14. □

Algorithm 6 Distilled Margin-based DRO with One-hot Validation Labels

```
**Inputs:** Teacher $p^t$, Student hypothesis class $\mathcal{F}$, Training set $S$, Validation set $S^{\text{val}}$, Step-size $\gamma \in \mathbb{R}_+$, Number of iterations $K$, Loss $\ell$
**Initialize:** Student $f^0 \in \mathcal{F}$, Multipliers $\lambda^0 \in \Delta_m$
**For** $k = 0$ to $K-1$
    $\tilde{\lambda}_j^{k+1} = \lambda_j^k \exp\big(\gamma \hat{R}_j\big), \forall j \in [m]$
        where $\hat{R}_j = \frac{1}{n^{\text{val}}} \frac{1}{\hat{\pi}_j^{\text{val}}} \sum_{(x,y)\in S^{\text{val}}} \ell(y, f^k(x))$ and $\hat{\pi}_j^{\text{val}} = \frac{1}{n^{\text{val}}} \sum_{(x,y)\in S^{\text{val}}} \mathbb{1}(y=j)$
    $\lambda_y^{k+1} = \frac{\tilde{\lambda}_y^{k+1}}{\sum_{j=1}^m \tilde{\lambda}_j^{k+1}}, \forall y$
    $f^{k+1} \in \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^n \mathcal{L}^{\text{mar}}\left(p^t(x_i), f(x_i); \frac{\lambda^{k+1}}{\hat{\pi}^t}\right)$   // Replaced with a few steps of SGD
**End For**
**Output:** $\overline{f}^s : x \mapsto \frac{1}{K}\sum_{k=1}^K f^k(x)$
```

B.3 DRO with One-hot Validation Labels

The updates on λ in Algorithm 2 use a validation set labeled by the teacher. One could instead perform these updates with a curated validation set containing the original one-hot labels. Each of these choices presents different merits. The use of a teacher-labeled validation set is useful in many real world scenarios where labeled data is hard to obtain, while unlabeled data abounds. In contrast, the use of one-hot validation labels, although more expensive to obtain, may make the student more immune to errors in the teacher's predictions, as the coefficients λs are now based on an unbiased estimate of the student's performance on each class.

Algorithm 6 contains a version of the margin-based DRO described in Section 3.5, where instead of teacher labels the original one-hot labels are used in the validation set.

Before proceeding to providing a convergence guarantee for this algorithm, we will find it useful to define the following one-hot metrics:

$$\phi_y^{\text{oh}}(f^s) = \frac{1}{\pi_y}\mathbb{E}_x\left[\eta_y(x)\,\ell\left(y, f^s(x)\right)\right]$$
$$\hat{\phi}_y^{\text{oh,val}}(f^s) = \frac{1}{\hat{\pi}_y}\frac{1}{n^{\text{val}}}\sum_{(x',y')\in S^{\text{val}}} \mathbb{1}(y'=y)\,\ell\left(y', f^s(x')\right).$$

Theorem 9. *Suppose $p^t \in \mathcal{F}$ and $\mathcal{F}$ is closed under linear transformations. Then the final scoring function $\overline{f}^s(x) = \frac{1}{K}\sum_{k=1}^K f^k(x)$ output by Algorithm 6 is of the form:*

$$\text{softmax}_y(\overline{f}^s(x')) \propto \overline{\lambda}_y p_y^t(x'), \qquad \forall (x', y') \in S,$$

where $\overline{\lambda}_y = \left(\prod_{k=1}^{K} \lambda^k_y/\pi^t_y\right)^{1/K}$. *Furthermore, suppose* ℓ *is the softmax cross-entropy loss in* ℓ^{xent}, $\ell(y, z) \le B$, *for some* $B > 0$, *and* $\max_{y\in[m]} \frac{1}{\pi_y} \le Z$, *for some* $Z > 0$. *Suppose for any* $\delta \in (0, 1)$, *the following holds: with probability at least* $1 - \delta$ *(over draw of* $S \sim D^n$*), for all* $f \in \mathcal{F}$,

$$\max_{y\in[m]} \left|\phi^{\mathrm{oh}}_y(f) - \hat{\phi}^{\mathrm{oh}}_y(f)\right| \le \Delta^{\mathrm{oh}}(n, \delta); \qquad \max_{y\in[m]} \left|\phi_y(f) - \hat{\phi}_y(f)\right| \le \Delta(n, \delta),$$

for some $\Delta^{\mathrm{oh}}(n, \delta), \Delta(n, \delta) \in \mathbb{R}_+$ *that is increasing in* $1/\delta$, *and goes to 0 as* $n \to \infty$. *Fix* $\delta \in (0, 1)$. *Then when the step size* $\gamma = \frac{1}{2BZ}\sqrt{\frac{\log(m)}{K}}$ *and* $n^{\mathrm{val}} \ge 8Z \log(2m/\delta)$, *with probability at least* $1 - \delta$ *(over draw of* $S \sim D^n$ *and* $S^{\mathrm{val}} \sim D^{n^{\mathrm{val}}}$*), for any* $\tau \in \mathbb{R}_+$,

$$L^{\mathrm{rob}}(\overline{f}^s) \le \min_{f\in\mathcal{F}} L^{\mathrm{rob}}(f) + \underbrace{2B \max_{y\in[m]} \mathbb{E}_x \left[\left|\tau \cdot \frac{p^t_y(x)}{\pi^t_y} - \frac{\eta_y(x)}{\pi_y}\right|\right]}_{\textit{Approximation error}}$$
$$+ \underbrace{2\tau \cdot \Delta^{\mathrm{oh}}(n^{\mathrm{val}}, \delta/2) + 2\Delta(n, \delta/2)}_{\textit{Estimation error}} + \underbrace{4BZ\sqrt{\frac{\log(m)}{K}}}_{\textit{EG convergence}}.$$

Comparing this to the bound in Theorem 4, we can see that there is an additional scaling factor τ against the teacher probabilities $p^t_y(x)$ and in the approximation error. When we set $\tau = 1$, the bound looks very similar to Theorem 4, except that the estimation error term Δ^{oh} now involves one-hot labels. Therefore the estimation error may incur a slower convergence with sample size as it no longer benefits from the lower variance that the teacher predictions may offer (see Appendix B.2 for details).

The τ-scaling in the approximation error also means that the teacher is no longer required to exactly match the (normalized) class probabilities $\eta(x)$. In fact, one can set τ to a value for which the approximation error is the lowest, and in general to a value that minimizes the upper bound in Theorem 9, potentially providing us with a tighter convergence rate than Theorem 4.

The proof of Theorem 9 is similar to that of Theorem 4, but requires a modified version of Lemma 13:

Lemma 15. Suppose the conditions in Theorem 4 hold. With probability $\le 1 - \delta$ (over draw of $S \sim D^n$ and $S^{\mathrm{val}} \sim D^{n^{\mathrm{val}}}$), at each iteration k and for any $\tau \in \mathbb{R}_+$,

$$\sum_{y=1}^{m} \lambda^{k+1}_y \phi^{\mathrm{oh}}_y(f^{k+1}) - \min_{f\in\mathcal{F}} \sum_{y=1}^{m} \lambda^{k+1}_y \phi^{\mathrm{oh}}_y(f) \le 2\tau \cdot \Delta(n, \delta) + 2B \max_{y\in[m]} \mathbb{E}_x \left[\left|\tau \frac{p^t_y(x)}{\pi^t_y} - \frac{\eta_y(x)}{\pi_y}\right|\right].$$

Furthermore, with the same probability, for any $\lambda \in \Delta_m$:

$$\left|\sum_{y=1}^{m} \lambda_y \hat{\phi}^{\mathrm{oh,val}}_y(f^{k+1}) - \sum_{y=1}^{m} \lambda_y \phi^{\mathrm{oh}}_y(f^{k+1})\right| \le \Delta^{\mathrm{oh}}(n^{\mathrm{val}}, \delta).$$

Proof. We first note from Lemma 12 that because $f^{k+1} \in \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}^{\text{mar}}\Big(p^t(x_i), f(x_i); \frac{\lambda^{k+1}}{\hat{\pi}}\Big)$, we have for the example-weighting $w_i = \tau, \forall i$:

$$\tau\sum_{y=1}^{m}\lambda_y^{k+1}\hat{\phi}_y(f^{k+1}) \,\leq\, \tau\sum_{y=1}^{m}\lambda_y^{k+1}\hat{\phi}_y(f),\ \forall f\in\mathcal{F}. \tag{B.12}$$

For a fixed iteration k, let us denote $\tilde{f} \in \operatorname*{argmin}_{f\in\mathcal{F}} \sum_{y=1}^{m}\lambda_y^{k+1}\phi_y(f)$. Then for the first part, we have for any $\tau \in \mathbb{R}_+$:

$$\begin{aligned}
&\sum_{y=1}^{m}\lambda_y^{k+1}\phi_y^{\text{oh}}(f^{k+1}) \,-\, \sum_{y=1}^{m}\lambda_y^{k+1}\phi_y^{\text{oh}}(\tilde{f})\\
&\leq \tau\left(\sum_{y=1}^{m}\lambda_y^{k+1}\phi_y(f^{k+1}) \,-\, \sum_{y=1}^{m}\lambda_y^{k+1}\phi_y(\tilde{f})\right) \,+\, \sum_{y=1}^{m}\lambda_y^{k+1}\left|\phi_y^{\text{oh}}(f^{k+1}) - \tau\phi_y(f^{k+1})\right|\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad + \sum_{y=1}^{m}\lambda_y^{k+1}\left|\phi_y^{\text{oh}}(\tilde{f}) - \tau\phi_y(\tilde{f})\right|\\
&\leq \tau\left(\sum_{y=1}^{m}\lambda_y^{k+1}\phi_y(f^{k+1}) \,-\, \sum_{y=1}^{m}\lambda_y^{k+1}\phi_y(\tilde{f})\right) \,+\, 2\max_{f\in\mathcal{F}}\sum_{y=1}^{m}\lambda_y^{k+1}\left|\phi_y^{\text{oh}}(f) - \tau\phi_y(f)\right|\\
&\leq \tau\left(\sum_{y=1}^{m}\lambda_y^{k+1}\phi_y(f^{k+1}) \,-\, \sum_{y=1}^{m}\lambda_y^{k+1}\phi_y(\tilde{f})\right) \,+\, 2\max_{f\in\mathcal{F}}\max_{\lambda\in\Delta_m}\sum_{y=1}^{m}\lambda\left|\phi_y^{\text{oh}}(f) - \tau\phi_y(f)\right|\\
&\leq \tau\left(\sum_{y=1}^{m}\lambda_y^{k+1}\phi_y(f^{k+1}) \,-\, \sum_{y=1}^{m}\lambda_y^{k+1}\phi_y(\tilde{f})\right) \,+\, 2\max_{f\in\mathcal{F}}\max_{y\in[m]}\left|\phi_y^{\text{oh}}(f) - \tau\phi_y(f)\right|\\
&\leq 2\tau\sup_{f\in\mathcal{F}}\max_{y\in[m]}\left|\hat{\phi}_y(f) - \phi_y(f)\right| \,+\, 2\max_{f\in\mathcal{F}}\max_{y\in[m]}\left|\phi_y^{\text{oh}}(f) - \tau\phi_y(f)\right|.
\end{aligned}$$

where the last inequality re-traces the steps in Lemma 13. Further applying the generalization bound assumed in Theorem 4, we have with probability $\leq 1-\delta$ (over draw of $S \sim D^n$), for all iterations $k \in [K]$ and any $\tau \in \mathbb{R}_+$,

$$\sum_{y=1}^{m}\lambda_y^{k+1}\phi_y^{\text{oh}}(f^{k+1}) \,-\, \sum_{y=1}^{m}\lambda_y^{k+1}\phi_y^{\text{oh}}(\tilde{f}) \,\leq\, 2\tau\Delta(n,\delta) + 2\max_{f\in\mathcal{F}}\max_{y\in[m]}\left|\phi_y^{\text{oh}}(f) - \tau\phi_y(f)\right|. \tag{B.13}$$

All that remains is to bound the second term in equation (B.13). For any $f \in \mathcal{F}$ and $y \in [m]$,

$$\left|\phi_y^{\text{oh}}(f) - \tau\phi_y(f)\right| \leq \left|\frac{1}{\pi_y}\mathbb{E}_x\left[\eta_y(x)\,\ell\left(y, f(x)\right)\right] \,-\, \frac{\tau}{\pi_y^t}\mathbb{E}_x\left[p_y^t(x)\,\ell\left(y, f(x)\right)\right]\right|$$

$$\begin{aligned}
&\leq \mathbb{E}_x\left[\left|\frac{1}{\pi_y}\eta_y(x)\,\ell\left(y, f(x)\right) \;-\; \frac{\tau}{\pi^t_y}p^t_y(x)\,\ell\left(y, f(x)\right)\right|\right]\\
&= \mathbb{E}_x\left[\left|\frac{1}{\pi_y}\eta_y(x) \;-\; \frac{\tau}{\pi^t_y}p^t_y(x)\right|\ell\left(y, f^s(x)\right)\right]\\
&\leq B\mathbb{E}_x\left[\left|\frac{\eta_y(x)}{\pi_y} \;-\; \tau\frac{p^t_y(x)}{\pi^t_y}\right|\right],
\end{aligned}$$

where we use Jensen's inequality in the second step, the fact that $\ell(y, z) \leq B$ is non-negative in the second step, and the fact that $\ell(y, z) \leq B$ in the last step. Substituting this upper bound back into equation (B.13) completes the proof of the first part.

The second part follows from a direct application of the bound on the per-class estimation error $\max_{y\in[m]}\left|\phi^{\text{oh}}_y(f) - \hat{\phi}^{\text{oh,val}}_y(f)\right|$. □

Proof of Theorem 9. The proof traces the same steps as Proposition 1 and Theorem 4, except that it applies Lemma 15 instead of Lemma 13.

Note that because $\min_{y\in[m]}\pi_y \geq \frac{1}{Z}$ and $n^{\text{val}} \geq 8Z\log(2m/\delta)$, we have by a direct application of Chernoff's bound (along with a union bound over all m classes) that with probability at least $1-\delta/2$:

$$\min_{y\in[m]}\hat{\pi}^{\text{oh,val}}_y \geq \frac{1}{2Z}, \forall y \in [m],$$

and consequently, $\hat{\phi}^{\text{oh,val}}_y(f) \leq 2BZ, \forall f \in \mathcal{F}$. The boundedness of $\hat{\phi}^{\text{oh,val}}_y$ will then allow us to apply standard convergence guarantees for exponentiated gradient ascent [Shalev-Shwartz et al., 2011]. For $\gamma = \frac{1}{2BZ}\sqrt{\frac{\log(m)}{K}}$, the updates on λ will give us:

$$\max_{\lambda\in\Delta_m}\frac{1}{K}\sum_{k=1}^{K}\sum_{y=1}^{m}\lambda_y\hat{\phi}^{\text{oh,val}}_y(f^k) \;\leq\; \frac{1}{K}\sum_{k=1}^{K}\sum_{y=1}^{m}\lambda^k_y\hat{\phi}^{\text{oh,val}}_y(f^k) \;+\; 4BZ\sqrt{\frac{\log(m)}{K}}$$

Applying the second part of Lemma 13 to each iteration k, we have with probability at least $1-\delta$:

$$\max_{\lambda\in\Delta_m}\frac{1}{K}\sum_{k=1}^{K}\sum_{y=1}^{m}\lambda_y\phi^{\text{oh}}_y(f^k) \;\leq\; \frac{1}{K}\sum_{k=1}^{K}\sum_{y=1}^{m}\lambda^k_y\phi^{\text{oh}}_y(f^k) \;+\; 4BZ\sqrt{\frac{\log(m)}{K}} \;+\; 2\Delta^{\text{oh}}(n^{\text{val}}, \delta/2),$$

and applying the first part of Lemma 13 to the RHS, we have with the same probability, for any $\tau \in \mathbb{R}_+$:

$$\begin{aligned}
\max_{\lambda\in\Delta_m}\frac{1}{K}\sum_{k=1}^{K}\sum_{y=1}^{m}\lambda_y\phi^{\text{oh}}_y(f^k) \leq \frac{1}{K}\sum_{k=1}^{K}\min_{f\in\mathcal{F}}\sum_{y=1}^{m}\lambda^k_y\phi^{\text{oh}}_y(f) \;+\; 4BZ\sqrt{\frac{\log(m)}{K}} \;+\; 2\Delta^{\text{oh}}(n^{\text{val}}, \delta/2)\\
+\; 2\tau\Delta(n, \delta/2) \;+\; 2B\max_{y\in[m]}\mathbb{E}_x\left[\left|\tau\frac{p^t_y(x)}{\pi^t_y} \;-\; \frac{\eta_y(x)}{\pi_y}\right|\right]
\end{aligned}$$

Algorithm 7 Distilled Margin-based DRO for Traded-off Objective

```
**Inputs:** Teacher $p^t$, Student hypothesis class $\mathcal{F}$, Training set $S$, Validation set $S^{\text{val}}$, Step-size $\gamma \in \mathbb{R}_+$, Number of iterations $K$, Loss $\ell$, Trade-off parameter $\alpha$
**Initialize:** Student $f^0 \in \mathcal{F}$, Multipliers $\lambda^0 \in \Delta_m$
**For** $k = 0$ to $K - 1$
    $\tilde{\lambda}_j^{k+1} = \lambda_j^k \exp\left(\gamma\alpha\hat{R}_j\right), \forall j \in [m]$  where $\hat{R}_j = \frac{1}{n^{\text{val}}}\frac{1}{\hat{\pi}_j^{t,\text{val}}}\sum_{(x,y)\in S^{\text{val}}} p_j^t(x_i)\,\ell(j, f^k(x))$
    $\lambda_y^{k+1} = \frac{\tilde{\lambda}_y^{k+1}}{\sum_{j=1}^m \tilde{\lambda}_j^{k+1}}, \forall y$
    $\beta_y^{k+1} = (1-\alpha)\frac{1}{m} + \alpha\lambda_y^{k+1}$
    $f^{k+1} \in \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \mathcal{L}^{\text{mar}}\left(p^t(x_i), f(x_i); \frac{\beta^{k+1}}{\hat{\pi}^t}\right)$   // Replaced with a few steps of SGD
**End For**
**Output:** $\overline{f}^s : x \mapsto \frac{1}{K}\sum_{k=1}^K f^k(x)$
```

$$\leq \min_{f\in\mathcal{F}} \frac{1}{K}\sum_{k=1}^K\sum_{y=1}^m \lambda_y^k\phi_y^{\text{oh}}(f) \,+\, 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta^{\text{oh}}(n^{\text{val}}, \delta/2)$$

$$+\, 2\tau\Delta(n, \delta/2) \,+\, 2B\max_{y\in[m]}\mathbb{E}_x\left[\left|\tau\frac{p_y^t(x)}{\pi_y^t} \,-\, \frac{\eta_y(x)}{\pi_y}\right|\right].$$

Using the convexity of $\phi(\cdot)$ in $f(x)$ and Jensen's inequality, we have that $\sum_{y=1}^m \lambda_y\phi_y(\overline{f}^s) \leq \frac{1}{K}\sum_{k=1}^K\sum_{y=1}^m \lambda_y\phi_y(f^k)$. We use this to further lower bound the LHS in terms of the averaged scoring function $\overline{f}^s(x) = \frac{1}{K}\sum_{k=1}^K f^k(x)$, and re-trace the steps in Theorem 4 to get"

$$\max_{y\in[m]} \phi_y^{\text{oh}}(\overline{f}^s) \leq \min_{f\in\mathcal{F}}\max_{y\in[m]} \phi_y^{\text{oh}}(f) \,+\, 4BZ\sqrt{\frac{\log(m)}{K}} \,+\, 2\Delta^{\text{oh}}(n^{\text{val}}, \delta/2)$$

$$+\, 2\tau\Delta(n, \delta/2) \,+\, 2B\max_{y\in[m]}\mathbb{E}_x\left[\left|\tau\frac{p_y^t(x)}{\pi_y^t} \,-\, \frac{\eta_y(x)}{\pi_y}\right|\right].$$

Noting that $L^{\text{rob}}(f) = \max_{y\in[m]} \phi_y^{\text{oh}}(f)$ completes the proof. □

B.4 DRO for Traded-off Objective

We present a variant of the margin-based DRO algorithm described in Section 3.5 that seeks to minimize a trade-off between the balanced and robust student objectives:

$$\hat{L}^{\text{tdf-d}}(f^s) = (1-\alpha)\hat{L}^{\text{bal-d}}(f^s) + \alpha\hat{L}^{\text{rob-d}}(f^s),$$

for some $\alpha \in [0, 1]$.

Expanding this, we have:

$$\begin{aligned}
L^{\text{tdf-d}}(f) &= (1-\alpha)\frac{1}{m}\sum_{y=1}^{m}\frac{1}{\hat{\pi}^t_y}\frac{1}{n}\sum_{i=1}^{n}p^t_y(x_i)\,\ell(y, f(x_i)) \;+\; \alpha\max_{y\in[m]}\sum_{y=1}^{m}\frac{1}{\hat{\pi}^t_y}\frac{1}{n}\sum_{i=1}^{n}p^t_y(x_i)\,\ell(y, f(x_i)) \\
&= (1-\alpha)\frac{1}{m}\sum_{y=1}^{m}\frac{1}{\hat{\pi}^t_y}\frac{1}{n}\sum_{i=1}^{n}p^t_y(x_i)\,\ell(y, f(x_i)) \;+\; \alpha\max_{\lambda\in\Delta_m}\sum_{y=1}^{m}\frac{\lambda_y}{\hat{\pi}^t_y}\frac{1}{n}\sum_{i=1}^{n}p^t_y(x_i)\,\ell(y, f(x_i)) \\
&= \max_{\lambda\in\Delta_m}\sum_{y=1}^{m}\left((1-\alpha)\frac{1}{m} + \alpha\lambda_y\right)\frac{1}{\hat{\pi}^t_y}\frac{1}{n}\sum_{i=1}^{n}p^t_y(x_i)\,\ell(y, f(x_i)).
\end{aligned}$$

The minimization of $L^{\text{tdf-d}}(f)$ over f can then be a cast as a min-max problem:

$$\min_{f:\mathcal{X}\to\mathbb{R}^m} L^{\text{tdf-d}}(f) = \min_{f:\mathcal{X}\to\mathbb{R}^m}\max_{\lambda\in\Delta_m}\sum_{y=1}^{m}\left((1-\alpha)\frac{1}{m} + \alpha\lambda_y\right)\frac{1}{\hat{\pi}^t_y}\frac{1}{n}\sum_{i=1}^{n}p^t_y(x_i)\,\ell(y, f(x_i)).$$

Retracing the steps in the derivation of Algorithm 2 in Section 3.5, we have the following updates on λ and f to solve the above min-max problem:

$$\begin{aligned}
\tilde{\lambda}^{k+1}_y &= \lambda^k_y \exp\left(\gamma\alpha\frac{1}{n\hat{\pi}^t_y}\sum_{i=1}^{n}p^t_y(x_i)\,\ell\left(y, f^k(x_i)\right)\right), \forall y \\
\lambda^{k+1}_y &= \frac{\tilde{\lambda}^{k+1}_y}{\sum_{j=1}^{m}\tilde{\lambda}^{k+1}_j}, \forall y \\
\beta^{k+1}_y &= (1-\alpha)\frac{1}{m} \,+\, \alpha\lambda^{k+1}_y \\
f^{k+1} &\in \operatorname*{argmin}_{f\in\mathcal{F}}\sum_{y\in[m]}\frac{\beta^{k+1}_y}{n\hat{\pi}^t_y}\sum_{i=1}^{n}p^t_y(x_i)\,\ell\left(y, f(x_i)\right),
\end{aligned}$$

for step-size parameter $\gamma > 0$. To better handle training of over-parameterized students, we will perform the updates on λ using a held-out validation set, and employ a margin-based surrogate for performing the minimization over f. This procedure is outlined in Algorithm 7.

Connection to post-hoc adjustment

The form of the student in Proposition 1 raises an interesting question. Instead of training an explicit student model, why not directly construct a new scoring model by making post-hoc adjustments to the teacher's predictions? Specifically, one could optimize over functions of the form $f^s_y(x) = \log(\gamma_y p^t_y(x))$, where the teacher p^t is fixed, and pick the coefficients $\gamma \in \mathbb{R}^m$ so that resulting scoring function yields the best worst-class accuracy on a held-out dataset. This simple *post-hoc adjustment* strategy may not be feasible if the goal is to distill to a student that is considerably smaller than the teacher. Often, this is the case in settings where distillation is used as a compression technique. Yet, this post-hoc method serves as good baseline to compare with.

B.5 Additional experiment details

This section contains further experiment details about the datasets, hyperparameters, and baselines.

Additional details about datasets

Building long tailed datasets

The long-tailed datasets were created from the original datasets following Cui et al. [2019] by downsampling examples with an exponential decay in the per-class sizes. As done by Narasimhan and Menon [2021], we set the imbalance ratio $\frac{\max_i P(y=i)}{\min_i P(y=i)}$ to 100 for CIFAR-10 and CIFAR-100, and to 83 for TinyImageNet (the slightly smaller ratio is to ensure that the smallest class is of a reasonable size). We use the long-tail version of ImageNet generated by Liu et al. [2017].

Dataset splits

The original test samples for CIFAR-10, CIFAR-10-LT, CIFAR-100, CIFAR-100-LT, TinyImageNet (200 classes), TinyImageNet-LT (200 classes), and ImageNet (1000 classes) are all balanced. Following Narasimhan and Menon [2021], we randomly split them in half and use half the samples as a validation set, and the other half as a test set. For the CIFAR and TineImageNet datasets, this amounts to using a validation set of size 5000. For the ImageNet dataset, we sample a subset of 5000 examples from the validation set each time we update the Lagrange multipliers in Algorithm 2.

In keeping with prior work Menon et al. [2021b], Narasimhan and Menon [2021], Lukasik et al. [2022], we use the same validation and test sets for the long-tailed training sets as we do for the original versions. For the long tailed training sets, this simulates a scenario where the training data follows a long tailed distribution due to practical data collection limitations, but the test distribution of interest still comes from the original data distribution. In plots, the "balanced accuracy" that we report for the long-tail datasets (e.g., CIFAR-10-LT) is actually the standard accuracy calculated over the balanced test set, which is shared with the original balanced dataset (e.g., CIFAR-10).

Both teacher and student were always trained on the same training set.

The CIFAR datasets had images of size 32 × 32, while the TinyImageNet and ImageNet datasets dataset had images of size 224 × 224.

These datasets do not contain personally identifiable information or offensive content. The CIFAR-10 and CIFAR-100 datasets are licensed under the MIT License. The terms of access for ImageNet are given at https://www.image-net.org/download.php.

Additional details about training and hyperparameters

Training details and hyperparameters

Temperature hyperparameters. We apply temperature scaling to the teacher scores on both the training set and validation set when training the student, i.e., compute $p^t(x) = \mathrm{softmax}(f^t(x)/\gamma)$, and vary the temperature parameter γ over a range of $\{1, 3, 5\}$. When training with teacher labels on the validation set (Algorithm 2), we vary the temperature parameters independently for the training set and the validation set. That is, we apply $p^t(x) = \mathrm{softmax}(f^t(x)/\gamma_{\text{train}})$ over the training set and $p^t(x) = \mathrm{softmax}(f^t(x)/\gamma_{\text{val}})$ over the validation set. When teacher labels are applied to the validation set, we additionally include a temperature of 0.1 on the teacher's validation set labels to approximate a hard thresholding of the teacher probabilities. Thus, the final hyperparameter search spaces are $\gamma_{\text{train}} \in \{1, 3, 5\}$, and $\gamma_{\text{val}} \in \{0.1, 1, 3, 5\}$.

Unless otherwise specified, in all tables, the temperature hyperparameters were chosen to achieve the best worst-class accuracy on the validation set. In all scatter plots such as Figure 3.1, for each α^t, α^s combination, temperature hyperparameters were selected to achieve the best worst-class accuracy on the validation set.

Learning rate hyperparameters. All models were trained using SGD with momentum of 0.9 [Lukasik et al., 2022, Narasimhan and Menon, 2021].

The learning rate schedule were chosen to mimic the settings in prior work Narasimhan and Menon [2021], Lukasik et al. [2022]. For CIFAR-10 and CIFAR-100 datasets, we ran the optimizer for 450 epochs, linearly warming up the learning rate till the 15th epoch, and then applied a step-size decay of 0.1 after the 200th, 300th and 400th epochs, as done by Lukasik et al. [2022]. For the long-tail versions of these datasets, we trained for 256 epochs, linearly warming up the learning rate till the 15th epoch, and then applied a step-size decay of 0.1 after the 96th, 192nd and 224th epochs, as done by Narasimhan and Menon [2021]. Similarly, for the TinyImageNet datasets, we train for 200 epochs, linearly warming up the learning rate till the 5th epoch, and then applying a decay of 0.1 after the 75th and 135th epochs, as done by Narasimhan and Menon [2021]. For ImageNet, we train for 90 epochs, linearly warming up the learning rate till the 5th epoch, then applying a decay of 0.1 after the 30th, 60th and 80th epochs, as done by Lukasik et al. [2022]. We used a batch size of 128 for the CIFAR-10 and the long-tailed TinyImageNet datasets [Narasimhan and Menon, 2021], a batch size of 512 for the balanced ImageNet dataset, a batch size of 2048 for the balanced TinyImageNet dataset, and a batch size of 1024 for other datasets Lukasik et al. [2022].

We apply an L_2 weight decay of 10^{-4} in all our SGD updates Lukasik et al. [2022]. This amounts to applying an L_2 *regularization* on the model parameters, and has the effect of keeping the model parameters (and as a result the loss function) bounded.

When training with the margin-based robust objective (see Algorithm 2), a separate step size α was applied for training the main model function f, and for updating the multipliers λ. We set α to 0.1 in all experiments.

Hardware. Model training was done using TPUv2.

Repeats

For all comparative baselines without distillation (Group DRO, Post shift, and all teachers alone), we provide average results over m retrained models ($m = 5$ for ImageNet / Tiny-ImageNet, or $m = 10$ for CIFAR datasets). For students on all CIFAR* datasets, unless otherwise specified, we train the teacher once and run the student training 10 times using the same arbitrarily chosen fixed teacher. We compute the mean and standard error of metrics over these $m = 10$ runs. For the resource-heavy TinyImageNet and ImageNet students, we reduce the number of repeats to $m = 5$. This methodology captures variation in the student retrainings while holding the teacher fixed. To capture the end-to-end variation in both teacher and student training, we include Appendix B.6 and Table B.4 which contains a rerun of the CIFAR experiments in Tables B.1 and B.2 using a distinct teacher for each student retraining. The overall best teacher/student objective combinations did not change for most datasets, with the only exception coming from a difference in the use of validation set labels.

Additional details about algorithms and baselines

Practical improvements to Algorithms 2–7

Algorithms 2–7 currently return a scorer that averages over all K iterates $\overline{f}^s(x) = \frac{1}{K}\sum_{k=1}^{K} f^k(x)$. While this averaging was required for our theoretical robustness guarantees to hold, in our experiments, we find it sufficient to simply return the last model f^K. Another practical improvement that we make to these algorithms following Cotter et al. [2019d], is to employ the 0-1 loss while performing updates on λ, i.e., set $\ell = \ell^{0\text{-}1}$ in the λ-update step. We are able to do this because the convergence of the exponentiated gradient updates on λ does not depend on ℓ being differentiable. This modification allows λs to better reflect the model's per-class performance on the validation sample.

Discussion on post-shifting baseline

We implement the post-shifting method in Narasimhan and Menon [2021] (Algorithm 3 in their paper), which provides for an efficient way to construct a scoring function of the form $f^s_y(x) = \log(\gamma_y p^t_y(x))$, for a fixed teacher p^t, where the coefficients $\gamma \in \mathbb{R}^m$ are chosen to maximize the worst-class accuracy on the validation dataset. Interestingly, in our experiments, we find this approach to do exceedingly well on the validation sample, but this does not always translate to good worst-class test performance. In contrast, some of the teacher-student combinations that we experiment with were seen to over-fit less to the validation sample, and as a result were able to generalize better to the test set. This could perhaps indicate that the teacher labels we use in these combinations benefit the student in a way that it improves its generalization. The variance reduction effect that Menon et al. [2021a] postulate may be one possible explanation for why we see this behavior.

B.6 Additional experimental results

This section contains additional experimental results.

Extended tables for objective combinations

We include extended tables comparing worst-class performance for different combinations of teacher and student objectives. The mean and standard errors are reported over repeat trainings as described in Appendix B.5.

Tables B.1 and B.2 are an extended version of Table 3.1 that includes standard errors for both worst-k accuracy and average accuracy.

Table B.3 includes similar comparisons when the student is compressed – that is, the student's architecture is smaller than the teacher's architecture.

Robust distillation with a onehot-labeled validation set

Tables B.1, B.2, and B.3 also include results when the robust student is trained using a validation set using onehot labels, as described in Appendix B.3. We report the accuracies for this robust student for different teachers trained with the standard, balanced, and robust objectives in the last rows of Tables B.1, B.2, and B.3 ($L^{\text{rob-d}}$ (one-hot val)). We compare these to the robust student trained using teacher labels on the validation set ($L^{\text{rob-d}}$ (teacher val)), which require less labeled data.

Perhaps surprisingly, it did not always benefit the robust student to utilize the true one-hot labels in the validation set. Instead, training the robust student with teacher labels on the validation set was often sufficient to achieve the best or close to the best worst-class performance. This is promising from a data efficiency standpoint, since it can be expensive to build up a labeled dataset for validation, especially if the training data is long-tailed.

Additional plots for all α^t, α^s combinations

Figure B.1 show accuracies for all α^t, α^s the equivalent of Figure 3.1 but for all datasets.

Table B.1: Worst-class accuracy comparison of self-distilled teacher/student combos on test for balanced datasets. The "none" row indicates the performance of the teacher alone. Worst-class accuracy is shown above, and average is accuracy shown in parentheses below. The combination with the best worst-class accuracy is in **bold**. We include results for the robust student using either a teacher labeled validation set ("teacher val"), or true one-hot class labels in the validation set ("one-hot val"), as outlined in Appendix B.3. Perhaps counterintuitively, the teacher with the best worst-class accuracy alone (the "none" row) did not always produce the student with the highest worst-class accuracy.

		CIFAR-10 Teacher Obj.		**CIFAR-100** Teacher Obj.		**TinyImageNet** Teacher Obj.	
		L^{std}	L^{rob}	L^{std}	L^{rob}	L^{std}	L^{rob}
Student Obj.	none	86.48 ± 0.32 (93.74 ± 0.05)	90.09 ± 0.22 (92.67 ± 0.09)	42.22 ± 0.90 72.42 ± 0.16	43.42 ± 1.03 68.81 ± 0.11	8.42 ± 1.88 (56.79 ± 0.33)	11.87 ± 1.74 (48.40 ± 0.15)
	$L^{\text{std-d}}$	87.66 ± 0.40 (94.34 ± 0.07)	90.12 ± 0.23 (94.07 ± 0.07)	43.81 ± 0.58 (74.61 ± 0.15)	$\mathbf{48.20} \pm 1.15$ (73.23 ± 0.07)	6.32 ± 2.31 (57.83 ± 0.13)	10.53 ± 1.49 (55.36 ± 0.16)
	$L^{\text{rob-d}}$ (teacher val)	$\mathbf{90.94} \pm 0.16$ (92.54 ± 0.05)	85.14 ± 0.47 (89.58 ± 0.11)	39.18 ± 1.58 (63.49 ± 0.29)	30.42 ± 1.30 (55.77 ± 0.39)	9.98 ± 1.87 (49.84 ± 0.21)	16.58 ± 1.23 (46.11 ± 0.37)
	$L^{\text{rob-d}}$ (one-hot val)	89.37 ± 0.17 (91.63 ± 0.06)	87.32 ± 0.21 (91.16 ± 0.10)	44.61 ± 1.55 (69.02 ± 0.30)	42.68 ± 0.74 (62.03 ± 0.24)	16.27 ± 0.43 (48.06 ± 0.24)	$\mathbf{17.36} \pm 1.32$ (43.92 ± 0.30)

Comparison to baselines of all Pareto efficient α^t, α^s

To supplement the comparison to baselines in Tables 3.2 and 3.3, Figures B.2, B.3, B.4, and B.5 show all Pareto efficient α^t and α^s combinations on test. Whereas only a single α^t, α^s combination was selected on the validation set and reported in Tables 3.2 and 3.3, Figures B.2, B.3, B.4, and B.5 show that there were many more combinations of α^t, α^s that could have Pareto dominated all baselines.

Figures B.2, B.3, B.4, and B.5 also give more insight into which values of α^t work best for different values of α^s. Whereas Figure 3.1 shows that α^s is highly correlated with average accuracy, the same is not true for α^t. Worst-class accuracy generally increases with α^s, but the teachers that achieve the Pareto efficient points all have $\alpha^t < 1$. This reveals counterintuitively that the teacher's worst-class accuracy is not a direct predictor of the robustness of a subsequent student. This couples with our theoretical understanding in Section 3.7, which showed that the ability of a teacher to train robust students is determined by the calibration of scores within each class.

Trading off average vs. worst-class accuracy. Figures B.2, B.3, B.4, and B.5 show that when we allow for more nuanced L^{tdf} objective combinations, the resulting models may have higher average accuracy and worst-class accuracy than standard distillation. Interestingly, the models with the most "even" trade-offs between average accuracy and worst-class accuracy tend to have low α^t (around 0.25) and low α^s (also around 0.25). Higher values of α^t tended

Table B.2: Worst-class accuracy comparison of self-distilled teacher/student combos on test for long-tailed datasets. The "none" row indicates the performance of the teacher alone. Worst-class accuracy is shown above (or worst-10 accuracy for TinyImageNet-LT), and average is accuracy shown in parentheses below. The combination with the best worst-class accuracy is in **bold**. We include results for the robust student using either a teacher labeled validation set ("teacher val"), or true one-hot class labels in the validation set ("one-hot val"), as outlined in Appendix B.3. Perhaps counterintuitively, the teacher with the best worst-class accuracy alone (the "none" row) did not always produce the student with the highest worst-class accuracy.

		CIFAR-10-LT Teacher Obj.			**CIFAR-100-LT** Teacher Obj.		
		L^{std}	L^{bal}	L^{rob}	L^{std}	L^{bal}	L^{rob}
	None	57.26 ± 0.55 (76.27 ± 0.20)	68.52 ± 0.52 (79.85 ± 0.20)	74.8 ± 0.30 (80.29 ± 0.12)	0.00 ± 0.00 (43.33 ± 0.16)	3.75 ± 0.62 (47.55 ± 0.17)	10.33 ± 0.82 (44.27 ± 0.13)
Student Obj.	$L^{\text{std-d}}$	36.67 ± 0.28 (69.5 ± 0.13)	66.96 ± 0.43 (79.25 ± 0.10)	71.15 ± 0.24 (80.95 ± 0.11)	0.00 ± 0.00 (43.86 ± 0.14)	2.39 ± 0.24 (48.95 ± 0.15)	7.32 ± 0.47 (47.93 ± 0.11)
	$L^{\text{bal-d}}$	71.23 ± 0.44 (80.5 ± 0.12)	70.52 ± 0.20 (81.12 ± 0.08)	72.96 ± 0.53 (80.71 ± 0.07)	4.39 ± 0.65 (50.4 ± 0.11)	7.08 ± 0.80 (50.1 ± 0.09)	7.19 ± 0.79 (47.51 ± 0.20)
	$L^{\text{rob-d}}$ (teacher val)	63.85 ± 0.21 (76.81 ± 0.08)	$\mathbf{75.56} \pm 0.19$ (80.81 ± 0.08)	69.21 ± 0.45 (76.72 ± 0.19)	9.05 ± 0.71 (33.75 ± 0.10)	12.52 ± 0.98 (34.05 ± 0.09)	10.32 ± 0.76 (36.83 ± 0.15)
	$L^{\text{rob-d}}$ (one-hot val)	73.59 ± 0.25 (77.92 ± 0.05)	75.43 ± 0.38 (79.02 ± 0.07)	74.7 ± 0.19 $(77.99 \pm 0.10$	12.28 ± 0.46 (30.79 ± 0.18)	11.94 ± 0.80 (29.8 ± 0.20)	$\mathbf{13.18} \pm 0.61$ $(31.88 \pm 0.20$

		TinyImageNet-LT Teacher Obj.		
		L^{std}	L^{bal}	L^{rob}
	None	0.00 ± 0.00 (33.15 ± 0.17)	2.11 ± 0.37 (35.96 ± 0.12)	4.92 ± 0.66 (27.23 ± 0.15)
Student Obj.	$L^{\text{std-d}}$	0.00 ± 0.00 (26.05 ± 0.18)	0.00 ± 0.00 (27.21 ± 0.15)	1.87 ± 0.23 (25.34 ± 0.13)
	$L^{\text{bal-d}}$	0.20 ± 0.18 (30.43 ± 0.06)	2.82 ± 0.14 (39.41 ± 0.15)	4.77 ± 0.41 (38.41 ± 0.15)
	$L^{\text{rob-d}}$ (teacher val)	0.00 ± 0.00 (22.66 ± 0.08)	4.93 ± 0.38 (35.43 ± 0.18)	3.32 ± 0.43 (25.11 ± 0.17)
	$L^{\text{rob-d}}$ (one-hot val)	1.55 ± 0.37 (21.59 ± 0.19)	6.11 ± 0.39 (28.24 ± 0.17)	$\mathbf{6.19} \pm 0.25$ (25.30 ± 0.18)

to lead to more extreme points on the trade-off curve, either with higher average accuracy at the expense of worst-class accuracy, or vice versa. Overall, the robust L^{tdf} combinations also Pareto dominated most of the baselines that all used the standard teacher. Together, these results highlight the fact that in robust distillation, the teacher's training objective is important and should be tailored to the desired final accuracy/robustness trade-off (perhaps using a held-out validation sample with some domain-specific criteria in practice). Figure

Table B.3: Comparison of ResNet-56→ResNet-32 distilled teacher/student combos on test on CIFAR datasets. Worst-class accuracy shown above, and average accuracy shown in parentheses below. The combination with the best worst-class accuracy is bolded. Mean and standard error are reported over 10 repeats. We include results for the robust student using either a teacher labeled validation set ("teacher val"), or true one-hot class labels in the validation set ("one-hot val"), as outlined in Section 3.5.

Student Obj. (ResNet-32)	**CIFAR-10** Teacher Obj.		**CIFAR-100** Teacher Obj.	
	L^{std}	L^{rob}	L^{std}	L^{rob}
$L^{\text{std-d}}$	86.4 ± 0.27 (93.73 ± 0.05)	89.56 ± 0.20 (93.38 ± 0.05)	41.82 ± 1.12 73.19 ± 0.10	$\mathbf{45.7} \pm 1.13$ 71.42 ± 0.22
$L^{\text{rob-d}}$ (teacher val)	$\mathbf{89.61} \pm 0.27$ (92.20 ± 0.08)	83.8 ± 0.95 (88.71 ± 0.24)	38.94 ± 2.61 (62.28 ± 0.40)	19.15 ± 0.00 (52.9 ± 0.00)
$L^{\text{rob-d}}$ (one-hot val)	87.92 ± 0.23 (90.89 ± 0.12)	86.57 ± 0.24 (90.54 ± 0.11)	33.19 ± 1.29 (57.43 ± 0.29)	41.23 ± 0.84 (61.14 ± 0.24)

Student Obj. (ResNet-32)	**CIFAR-10-LT** Teacher Obj.			**CIFAR-100-LT** Teacher Obj.		
	L^{std}	L^{bal}	L^{rob}	L^{std}	L^{bal}	L^{rob}
$L^{\text{std-d}}$	57.23 ± 0.53 (75.76 ± 0.12)	66.80 ± 0.25 (78.99 ± 0.06)	72.36 ± 0.39 (80.74 ± 0.09)	0.00 ± 0.00 (44.33 ± 0.11)	1.38 ± 0.39 (47.28 ± 0.13)	7.99 ± 0.48 (47.34 ± 0.08)
$L^{\text{bal-d}}$	71.37 ± 0.50 (81.13 ± 0.12)	71.00 ± 0.45 (81.12 ± 0.15)	72.17 ± 0.40 (79.91 ± 0.08)	3.57 ± 0.58 (49.21 ± 0.10)	4.28 ± 0.45 (46.56 ± 0.13)	5.58 ± 0.53 (48.58 ± 0.09)
$L^{\text{rob-d}}$ (teacher val)	64.1 ± 0.36 (76.34 ± 0.12)	73.51 ± 0.33 (80.10 ± 0.10)	69.90 ± 0.42 (76.37 ± 0.14)	10.24 ± 0.71 (33.55 ± 0.16)	$\mathbf{13.41} \pm 0.72$ (33.37 ± 0.17)	11.27 ± 0.61 (36.14 ± 0.19)
$L^{\text{rob-d}}$ (one-hot val)	72.65 ± 0.27 (77.69 ± 0.11)	74.39 ± 0.34 (78.68 ± 0.16)	$\mathbf{74.45} \pm 0.26$ $(77.97 \pm 0.10$	10.93 ± 0.65 (29.48 ± 0.22)	12.2 ± 0.65 (30.27 ± 0.18)	12.93 ± 0.62 $(31.83 \pm 0.17$

B.6 confirms that these results also hold up in a compression setting, where the compressed models can actually even beat their larger teachers.

Different teachers on repeat trainings

Distillation experimental results in the main paper use the same teacher for all repeat trainings of the student. This captures the variance in the student training process while omitting the variance in the teacher training process. To capture the variance in the full training pipeline, we ran an additional set of experiments where students were trained on different retrained teachers, rather than on the same teacher. We report results on all CIFAR datasets in Table B.4. The best teacher/student combinations are identical for all datasets except for CIFAR-10-LT, for which the best teacher/student combinations from Table B.4 and Tables B.1 and B.2 were both a robust student trained with a balanced teacher, and only differed in

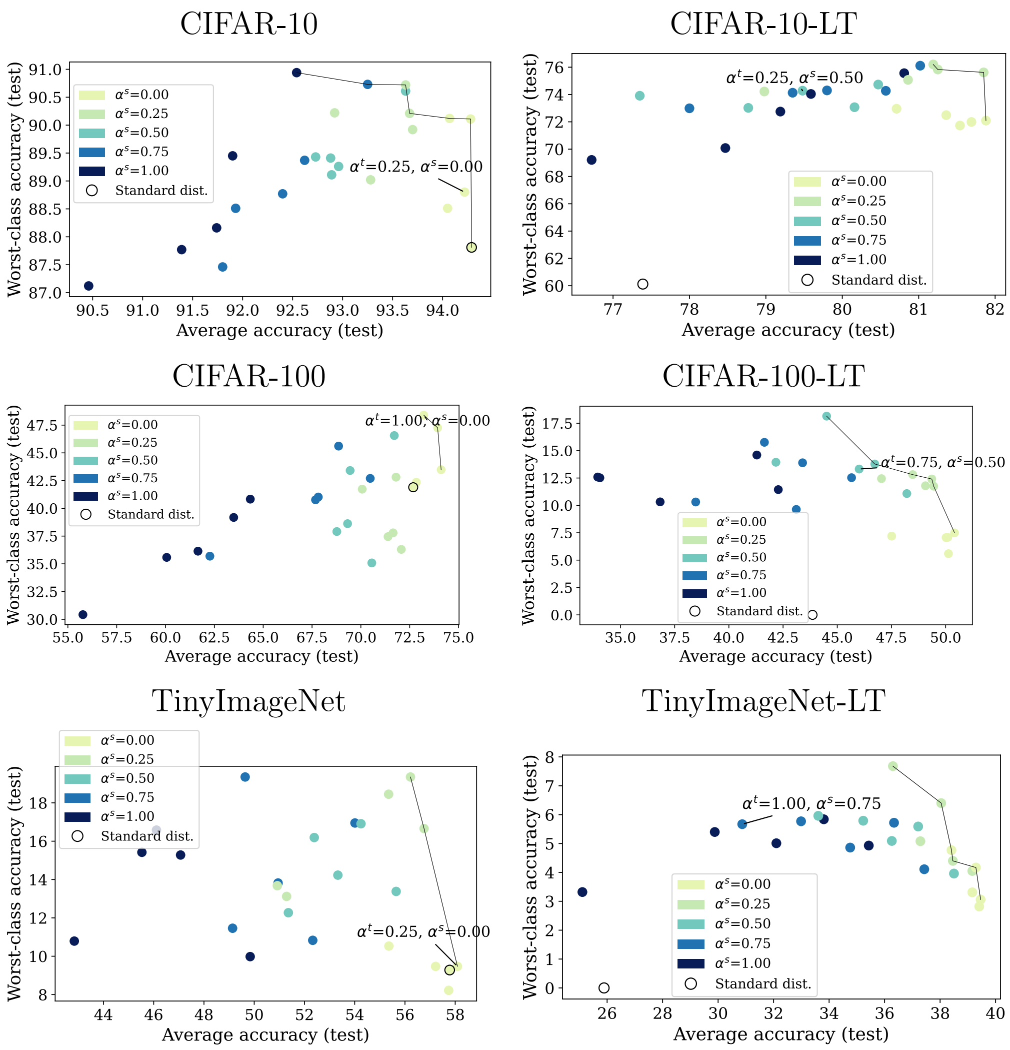
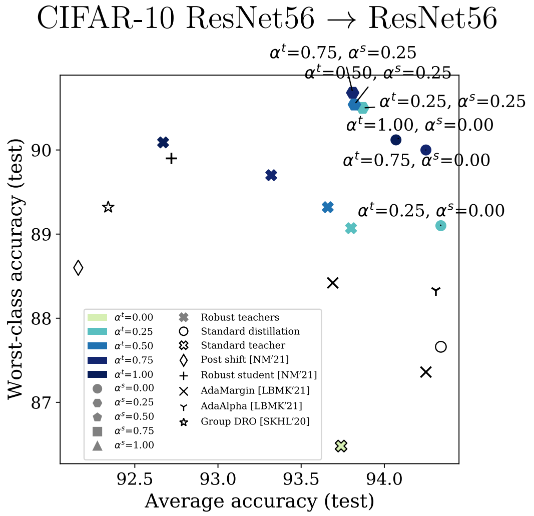


Figure B.1: All α^t, α^s combinations for all datasets on test. The black line traces out the Pareto frontier. Average accuracy is roughly determined by α^s. The labeled point corresponds to the "best" combination selected in Tables 3.2 and 3.3 based on validation criteria, but other domain-specific tradeoff criteria could yield any of these other points.

whether the validation set contained teacher labels or one-hot labels ($L^{\text{bal}}/L^{\text{rob-d}}$ (one-hot val) in Table B.4 vs. $L^{\text{bal}}/L^{\text{rob-d}}$ (teacher val) in Tables B.1 and B.2). Note that the first and second rows of Tables B.1 and B.2 are already averaged over m retrained teachers ($m = 5$ for TinyImageNet, or $m = 10$ for CIFAR datasets), and those same m teachers are used in the repeat trainings in Table B.4.

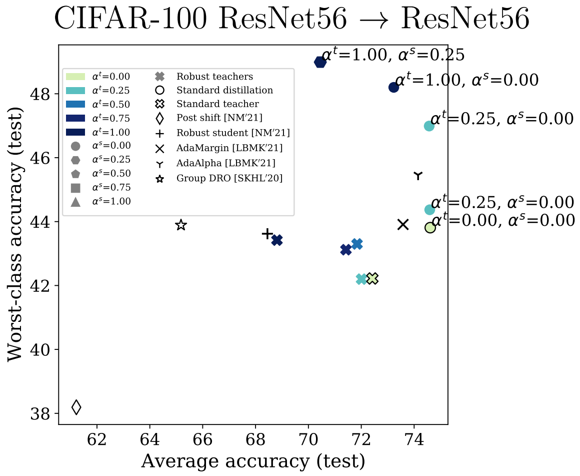


Pareto efficient robust distillation results (test)

α^t	α^s	Average acc.	Worst-class acc.
0.75	0.25	93.81 ± 0.07	90.68 ± 0.20
0.50	0.25	93.82 ± 0.09	90.54 ± 0.22
0.25	0.25	93.87 ± 0.08	90.50 ± 0.18
1.00	0.00	94.07 ± 0.07	90.12 ± 0.23
0.75	0.00	94.25 ± 0.05	90.00 ± 0.17
0.25	0.00	94.34 ± 0.06	89.10 ± 0.31

Baseline results (test)

Baseline	Average acc.	Worst-class acc.
Standard distillation	94.34 ± 0.07	87.66 ± 0.40
Post shift [NM'21]	92.16 ± 0.18	88.60 ± 0.35
Robust student [NM'21]	92.72 ± 0.05	89.90 ± 0.21
AdaMargin [LBMK'21]	93.69 ± 0.06	88.42 ± 0.36
AdaAlpha [LBMK'21]	94.31 ± 0.01	88.33 ± 0.14
Group DRO [SKHL'20]	92.34 ± 0.07	89.32 ± 0.21



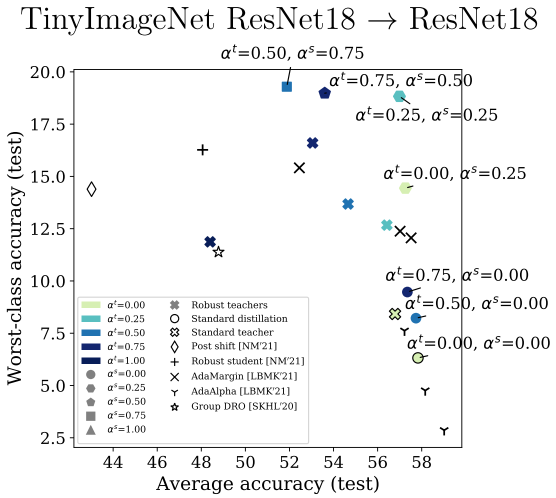
Pareto efficient robust distillation results (test)

α^t	α^s	Average acc.	Worst-class acc.
1.00	0.25	70.45 ± 0.16	48.99 ± 0.72
1.00	0.00	73.23 ± 0.07	48.20 ± 1.15
0.25	0.00	74.57 ± 0.12	46.99 ± 1.09
0.25	0.00	74.59 ± 0.09	44.37 ± 0.58
0.00	0.00	74.61 ± 0.15	43.81 ± 0.58

Baseline results (test)

Baseline	Average acc.	Worst-class acc.
Standard distillation	74.61 ± 0.15	43.81 ± 0.58
Post shift [NM'21]	61.22 ± 0.36	38.19 ± 0.40
Robust student [NM'21]	68.45 ± 0.13	43.62 ± 1.27
AdaMargin [LBMK'21]	73.58 ± 0.11	43.91 ± 1.11
AdaAlpha [LBMK'21]	74.15 ± 0.08	45.46 ± 0.67
Group DRO [SKHL'20]	65.18 ± 0.08	43.89 ± 1.12

Figure B.2: Tradeoffs in worst-class test accuracy vs. average test accuracy for CIFAR-10 and CIFAR-100 distilling from ResNet-56 to ResNet-56. All baseline results that require a teacher use the "standard teacher" (trained using L^{std}), as done in the original papers. For methods run multiple times with multiple hyperparameters (e.g. temperatures), all Pareto efficient results are shown in the plot, but the tables show only the baseline results with the best worst-class accuracy (on the validation set). The highlighted row indicates the model with the highest worst-class accuracy that also achieves at least as high average accuracy as *standard distillation* (within error margins).



Pareto efficient robust distillation results (test)

α^t	α^s	Average acc.	Worst-class acc.
0.50	0.75	51.88 ± 0.18	19.29 ± 1.27
0.75	0.50	53.60 ± 0.31	18.98 ± 0.86
0.25	0.25	56.99 ± 0.14	18.83 ± 0.85
0.00	0.25	57.26 ± 0.15	14.44 ± 0.91
0.75	0.00	57.35 ± 0.17	9.47 ± 1.76
0.50	0.00	57.74 ± 0.20	8.22 ± 1.09

Baseline results (test)

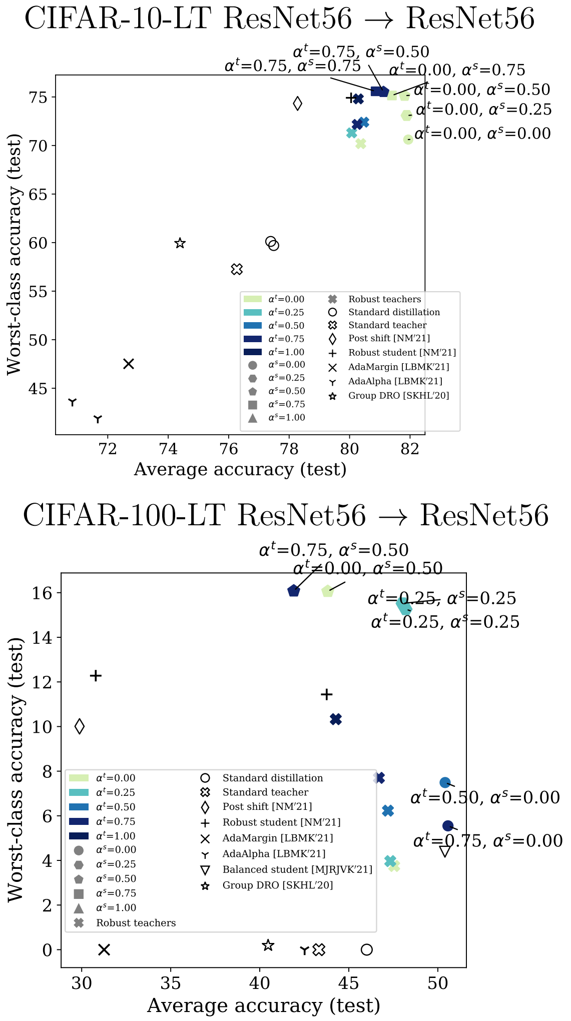
Baseline	Average acc.	Worst-class acc.
Standard distillation	57.83 ± 0.13	6.32 ± 2.31
Post shift [NM'21]	43.02 ± 0.79	14.39 ± 1.13
Robust student [NM'21]	48.06 ± 0.24	16.27 ± 0.43
AdaMargin [LBMK'21]	52.45 ± 0.08	15.41 ± 0.71
AdaAlpha [LBMK'21]	57.22 ± 0.08	7.62 ± 2.17
Group DRO [SKHL'20]	48.78 ± 0.21	11.38 ± 1.79

Figure B.3: Tradeoffs in worst-class test accuracy vs. average test accuracy for TinyImageNet distilling from ResNet-18 to ResNet-18. All baseline results that require a teacher use the "standard teacher" (trained using L^{std}), as done in the original papers. For methods run multiple times with multiple hyperparameters (e.g. temperatures), all Pareto efficient results are shown in the plot, but the tables show only the baseline results with the best worst-class accuracy (on the validation set). The highlighted row indicates the model with the highest worst-class accuracy that also achieves at least as high average accuracy as *standard distillation* (within error margins).

AdaAlpha and AdaMargin comparisons with different teachers

We include and discuss additional comparisons to the AdaMargin and AdaAlpha methods Lukasik et al. [2022], which each define additional ways to modify the student training algorithm (see Section 3.6). In Tables 3.2 and 3.3, we show results with each of these methods using the standard teacher, as done in the original paper. However, in this section we extend these results by also applying AdaMargin and AdaAlpha with different teachers trained with the robust and balanced objectives. Table B.5 compares the results of AdaMargin and AdaAlpha for these different teachers under the same self distillation setup as Tables B.1 and B.2.

Overall, the use of a robust teacher leads to marked improvements for students trained by AdaMargin and AdaAlpha. For the balanced datasets, AdaMargin was competitive with the robust and standard students: on CIFAR-100 and TinyImageNet, AdaMargin combined with the robust teacher and the standard teacher (respectively) achieved worst-class accuracies that are statistically comparable to the best worst-class accuracies in Tables B.1 and B.2. However, on the long-tailed datasets, AdaAlpha and AdaMargin did not achieve worst-class accuracies as high as other teacher/student combinations. This suggests that the AdaMargin method can work well on balanced datasets in combination with a robust teacher, but other


CIFAR-10-LT ResNet56 → ResNet56

CIFAR-100-LT ResNet56 → ResNet56

Pareto efficient robust distillation results (test)

α^t	α^s	Average acc.	Worst-class acc.
0.75	0.75	80.86 ± 0.09	75.58 ± 0.17
0.75	0.50	81.12 ± 0.11	75.52 ± 0.22
0.00	0.75	81.40 ± 0.10	75.15 ± 0.38
0.00	0.50	81.82 ± 0.11	75.13 ± 0.24
0.00	0.25	81.89 ± 0.08	73.09 ± 0.32
0.00	0.00	81.94 ± 0.16	70.61 ± 0.39

Baseline results (test)

Baseline	Average acc.	Worst-class acc.
Standard distillation	77.39 ± 0.10	60.12 ± 0.56
Post shift [NM'21]	78.28 ± 0.05	74.33 ± 0.09
Robust student [NM'21]	80.05 ± 0.13	74.91 ± 0.24
AdaMargin [LBMK'21]	72.69 ± 0.24	47.52 ± 0.95
AdaAlpha [LBMK'21]	70.83 ± 0.28	43.64 ± 1.09
Group DRO [SKHL'20]	74.39 ± 0.17	59.93 ± 0.59

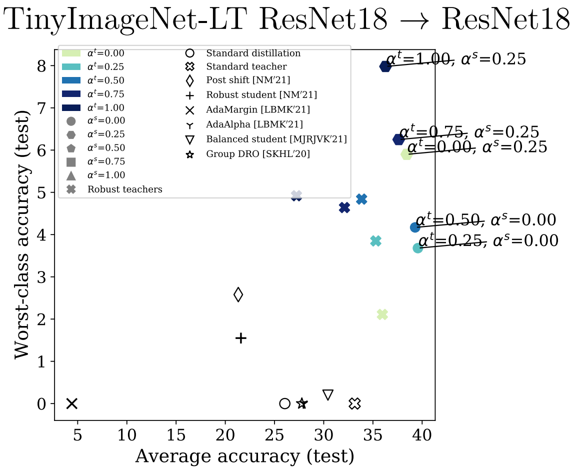
Pareto efficient robust distillation results (test)

α^t	α^s	Average acc.	Worst-class acc.
0.75	0.50	41.91 ± 0.15	16.08 ± 0.52
0.00	0.50	43.82 ± 0.14	16.06 ± 0.89
0.25	0.25	48.01 ± 0.09	15.52 ± 0.41
0.25	0.25	48.20 ± 0.11	15.26 ± 0.73
0.50	0.00	50.41 ± 0.11	7.49 ± 0.72
0.75	0.00	50.57 ± 0.18	5.55 ± 0.54

Baseline results (test)

Baseline	Average acc.	Worst-class acc.
Standard distillation	46.01 ± 0.16	0.00 ± 0.00
Post shift [NM'21]	29.88 ± 0.61	10.01 ± 0.72
Robust student [NM'21]	30.79 ± 0.18	12.28 ± 0.46
AdaMargin [LBMK'21]	31.26 ± 0.21	0.00 ± 0.00
AdaAlpha [LBMK'21]	42.52 ± 0.08	0.00 ± 0.00
Balanced student [MJRJVK'21]	50.40 ± 0.12	4.39 ± 0.66
Group DRO [SKHL'20]	40.47 ± 0.17	0.19 ± 0.17

Figure B.4: Tradeoffs in worst-class test accuracy vs. average test accuracy for CIFAR-10-LT and CIFAR-100-LT under self-distillation. All baseline results that require a teacher use the "standard teacher" (trained using L^{std}), as done in the original papers. For methods run multiple times with multiple hyperparameters (e.g. temperatures), all Pareto efficient results are shown in the plot, but the tables show only the baseline results with the best worst-class accuracy (on the validation set). The highlighted row indicates the model with the highest worst-class (or worst-10) accuracy that also achieves at least as high average accuracy as *standard distillation* (within error margins). Note that L^{tdf} mixes between L^{bal} and L^{rob}.


TinyImageNet-LT ResNet18 → ResNet18

Pareto efficient robust distillation results (test)

α^t	α^s	Average acc.	Worst-10 acc.
1.00	0.25	36.28 ± 0.17	7.98 ± 0.21
0.75	0.25	37.62 ± 0.15	6.25 ± 0.12
0.00	0.25	38.44 ± 0.13	5.90 ± 0.45
0.50	0.00	39.29 ± 0.09	4.17 ± 0.34
0.25	0.00	39.57 ± 0.06	3.68 ± 0.30

Baseline results (test)

Baseline	Average acc.	Worst-10 acc.
Standard distillation	26.05 ± 0.18	0.00 ± 0.00
Post shift [NM'21]	21.32 ± 0.49	2.58 ± 0.42
Robust student [NM'21]	21.59 ± 0.19	1.55 ± 0.37
AdaMargin [LBMK'21]	4.41 ± 0.09	0.00 ± 0.00
AdaAlpha [LBMK'21]	27.95 ± 0.14	0.00 ± 0.00
Balanced student [MJRJVK'21]	30.43 ± 0.06	0.20 ± 0.18
Group DRO [SKHL'20]	27.78 ± 0.13	0.00 ± 0.00

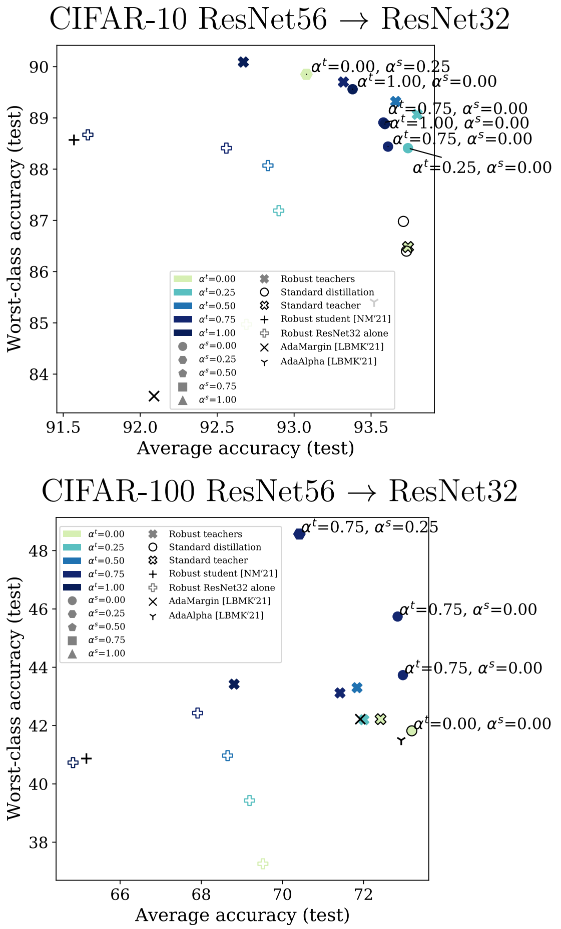
Figure B.5: Tradeoffs in worst-class test accuracy vs. average test accuracy for TinyImageNet-LT under self-distillation. All baseline results that require a teacher use the "standard teacher" (trained using L^{std}), as done in the original papers. For methods run multiple times with multiple hyperparameters (e.g. temperatures), all Pareto efficient results are shown in the plot, but the tables show only the baseline results with the best worst-class accuracy (on the validation set). The highlighted row indicates the model with the highest worst-10 accuracy that also achieves at least as high average accuracy as *standard distillation* (within error margins). Note that L^{tdf} mixes between L^{bal} and L^{rob}.

combinations of standard/balanced/robust objectives are valuable for long-tailed datasets.

Relative to each other, AdaMargin usually achieved higher worst-class accuracy than AdaAlpha, whereas AdaAlpha often achieved higher average accuracy.

Group DRO comparison

Sagawa et al. [2020a] propose a group DRO algorithm to improve long tail performance without distillation. In this section we present additional experimental comparisons to Algorithm 1 from Sagawa et al. [2020a]. This differs from our robust optimization methodology in Section 3.5 in two key ways: *(i)* we apply a margin-based surrogates of Menon et al. [2021b], and *(ii)* we use a validation set to update the Lagrange multipliers λ in Algorithm 7. Table B.6 shows results from running group DRO directly as specified in Algorithm 1 in Sagawa et al. [2020a], as well as a variant where we use the validation set to update Lagrange multipliers in group DRO (labeled as "with vali" in Table B.6). Table B.6 shows that this latter variant "with vali" performs better than the original version without a validation set; thus, for the results in Figures B.2, B.3, B.4, and B.5, we report these better results marked in Table B.6 as "with vali." Overall, this comparison shows that L^{rob} is comparable to group DRO, and that robust distillation protocols can outperform group DRO alone.



Pareto efficient robust distillation results (test)

α^t	α^s	Average acc.	Worst-class acc.
0.00	0.25	93.08 ± 0.07	89.85 ± 0.22
1.00	0.00	93.38 ± 0.05	89.56 ± 0.20
0.75	0.00	93.58 ± 0.09	88.91 ± 0.25
1.00	0.00	93.59 ± 0.06	88.88 ± 0.36
0.75	0.00	93.61 ± 0.05	88.44 ± 0.33
0.25	0.00	93.74 ± 0.07	88.41 ± 0.32

Baseline results (test)

Baseline	Average acc.	Worst-class acc.
Standard distillation	93.71 ± 0.05	86.98 ± 0.36
Robust student [NM'21]	91.57 ± 0.08	88.57 ± 0.18
AdaMargin [LBMK'21]	92.09 ± 0.09	83.57 ± 0.64
AdaAlpha [LBMK'21]	93.52 ± 0.11	85.41 ± 0.45

Pareto efficient robust distillation results (test)

α^t	α^s	Average acc.	Worst-class acc.
0.75	0.25	70.42 ± 0.14	48.57 ± 0.55
0.75	0.00	72.84 ± 0.22	45.74 ± 1.57
0.75	0.00	72.97 ± 0.18	43.73 ± 1.72

Baseline results (test)

Baseline	Average acc.	Worst-class acc.
Standard distillation	73.19 ± 0.10	41.82 ± 1.12
Robust student [NM'21]	65.17 ± 0.11	40.87 ± 0.89
AdaMargin [LBMK'21]	71.92 ± 0.17	42.22 ± 1.65
AdaAlpha [LBMK'21]	72.93 ± 0.09	41.50 ± 1.14

Figure B.6: Tradeoffs in worst-class test accuracy vs. average test accuracy for CIFAR-10 and CIFAR-100 distilling from ResNet-56 to ResNet-32. All baseline results that require a teacher use the "standard teacher" (trained using L^{std}), as done in the original papers. For methods run multiple times with multiple hyperparameters (e.g. temperatures), all Pareto efficient results are shown in the plot, but the tables show only the baseline results with the best worst-class accuracy (on the validation set). The highlighted row indicates the model with the highest worst-class accuracy that also achieves at least as high average accuracy as *standard distillation* (within error margins).

Table B.4: Comparison using different teachers for student retrainings for self-distilled teacher/student combos on test. For each student/teacher objective pair, we train $m = 10$ students total on each of $m = 10$ distinct retrained teachers. For comparability, the same set of m teachers is used for each student. This differs from Table 3.1 in that in Table 3.1, the students are retrained on each repeat using the same teacher (arbitrarily selected). Otherwise, setups are the same as in Table 3.1.

		CIFAR-10 Teacher Obj.		**CIFAR-100**Teacher Obj.	
		L^{std}	L^{rob}	L^{std}	L^{rob}
Student Obj.	$L^{\text{std-d}}$	87.09 ± 0.51 (93.78 ± 0.22)	89.68 ± 0.20 (93.74 ± 0.07)	44.21 ± 0.57 74.6 ± 0.11	$\mathbf{47.79} \pm 0.82$ 73.48 ± 0.11
	$L^{\text{rob-d}}$ (teacher val)	$\mathbf{90.62} \pm 0.19$ (92.58 ± 0.08)	87.12 ± 0.38 (90.46 ± 0.08)	39.7 ± 1.32 (64.28 ± 0.41)	31.09 ± 1.21 (55.39 ± 0.28)
	$L^{\text{rob-d}}$ (one-hot val)	88.15 ± 0.66 (91.03 ± 0.47)	86.44 ± 0.52 (90.16 ± 0.42)	39.44 ± 0.94 (61.23 ± 0.36)	39.65 ± 0.59 (60.89 ± 0.29)

		CIFAR-10-LT Teacher Obj.			**CIFAR-100-LT** Teacher Obj.		
		L^{std}	L^{bal}	L^{rob}	L^{std}	L^{bal}	L^{rob}
Student Obj.	$L^{\text{std-d}}$	60.12 ± 0.56 (77.39 ± 0.10)	66.13 ± 0.47 (79.16 ± 0.20)	69.75 ± 0.52 (80.73 ± 0.08)	0.00 ± 0.00 (45.84 ± 0.13)	1.41 ± 0.41 (49.67 ± 0.20)	9.17 ± 0.74 (48.55 ± 0.14)
	$L^{\text{bal-d}}$	72.41 ± 0.52 (81.97 ± 0.11)	71.49 ± 0.30 (81.20 ± 0.15)	71.70 ± 0.33 (80.29 ± 0.11)	5.83 ± 0.54 (50.58 ± 0.15)	5.94 ± 0.50 (50.85 ± 0.14)	8.37 ± 0.72 (48.16 ± 0.20)
	$L^{\text{rob-d}}$ (teacher val)	62.77 ± 0.58 (77.18 ± 0.15)	73.09 ± 0.34 (80.03 ± 0.22)	68.04 ± 0.47 (75.36 ± 0.25)	10.53 ± 0.76 (33.69 ± 0.14)	12.04 ± 0.89 (34.08 ± 0.12)	9.66 ± 1.15 (37.10 ± 0.15)
	$L^{\text{rob-d}}$ (one-hot val)	$\mathbf{75.10} \pm 0.36$ (79.27 ± 0.13)	$\mathbf{75.10} \pm 0.50$ (79.07 ± 0.20)	74.16 ± 0.34 $(78.11 \pm 0.14$	10.74 ± 0.44 (30.36 ± 0.39)	11.95 ± 0.69 (31.00 ± 0.16)	$\mathbf{12.87} \pm 0.81$ $(31.62 \pm 0.34$

Table B.5: Results for AdaAlpha and AdaMargin baselines for different teachers under self-distillation. For all CIFAR datasets, self-distillation is done from ResNet56 → ResNet56. For TinyImageNet, self-distillation is done from ResNet18 → ResNet18. Worst-class accuracy shown above (or worst-10 accuracy for TinyImageNet-LT), and average accuracy is shown in parentheses below. The temperature hyperparameter was tuned to maximize worst-class accuracy on the held-out validation set. Mean and standard error are reported over 5 repeats for all datasets.

	CIFAR-10 Teacher Obj.		**CIFAR-100** Teacher Obj.		**TinyImageNet** Teacher Obj.	
	L^{std}	L^{rob}	L^{std}	L^{rob}	L^{std}	L^{rob}
Ada	88.33 ± 0.14	89.96 ± 0.44	43.50 ± 0.62	45.59 ± 0.82	11.11 ± 1.29	16.58 ± 1.67
Alpha	(94.31 ± 0.01)	(93.97 ± 0.07)	73.96 ± 0.09	71.42 ± 0.14	61.13 ± 0.09	56.84 ± 0.15
Ada	87.36 ±0.06	90.37±0.26	43.91 ±1.11	47.78 ±0.96	18.17 ±3.89	17.84 ±1.77
Margin	(94.25 ± 0.02)	(94.02 ± 0.12)	(73.58 ± 0.11)	(70.92 ± 0.09)	(61.3 ± 0.28)	(55.77 ± 0.32)

	CIFAR-10-LT Teacher Obj.			**CIFAR-100-LT** Teacher Obj.		
	L^{std}	L^{bal}	L^{rob}	L^{std}	L^{bal}	L^{rob}
Ada	41.90 ± 0.44	66.23 ± 0.39	71.17 ± 0.32	0.00 ± 0.00	1.46 ± 0.61	9.15 ± 0.54
Alpha	(71.67 ± 0.08)	(77.87 ± 0.16)	(79.66 ± 0.13)	(42.52 ± 0.08)	(45.44 ± 0.14)	(45.64 ± 0.11)
Ada	47.52 ±0.95	66.74 ±0.35	70.33 ±0.50	0.00 ±0.00	0.00 ±0.00	12.46 ±0.36
Margin	(72.69 ± 0.24)	(78.20 ± 0.09)	(78.87 ± 0.12)	(31.26 ± 0.21)	(34.06 ± 0.12)	(42.90 ± 0.07)

	TinyImageNet-LT Teacher Obj.		
	L^{std}	L^{bal}	L^{rob}
Ada	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Alpha	(28.14 ± 0.12)	(0.50 ± 0.00)	(0.50 ± 0.00)
Ada	0.00 ± 0.00	0.00 ± 0.00	0.41 ± 0.17
Margin	(9.18 ± 0.09)	(7.92 ± 0.10)	(23.08 ± 0.15)

Table B.6: Results from comparison to group DRO (Algorithm 1 in Sagawa et al. [2020a]) without distillation. "No vali" uses the training set to update group Lagrange multipliers, as done originally by Sagawa et al. [2020a]. "With vali" uses the validation set to compute group Lagrange multipliers as done in all other experiments in our paper. Worst-class accuracy is shown above, and balanced accuracy is shown in parentheses below. Mean and standard error are shown over 5 repeats.

CIFAR-10 group DRO		**CIFAR-100** group DRO		**TinyImageNet** group DRO	
No vali	With vali	No vali	With vali	No vali	With vali
86.65 ±0.49	89.32 ±0.21	40.35 ±1.18	43.89 ±1.12	0.00 ±0.00	9.17 ±1.55
(93.61 ± 0.09)	(92.34 ± 0.07)	70.25 ± 0.17	65.18 ± 0.08	(6.55 ± 0.41)	(47.67 ± 0.22)

CIFAR-10-LT group DRO		**CIFAR-100-LT** group DRO		**TinyImageNet-LT** group DRO	
No vali	With vali	No vali	With vali	No vali	With vali
51.59 ±2.49	59.93 ±0.59	0.00 ±0.00	0.19 ±0.17	0.00 ±0.00	0.00 ±0.00
(71.94 ± 0.75)	(74.39 ± 0.17)	(39.81 ± 0.23)	(40.47 ± 0.17)	(9.79 ± 0.40)	(22.49 ± 0.10)

Appendix C

Deferred Proofs and Discussion for Chapter 4

C.1 Proofs

Lemma 2. *Let $(\Omega, \mathcal{F})$ be a measurable space with a regular conditional probability property, and let $X : \Omega \to \mathbb{R}^D$, $Z : \Omega \to \mathbb{R}$ be $\mathcal{F}$-measurable random variables. Suppose P_j and P_k are σ-finite probability measures on $(\Omega, \mathcal{F})$, where P_j denotes the conditional probability measure of X given that $Z = j$, and P_k denote the same for $Z = k$, and P_j is absolutely continuous with respect to P_k. Let $f : \mathbb{R}^D \times \mathbb{R} \to \mathbb{R}$ be defined as in Section 4.3, and $f(x, z) \geq 0$ for all $x \in \mathbb{R}^D$, $z \in \mathbb{R}$. If the function f satisfies monotonicity in the second argument such that $f(x, j) \leq f(x, k)$ for all $x \in \mathbb{R}^D$ and for $j \leq k$, and if the Radon Nikodym derivative $\frac{dP_j}{dP_k}$ is bounded almost everywhere with respect to P_k by a finite constant $C > 0$, then*

$$E[f(X, Z)|Z = j] \leq CE[f(X, Z)|Z = k].$$

Proof. Under Lemma 2's assumptions,

$$\begin{aligned}
E[f(X, Z)|Z = j] &= \int_{\mathbb{R}^D} f(x, j) dP_j \\
&\leq \int_{\mathbb{R}^D} f(x, k) dP_j \\
&= \int_{\mathbb{R}^D} f(x, k) \frac{dP_j}{dP_k} dP_k \\
&\leq C \int_{\mathbb{R}^D} f(x, k) dP_k \\
&= CE[f(X, Z)|Z = k].
\end{aligned}$$

The second inequality follows from monotonicity, and the third by the Radon Nikodym theorem since $P_j << P_k$. □

Lemma 3. *Let $f : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$, where $\mathcal{X} \subseteq \mathbb{R}^D$, $\mathcal{Z} \subseteq \mathbb{R}$. Assume that $\mathcal{X}, \mathcal{Z}$ are both finite, with $X \in \mathcal{X}$, $Z \in \mathcal{Z}$. Let $\tilde{f}$ be the projection of f onto the set of functions over $\mathcal{X} \times \mathcal{Z}$ that are monotonic with respect to Z such that for $j \leq k$, $f(x, j) \leq f(x, k)$. For $z_{(i)} \in \mathcal{Z}$, let $z_{(1)} \leq z_{(2)} \leq ... \leq z_{(|\mathcal{Z}|)}$. Define the average statistical parity violation:*

$$R_f \triangleq \sum_{i=1}^{|\mathcal{Z}|} \frac{E[f(X,Z)|Z = z_{(i)}] - E[f(X,Z)|Z = z_{(i+1)}]}{|\mathcal{Z}|}$$

Then $R_{\tilde{f}} \leq R_f$.

Proof. Let $\tilde{f} : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ be the projection of f onto the class of functions monotonic in the second argument, defined as follows:

$$\begin{aligned} \tilde{f} = \arg\min_{f'} & \ ||f - f'|| \\ \text{s.t. } & f'(x, j) \leq f'(x, k) \ \forall j, k \in \mathcal{Z}; j \leq k \end{aligned} \tag{C.1}$$

where

$$||f - f'||^2 = \sum_{x \in \mathcal{X}, z \in \mathcal{Z}} (f(x, z) - f'(x, z))^2.$$

The projection $\tilde{f}$ can be computed in $O(|\mathcal{X}||\mathcal{Z}|)$ time using the pool-adjacent-violators algorithm from isotonic regression [Ayer et al., 1955, JB, 1964], since a one dimensional projection can be done independently in $O(|\mathcal{Z}|)$ time for each $x \in \mathcal{X}$.

R_f is a telescoping sum:

$$R_f = \frac{E[f(X,Z)|Z = z_{(1)}] - E[f(X,Z)|Z = z_{(|\mathcal{Z}|)}]}{|\mathcal{Z}|}$$

For discrete X and Z, we have

$$E[f(X,Z)|Z = j] = \sum_{x \in \mathcal{X}} f(x, j) P(X = x|Z = j)$$

which implies

$$\begin{aligned} R_f = \frac{1}{|\mathcal{Z}|} \sum_{x \in \mathcal{X}} \Big(& f(x, z_{(1)}) P(X = x|Z = z_{(1)}) \\ & - f(x, z_{(|\mathcal{Z}|)}) P(X = x|Z = z_{(|\mathcal{Z}|)}) \Big). \end{aligned}$$

We now show that $\tilde{f}(x, z_{(1)}) \leq f(x, z_{(1)})$, and $\tilde{f}(x, z_{(|\mathcal{Z}|)}) \geq f(x, z_{(|\mathcal{Z}|)})$: Suppose $\tilde{f}(x, z_{(1)}) > f(x, z_{(1)})$. Then we can set $\tilde{f}'(x, z_{(1)}) = f(x, z_{(1)})$ without violating the monotonicity constraints, and $||f - \tilde{f}'|| < ||f - \tilde{f}||$, which contradicts that $\tilde{f}$ solves (C.1). A similar argument can be made for $z_{(|\mathcal{Z}|)}$.

Since $\tilde{f}(x, z_{(1)}) \leq f(x, z_{(1)})$ and $\tilde{f}(x, z_{(|\mathcal{Z}|)}) \geq f(x, z_{(|\mathcal{Z}|)})$, we have

$$\begin{aligned} f(x, z_{(1)})&P(X = x|Z = z_{(1)}) \\ &\quad - f(x, z_{(|\mathcal{Z}|)})P(X = x|Z = z_{(|\mathcal{Z}|)}) \\ \geq \tilde{f}(x, z_{(1)})&P(X = x|Z = z_{(1)}) \\ &\quad - \tilde{f}(x, z_{(|\mathcal{Z}|)})P(X = x|Z = z_{(|\mathcal{Z}|)}) \end{aligned}$$

Since the above inequality is true for all x, it holds for the sum over $x \in \mathcal{X}$, therefore $R_{\tilde{f}} \leq R_f$. □

Lemma 4. *Suppose X is a continuous (or with a straightforward extension, discrete) random variable, and let $\mathcal{S}$ be a nonempty set such that for all $x \in \mathcal{S}$, the joint probability density values $p_{X,\hat{Y}|Z=z}(x, 1) > 0$ for $z = j, k$. Suppose we have monotonicity where $f(x, j) \leq f(x, k)$ for $j \leq k$ for all $x \in \mathcal{S}$. For a binary classifier this implies $P(\hat{Y} = 1|X = x, Z = j) \leq P(\hat{Y} = 1|X = x, Z = k)$. Then we can bound one-sided statistical parity as follows:*

$$\frac{P(\hat{Y} = 1|Z = j)}{P(\hat{Y} = 1|Z = k)} \leq \inf_{x \in \mathcal{S}} \frac{p_{X|Z=j}(x) p_{X|\hat{Y}=1,Z=k}(x)}{p_{X|Z=k}(x) p_{X|\hat{Y}=1,Z=j}(x)}$$

Proof. Fix $x \in \mathcal{S}$. By Bayes' theorem and monotonicity,

$$\begin{aligned} &P(\hat{Y} = 1|Z = j) \\ &\quad = P(\hat{Y} = 1|X = x, Z = j) \frac{p_{X|Z=j}(x)}{p_{X|\hat{Y}=1,Z=j}(x)} \\ &\quad \leq P(\hat{Y} = 1|X = x, Z = k) \frac{p_{X|Z=j}(x)}{p_{X|\hat{Y}=1,Z=j}(x)} \\ &\quad = P(\hat{Y} = 1|Z = k) \frac{p_{X|\hat{Y}=1,Z=k}(x)}{p_{X|Z=k}(x)} \frac{p_{X|Z=j}(x)}{p_{X|\hat{Y}=1,Z=j}(x)} \end{aligned}$$

Since the inequality holds for all $x \in \mathcal{S}$, the tightest bound holds for the infimum. □

Lemma 5. *Let $Y \in \{0, 1\}$ be a random variable representing the target. Let $\mathcal{S}$ be a nonempty set such that for all $x \in \mathcal{S}$, the following joint probability density values are non-zero for $z = j, k$: $p_{X,Y,\hat{Y}|Z=z}(x, 1, 1) > 0$ and $p_{X,Y|\hat{Y}=1,Z=z}(x, 1) > 0$. Then,*

$$\frac{P(\hat{Y} = 1|Y = 1, Z = j)}{P(\hat{Y} = 1|Y = 1, Z = k)} \leq \inf_{x \in \mathcal{S}} \frac{c_j(x)}{c_k(x)}$$

$$\textit{where } c_z(x) = \frac{p_{X|Z=z}(x)P(Y=1|\hat{Y}=1,Z=z)}{p_{X|\hat{Y}=1,Z=z}(x)P(Y=1|Z=z)}$$

Proof. Let $\mathcal{S}$ be a nonempty set such that for all $x \in \mathcal{S}$, the following joint probability density values are non-zero for $z = j, k$:

$p_{X,Y,\hat{Y}|Z=z}(x,1,1) > 0$ and $p_{X,Y|\hat{Y}=1,Z=z}(x,1) > 0$

Fix $x \in \mathcal{S}$.

Suppose we have a monotonic binary classifier, where $P(\hat{Y}=1|X=x,Z=j) \leq P(\hat{Y}=1|X=x,Z=k)$ for $j \leq k$.

By Bayes' theorem, we have

$P(Y=1|Z=j)P(\hat{Y}=1|Y=1,Z=j)p_{X|Y=1,\hat{Y}=1,Z=j}(x)$
$= p_{X|Z=j}(x)P(\hat{Y}=1|X=x,Z=j)P(Y=1|X=x,\hat{Y}=1,Z=j)$

and $p_{X|\hat{Y}=1,Z=j}(x)P(Y=1|X=x,\hat{Y}=1,Z=j)$
$= p_{X|Y=1,\hat{Y}=1,Z=j}(x)P(Y=1|\hat{Y}=1,Z=j)$

Let $c_z(x) = \frac{p_{X|Z=z}(x)P(Y=1|\hat{Y}=1,Z=z)}{p_{X|Y=1,Z=z}(x)P(Y=1|Z=z)}$. This is well defined for $x \in \mathcal{S}$.

Combining both applications of Bayes' theorem and the monotonicity assumption:

$$\begin{aligned}
P(\hat{Y}=1 \quad & |Y=1,Z=j) \\
= \quad & \frac{p_{X|Z=j}(x)P(Y=1|X=x,\hat{Y}=1,Z=j)}{P(Y=1|Z=j)p_{X|Y=1,\hat{Y}=1,Z=j}(x)} \\
& * P(\hat{Y}=1|X=x,Z=j) \\
= \quad & \frac{p_{X|Z=j}(x)P(Y=1|\hat{Y}=1,Z=j)}{P(Y=1|Z=j)p_{X|\hat{Y}=1,Z=j}(x)} \\
& * P(\hat{Y}=1|X=x,Z=j) \\
= \quad & c_j(x)P(\hat{Y}=1|X=x,Z=j) \\
\leq \quad & c_j(x)P(\hat{Y}=1|X=x,Z=k) \\
= \quad & \frac{c_j(x)}{c_k(x)}P(\hat{Y}=1|Y=1,Z=k)
\end{aligned}$$

Since this holds for all $x \in \mathcal{S}$, it holds for the infimum. □

C.2 Counterexamples

To supplement Section 4.6, we give various counterexamples showing that certain relations between *statistical parity* and monotonicity do not hold.

Monotonicity does not imply statistical parity.

We show that monotonic function f may violate *one-sided statistical parity* by an example that illustrates Simpson's paradox. Suppose $X \in \{0, 1\}$, where $X = 1$ means a law student passed the bar and $X = 0$ means the student did not. Let $Z \in \{0, 1, 2, 3\}$ be the poverty level of the student, where $Z = 3$ represents the highest poverty level. Suppose $f(X, Z)$, or the admissions score, is monotonic in Z and takes the values shown in Fig. C.1. Suppose that the distributions $P(X = x|Z = z)$ are given by figure C.2. Then the maximum *one-sided statistical parity* violation is

$$\begin{aligned}
&E[f(X,Z)|Z=1] - E[f(X,Z)|Z=2]\\
&= f(0,1)P(X=0|Z=1) - f(0,2)P(X=0|Z=2)\\
&\quad - f(1,1)P(X=0|Z=1) + f(1,2)P(X=0|Z=2)\\
&= 1.5(0.9) - 1.5(0.1)\\
&= 1.2.
\end{aligned}$$

Thus, there is a positive one-sided satistical parity violation even though $f(X, Z)$ is monotonic in Z. This violation comes from the fact that even though $f(0, 1) \leq f(0, 2)$, this is outweighed by the fact that $P(X = 0|Z = 1) \geq P(X = 1|Z = 2)$. This illustrates that for a monotonic function, the *statistical parity* violation depends on the conditional probabilities $P(X = x|Z = z)$, and indeed Lemma 2 bounds the one-sided statistcal parity violation by a ratio of conditional probabilities.

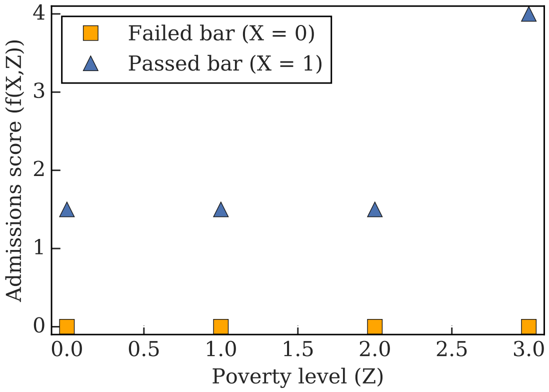


Figure C.1: Monotonic admissions scores for Counterexamples C.2 and C.2.

Statistical Parity does not imply a bound on monotonicity violations.

We show that the converse of Lemma 2 does not hold: a model that satisfies *statistical parity* may have arbitrarily high monotonicity violations regardless of the likeihood ratio C. Suppose

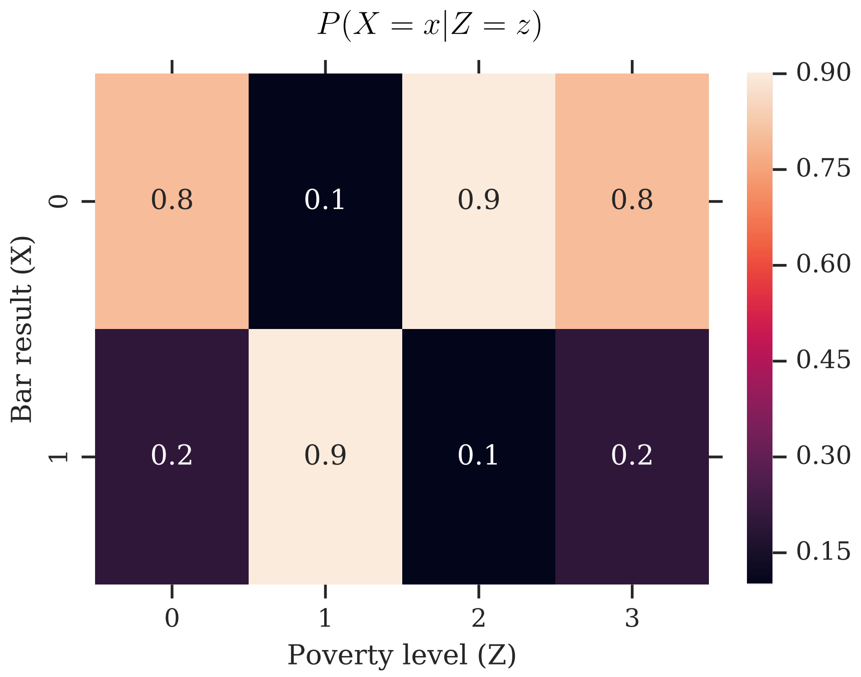


Figure C.2: Distribution of X, Z for Counterexamples C.2 and C.2. The displayed values are $P(X = x|Z = z)$ for $X \in \{0, 1\}$ and $Z \in \{0, 1, 2, 3\}$.

the distribution of men and women for a given height x is equal for all heights, such that $C = 1$. Suppose that *statistical parity* is satisfied such that men and women were equally likely to be selected for a sports team on average.*Statistical parity* could hold if the model accepted all men over some height h that splits the population in half (say $h = 5'8''$), and accepted all women under height h. But then for a height less than h, $P(\hat{Y} = 1|Z = \text{female}) = 0$ while $P(\hat{Y} = 1|Z = \text{male}) = 1$, and for height over h, $P(\hat{Y} = 1|Z = \text{male}) = 0$ while $P(\hat{Y} = 1|G = \text{female}) = 1$. Therefore, neither a positive nor a negative monotonicity constraint holds: there is no constant $C' > 0$ such that $P(\hat{Y} = 1|X = x, Z = \text{male}) \leq C'P(\hat{Y} = 1|X = x, Z = \text{female})$ or $P(\hat{Y} = 1|X = x, Z = \text{female}) \geq C'P(\hat{Y} = 1|X = x, Z = \text{male})$ for all x.

Monotonic projection can be more unfair in the worst case.

While Lemma 3 shows that projecting a function onto monotonicity constraints cannot increase the *average one-sided statistical parity* violation, it can increase violations *in the worst case*. Consider a continuation of the example from C.2, but this time let $f(X, Z)$ be defined by Fig. C.3, and let $\tilde{f}(X, Z)$ be defined by Fig. C.1. In this case, Fig. C.1 is the monotonic projection of Fig. C.3. Then the worst case *statistical parity* violation for the monotonic projection $\tilde{f}$ is *higher* than the worst case *statistical parity* violation for the non-monotonic f:

$$\begin{aligned}&E[\tilde{f}(X,Z)|Z=1]-E[\tilde{f}(X,Z)|Z=2]\\&=1.5(0.9)-1.5(0.9)\\&=1.2\end{aligned}$$

$$\begin{aligned}&E[f(X,Z)|Z=1]-E[f(X,Z)|Z=2]\\&=1.0(0.9)-0.5(0.9)\\&=0.85\end{aligned}$$

For a given pair j, k, as long as $\tilde{f}(x,j) \leq f(x,j)$ and $\tilde{f}(x,k) \geq f(x,k)$, then the violation

$$R_f(j,k) = E[f(X,Z)|Z=j] - E[f(X,Z)|Z=k]$$

will not be worse for the monotonic projection $\tilde{f}$: $R_{\tilde{f}}(j,k) \leq R_f(j,k)$. Lemma 3 holds because the inequalities $\tilde{f}(x,j) \leq f(x,j)$ and $\tilde{f}(x,k) \geq f(x,k)$ hold for $j = z_{(1)}$ and $k = z_{(|\mathcal{Z}|)}$, but this counterexample exists because those inequalities do not necessarily hold for any other pairs j, k in between.

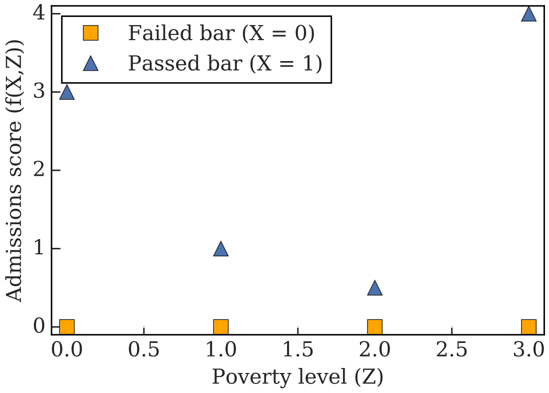


Figure C.3: Nonmonotonic admissions scores for Counterexample C.2.

C.3 Tradeoff between likelihood ratios in Lemma 4

The bound in Lemma 4 contains two likelihood ratios: $\frac{p_{X|Z=j}(x)}{p_{X|Z=k}(x)}$ and $\frac{p_{X|\hat{Y}=1,Z=k}(x)}{p_{X|\hat{Y}=1,Z=j}(x)}$. When the first likelihood ratio is low, the second inverse likelihood ratio may be high. For example, suppose Z is an individual's poverty level (j being low poverty and k being high poverty), X is the number of extracurricular activities the individual is involved in, and $\hat{Y}=1$ means the individual is accepted into university. Suppose all individuals with above a certain number of extracurricular activities is accepted. Then the first likelihood ratio could be low when the number of extracurricular activities X is low. Similarly, the likelihood that a high poverty individual accepted into university has a low number of extra curricular activities is probably also higher than the likelihood that a low poverty individual accepted into university has a low number of extracurricular activities. This implies that the second inverse likelihood ratio would be high, thus trading off with the first likelihood ratio.

C.4 Further Analysis of Law School Admissions Experiments

Figure C.4 shows the distribution of the LSAT scores, undergraduate GPA, and bar exam outcomes. Examples where the bar exam outcome was missing were omitted in our experiments.

C.5 Further Analysis of Funding Proposals Experiments

Figure C.5 gives a histogram of the four different poverty levels, which are ordinal with level 3 being the most impoverished.

Figure C.6 *(top)* shows the training examples' average number of exciting projects, where the error bars show the standard error of the mean. The poverty level feature ranges from 0 to 3, with 0 denoting low poverty and 3 denoting the highest poverty level. For ease of visualization, we show the quartiles of the students-reached feature.

Figure C.6 *(middle)* shows the predicted probability that a project is exciting for a GAM model without the proposed ethical constraints. The model gives lower scores to poverty level 2 (poorer schools) than to poverty level 1 (richer schools) for every quartile of students reached. The model also gives higher scores for project that reach 30-100 students tahn to projects that reach 100+ students.

Figure C.6 *(bottom)* shows that training with an ethical monotonicity shape constraint works: at the same poverty level, projects that affect more students are given a higher score. For the same quartile of students reached, the score also does not decrease for higher poverty levels.

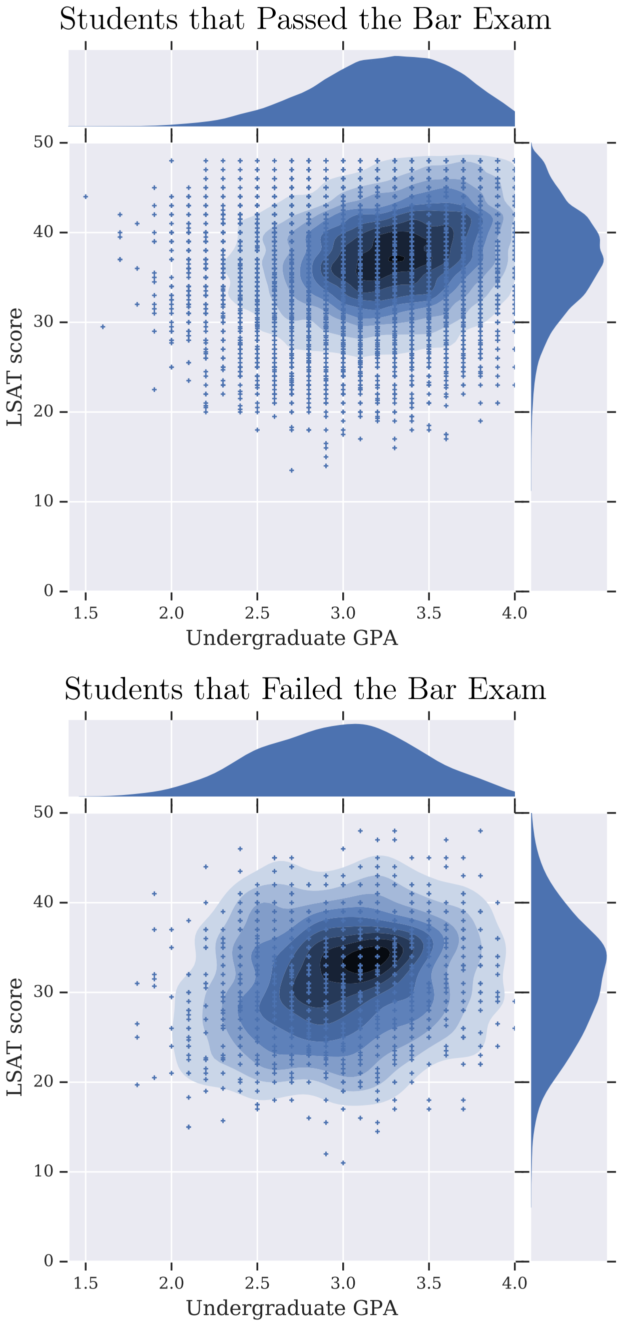


Figure C.4: Distribution over the full Law School Admissions dataset of undergraduate GPA and LSAT score students for students that passed the bar exam *(top)* and students that failed the bar exam *(bottom)*. The dataset consists of 94.86% students that passed the bar exam.

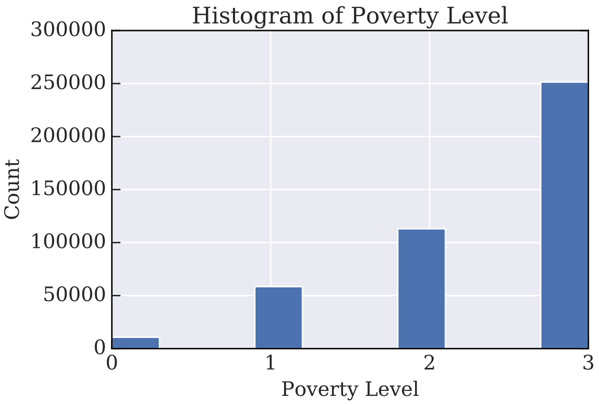
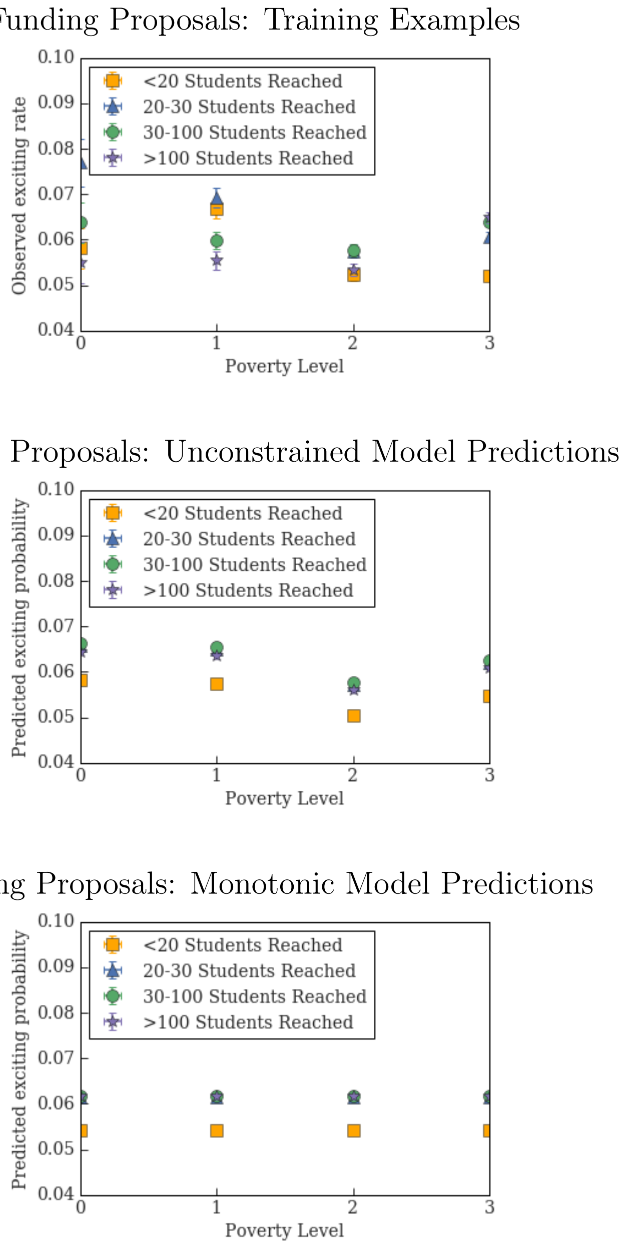


Figure C.5: Histogram of the poverty level feature from the Funding Proposals dataset. 0 represents lowest poverty and 3 represents highest poverty.



Funding Proposals: Training Examples

Funding Proposals: Unconstrained Model Predictions

Funding Proposals: Monotonic Model Predictions

Figure C.6: *(top)* Plot of the observed rate of exciting projects (mean number of exciting projects) as a function of each project's poverty level and number of students reached. Error bars show the standard deviation. *(middle)* Unconstrained model predictions. *(bottom)* Shape-constrained model predictions.

Part II

Metrics in Multi-Stakeholder Systems

Chapter 5

On Counterfactual Metrics for Social Welfare: Incentives, Ranking, and Information Asymmetry

5.1 Introduction

As machine learning (ML) is increasingly deployed in dynamic social systems with asymmetries in power and information between stakeholders, a core challenge is that the *metrics* that are optimized do not always align with social welfare. From education to healthcare to recommender systems, it has been repeatedly shown that the impact of the ability of modern ML to optimize arbitrarily complex objectives is often limited by the difficulty in choosing *what* to optimize [Liu et al., 2023, McNee et al., 2006, Obermeyer et al., 2019].

One particularly consequential example of this gap is the incentive misalignment that occurs when *average treated outcome* measures are used as accountability and ranking metrics. This has led to measurable societal harm when those being ranked may selectively choose whom to treat. As an illustrative example, Dranove et al. [2003] showed that the publication of hospital mortality rate metrics by the New York Health Department led to dramatically worse outcomes for severely ill patients. Specifically, hospitals had an incentive to selectively treat the healthiest patients, rather than those who would benefit most from treatment. Today, the Centers for Medicare and Medicaid Services (CMS) continues to invest billions of dollars in the development of quality metrics [Wadhera et al., 2020, Casalino et al., 2016]. In addition to determining direct provider compensation [Institute, 2022], these metrics also feed into large scale *ranking systems* such as the US News and World Report and the LeapFrog Hospital Safety Score [Rosenberg, 2013, Health and Agency, 2022]. At the same time, studies have continued to question the relationship between these metrics and patient outcomes [see, e.g., Glance et al., 2021, Gonzalez and Ghaferi, 2014, Ryan et al., 2009, Hwang et al., 2014, Jha et al., 2008, Smith et al., 2017].

In this work, we directly study the incentive misalignment when *average treated outcomes* are used as quality metrics. To mitigate this misalignment, we propose alternative metrics that have a foundation in causal inference. Specifically, given counterfactual estimates of patient outcomes, we outline effective usage of these estimates and discuss prevailing limitations when providers may engage in strategic behavior.

More generally, our analysis applies to any environment where metrics are learned from data over which an agent controls treatment selection. We refer to healthcare as a running example where these effects are well documented. However, other domains subject to this phenomenon include education, where student outcome metrics affect school rankings, funding, and accreditation [Koretz, 2017]; and online ranking platforms for commercial businesses like restaurants that may exercise some screening over their customers.

To study the welfare effects that arise from this dynamic interaction between quality metrics and hospitals, we employ a *principal-agent model* where the *principal* chooses a quality metric as a reward function, and the *agent* responds by optimizing this reward function to the best of their private abilities and information. To analyze the welfare effects of metric choices, we apply a causal framework similar to that of policy learning, where the goal is to allocate treatments that maximize the total positive effects over a population relative to treating no one [Manski, 2008]. Our key ingredient is that the principal only has indirect control of the implemented policy—they may design a metric that shapes an agent's reward and hence behavior but cannot directly choose an agent's policy.

Contributions. We show that *average treated outcome* metrics incur unbounded regret by this definition of welfare, and show that regret can be reduced by *(i)* accounting for counterfactual untreated outcomes and *(ii)* considering total welfare instead of average welfare among treated patients. Applying these two simple insights yields an optimal functional form for a quality measure that achieves zero regret as long as the principal can learn the mean conditional untreated potential outcomes. We refer to this as rewarding the *total treatment effect.* Connecting the proposed counterfactual metric to practice, we discuss two issues that arise when operationalizing the metric in real applications. First, we study the complications that arise when the total treatment effect is used to *rank* different agents that might serve different treatment populations. Second, we consider practical issues of *information asymmetry*, where the agent might observe more features about each patient than the principal. *Even an unbiased estimate of the counterfactual untreated outcome is not sufficient to maximize patient welfare when information asymmetries remain.* In addition to giving theoretical regret bounds, we also empirically show that it is not always better for a principal to condition on all known features, as this can *amplify regret.* Our model yields new connections between information asymmetry in the principal-agent model and unobserved heterogeneity in causal inference.

5.2 Related work

Our work combines technical structure from policy learning and contract theory to analyze misalignments in accountability metrics, which have long been critiqued in the social sciences.

Policy learning. The evaluation of treatment policies by their causal effects is well established in policy evaluation and policy learning [Manski, 2008, Hirano and Porter, 2009, Stoye, 2009, Athey and Wager, 2021]. We directly apply the same measures of utility and regret in this work. Building on these measures, we consider a setting where a principal (regulatory agency) can only indirectly affect the policy through measuring an agent (hospital). We also note that our formulation is not the only way in which strategic behavior may arise. For example, Sahoo and Wager [2022] model patients' strategic responses when the hospital learns a policy with limited treatment capacity. Combining this model with ours would produce an interesting pipeline analysis.

Contract theory. The principal-agent model is well established in economics as a way to model incentives and equilibrium dynamics when a *principal* sets a contract with rewards as a function of actions, and an *agent* decides which action to take based on private information and costs [Laffont and Martimort, 2009, Gibbons et al., 2013, Milgrom and Roberts, 1992]. Contract theory provides a structure to analyze moral hazard, where the structure of the contract and information asymmetry may lead to misalignment between the principal and agents' incentives [Arrow, 1963, Holmstrom and Milgrom, 1991]. More recently, there is a growing recognition of the importance of the algorithmic and statistical aspects of contract theory [e.g., Carroll, 2015, Tetenov, 2016, Spiess, 2018, Dütting et al., 2019, Dütting et al., 2020, Bates et al., 2022, Alon et al., 2022]. Our work contributes to this line of thought, linking moral hazard from contract theory with causal inference, and showing how incentive and statistical considerations jointly guide the choice of accountability metrics.

Specifically, we use this framework to analyze misalignments under a particular form of action and information asymmetry where the agent controls treatment selection, and might know more about each treatment unit than the principal. Lazzarini et al. [2022] applied contract theory to analyze *why* regulatory agencies might lean towards outcome-based contracts (such as the average treated outcome) rather than counterfactual assessment. Our model differs in both goal and setup: while they model the agent's effort levels, our model considers welfare effects when the agent has general control over all treatment assignments.

Strategic classification. The framework of *performative prediction* describes the distribution shifts and distortion in validity that result from agent strategy in response to ML-driven decision making [Perdomo et al., 2020, Hardt et al., 2016a]. Related lines of work in strategic classification have considered whether classifiers incentivize agent improvement [Ahmadi et al., 2022, Kleinberg and Raghavan, 2020, Bechavod et al., 2020, Haghtalab et al., 2020], drawing connections to causality [Miller et al., 2020, Shavit et al., 2020]. We consider a specific strategic structure to improve metrics when agents manipulate treatment policies.

Accountability, auditing, and measurement. Historically, the mismatch between accountability metrics and welfare has been well documented across domains, from monetary policy to education [Goodhart, 1984, Campbell, 1979, Muller, 2019, Koretz, 2017, Mau, 2019].

Extensive work in the social sciences has critically examined accountability and auditing practices [Power, 1994, Strathern, 1997, Hoskin, 1996, Rothstein, 2008]. Our work builds on these qualitative insights by modeling a particular misalignment that occurs when the measured party has selection power over treatment allocations, also known as "creaming" [Lacireno-Paquet et al., 2002]. Measurement theory has also given both qualitative and statistical tools for understanding the validity of measurements [Bandalos, 2018], with recent extensions to fair ML [Jacobs and Wallach, 2021]. Guerdan et al. [2023] apply counterfactual modeling to estimate the measurement error when proxy outcomes are used to guide treatment decisions. Complementing this focus on measurement validity, we model the treatment incentives induced when measurements are used to reward agents, and study the resulting welfare effects.

5.3 Principal-Agent Model

To analyze the incentives and welfare effects that arise from quality metrics, we define a principal-agent model where we will refer to the organization that collects the metrics and pays or ranks the providers as the *principal*, and the healthcare providers as *agents*. First, the principal specifies a function for rewarding the agents based on their actions and the observed outcomes. In response, the agent allocates treatment decisions with knowledge of this reward function. Our focus is on designing reward functions with high utility for the principal, which corresponds to social welfare.

Formal model

We suppose that a single agent has access to independently and identically distributed samples of characteristics $X_i \in \mathcal{X}$ for $i \in \{1, ..., n\}$ treatment units (referred to generally as the *treatment population*). The agent assigns a binary treatment according to treatment rule $\pi : \mathcal{X} \to [0, 1]$, and we use $T_i^\pi \in \{0, 1\}$ to denote a Bernoulli random variable indicating the realizations of the treatment rule: $\pi(X_i) = P(T_i^\pi = 1|X_i)$. This framework allows for T^π to be either stochastic or deterministic.

To model patient outcomes from treatment, we apply the potential outcomes framework [Neyman, 1923, Rubin, 1974]. Let $Y_i(t)$ be the potential outcome if the patient had received treatment $t \in \{0, 1\}$. Let $Y_i \in \mathbb{R}$ denote the observed outcome under treatment assignments $Y_i = Y_i(T_i^\pi)$ under the Stable Unit Treatment Value Assumption (SUTVA) [Rubin, 1980], which implies the consistency and non-interference assumptions. Let $\tau(X) = E[Y_i(1) - Y_i(0)|X]$ denote the conditional average treatment effect given covariates X, and let $\mu_t(X) = E[Y_i(t)|X]$ denote the conditional mean of the potential outcome under treatment t.

We suppose that the principal observes $\{X_i, T_i^\pi\}_{i=1}^n$ (also denoted $\mathbf{X}, \mathbf{T}^\pi$) for all units, and Y_i for all units for which $T_i^\pi = 1$ (denoted $\mathbf{Y}$). The principal must then choose a reward function $w : \mathcal{X}^n \times \{0, 1\}^n \times \mathbb{R}^n \to \mathbb{R}$ with which to reward the agent.

Turning to our behavioral assumption, we consider an agent that is risk neutral and maximizes their expected reward. That is, the agent chooses the treatment rule π from their set of possible treatment rules Π that maximizes $E[w(\mathbf{X}, \mathbf{T}^{\pi}, \mathbf{Y})]$. Let

$$\pi^{w} \in \arg\max_{\pi \in \Pi} E[w(\mathbf{X}, \mathbf{T}^{\pi}, \mathbf{Y})]$$

denote this best response. In maximizing this expected reward, we assume that the agent knows $\mu_t(x)$ for all $x \in \mathcal{X}$ and $t \in \{0, 1\}$. Note that there is no explicit model of the agent's cost here to keep the focus on incentives induced by accountability metrics alone, and all budget constraints are contained in the agent's feasible set of treatment rules Π. We further discuss these assumptions and possible extensions in the Appendix.

Total welfare and regret

Following existing work in policy evaluation [Manski, 2008], for a given treatment rule π, we define the total effect of treatment on welfare as $V(\pi) = E[Y_i(T_i^{\pi}) - Y_i(0)]$. $V(\pi)$ is the utility of treatment rule π relative to the alternative outcomes under no treatment [Manski, 2008, Athey and Wager, 2021]. Thus, maximizing $V(\pi)$ also maximizes total welfare $E[Y]$ compared to treating no one. As also done in policy learning [Athey and Wager, 2021], to evaluate different quality measures chosen by the principal, we define the regret for a given policy π compared to the best feasible policy in Π to be $R(\pi) = \max_{\tilde{\pi} \in \Pi} V(\tilde{\pi}) - V(\pi)$. We compare different choices of quality measures w by analyzing the effect on total welfare for the induced treatment rule π. In other words, the principal's goal is to choose a reward function w that leads the agent to best respond with a treatment rule with minimal regret, $R(\pi^{w})$.

5.4 Comparisons of Quality Metrics

Using a principal-agent model, we formally compare different choices of quality metrics w by analyzing the regret for the induced treatment rule π given by the agent's best response. Starting with the *status quo* of rewarding the average treated outcome, we show that this incurs unbounded regret. We reduce this to two main problems with the metric: *(i)* lack of accounting for untreated outcomes, and *(ii)* rewarding an average effect instead of a total effect. Addressing each of these in turn, we show that the regret for rewarding the *average treatment effect on the treated (ATT)* is bounded but can still be high, and that rewarding the *total treatment effect* finally achieves zero regret.

Status quo: average treated outcome

We begin by analyzing regret under the current common quality measure that rewards the average treated outcome, as done by the mortality measures in the New York and Philidelphia health departments in the 1990s analyzed by [Dranove et al., 2003], and in many CMS quality measures [Institute, 2022]. Lazzarini et al. [2022] also refers to these as "outcome contracts." The reward function for the *average treated outcome (ATO)* takes the following form:

Reward Function 1 (ATO).

$$w_{\text{ATO}}(\mathbf{X}, \mathbf{T}^{\pi}, \mathbf{Y}) = \begin{cases} \frac{\sum_{i=1}^{n} Y_i T_i^{\pi}}{\sum_{i=1}^{n} T_i^{\pi}} & \sum_{i=1}^{n} T_i^{\pi} > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The agent's unconstrained best response is

$$\pi^{w_{\text{ATO}}}(x) = \mathbb{1}(x \in \arg\max_{x} \mu_1(x) \text{ and } \mu_1(x) > 0).$$

Proposition 2 (ATO Regret)**.** If the conditional mean untreated potential outcomes $\mu_0(x)$ are unbounded, then the regret for the reward function w_{ATO} may be arbitrarily large.

Intuitively, there are at least two failure modes that can lead to this unbounded regret. First, this reward leads the agent to ignore higher treatment effects of the patients with a lower treated outcome, such as sicker patients with higher mortality probability but more benefit from surgery. This matches the findings from Dranove et al. [2003]. Second, w_{ATO} rewards agents that treat the patients with a higher treated outcome, even though treatment actually harms those patients, such as healthier patients who might incur more risks or side effects from treatment.

More broadly, there are two main problems with the construction of the ATO metric that lead to this unbounded regret. First, the lack of accounting for counterfactual outcomes leads to the two failure modes above. Second, the measure of an average outcome instead of a total outcome means that the agent will only treat the *single* patient with the covariate value x that *maximizes* μ_1. We next analyze several reasonable modifications to reward functions that address each of these problems.

Accounting for counterfactuals

When the principal operates with full information of the agent's selection covariates X_i, then regret can be reduced if they also have access to an unbiased estimator of the mean conditional untreated potential outcome.

Assumption 1. The principal accesses an estimator $\hat{\mu}_0(x)$ which is unbiased: $E[\hat{\mu}_0(x)] = \mu_0(x) \ \forall x \in \mathcal{X}$.

In general, obtaining an unbiased estimator $\hat{\mu}_0(x)$ can be difficult, but circumstances under which causal inference can be reliably conducted are well understood [Hernan and Robins, 2020]. As a concrete example, suppose the principal has access to an auxiliary data source $\{X'_j, T'_j, Y'_j\}_{j=1}^{m}$, with outcomes for untreated patients with $T'_j = 0$, collected from clinical trials or observational data. In addition to standard assumptions for identification of

$\mu'_0(x) = E[Y'_j(0)|X'_j = x]$, the principal may use this data if the distribution of the conditional untreated potential outcome is also the same, $\mu'_0(x) = \mu_0(x)$, and the support of X'_j covers the support of X_i. The difficulty of obtaining such a dataset is lessened by the fact that the principal does not require identification of the *treated* potential outcome $\mu'_1(x)$ or the conditional average treatment effect $\tau'(x)$, so the treatment need not be the same. Access to additional scientific knowledge (in the form of, e.g., a more intricate structural model or functional form assumptions) can also aid in the estimation of $\mu_0(x)$. Medical research continues to develop patient risk scores using combinations of such methods [Sullivan et al., 2004, Jones and Cossart, 1999] and evaluate their validity [Kaafarani et al., 2011, Janssens, 2019].

Under the perhaps optimistic assumption of access to an unbiased estimator $\hat{\mu}_0(x)$, this work focuses on effective ways for the principal to apply this estimator. It turns out that even given this unbiased estimate, principal-agent incentive misalignment can still occur, and there are still pitfalls with its downstream usage. We begin by showing how to effectively incorporate this estimator into the reward function. In later sections, we discuss effective usage in a ranking context and regret bounds under information asymmetry.

Given $\hat{\mu}_0(x)$, the principal can modify w_{ATO} by directly subtracting an estimate of the untreated potential outcomes, and thus reward the *average treatment effect on the treated (ATT)*:

Reward Function 2 (ATT).

$$w_{\text{ATT}}(\mathbf{X}, \mathbf{T}^\pi, \mathbf{Y}) = \begin{cases} \frac{\sum_{i=1}^n (Y_i - \hat{\mu}_0(X_i))T_i^\pi}{\sum_{i=1}^n T_i^\pi} & \sum_{i=1}^n T_i^\pi > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The agent's unconstrained best response is

$$\pi^{w_{\text{ATT}}}(x) = \mathbb{1}(x \in \arg\max_x \tau(x) \text{ and } \tau(x) > 0).$$

The resulting regret is bounded, but still not zero.

Proposition 3 (ATT Regret)**.** If $\hat{\mu}_0(x)$ is unbiased and π is unconstrained, then the regret for the reward function w_{ATT} is upper bounded as
$R(\pi^{w_{\text{ATT}}}) \leq \max_{\pi \in \Pi} V(\pi)$.

Now that the reward function accounts for the counterfactual untreated outcome, Proposition 3 shows that the regret cannot exceed the maximum utility. This is notably not true of the w_{ATO}, where the regret can be arbitrarily high due to the agent sometimes treating those with a negative treatment effect. Still, while accounting for untreated potential outcomes avoids treating those with a negative treatment effect, the ATT as a reward function still suffers from misalignment with total welfare due to the fact that it rewards the *average* effect rather than the *total* effect. This means that in the best response, the agent only treats patients with the single value x with maximum treatment effect $\tau(x)$.

Rewarding total effects

To expand the agent's treatments to cover *all* individuals who would benefit, we modify the above reward function by simply removing the denominator, thus rewarding a *total* effect instead of an *average* effect. This yields a reward function for the *total treatment effect (TT)*.

Reward Function 3 (TT).

$$w_{\text{TT}}(\mathbf{X}, \mathbf{T}^{\pi}, \mathbf{Y}) = \sum_{i=1}^{n}(Y_i - \hat{\mu}_0(X_i))T_i^{\pi}$$

The agent's unconstrained best response is

$$\pi^{w_{\text{TT}}}(x) = \mathbb{1}(\tau(x) > 0),$$

which yields zero regret.

Proposition 4 (TT Regret)**.** If $\hat{\mu}_0(x)$ is unbiased, then the regret is $R(\pi^{w_{\text{TT}}}) = 0$.

The regret is zero regardless of the feasible set Π. Thus, with two modifications to the status quo w_{ATO}, a quality measure w_{TT} can be constructed that is aligned with total welfare, as long as the principal has access to an unbiased estimator $\hat{\mu}_0(x)$.

5.5 Ranking With Multiple Agents

In Section 5.4 we've shown that rewarding the total treatment effect leads the agent to maximize total welfare. While this theory applies cleanly in isolation, in real systems, quality measures are often further employed to *rank* hospitals [Rosenberg, 2013, Health and Agency, 2022, Smith et al., 2017]. It turns out that even w_{TT} can exhibit problems as a ranking measure when different hospitals have different treatment population sizes and distributions. To apply reward functions as ranking measures, we show that the total treatment effect can be modified to a more general form that allows for reweighting by covariates X while still preserving incentive alignment. By reweighting the reward function relative to a reference covariate distribution, we show that the resulting quality measure leads to better hospitals receiving better rankings.

Notationally, discussion of ranking requires extending our setting to account for multiple agents. Suppose each agent $k \in \{1, ..., K\}$ observes its own sample of n_k patients with covariates drawn i.i.d. from distribution P_{X^k} with the same support $\mathcal{X}$. Each agent has different treatment effects, denoted by $\mu_t^k(x)$ and $\tau^k(x)$. For rankings to be meaningful, we assume that the untreated potential outcome is the same for all k: $\mu_0^k(x) = \mu_0(x)$ for all k. In short, one provider not treating a patient is equivalent to another provider not treating the same patient. Extending the principal's action space to the multi-agent ranking setting, the principal publishes score functions $\{w_k\}_{k=1}^K$, and each agent k best responds individually with their own treatment policy π_k. The agents are then ranked from highest to lowest score function values. We expand this notation in the Appendix.

Defining desirable ranking properties

For any regulatory agency or potential patient that would utilize these rankings, a clear desirable property would be that *better hospitals should be ranked higher*. From the agents' perspectives, this property may also make the scores feel more "fair." We formally define this property with two different degrees of strictness for the meaning of "better."

First, we define "better" as an agent having *uniformly* higher treatment effects for all possible covariate values x, such that any patient would be better off being treated by this agent.

Definition 3 (Uniform Rank Preservation)**.** A set of score functions $\{w_k\}_{k=1}^K$ preserves treatment effect ordering uniformly over $\mathcal{X}$ if for all $j, k \in \{1, ..., K\}$,

$$\tau^j(x) \geq \tau^k(x) \ \forall x \in \mathcal{X} \implies \max_{\pi_j \in \Pi_j} E[w_j(\mathbf{X}^k, \mathbf{T}^{\pi_j}, \mathbf{Y}^j)] \geq \max_{\pi_k \in \Pi_k} E[w_k(\mathbf{X}^k, \mathbf{T}^{\pi_k}, \mathbf{Y}^k)].$$

A more relaxed version of the uniform rank preservation requirement is one where an agent is "better" if it has higher treatment effects on *average* over a reference covariate population P_{X_0}.

Definition 4 (Relative Rank Preservation)**.** A set of score functions $\{w_k\}_{k=1}^K$ preserves treatment effect ordering relative to a reference population P_{X_0} with support $\mathcal{X}$ if for all $j, k \in \{1, ..., K\}$,

$$E[\tau^j(X_0)] \geq E[\tau^k(X_0)] \implies \max_{\pi_j \in \Pi_j} E[w_j(\mathbf{X}^k, \mathbf{T}^{\pi_j}, \mathbf{Y}^j)] \geq \max_{\pi_k \in \Pi_k} E[w_k(\mathbf{X}^k, \mathbf{T}^{\pi_k}, \mathbf{Y}^k)].$$

This relative definition requires explicitly defining a reference population, which calls for careful consideration of policy goals and societal needs. Any set of scores will implicitly prioritize some populations, and calling attention to this as an explicit part of the ranking properties induced by quality measures can help policymakers more intentionally align their choices with policy goals.

Satisfying desirable ranking properties

We now formally show that w_{TT} as written does not directly satisfy these ranking properties.

Proposition 5. If w_k is directly given by w_{TT} for each agent k, then both ranking properties in Definitions 3 and 4 will be violated.

Intuitively, this breaks down because w_{TT} is subject to two auxiliary effects on top of the treatment effects. First, agents with a larger treatment population n_k (e.g., larger hospitals) will have higher rankings even with the same conditional average treatment effects. Second, agents with different distributions of covariates P_{X^k} but the same conditional average treatment effects will also end up with different rankings if some covariate values are "easier" to treat than others.

To mitigate these auxiliary effects, we show that there exists a general modular form of w_k that preserves the zero regret property for individual agent best responses. This general form can then be tailored to satisfy desirable ranking properties. In particular, consider the following weighted total treatment effect reward function:

Reward Function 4 (Weighted TT).

$$w^g_{\text{TT}}(\mathbf{X}, \mathbf{T}^\pi, \mathbf{Y}) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_0(X_i)) T^\pi_i g(X_i).$$

Any reward function in this family induces the desired agent best response.

Theorem 10 (Incentive Alignment). *Suppose $\hat{\mu}_0(x)$ is unbiased, and π is unconstrained. For any function $g : \mathcal{X} \to \mathbb{R}^+$, w^g_{TT} yields an agent best response with regret $R(\pi^{w^g_{TT}}) = 0$.*

Theorem 10 shows that reweighting the reward function by any function of the covariates X does not hurt incentive alignment. Thus, the principal may choose functions g_k for each agent to achieve desirable ranking properties. Specifically, setting g_k to reweight each agent's covariate distribution to the reference distribution P_{X_0} satisfies both ranking properties in Definitions 3 and 4.

Theorem 11 (Ranking Desiderata Satisfied). *Let P_{X^k} be absolutely continuous with respect to P_{X_0}, and let $g_k = \frac{1}{n_k} \frac{dP_{X_0}}{dP_{X^k}}$ be the normalized Radon–Nikodym derivative of the reference distribution P_{X_0} with respect to agent k's covariate distribution P_{X^k}. Then setting w_k to be $w^{g_k}_{TT}$ for agent k's treatment population satisfies both ranking properties in Definitions 3 and 4 as long as Π_k is unconstrained and treatment effects are nonnegative, $\tau^k(x) \geq 0$, for all $k \in \{1, ..., K\}$.*

Theorem 11 shows that a simple distributional reweighting can achieve the desirable ranking properties that preserve treatment effect ordering both uniformly over all $\mathcal{X}$ and relatively to some reference population P_{X_0}. In practice, if the exact Radon–Nikodym derivatives are not known, the importance sampling literature contains many techniques for estimating expectations with distributional reweighting [Owen, 2013]. Overall, this simple reweighting modification of the total treatment effect score function addresses important policy considerations when quality measures are used for ranking.

5.6 Information Asymmetry

So far, the incentive alignment and ranking properties have relied on the assumption that both principal and agent operate with the same covariate information X. In practice, Dranove et al. [2003] remark that "providers may be able to improve their ranking by selecting patients on the basis of characteristics that are unobservable to the analysts but predictive of good outcomes." Under such information asymmetry, we show that even the optimistic assumption of an unbiased estimator $\hat{\mu}_0(x)$ is not enough to guarantee zero regret. We both upper and lower bound regret in terms of the additional *heterogeneity* observed by the agent.

Suppose the agent observes additional covariates $U_i \in \mathcal{U}$, and selects a treatment rule $\pi : \mathcal{X}, \mathcal{U} \to [0, 1]$. Suppose the principal still observes only $\{X_i, T_i^\pi, Y_i\}_{i=1}^n$, and chooses a reward function $w(\mathbf{X}, \mathbf{T}^\pi, \mathbf{Y})$ that does not depend on U_i. Let $\mu_t(X, U) = E[Y_i(t)|X, U]$ and $\tau(X, U) = E[Y_i(1) - Y_i(0)|X, U]$. The utility and regret are still defined as in Section 5.3.

Applying the optimal reward function from the full information setting in Section 5.4, suppose the principal rewards the agent with the total treatment effect w_{TT}. As in Section 5.4, suppose the principal applies an unbiased estimator $\hat{\mu}_0(X)$ of the untreated potential outcome conditional on X, with $E[\hat{\mu}_0(x)] = \mu_0(x)\ \forall x \in \mathcal{X}$. Note that the principal identifies $\mu_0(X)$, but *not* $\mu_0(X, U)$. We show that the regret is bounded if the effect of the agent's private information U_i on the untreated potential outcomes is bounded.

Assumption 2 (Bounded Heteogeneity)**.** The effect of U_i on the conditional untreated potential outcome is bounded as $E[|\mu_0(X_i) - \mu_0(X_i, U_i)|] \leq \gamma_{\text{marg}}$.

Note that $\mu_0(X_i) = E[\mu_0(X_i, U_i)|X_i]$. Thus, the heterogeneity bound is akin to bounding a statistical error between the conditional untreated potential outcome $\mu_0(X_i, U_i)$ known to the agent, and "marginal" $\mu_0(X_i)$ estimated by the principal. We can both upper and lower bound the regret in terms of this error:

Theorem 12 (Regret With Information Asymmetry)**.** *Suppose $\hat{\mu}_0(x)$ is unbiased. Under Assumption 2, the regret is upper bounded as $R(\pi^{w_{TT}}) \leq 2\gamma_{marg}$.*

This upper bound is tight up to a linear constant.

Proposition 6. $\forall \varepsilon > 0$, there exist distributions of $X_i, U_i, Y_i(0), Y_i(1)$ wherein $R(\pi^{w_{\text{TT}}}) \geq \gamma_{\text{marg}} - \varepsilon$.

Thus, this notion of *heterogeneity* is key to determining regret under information asymmetry. As a realistic example of cases when Assumption 2 might be satisfied, studies of cardiovascular disease risk have shown that "the magnitude of risk related to smoking is far larger than any ostensible benefit related to moderate drinking" [Mukamal, 2006]. Thus, if U were some attribute for which the relative effect on top of X was small, then γ_{marg} would be small. On the other hand, Rodgers et al. [2019] report that sex hormones and diabetes have compounding effects on cardiovascular disease risk. If X and U have strong compounding effects on $Y(0)$, then γ_{marg} could be large.

Information asymmetry and confounding

Information asymmetry relates closely to the possibility of confounding bias in the estimator $\hat{\mu}_0(x)$. First, the agent's knowledge of U could mean that the data source from which the principal estimated $\hat{\mu}_0(x)$ was also confounded by U. In this case, the literature on sensitivity analysis and policy learning with unobserved confounding proposes a range of robust estimates for $\mu_0(x)$ [see, e.g., Yadlowsky et al., 2022, Kallus and Zhou, 2018]. In our setting, robust estimation of $\mu_0(x)$ is not enough, since the agent's treatment rule can depend on U. Still, the minimax techniques from these works may be a useful avenue for designing future robust reward functions w. Second, information asymmetry can exacerbate confounding if the agent were able to directly affect the principal's estimator $\hat{\mu}_0(x)$, which may happen if, e.g., the principal were to estimate $\hat{\mu}_0(x)$ from the agent's untreated units with $T_i^\pi = 0$. We discuss this case in detail in the Appendix, and show that a stronger assumption yields a similar bound to Theorem 14. Most importantly, our key result is that even without confounding, information asymmetry *still* causes problems via the agent's ability to discriminate on U.

5.7 Experiments

We turn to several clinical datasets to evaluate the welfare impacts of different quality metrics under different conditions of information asymmetry. We show empirically that the regret incurred by w_{ATO} can be high. We also show that under information asymmetry, regret can be *amplified* if the principal estimates $\hat{\mu}_0(x)$ conditioned on some subsets of features. With careful feature selection for x, regret may be reduced.

Horse Colic dataset

The Horse Colic dataset from the UCI repository [Dua and Graff, 2017] contains $n = 300$ horse colic cases. Horses were either treated with surgery ($T = 1$) or not ($T = 0$), with a treatment rate of 0.6. The 20 covariates include each horse's age and presenting symptoms such as abdominal distension, pulse, blood test results, etc. (see Appendix for a full list). Our outcome of interest Y is whether the horse lived ($Y = 1$) or died ($Y = -1$). To approximate the mean conditional potential outcomes, we apply a logistic model with interaction terms between the treatment and covariates: $P(Y(t) = 1|X = x) = \sigma(\beta_0 + \beta_1 x + \beta_2 t + \beta_3 xt)$. We estimate the parameters β on the dataset using logistic regression, and take these as given to produce $\mu_0(x)$ and $\mu_1(x)$. The fitted $\mu_0(x)$ and $\mu_1(x)$ show 62 horses benefiting from surgery and 146 being better off without surgery. On the horses that would benefit, the average benefit was 0.147, which is fairly significant. The clinical validity of these estimated potential outcomes cannot be verified from this data alone, and we instead take these estimates as synthetic potential outcomes.

International Stroke Trial dataset

We also consider data from the International Stroke Trial Collaborative Group [1997], which was a randomized trial studying the effects of drug treatments in acute stroke. Kallus and Zhou [2018] studied this dataset in a different policy learning setting, and we apply a similar setup by comparing treating with high doses of heparin and aspirin ($T = 1$) with aspirin alone ($T = 0$). This leaves $n = 7264$ patients and a treatment rate of 0.33. The 20 covariates include each patient's age, sex, and clinical symptoms such as prior stroke types and complications. Like Kallus and Zhou [2018], we consider a scalarized outcome score $Y \in [-4, 3]$ that accounts for patient outcomes including death, recovery, and side effects at 14 days and 6 months after treatment (details in the Appendix). We approximate the mean conditional potential outcomes using a linear model with interaction terms between treatment and covariates: $E[Y(t)|X = x] = \beta_0 + \beta_1 x + \beta_2 t + \beta_3 xt$. We fit an OLS estimate of β, and use the resulting $\mu_0(x)$ and $\mu_1(x)$ functions as synthetic mean conditional potential outcomes. The fitted $\mu_0(x)$ and $\mu_1(x)$ showed 1360 patients benefiting from heparin and 5904 being better off without heparin. On the patients that would benefit, the average benefit of treatment was 0.025. This tracks with the study's findings that the benefit of heparin was non-significant and inconclusive.

Table 5.1: Utility and regret comparisons for different reward functions. For each reward function w, we report utility $V(\pi^w)$, regret $R(\pi^w)$, and the realized treatment rate $P(T^{\pi^w} = 1)$.

	Horse Colic dataset			Stroke Trial dataset		
Reward function	Utility	Regret	Treat rate	Utility	Regret	Treat rate
w_{ATO}	0.00000	0.1470	0.1922	0.00004	0.0251	0.0001
w_{ATT}	0.00784	0.1391	0.0039	0.00013	0.0250	0.0001
w_{TT}	0.14695	0.000	0.2431	0.02518	0.000	0.1872
w_{TT} (no info)	0.09761	0.0493	0.6275	-0.04888	0.0741	0.4829
w_{TT} (demographic info)	0.09761	0.0493	0.6275	-0.06392	0.0891	0.5041

Results and discussion

We first compare the reward functions from Section 5.4 by calculating the utility and regret empirically over the dataset using the synthetic potential outcomes. Table 5.1 shows that for both datasets, the utility for w_{TT} is positive and higher for the Horse Colic dataset than for the Stroke Trial dataset, which tracks with the clinical finding that the heparin treatment did not have significant effect. The utility for w_{ATO} is close to zero for both datasets.

Next, we consider the effect of information asymmetry on regret. The last two rows of Table 5.1 show the regret when the principal applies w_{TT}, but either observes no covariates ("no info") and identifies $E[Y(0)]$, or only observes age and/or sex ("demographic info") and identifies $E[Y(0)|\text{demographics}]$. While conditioning on age has no effect on the Horse Colic dataset, on the Stroke Trial dataset, conditioning on demographics actually *hurts* utility compared to using "no info." This suggests that it is not always better for the principal to condition on all known information, and thus policymakers should exercise caution in designing stratified quality metrics. For example, CMS currently measures age-specific kidney transplant rates for organ procurement organizations (OPOs) [QCOR, 2023], and our findings question the value of such incomplete stratification under OPO treatment selection. Theoretically, this finding structurally mirrors the phenomenon of *bias amplification* when estimating causal effects with unobserved confounding, where conditioning on more observed features can actually increase bias [Pearl, 2012]. Here we observe *regret amplification*, and we encourage replication of similar analyses by regulatory agencies with internal data sources.

We also study a more continuous spectrum of information asymmetry by showing regret as the principal accumulates increasingly large subsets of the available features. Relating this to Section 5.6, we first sort the features in ascending order of feature importance, as measured by estimating γ_{marg} when the principal knows only the individual feature. Then, Figure 5.1 shows regret when the principal knows increasingly large feature subsets, building up starting from the most "important" feature. Regret is reduced significantly after accounting for less than half of the features. In practice, the principal would not know true values of γ_{marg}. However, approximations of γ_{marg} may serve as a reasonable heuristic for feature selection when a regulator has a large set of known covariates, but needs to prune them for interpretability or cost.

5.8 Conclusions

We have studied the harm to social welfare that occurs when accountability metrics are not aligned with social utility. Even under optimistic assumptions about the availability of an unbiased counterfactual estimate, the potential for regret still exists under information asymmetry with treatment effect heterogeneity. Given the compounded difficulty of estimating causal effects on top of our consideration of treatment incentives and ranking, we recommend that designers exercise caution, humility, and vigilance in their construction of metrics. The task is difficult, but we have established the contours of one potentially fruitful approach.

Future work

There are many important avenues of future work extending from our framework. Information asymmetry presents a prevailing challenge in which we provided bounds on regret, and further analysis on how to improve these regret bounds using ideas from contract design or robust policy learning would be promising and impactful. There is also significant room for modeling extensions. While we considered the simplest modeling framework that could capture the incentive effects of quality measures on treatment selection, there are many other significant factors to consider in practice, including uncertainty and variation in treatment costs, competition between agents, and the ability of treatment units (e.g., patients) to decide where to seek treatment. Analyzing end-to-end regret with these additional factors would be valuable future work.

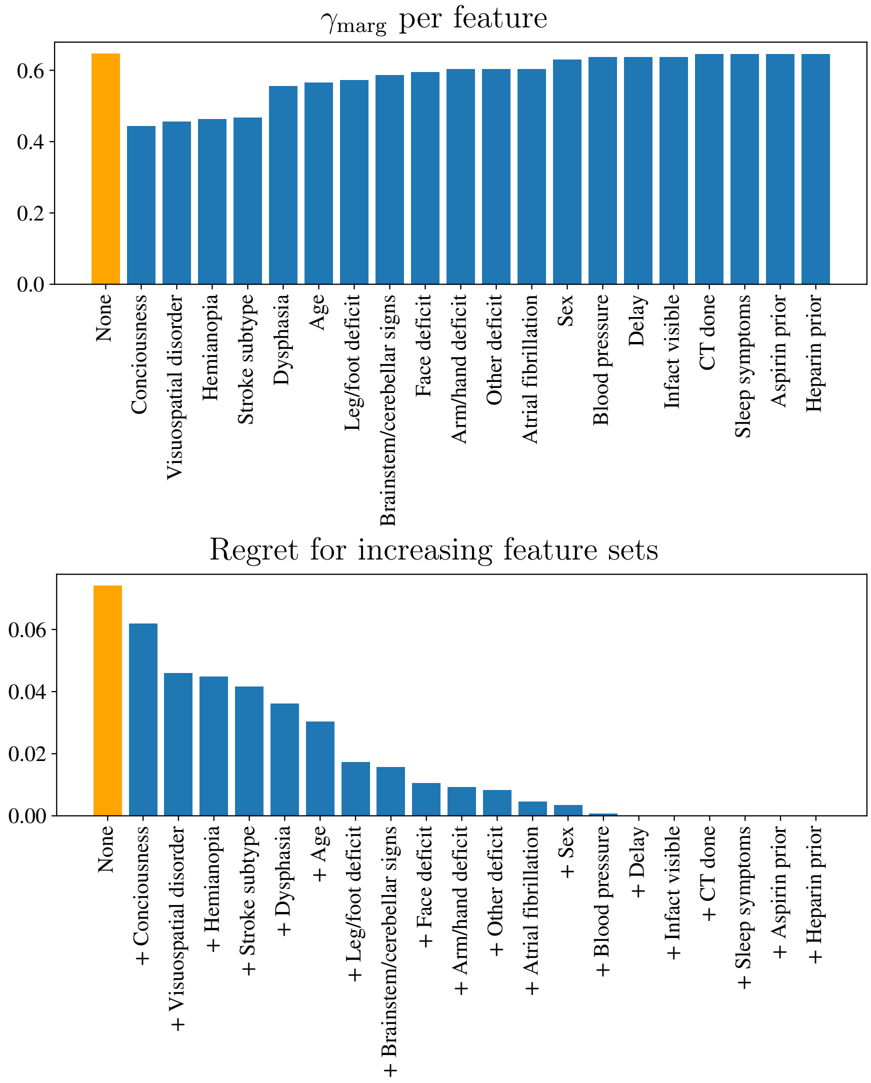


Figure 5.1: Regret under fine-grained information asymmetry on the Stroke Trial dataset (Horse Colic in the Appendix). The *top* plot shows γ_{marg} values if the principal only knows each individual feature. The *bottom* plot shows regret as the principal accumulates features from the left (most important).

Chapter 6

Information Elicitation in Agency Games

6.1 Introduction

The rise of algorithmic and data-driven decision-making has come with an accompanying rise in reliance on numerical measurements of performance. These numerical metrics drive actions at scales ranging from that of individual workers such as teachers [Koretz, 2017], to entire institutions such as hospitals [Muller, 2019], and can have far-reaching impact on social welfare. The problem of designing and choosing evaluation metrics thus continues to be both highly consequential and challenging for all types of organizations, from government agencies to companies.

For example, consider the evaluation-metric design problem faced by a company with a research division. To monitor the health of the organization and guide compensation, metrics are collected to evaluate the research division's productivity. To choose these metrics, company leadership may start with metrics obtained from ad hoc brainstorming or prior experience, taking the form of simple measures such as the number of academic publications, patents, conference presentations, and citations to all of these, along with collaborations with product teams. Such measures may not, however, fully capture the performance of the research division—in fact, they will almost inevitably omit some important factors. The challenge in designing a monitoring system is thus not only in developing methods for measuring performance and optimizing a given set of metrics, but more broadly how to bring new metrics to the fore that were previously unknown and are highly relevant.

This frames our central question of *how to improve a firm's specification of metrics*. The perspective underlying this work is that while company leadership may operate under incomplete information, this does not mean that better information does not exist elsewhere in the firm. In fact, often, agents being evaluated may have better information not only about their effort, but also about how to measure outcomes in a more effective, less gameable way. In the above example, researchers might know of other metrics capturing the quality of their work that company leadership had overlooked, like the comparative quality of publication venues. Thus, we model and analyze situations under which *an agent being evaluated might* *be willing to share information with the principal to improve the metrics that are used for evaluation.* A key strategic dependence in this work is the relationship between the informant (the agent) and the metrics—we consider an agent's incentives to reveal information about metrics that are directly used to evaluate them.

The incomplete nature of metrics has been discussed in the seminal work of Holmstrom and Milgrom [1991] on contracts with multidimensional tasks, where a principal may observe only a subset of dimensions that are relevant to their value or the agent's cost. When the principal is aware of which dimensions are missing, Holmstrom and Milgrom [1991] show how to optimally reward the observed dimensions given properties of the agent's cost structures. The distinguishing feature of our setting is that we consider the unobserved dimensions to be *unknown unknowns*. We focus on an information transfer mechanism where the agent has the power to possibly reveal these hidden dimensions to the principal.

Our question also fits into a broad class of problems on information design and persuasion games that studies the welfare effects of different information structures [Kamenica and Gentzkow, 2011, Milgrom and Roberts, 1986]. Our model treats an agent as a sender and principal as a receiver, and can be seen as applying specific restrictions to the class of information-revelation strategies that the agent can choose from in order to capture properties of the design of metrics in agency problems faced by firms. Specifically, we consider a setting where the agent's information revelation strategy cannot depend on the value of the realized signal. Thus, we seek to model an agent's choice of whether or not to reveal the *observability* of a metric, rather than a realized signal value.

To yield initial tractable insights to this metric discovery problem, we examine an agent's incentives for information sharing through the lens of an *agency game with information transfer*. We build on a classical agency game where a principal contracts an agent to complete a task, and the principal only has partial information about the agent's costs when setting a contract. To capture the agent's additional information and opportunity to improve the metrics by which they are evaluated, our model supposes that the agent is privately aware of additional variables that correlate with their cost of task completion, and further has the opportunity to reveal these additional variables to the principal prior to the design of the contract. We analyze when the agent would prefer for the contract to depend on these cost-correlated variables compared to an agnostic contract. Importantly, the agent must decide whether to reveal their cost-correlated variables to the principal prior to realizing the values of those variables or their true costs.

While the principal is always better off when having more information about the agent's costs, the incentives for the agent are more nuanced—revealing information reduces the amount of information rent the agent can extract when their costs turn out to be low. However, better information also means more high-cost jobs will go forward that the agent otherwise would not have accepted, as the optimal contract would adequately reward high-cost tasks.

We first consider whether the agent prefers to reveal or conceal the observability of a cost-correlated environmental variable. We show that the agent prefers to reveal the variable if conditioning on this cost-correlated variable reveals a strong enough differentiation between high and low costs of task completion. Next, we expand the agent's action space to include the ability to *garble* or add noise to their information before it passes to the principal. For example, in modern online platforms, avenues for introducing this type of noise include third party clients or policies requiring differential privacy [Dwork et al., 2006]. We show that under a fairly wide set of conditions, the agent may prefer to reveal a garbled signal over both fully concealing and fully revealing the original variable. This suggests that a noisy information transfer mechanism can yield superior equilibria over simpler or more restrictive alternatives.

Contributions

Our contributions can be summarized as follows:

1. We introduce a model for the process of discovery of unknown metrics, in the form of an agency game with information transfer.

2. We present sufficient conditions under which an agent would prefer revealing a metric to a principal over keeping the metric concealed; and vice versa.

3. We further relax the agent's action space to include the ability to reveal *garbled* information, where the choice of the amount of garbling interpolates between concealment and revelation. In this setting, we give sufficient conditions for the agent to prefer garbling over full revelation.

4. We analyze the consequences of information revelation on principal utility and total welfare, leveraging connections between our model and price discrimination.

Related Work

Our model builds on literature from contract design and agency games. It also fits into a large literature on information design, and can be seen as a specific structure of information design problem. Our model also overlaps with price discrimination, and we bring techniques from these literatures to a new motivation of discovering metrics.

Agency games and contract design We build from the well-established contract design problem of Laffont and Tirole [1986], which concerns a principal's design of a contract when an agent's effort and cost type are privately held by the agent. Key to this setting are asymmetries in *values* and *information* between the principal and agent, and the literature explores issues of moral hazard and adverse selection that arise from these asymmetries [Laffont and Martimort, 2009, Gibbons et al., 2013, Milgrom and Roberts, 1992].

A fundamental result regarding signaling incentives in contract theory is Holmström [1979]'s sufficient statistic theorem, which showed that it benefits an agent for the contract to be conditioned on any information that is independently informative of the agent's effort.

Our setting models an agent's incentive to share information about its cost type, which yields different results from the analysis of effort signaling, and is closer to some analyses of persuasion games which we discuss in more detail below. Milgrom [1981] also classifies the "favorableness" of signals to an agent, presenting monotonicity properties that we also leverage in this work.

Persuasion games and information design The use of stakeholder-supplied information for decision-making was seminally introduced by Milgrom and Roberts [1986] in the form of *persuasion games*. Milgrom and Roberts [1986] give an example of understanding a buyer's purchasing strategy in the face of a seller who can send a quality signal about their product. While this broad motivation of information transfer from interested parties is very close to our motivation of learning business metrics from evaluated agents, the game that we present differs in two main ways from the particular setting in Milgrom and Roberts [1986].

First, we restrict the agent's information revelation strategy for a given variable to not be able to depend on the realized value of the variable. In other words, the agent must commit to fully revealing a variable in advance of realizing the variable's value. In this sense, our setting can be viewed as a constrained persuasion game. The purpose of this restriction is to bring our setting closer to a problem of variable *discovery*, where the fundamental problem is a principal's lack of awareness of a variable's observability, rather than signaling [Spence, 1978], where the uncertainty is about the variable's realized value.

The second main distinction is in the direction of information transfer. In our setting, the "decision-maker" is the agent, who decides whether or not to complete a task. We consider the agent's incentive to reveal information about their decision-making parameters (their cost type) to a principal who designs a payment contract. This second distinction is more technical and less important for the fundamental motivation of revealing observability instead of value.

More broadly, Bayesian persuasion and information design provides a general framework for analyzing the effects of the distribution of information on the outcomes of a game [Kamenica, 2019, Bergemann and Morris, 2019]. Many works have cast this general form to analyze applications from grading systems in schools to courtroom evidence policies [Kamenica and Gentzkow, 2011, Boleslavsky and Cotton, 2015, Ostrovsky and Schwarz, 2010]. Our model can be seen as a specific instantiation of an information design problem where the *sender* is the agent, the *receiver* is the principal, and the contracting relationship determines the principal's and agent's action spaces and equilibria. We impose a constraint over the sender's information transfer policy in order to capture the incentives for an agent to reveal *observability* of a variable to the principal, whereas prior work has focused on revealing the value of a signal (similarly to Milgrom and Roberts [1986]). Also built into this constraint is that the agent cannot lie about their signal, as our model is motivated by understanding settings where a principal has powerful data collection and verification capabilities.

Price discrimination Our analysis framework of an agency game with information transfer has analogies with classical price discrimination, and thereby opens new avenues for applying the well-developed tools from price discrimination to understanding the metric design problem. It also uncovers new challenges that extend existing price discrimination perspectives. This connection has also been discussed by Bergemann et al. [2015], who define a mapping from third degree price discrimination onto the class of agency problems and establish the existence of market segmentations that achieve all possible trade-offs between consumer and producer surplus within some basic constraints. Our work complements this analysis—instead of analyzing all possible segmentations, we consider the agent's information-revelation incentives as a function of the properties of a specific segmentation induced by some cost-correlated variable which is initially only available to the agent.

Our garbling setting also notably differs from the models of price discrimination with transportation costs and arbitrage [Wright, 1993] or restricted price discrimination [Aguirre et al., 2010]. While both garbling and restricted price discrimination mechanisms effectively interpolate between full discrimination and no discrimination, we prove a substantive difference between these mechanisms by showing that while the agent might prefer some intermediate amount of garbling over both full revelation or hiding, the agent will always prefer either full discrimination or no discrimination over any intermediate restriction over the amount of allowed price discrimination. We give a more detailed comparison of our garbling model to price discrimination in Section 6.4.

Sunspots and correlated equilibria Our model is also connected to the so-called sunspots literature [Woodford, 1990, Howitt and McAfee, 1992]. In this literature, there are multiple steady-state equilibria and an observable random variable produces a correlated equilibrium, with agents conditioning their behavior on this otherwise extraneous variable not because it matters to payoffs, but because it predicts others' behaviors. The literature is known as sunspots because, during the 19th century, some people believed sunspot activity predicted agricultural yields [Jevons, 1884], and while it did not, it could predict the behavior of commodity traders who believed it did. This literature is the polar opposite of the problem we study: sunspots are a known, extraneous variable believed to be relevant, while we study an unknown (to the principal) relevant variable.

Preferences for privacy Our model also interacts with literatures on privacy and information-hiding, by highlighting a simple situation where agents experience conflicting incentives both for and against sharing information. The result can be that agents may prefer partial sharing, which echoes the subtleties that arise elsewhere when privacy and behavior interact [Cummings et al., 2016].

6.2 Model Setup: Agency Game with Information Transfer

To gain insight into the incentives surrounding the discovery of metrics, we start with a standard agency game in which a principal contracts an agent to complete a task. In such a game, we ask, *when does the agent have an incentive to reveal observability of a cost-correlated variable to the principal?*

Specifically, suppose a principal contracts the agent to complete a task, where an agent may exert binary effort. Suppose the principal receives value b if the agent exerts effort and completes the task, and zero otherwise (we assume that the task is completed deterministically if the agent exerts effort). Suppose the agent incurs cost $C \in \mathbb{R}_+$ for exerting effort. The exact cost is unobserved to the principal, but the principal is aware of a prior distribution over agent's cost type, denoted by the random variable C. The agent observes both the proposed transfer and their realized cost type before deciding whether or not to exert effort. This maps onto the well-established agency game setup where the principal must design a contract when the agent's cost and effort are private [Laffont and Tirole, 1986].

To model a setting where the agent might possess additional information, suppose the agent is aware of an environmental variable $X \in \mathcal{X}$ which is correlated with cost C. The question of knowledge of X becomes significant when X is correlated with C.

In introducing this correlation into the agency game, we assume that X is an *unknown unknown* to the principal. Prior to the principal's design of the contract, the agent has a choice of whether or not to inform the principal of the environmental variable X, which entails revealing both the *observability* of X at the time of the design of the contract, and the realized *value* of X at the execution of the contract (which the principal can verify). If the agent chooses not to inform the principal of the observability of X, then the rest of the game proceeds as a standard agency game with private cost: the principal designs a contract based on their knowledge of the prior distribution C. If the agent chooses to inform the principal of X, then the principal can offer a contract that conditions on the realized value of X.

The timing of the full agency game with the possibility of agent information transfer is summarized in Figure 6.1. The text in black matches the standard agency game [Laffont and Tirole, 1986], and the text in gray represents additional elements introduced by our model.

In our simplified environment involving binary effort, the principal's contract design problem when X is concealed reduces to choosing a single transfer amount p where the agent is paid p if the task is completed, and zero otherwise. If X is revealed, then the principal offers a contract with distinct transfers $\rho(x)$ for different realized values of $x \in \mathcal{X}$. At the time of the execution of the contract, the agent receives transfer $\rho(x)$ if the task is completed and $X = x$; and zero otherwise. We assume that the principal still receives the same value b if the task is completed, regardless of X.

The key question in this work concerns the agent's decision of whether or not to inform the principal of the existence of the environmental variable X at time $t = 1$. In Section 6.3, we consider this as a binary decision of whether or not to reveal X; we will later relax

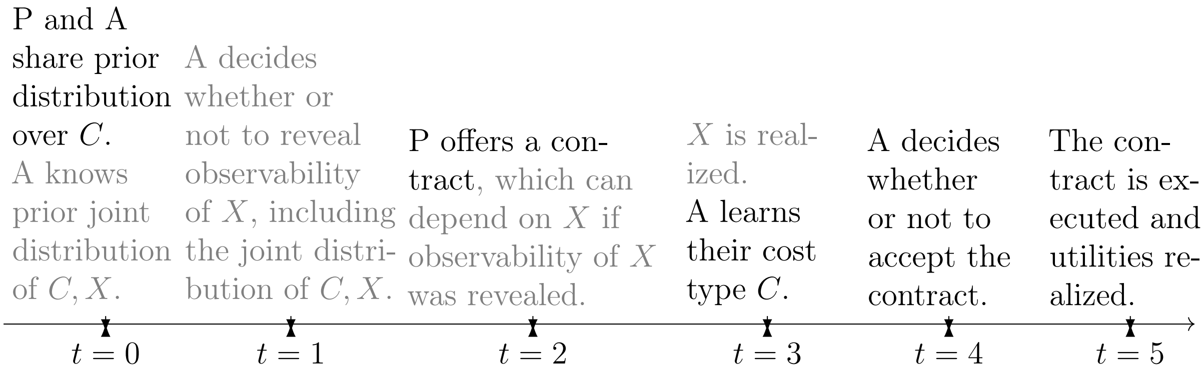

P and A share prior distribution over C.
A knows prior joint distribution of C, X.
A decides whether or not to reveal observability of X, including the joint distribution of C, X.
P offers a contract, which can depend on X if observability of X was revealed.
X is realized.
A learns their cost type C.
A decides whether or not to accept the contract.
The contract is executed and utilities realized.
$t = 0$
$t = 1$
$t = 2$
$t = 3$
$t = 4$
$t = 5$

Figure 6.1: Timing of the agency game with information transfer between principal (P) and agent (A).

this in Section 6.4 to expand the agent's action space to reveal a garbled version of X, thus interpolating between the concealed and revealed settings.

Notation

Let $F(c) = \mathbb{P}(C \leq c)$ denote the cumulative distribution function (CDF) of the prior cost distribution C. We also define $F_x(c) = \mathbb{P}(C \leq c | X = x)$ as the CDF of the conditional distribution of C given X. Throughout, we will assume that the cost C is a continuous random variable with density $f(c) = F'(c)$ and with conditional density $f_x(c) = F'_x(c)$.

We will use Π to denote the principal's utility, V to denote the agent's utility, and W to denote total welfare, defined as the sum of the principal's and agent's utilities.

Let $\mathbb{1}(\cdot)$ denote an indicator function with $\mathbb{1}(x \in S) = \begin{cases} 1 & \text{if } x \in S \\ 0 & \text{otherwise.} \end{cases}$

All proofs are given in the Appendix.

6.3 Welfare Effects of Information Revelation

We now consider the agent's decision of whether to conceal or reveal the existence X at time $t = 1$ in the agency game with information transfer outlined in Figure 6.1. To analyze this decision, we compare the agent's utility in the information-agnostic contract that results from keeping X concealed and in the informed contract that depends on X. We also analyze the consequences of the resulting decision on the principal's utility and on total welfare.

As an overview, we give sufficient conditions for the agent to prefer to conceal, and sufficient conditions for the agent to prefer to reveal X. In general, the principal always prefers the revealed setting. Finally, we analyze the consequences of revelation on total welfare, connecting to analogous results from price discrimination.

Utilities for Principal and Agent

We begin by outlining the contract and equilibrium behavior of the principal and agent when the contract is agnostic of X, which we refer to as the *concealed information* setting. Then, we outline the contract and equilibrium behavior when the principal can condition on X to determine payments to the agent, which we refer to as the *revealed information* setting. We assume that both principal and agent are risk neutral throughout.

Concealed information contract

If information is not revealed, then the rest of the game proceeds as a standard agency game with private cost, where the principal chooses a single transfer p based on their knowledge of the prior distribution over the agent's cost C. The agent's optimal policy at the execution of the contract is to exert effort if their realized cost is less than or equal to the payment. Thus, the agent's best response to the principal's choice of transfer p is to exert effort with probability $F(p) = \mathbb{P}(C \leq p)$ (the set where $C = p$ has measure zero).

For a given choice of transfer p, the principal's expected utility when X is hidden is given by

$$\Pi_{\text{con}}(p) \coloneqq F(p)(b - p). \tag{6.1}$$

The utility of the agent under the principal's choice of transfer p is given by

$$V_{\text{con}}(p) \coloneqq \mathbb{E}[(p - C)\,\mathbb{1}(C < p)]. \tag{6.2}$$

The principal moves first and chooses $p^* \in \arg\max_{p \geq 0} \Pi_{\text{con}}(p)$. The agent's utility at equilibrium is then $V_{\text{con}}(p^*)$.

Revealed information contract

If the environmental variable X is revealed at time $t = 1$, then the principal has the ability to instead choose a transfer which depends the value of X, denoted as $\rho : \mathcal{X} \to \mathbb{R}_+$. Since the agent is aware of the values of both C and X before deciding whether to exert effort, and there is no possibility to lie about X, the agent's optimal policy is to exert effort if their realized cost is less than or equal to the transfer given the realized X value. Therefore, the agent's best response to the principal's transfer function $\rho(\cdot)$ is to exert effort with probability $F_x(\rho(x))$ when $X = x$. The principal's expected utility in this revealed setting is given by

$$\Pi_{\text{rev}}(\rho) \coloneqq \mathbb{E}[F_X(\rho(X))(b - \rho(X))] = \mathbb{E}[\Pi_X(\rho(X))],$$

where $\Pi_x(p) \coloneqq F_x(p)(b - p)$.

The utility of the agent under the principal's choice of transfer function $\rho(\cdot)$ is then

$$V_{\text{rev}}(\rho) \coloneqq \mathbb{E}[(\rho(X) - C)\,\mathbb{1}(C < \rho(X))] = \mathbb{E}[V_X(\rho(X))],$$

where $V_x(p) \coloneqq \mathbb{E}[(p - c)\,\mathbb{1}(C < p) | X = x]$.

For ease of exposition, our analysis will focus on a one-dimensional binary feature: $\mathcal{X} = \{0, 1\}$, where $X = 1$ with probability θ. We frame the principal's decision problem as that of choosing $\rho(0) = p_0$ and $\rho(1) = p_1$. Further extensions to higher cardinalities or continuous X may be of practical interest, though we expect our qualitative insights to extend to these settings.

The principal's expected utility becomes

$$\Pi_{\text{rev}}(p_0, p_1) := (1 - \theta)\Pi_0(p_0) + \theta\Pi_1(p_1),$$

and the agent's utility becomes

$$V_{\text{rev}}(p_0, p_1) := (1 - \theta)V_0(p_0) + \theta V_1(p_1).$$

The principal moves first and chooses $p_0^*, p_1^* \in \arg\max_{p_0,p_1} \Pi_{\text{rev}}(p_0, p_1)$, resulting in an equilibrium where the agent's utility is $V_{\text{rev}}(p_0^*, p_1^*)$.

Without loss of generality, we will refer to the situation when $X = 1$ as the "stronger" situation with generally lower cost for effort. That is, $p_0^* > p_1^*$.

In the rest of this section, we will analyze the effects of information revelation on the agent's utility, principal's utility, and total welfare. Specifically, we will compare $V_{\text{con}}(p^*)$ to $V_{\text{rev}}(p_0^*, p_1^*)$ to understand the agent's incentive to reveal X. We will also analyze the consequences of the agent's revelation decision on the principal's utility, as well as consequences on total welfare.

Agent's Revelation Incentives

The central goal in this work is to analyze the circumstances under which the agent would prefer to either conceal or reveal the existence of X at time $t = 1$. Thus, to begin, we will analyze situations that lead to $V_{\text{con}}(p^*)$ being higher than $V_{\text{rev}}(p_0^*, p_1^*)$ or vice versa, focusing on properties of the distributions F, F_0, and F_1. In the analogy to price discrimination, we may think of $F(p)$ as the task completion "quantity" as a function of price p, or the proportion of agents drawn uniformly at random from a population with costs distributed as C that would complete the task for price p.

We present sufficient conditions on the distributions F_0, F_1 for the agent to either prefer to conceal or to prefer to reveal. We then build intuition for these conditions through an example using exponential and Weibull distributions.

Concealment condition with one zero-cost type

First, we present a sufficient condition for the agent to prefer to conceal X when one of the agent types is anchored at zero. That is, $X = 1$ implies that the agent incurs zero cost. Proposition 7 gives a sufficient condition on F_0 for the agent to prefer for the environmental variable X to remain concealed.

Proposition 7 (Sufficient concealment condition with zero-cost type)**.** Suppose F_0 is a concave and continuously differentiable CDF. Suppose $C|X = 1$ takes value 0 with probability 1. Suppose the ratio $\frac{F_0(p)}{f_0(p)}$ is strictly monotone increasing for $p > 0$. Then $V_{\text{con}}(p^*) > V_{\text{rev}}(p_0^*, p_1^*)$ if

$$\theta > (1-\theta)\frac{1}{\eta((1-\theta)p_0^*)} - \frac{1}{\eta_0\left(p_0^*\right)}, \tag{6.3}$$

where $\eta(p) = \frac{p(1-\theta)f_0(p)}{(1-\theta)F_0(p)+\theta}$ and $\eta_0(p) = \frac{pf_0(p)}{F_0(p)}$ are the respective price elasticities for task completion quantity for the mixture distribution C and the conditional distribution F_0.

Qualitatively, the elasticity η captures the sensitivity of the task completion to price. Thus, the inequality in equation (6.3) corresponds to a scenario when the sensitivity of the task completion to price when $X = 0$ does not differ too strongly from that of the concealed setting. For example, this arises when F_0 is close to the constant function 1. We give another example using the exponential distribution in Section 6.3 below, where equation (6.3) holds if the mean of F_0 is low enough. In summary, the agent prefers concealment if the higher cost type still has relatively low cost.

Concealment and revelation conditions under a decreasing ratio assumption

To give additional sufficient conditions for concealment and revelation beyond the anchored setting with one zero-cost type, we apply an analysis technique similar to that of Aguirre et al. [2010], who analyzed the effects of third degree monopoly price discrimination on total welfare.

Suppose the principal, on knowing X, is constrained to choose transfers p_0, p_1 subject to the constraint that $p_0 - p_1 < r$ for some $r \geq 0$. Let $p_0(r), p_1(r)$ denote the principal's optimal transfers under this constraint:

$$\begin{aligned} p_0(r), p_1(r) \in \arg\max_{p_0,p_1} \quad & \Pi_{\text{rev}}(p_0, p_1) \\ \text{s.t.} \quad & p_0 - p_1 \geq r. \end{aligned} \tag{6.4}$$

For notational convenience, let $V_{\text{const}}(r) \coloneqq V_{\text{rev}}(p_0(r), p_1(r))$. In a similar structure to Aguirre et al. [2010], the results in this section come from considering the "marginal effect of relaxing the constraint" on the agent's value.

Under the following closely analogous assumptions to those invoked by Aguirre et al. [2010], we derive properties of $V_{\text{const}}(r)$.

Assumption 3 (Concave principal utility)**.** The principal's utility in each realized environment is strictly concave: $\Pi_0''(p) < 0$, $\Pi_1''(p) < 0$.

Assumption 4 (Decreasing ratio condition (DRC))**.** The ratios $\frac{V_0'(p)}{\Pi_0''(p)}$ and $\frac{V_1'(p)}{\Pi_1''(p)}$ are both decreasing in p.

Assumption 4 is analogous to the "increasing ratio condition" assumption from Aguirre et al. [2010], which instead has the derivative of total welfare in the numerator. Our analysis naturally extends this to focus on agent utility. Assumption 4 holds in almost the same set of conditions as the assumption on total welfare from Aguirre et al. [2010], and we discuss the subtleties of the differences between these assumptions in Appendix E.1.

Lemma 16. Under Assumptions 3 and 4, $V_{\text{const}}(r)$ is strictly quasi-convex for $r \in [0, p_0^* - p_1^*]$. That is, if there exists $\hat{r} \in [0, p_0^* - p_1^*]$ such that $V_{\text{const}}(\hat{r}) = 0$, then $V''_{\text{const}}(\hat{r}) > 0$.

The strict quasi-convexity of the agent's utility in r makes it possible to derive sufficient conditions for revelation and concealment by differentiating V_{const} and evaluating the sign of the derivative at extreme values of r. Adapting this machinery from Aguirre et al. [2010], but focusing on the agent's value instead of total welfare, we give such sufficient conditions for the agent to prefer concealing or revealing X below.

Proposition 8 (Sufficient concealment condition under DRC)**.** Under Assumptions 3 and 4, $V_{\text{con}}(p^*) > V_{\text{rev}}(p_0^*, p_1^*)$ if

$$\frac{(1-\theta)(b-p_0^*)}{2-\sigma_0(p_0^*)} < \frac{\theta(b-p_1^*)}{2-\sigma_1(p_1^*)}, \tag{6.5}$$

where $\sigma_x(p) = \frac{F_x(p) f_x'(p)}{f_x^2(p)}$ is the curvature of the inverse of the task completion quantity function $F_x(p)$.

Proposition 8 implies that a high enough difference in curvature between $\sigma_0(p_0^*)$ and $\sigma_1(p_1^*)$ implies that the agent will prefer the concealed contract over the revealed contract. That is, inverse task completion quantity when $X = 0$ is more convex than the inverse task completion quantity when $X = 1$ at the revealed transfers p_0^*, p_1^*. This is exactly the flipped version of the condition in Aguirre et al. [2010]'s Proposition 2, which implied that total welfare is higher under price discrimination. We next give a sufficient condition for the agent to prefer revelation.

Proposition 9 (Sufficient revelation condition under DRC)**.** Under Assumptions 3 and 4, $V_{\text{con}}(p^*) < V_{\text{rev}}(p_0^*, p_1^*)$ if

$$\frac{2 + L(p^*)\alpha_1(p^*)}{2 + L(p^*)\alpha_0(p^*)} > \frac{\theta F_1(p^*)/f_1(p^*)}{(1-\theta)F_0(p^*)/f_0(p^*)}, \tag{6.6}$$

where $L(p) = \frac{b-p}{p}$ is the Lerner index [Lerner, 1995], and $\alpha_x(p) = \frac{-p f_x'(p)}{f_x(p)}$ is the curvature of the task completion quantity function $F_x(p)$,

Intuitively, Proposition 9 says that if the curvatures of F_0 and F_1 are different enough (relative to the ratio of the CDFs themselves), then the agent will prefer to reveal the environmental variable X.

Remark An important property analyzed by Milgrom [1981] is the monotone likelihood ratio property (MLRP), which here would say that $\frac{f_0(c)}{f_1(c)}$ is increasing for all c. The MLRP would imply that both sides of the inequality in equation (6.6) are greater than 1. However, this does not necessarily imply an order between these ratios, and there exist distributions that satisfy the MLRP that yield either of the inequality directions above. We give a specific example of this using the Weibull distribution in Section 6.3 below.

Example: Exponential and Weibull Distributions

To concretely illustrate the conditions in Propositions 7, 8, and 9, we parameterize the conditional cost distributions using the exponential distribution and the more general Weibull distribution.

First, to illustrate the condition in Proposition 7, let $C|X = 0 \sim \text{Exp}(\frac{1}{\lambda_0})$, where λ_0 represents the scale parameter and is also the mean of the distribution. Specifically,

$$F_0(c) = \begin{cases} 1 - e^{-\frac{1}{\lambda_0}c} & c \geq 0 \\ 0 & c < 0. \end{cases} \tag{6.7}$$

Then the condition in Equation (6.3) is equivalent to $\lambda_0 < b\psi(\theta)$, where $\psi(\theta) = \frac{\theta}{\left(\left(\frac{1}{1-\theta}\right)^{\frac{1}{\theta}} - \left(\frac{1}{1-\theta}\right)^{\frac{1-\theta}{\theta}} - \theta\right)}$ is monotone decreasing function bounded between 0 and 1 for $\theta \in [0, 1]$. Thus, as long as the average cost λ_0 is less than a θ-dependent scaling of the task completion value b, the condition in Proposition 7 holds. In other words, if the average cost when $X = 0$ is not too high, then the agent will prefer concealment.

Beyond fixing F_1 at zero cost, let C be a mixture of exponential distributions with $C|X = 0 \sim \text{Exp}(\frac{1}{\lambda_0})$ and $C|X = 1 \sim \text{Exp}(\frac{1}{\lambda_1})$, where

$$F_x(c) = \begin{cases} 1 - e^{-\frac{1}{\lambda_x}c} & c \geq 0 \\ 0 & c < 0. \end{cases} \tag{6.8}$$

Figure 6.2 plots the difference $V_{\text{rev}}(p_0^*, p_1^*) - V_{\text{con}}(p^*)$ for all $\lambda_0, \lambda_1 \in [0, b]$. As seen in Proposition 7, if $\lambda_1 = 0$, then the agent prefers to hide if λ_0 is sufficiently low. This continues to hold for λ_1 sufficiently close to 0. More generally, Figure 6.2 shows that the agent prefers to reveal if the means λ_0, λ_1 are sufficiently far apart.

For the exponential mixture, the inequalities in Propositions 8 and 9 do not hold for any combinations of λ_0, λ_1. Thus, the condition in Proposition 7 covers cases not covered by Proposition 8. However, we see Propositions 8 and 9 take effect for the more general Weibull distribution, with

$$F_x(c) = \begin{cases} 1 - e^{\left(-\frac{1}{\lambda_x}c\right)^{k_x}} & c \geq 0 \\ 0 & c < 0. \end{cases} \tag{6.9}$$

For example, for a fixed λ_0, λ_1, increasing $k = k_0 = k_1$ increases the difference in curvature between F_0 and F_1 at p^*, yielding a set of values of $k > 1$ in which the condition in Proposition

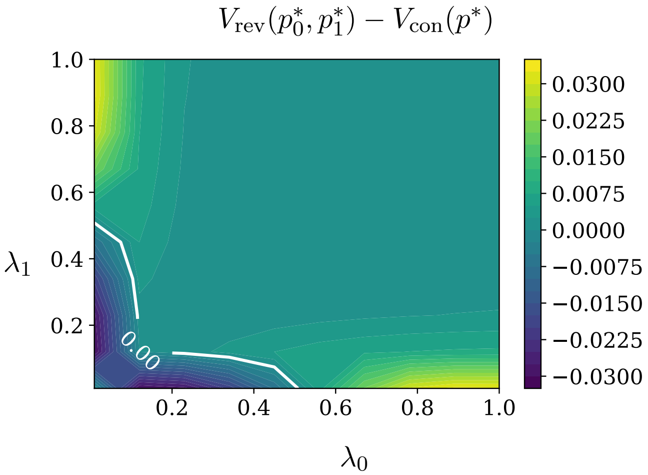


Figure 6.2: Difference between agent's utility in the revealed setting and concealed settings when C is distributed as a mixture of exponentials. For each pair (λ_0, λ_1), a positive value indicates that the agent prefers revelation, and a negative value indicates that the agent prefers concealment. The contour line shows all (λ_0, λ_1) for which $V_{\text{rev}}(p_0^*, p_1^*) - V_{\text{con}}(p^*) = 0$. The parameters $b = 1$ and $\theta = \frac{1}{2}$ are fixed, and λ_0 and λ_1 are varied up to b.

9 holds. For Proposition 8, the condition holds if k_1 is sufficiently small, and k_0 is sufficiently large for $\lambda_0 > \lambda_1$.

Principal's Revelation Preferences

While the agent might sometimes prefer the hidden setting over the revealed setting, we next show that the principal always prefers revelation. First, Lemma 17 shows that the principal is never worse off under revelation.

Lemma 17 (Principal prefers revelation)**.** Revealing X never decreases the value of the principal: $\Pi_{\text{rev}}(\rho^*) \geq \Pi_{\text{con}}(p^*)$, where $\rho^* \in \arg\max_\rho \Pi_{\text{rev}}(\rho)$. Revealing X strictly increases the value of the principal only if X and C are not independent.

The principal strictly benefits from information revelation if the monotone likelihood ratio property (MLRP) is satisfied between the revealed distributions.

Assumption 5 (Monotone likelihood ratio property (MLRP) [Milgrom, 1981])**.** The ratio $\frac{f_0(c)}{f_1(c)}$ is strictly increasing in c.

Lemma 18 (Principal strictly benefits from revelation)**.** Let F_0 and F_1 be continuously differentiable CDFs. If the MLRP holds (Assumption 5), then the principal strictly benefits when X is revealed: $\Pi_{\text{rev}}(p_0^*, p_1^*) > \Pi_{\text{con}}(p^*)$.

Intuitively, as the first mover, the principal will never be hurt by having additional freedom to condition on X when selecting prices that maximize their utility. Lemma 18 gives the MLRP as a sufficient condition for revelation of X to yield a strict benefit for the principal.

Total Welfare Consequences

We now consider whether revealing X increases total welfare, or the sum of utilities of the principal and agent. When X is concealed, total welfare is given by

$$W_{\text{con}}(p) \coloneqq V_{\text{con}}(p) + \Pi_{\text{con}}(p),$$

and when X is revealed, total welfare is given by

$$W_{\text{rev}}(p_0, p_1) \coloneqq V_{\text{rev}}(p_0, p_1) + \Pi_{\text{rev}}(p_0, p_1).$$

The question of whether price discrimination increases total welfare has been well studied [Varian, 1985]. Our comparison of the concealed vs. revealed contracts has a direct isomorphism with third degree monopoly price discrimination. Specifically, our concealed setting corresponds to monopoly pricing without price discrimination, with the seller acting as principal and buyer acting as agent. Our revealed setting corresponds to a monopoly seller enacting third degree price discrimination over markets segmented by X.

Thus, with minor adjustments, we can apply results from the price discrimination literature that characterize the effects of third degree monopoly price discrimination on total welfare. Mirroring Varian [1985]'s seminal work, Lemma 19 shows that total welfare increases only if the quantity of tasks completed also increases in the revealed setting compared to the concealed setting.

Lemma 19. Total welfare increases under revelation $(W_{\text{rev}}(p_0^*, p_1^*) - W_{\text{con}}(p^*))$ only if task completion quantity increases under revelation,

$$F(p^*) < (1-\theta)F_0(p_0^*) + \theta F_1(p_1^*).$$

Example when total welfare decreases

It is still possible for total welfare to decrease under information revelation. Mirroring an example from Varian [1985], we provide an illustrative example here where total welfare decreases when task completition quantity does not increase.

Suppose $C|X = 1 \sim \text{Unif}(0, 1)$, and $C|X = 0 \sim \text{Unif}(\frac{1}{2}, \frac{3}{2})$. Suppose $\theta = \frac{1}{2}$. Suppose $b = 1$. Then the optimal payments for each of F, F_0, F_1 all fall in the "interior" of $F(x)$:

$$\frac{1}{2} \leq p_1^* < p^* < p_0^* \leq 1.$$

When the solutions all fall in the interior, we have $p^* = \frac{p_1^* + p_0^*}{2}$, and $F(p^*) = \frac{F_1(p_1^*) + F_0(p_0^*)}{2}$. This now violates the necessary condition in Lemma 19, since the output does not increase under revelation, but the payments change. Total welfare decreases as long as $p_0^* \neq p_1^*$.

6.4 Information Revelation With Garbling

So far, we have compared the settings when the environmental variable X is either concealed or and revealed, focusing on the agent's incentive to induce each setting at time $t = 1$. We now generalize the agent's action space to instead be able to reveal a *garbled* version of the variable X. From a modeling standpoint, the agent's garbling action space interpolates between the concealed and revealed settings. In this work, we consider a *randomized response* garbling mechanism which has been applied in many settings, from statistical informativeness [Blackwell, 1951] to survey experiment design [Warner, 1965] to differential privacy [Dwork et al., 2006, Kasiviswanathan et al., 2011].

As an overview of results in this section, we show that there exist conditions under which the agent would prefer to garble over both concealment and revelation. Thus, having the option to garble can benefit the agent, and even induce the agent to reveal more information when they would otherwise fully opt to conceal X. We also analyze the effects of garbling on the principal's utility and total welfare.

Model Setup: Information Transfer with Garbling

Suppose the agent has the option to present the principal with a different variable Y with random noise on top of X, defined as

$$Y = \begin{cases} X & \text{w.p. } \varepsilon \\ \xi & \text{w.p. } 1-\varepsilon \end{cases} = \begin{cases} X & \text{w.p. } \frac{1+\varepsilon}{2} \\ \neg X & \text{w.p. } \frac{1-\varepsilon}{2}, \end{cases} \tag{6.10}$$

where $\xi \sim \text{Bernoulli}(\gamma)$ is independent of X and C. This work considers this randomized response mechanism for binary X, though further extensions with different noise models can be made for continuous X.

The game with garbling proceeds as before, but the agent selects $\varepsilon \in [0, 1]$ at time $t = 1$. The full timing is outlined in Figure 6.3. The agency game with garbled information transfer is a generalization of the previous game: in the previous timing in Figure 6.1, the agent's choice at $t = 1$ would be equivalent to selecting ε from a more restricted set $\{0, 1\}$.

It is assumed that the principal, on knowing Y, cannot back out the previous value of X. The principal treats the variable Y as an environmental variable with prior joint distribution Y, C, and proceeds to design the optimal contract conditioning on Y. The principal does not have access to ε when designing the contract—only the variable Y and its potential values.

The game proceeds as delineated for the revealed information contract in Section 6.3, but using the variable $Y \in \mathcal{Y}$ instead of X. Specifically, the principal designs a contract with transfer function $\rho : \mathcal{Y} \to \mathbb{R}_+$, and upon execution of the contract, the agent receives transfer $\rho(y)$ if they exert effort and $Y = y$. For binary X and Y following equation (6.10), the principal chooses $\rho(0) = p_0$ and $\rho(1) = p_1$.

Let the conditional distribution of the cost C given Y be distributed with CDF $\mathbb{P}(C \leq c|Y = y) = G_y(c)$. We focus our analysis on a simple randomized response noise model that

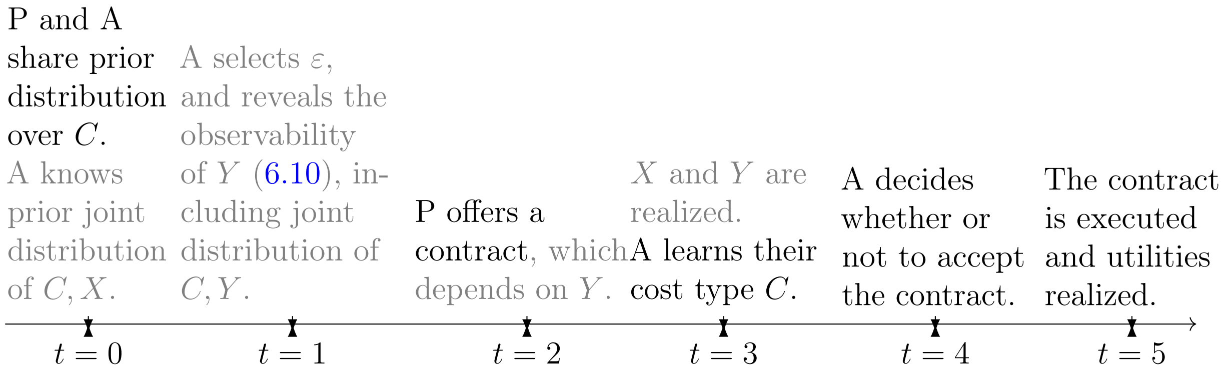

P and A share prior distribution over C.
A knows prior joint distribution of C, X.
A selects ε, and reveals the observability of Y (6.10), including joint distribution of C, Y.
P offers a contract, which depends on Y.
X and Y are realized.
A learns their cost type C.
A decides whether or not to accept the contract.
The contract is executed and utilities realized.
$t = 0$
$t = 1$
$t = 2$
$t = 3$
$t = 4$
$t = 5$

Figure 6.3: Timing of the agency game with *garbled* information transfer between principal (P) and agent (A).

does not change the marginal distribution over X, setting $\gamma = \theta = \frac{1}{2}$. In this case, we have

$$G_0(c) = \frac{1+\varepsilon}{2} F_0(c) + \frac{1-\varepsilon}{2} F_1(c), \quad \text{and} \quad G_1(c) = \frac{1+\varepsilon}{2} F_1(c) + \frac{1-\varepsilon}{2} F_0(c).$$

Similar results can be derived for general γ, θ, and there is further room to explore the interaction between these parameters in future work.

As in Section 6.3, the utility of the principal under revelation of a given garbled variable Y defined by equation (6.10) is given by

$$\Pi_{\text{garb}}(p_0, p_1) = (1-\theta) E[(b - p_0)\,\mathbb{1}(C < p_0) | Y = 0] + \theta E[(b - p_1)\,\mathbb{1}(C < p_1) | Y = 1],$$

and the utility of the agent is given by

$$V_{\text{garb}}(p_0, p_1) = (1-\theta) E[(p_0 - C)\,\mathbb{1}(C < p_0) | Y = 0] + \theta E[(p_1 - C)\,\mathbb{1}(C < p_1) | Y = 1].$$

The principal moves first and chooses $p_0(\varepsilon), p_1(\varepsilon) \in \arg\max_{p_0, p_1} \Pi_{\text{garb}}(p_0, p_1)$. The agent's resulting utility is $V_{\text{garb}}(p_0(\varepsilon), p_1(\varepsilon))$.

To simplify notation, we make the following abbreviations in the rest of this section: $\Pi_{\text{garb}}(\varepsilon) = \Pi_{\text{garb}}(p_0(\varepsilon), p_1(\varepsilon))$, $V_{\text{garb}}(\varepsilon) = V_{\text{garb}}(p_0(\varepsilon), p_1(\varepsilon))$.

Garbling and Prices

First, we analyze the effects of the information revelation amount ε on the equilibrium prices set by the principal. Specifically, we show that both $p_0(\varepsilon), p_1(\varepsilon)$ are monotone with respect to ε.

Lemma 20 (Monotonic price changes)**.** Suppose $\gamma = \theta = \frac{1}{2}$. Suppose the principal's utility is strictly concave as a function of price (Assumption 3). Suppose the MLRP holds (Assumption 5). Then $p_0'(\varepsilon) > 0$ and $p_1'(\varepsilon) < 0$ for all $\varepsilon \in [0, 1]$.

Lemma 20 acts as a sanity check that as the principal is aware of more information, the degree of differentiation between prices also increases. Furthermore, with less noise, the price in the higher-cost environment will only increase, and the price in the lower-cost environment will only decrease.

Agent's Garbling Incentives

For the agent, there is no clear ordering between $V_{\text{garb}}(p_0(\varepsilon), p_1(\varepsilon))$, $V_{\text{con}}(p^*)$, and $V_{\text{rev}}(p_0^*, p_1^*)$. Garbling can be better for the agent than concealment, even when hiding is preferred over revelation, $V_{\text{con}}(p^*) > V_{\text{rev}}(p_0^*, p_1^*)$. Garbling can also be better than revelation, even when revelation is preferred over hiding, $V_{\text{con}}(p^*) < V_{\text{rev}}(p_0^*, p_1^*)$. We illustrate both of these cases using exponential distributions in Section 6.4, as previously done for the full concealment and revelation comparison in Section 6.3.

To give a more general theoretical characterization of the agent's garbling incentives, we give sufficient conditions for the agent to prefer a nonzero amount of garbling over full revelation. This can be seen as a softer version of the previous analysis of when the agent prefers full concealment over full revelation.

Garbling condition under one zero-cost type

As in Section 6.3, we first analyze the restricted case where one of the agent types is anchored at zero-cost: suppose $C|X = 1$ takes value 0 with probability one. Proposition 10 gives a sufficient condition for the agent to prefer a non-zero amount of garbling over full revelation.

Proposition 10 (Sufficient garbling condition with zero-cost type)**.** Suppose $\gamma = \theta = \frac{1}{2}$. Suppose $C|X = 1$ takes value 0 with probability 1. Suppose F_0 is continuously differentiable and $f_0(c)$ is bounded. $V_{\text{garb}}(\varepsilon)$ is maximized at $\varepsilon^* < 1$ if

$$\frac{b - p_0^*}{2 - \sigma_0(p_0^*)} < g_0(p_0^*), \tag{6.11}$$

where $g_0(p) = \int_0^p (1 - F_0(c))dc$ is the restricted mean cost of task completion, and $\sigma_0(p) = \frac{F_0(p) f_0'(p)}{f_0(p)^2}$ is the curvature of the inverse quantity function.

Notably, the inequality in equation (6.11) captures distributions that are not captured by Proposition 7. Thus, comparing Proposition 10 to Proposition 7 shows that the agent might want to garble, even if they may not always want to hide. In fact, the condition in equation (6.11) is quite general, and we show in the example in Section 6.4 below that equation (6.11) applies to any log-concave Weibull distribution.

General garbling condition

Generalizing beyond the anchored setting, Proposition 11 gives a sufficient condition for the agent to prefer garbling over revelation for general F_0, F_1, which depends on similar identities. First, we generalize the restricted mean cost function $g_0(p)$ to a comparison of agent utilities.

Definition 5 (Agent utility dominance)**.** Let $\Delta(p_0, p_1) \coloneqq (V_1(p_0) - V_1(p_1)) - (V_0(p_0) - V_0(p_1))$ denote the difference in sensitivities to the price change from p_0 to p_1 in each environment.

When F_1 exhibits first order stochastic dominance over F_0, we have that $\Delta(p_0, p_1) > 0$ for $p_0 > p_1$. The greater the dominance of $V_1(p)$ over $V_0(p)$ for all p, the greater the difference Δ. Thus, we refer to $\Delta(p_0, p_1)$ as *agent utility dominance.* Using this definition, we now generalize the sufficient garbling condition from the anchored setting.

Proposition 11 (Sufficient garbling condition)**.** Suppose $\gamma = \theta = \frac{1}{2}$. Suppose F_0, F_1 are continuously differentiable. $V_{\text{garb}}(\varepsilon)$ is maximized at $\varepsilon^* < 1$ if

$$-\Pi_1'(p_0^*)\left(\frac{b - p_0^*}{2 - \sigma_0(p_0^*)}\right) - \Pi_0'(p_1^*)\left(\frac{b - p_1^*}{2 - \sigma_1(p_1^*)}\right) < \Delta(p_0^*, p_1^*), \tag{6.12}$$

where $\sigma_x(p) = \frac{F_x(p) f_x'(p)}{f_x(p)^2}$ is the curvature of the inverse quantity function.

The left hand side of the inequality in equation (6.12) is a weighted version of the difference $\frac{b-p_0^*}{2-\sigma_0(p_0^*)} - \frac{b-p_1^*}{2-\sigma_1(p_1^*)}$, which arises repeatedly in Aguirre et al. [2010]'s analysis of the effects of price discrimination on total welfare, and also previously arose in Proposition 8. In cases where $\Pi_0'(p_1^*) > -\Pi_1'(p_0^*)$, the condition in Proposition 8 (equation (6.3)) would imply the condition in Proposition 11 (equation (6.12)), since $\Delta(p_0^*, p_1^*) \geq 0$.

Example: Exponential and Weibull Distributions

We first illustrate the condition in Proposition 10 using an exponential distribution. Suppose $C|X = 1$ takes value 0 with probability one, and suppose $C|X = 0 \sim \text{Exp}(\frac{1}{\lambda_0})$, with F_0 defined as in equation (6.7). In this case, $g_0(p_0^*) = \lambda_0 F_0(p_0^*)$, and $\frac{b-p_0^*}{2-\sigma_0(p_0^*)} = \lambda_0 F_0(p_0^*) \frac{1}{2-F_0(p_0^*)}$. Therefore, the inequality in equation (6.11) holds for all $\lambda_0 > 0$.

The significance of this example is that if one agent type is anchored at 0, and the non-zero-cost environment induces an exponential distribution, the agent will *always* have an incentive to garble, regardless of the mean of the non-zero-cost distribution. Consider this in comparison to the exponential example from Section 6.3, where the condition in Proposition 7 showed that agent prefers to fully conceal X when λ_0 is small enough.

For a Weibull distribution with F_0 given by equation (6.9), we have for $k_0 \geq 1$,

$$g_0(p) = \frac{\lambda}{k_0}\left(\Gamma\left(\frac{1}{k_0}\right) - \Gamma\left(\frac{1}{k_0}, \frac{p^{k_0}}{\lambda_0^{k_0}}\right)\right).$$

If $k_0 \geq 1$, then the inequality in equation (6.11) from Proposition 10 holds for all λ_0. This encompasses all log-concave Weibull distributions. If $k_0 < 1$, then equation (6.11) does not necessarily hold, and fully flips for $k_0 < 0.5$.

Similarly to Figure 6.2, we can also consider simulations beyond the zero-cost anchored setting by considering all combinations of λ_0, λ_1 when $C|X = 0 \sim \text{Exp}(\frac{1}{\lambda_0})$ and $C|X = 1 \sim \text{Exp}(\frac{1}{\lambda_1})$, with F_x given by equation (6.8). Figure 6.4 illustrates the combinations of λ_0, λ_1 for which the agent prefers some amount of garbling over full revelation. Figure 6.4 shows that $V'_{\text{garb}}(1) < 0$ as long one of the conditional means λ_0 or λ_1 is small enough. That is, as long as one of the revealed settings has low enough cost, the agent always prefers to garble, regardless of how high the cost of the other setting goes. This contrasts the revelation example in Section 6.3, where even if λ_1 is close to 0, high enough λ_0 leads to the agent being willing to reveal.

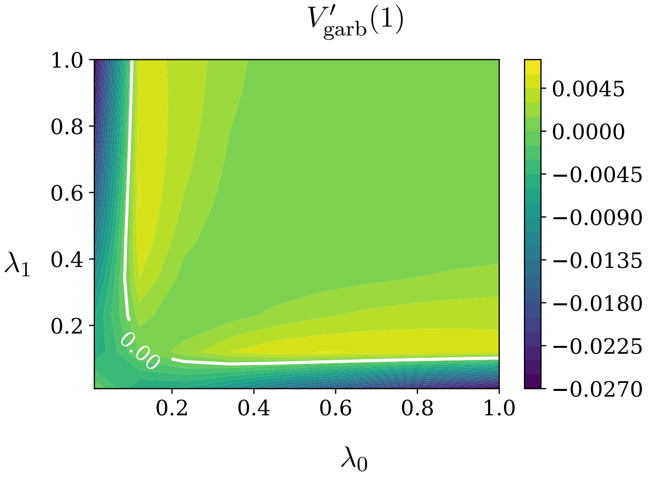


Figure 6.4: Plot of $V'_{\text{garb}}(1)$ for a mixture of exponential distributions with means λ_0, λ_1. A negative value indicates that the agent prefers some amount of garbling over full revelation. As long as one of the revealed settings has low enough average cost, the agent always prefers some amount of garbling over full revelation, regardless of how high the mean is of the other setting.

Finally, Figure 6.5 show a case in which the agent would prefer to reveal some amount of information via garbling over both concealment and revelation, but if *not* given the option to garble, then they would otherwise prefer concealment over revelation in the game from Figure 6.1.

Principal's Garbling Preferences

Similarly to Section 6.3, the principal always prefers for more information to be revealed. In fact, the garbling parameter ε directly interpolates between the concealed and revealed utilities for the principal.

Lemma 21. $\Pi_{\text{con}}(p^*) \leq \Pi_{\text{garb}}(p_0(\varepsilon), p_1(\varepsilon)) \leq \Pi_{\text{rev}}(p_0^*, p_1^*)$ for all ε.

Also similarly to the comparison of full concealment and revelation settings, revealing less noisy information yields a strict improvement in the principal's utility if the MLRP holds.

Lemma 22 (Strict principal improvement)**.** Suppose $\gamma = \theta = \frac{1}{2}$. Suppose the principal's utility is strictly concave (Assumption 3). If the MLRP holds (Assumption 5), then $\Pi'_{\text{garb}}(\varepsilon) > 0$ for all $\varepsilon \in [0, 1]$.

Garbling and Total Welfare

We now discuss the relationship between the agent's the choice of garbling amount ε and total welfare, $W_{\text{garb}}(p_0, p_1) := \Pi_{\text{garb}}(p_0, p_1) + V_{\text{garb}}(p_0, p_1)$. For notational convenience, let $W_{\text{garb}}(\varepsilon) := W_{\text{garb}}(p_0(\varepsilon), p_1(\varepsilon))$.

First, we show that relative to the fully concealed setting, the marginal effect of revealing any information on total welfare is initially positive relative to full concealment.

Lemma 23 (More information initially increases total welfare)**.** $W'_{\text{garb}}(0) \geq 0$. The inequality is strict if $\Pi'_{\text{garb}}(0) > 0$, which is true under strict concavity of $\Pi_0(p), \Pi_1(p)$ (Assumption 3) and the MLRP (Assumption 5).

Second, we also know that the *optimal* amount of garbling chosen by the agent also increases total welfare over the fully concealed setting.

Lemma 24 (Optimal garbling increases welfare over concealment)**.** Let $\varepsilon^* \in \arg\max_{\varepsilon \in [0,1]} V_{\text{garb}}(\varepsilon)$. Then $W_{\text{garb}}(\varepsilon^*) \geq W_{\text{garb}}(0)$.

Intuitively, Lemma 24 follows from the fact that the principal is never hurt by additional information. Given that the agent benefits from their optimal garbling choice, total welfare must increase.

The increase in total welfare is important from a mechanistic standpoint: consider a principal deciding whether or not to allow for agent garbling in an information transfer mechanism. There exist cases when the agent would reveal some amount of garbled information in the game, but would choose to conceal if only allowed to choose between concealment and revelation (see Figure 6.5 for an example). Thus, the garbling mechanism leads to higher total welfare in those situations.

While the agent's chosen amount of garbling ε^* improves total welfare over concealment, the question remains of where ε^* falls relative to the optimal amount of garbling that maximizes total welfare. Lemma 25 shows that the optimal amount of garbling that maximizes total welfare must necessarily reveal at least as much information as the optimal amount of garbling chosen by the agent. Under Assumptions 3 and 5, if $\varepsilon^* < 1$, then the ε that maximizes total welfare is strictly higher than the ε that maximizes agent utility.

Lemma 25 (More information increases total welfare relative to agent optimal garbling)**.** $W'_{\text{garb}}(\varepsilon) \geq V'_{\text{garb}}(\varepsilon)$ for all $\varepsilon \in [0, 1]$. The inequality is strict if $\Pi'_{\text{garb}}(0) > 0$, which is true under strict concavity of $\Pi_0(p), \Pi_1(p)$ (Assumption 3) and the MLRP (Assumption 5).

Garbling vs. Restricted Price Discrimination

Our garbling model expands the agent's action space from a binary choice between full concealment and full revelation for a given X to a continuous choice of revealing a garbled version parameterized by ε. Thus, ε interpolates between the full concealment and full revelation settings.

Garbling is not the only way to interpolate between the full concealment and full revelation settings. In the price discrimination literature, there is an established model that interpolates between full price discrimination and no price discrimination by restricting that price difference between market segments can be no greater than some parameter r. Wright [1993] models this restriction as arising from a "cost of transport" or arbitrage between two markets. Aguirre et al. [2010] apply this interpolation by analyzing the marginal effect of r on total welfare. We refer to this interpolation using r as a *restricted price discrimination* model. We also leveraged this technique to analyze the agent's utility in Section 6.3.

We now discuss in detail how the garbling model that we have introduced compares with this restricted price discrimination model. Specifically, we consider how the interpolation between concealment and revelation introduced through varying ε in our garbling model compares to interpolation using a constraint parameter r.

In fact, the trajectory of the principal and agents' utilities as r varies is different from the trajectory of the principal and agents' utilities as ε varies. Most importantly to our setting, there is a qualitative difference between the functions $V_{\text{const}}(r)$ and $V_{\text{garb}}(\varepsilon)$. Lemma 16 shows that the value of r that maximizes $V_{\text{const}}(r)$ always corresponds with either full concealment or full revelation. However, under the same conditions, the value of ε that maximizes $V_{\text{garb}}(\varepsilon)$ is *not* always at the extremes, and is often somewhere in between 0 and 1. This is significant in our setting since the agent's power to choose ε is directly built into the game, and the existence of an optimal $\varepsilon \in (0, 1)$ means that the agent benefits from the additional degree of freedom in their action space.

To visualize this difference between these interpolation methods, we can further map the combinations of principal and agent value onto the surplus triangle from Bergemann et al. [2015]. Figure 6.5 shows an example where there exists an intermediate value ε that the agent prefers over both concealment and revelation. In summary, both this example and Lemma 16 show that while there sometimes exist intermediate values ε that the agent prefers over both concealment and revelation, this is notably *not* true for intermediate restrictions r to the amount of price discrimination.

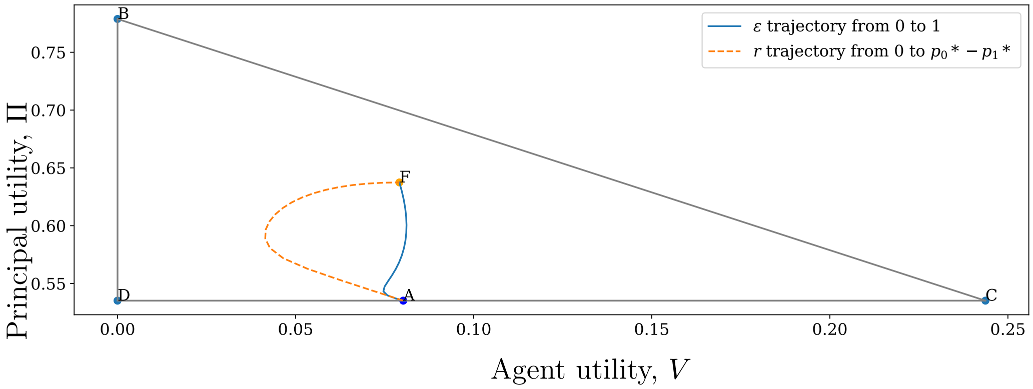


Figure 6.5: Trajectories of principal and agent utilities over ε and r, mapped onto the triangle of possible combinations of principal and agent utilities from Bergemann et al. [2015]. Here, cost is distributed as a mixture of exponentials with $C|X = 1 \sim \text{Exp}(\frac{1}{\lambda_1})$, $C|X = 0 \sim \text{Exp}(\frac{1}{\lambda_0})$, with $\lambda_0 = 0.5$, $\lambda_1 = 0.01$. The point A corresponds to the concealed setting $(V_{\text{con}}(p^*), \Pi_{\text{con}}(p^*))$, and the point F corresponds to the revealed setting $(V_{\text{rev}}(p_0^*, p_1^*), \Pi_{\text{rev}}(p_0^*, p_1^*))$. The solid blue line shows all combinations of $\Pi_{\text{garb}}(\varepsilon), V_{\text{garb}}(\varepsilon)$ for $\varepsilon \in [0, 1]$. The dashed orange line shows all combinations of $V_{\text{const}}(r), \Pi_{\text{const}}(r)$ for $r \in [0, p_0^* - p_1^*]$. First, note that the agent's utility at A is higher than at F, so the agent prefers concealment over revelation for this particular X. However, there exists a point along the ε trajectory in which $V_{\text{garb}}(\varepsilon)$ achieves higher agent utility than the point A. However, this is *not* true of the r trajectory. In general, Lemma 16 shows that intermediate values of r will always be dominated by either the fully concealed or fully revealed settings.

6.5 Conclusions and Future Work

We have presented a model in which an agent decides whether or not to reveal the observability of an environmental variable to a principal. We first considered a simple action space where the agent must choose between concealing or revealing a given variable X, and we later relaxed this to a continuous action space where the agent can choose to reveal a *garbled* version of the given environmental variable X. In both cases, we gave sufficient conditions on the conditional cost distributions for the agent to prefer revelation over concealment, concealment over revelation, and garbling over revelation. Our agency game has a price discrimination analog, which makes it possible to leverage existing results from price discrimination to analyze total welfare; however, our model also offers a qualitatively different perspective and interpolation method via garbling.

More broadly, this work was motivated by analyzing a mechanism by which one might discover *unknown unknowns*. Even as data and computational methods become increasingly sophisticated and widely available, this problem of discovery of *which* metrics or variables to analyze continues to permeate the natural sciences, social sciences, and engineering. In this work, we've explored one avenue for discovery of metrics in a setting of information asymmetry where relevant environmental variables are unknown to a principal, but known to an evaluated agent. There are other possibilities for formulating the question of *who* holds relevant information, and *when* they would be willing to share it. For example, one could model incentives for third-party individuals to offer new metrics, or perhaps bi-directional information transfer where a principal and agent both hold distinct information. The key element of our work that may be worth retaining in alternative information design frameworks is the property that the variable itself may be unknown to the information receiver.

Going beyond information design, our broader motivation is to apply these insights to help design mechanisms that would benefit all players. For example, if sharing information would increase total welfare, but decrease an agent's private utility, then perhaps a wealth transfer mechanism exists to adequately compensate the agent for revealing the information. Or, perhaps there exist auction or tournament frameworks that may further encourage information sharing. Finally, a particularly interesting avenue for future work would be to consider the possibility of cooperation and competition between multiple agents in this agency game with information transfer.

Appendix D

Deferred Proofs and Discussion for Chapter 5

D.1 Estimating the Untreated Potential Outcome

Our main paper considers incentive problems under the assumption that the principal has access to measures of patient risk through an unbiased estimator of the untreated potential outcome, $\hat{\mu}_0(x)$ with $E[\hat{\mu}_0(x)] = \mu_0(x)$.

Obtaining such an unbiased estimator can be difficult. Still, the conditions under which causal estimation can be done are well understood. In this section, we give some examples of sufficient conditions for identification of $\mu_0(x)$. This is not an exhaustive list, but rather a starting point for analysis of existing data sources.

Examples of data sources

We list two examples of data sources that may contribute to building the estimator $\hat{\mu}_0$, along with sufficient conditions for identification. These may not always be realistic, and are also not the only possible conditions for identification. In reality, practitioners may also be able to leverage knowledge of a more detailed structure causal model, or functional form assumptions (see Hernan and Robins [2020] for a more thorough coverage of methodologies and assumptions).

1. **Auxiliary untreated data.** Suppose the principal has access to an auxiliary dataset $\{X'_j, T'_j, Y'_j\}_{j=1}^m$. Then the principal may produce an unbiased estimator $\hat{\mu}_0(x)$ for $\mu_0(x)$ from this auxiliary dataset if:

   a) $\mu'_0(x) = E[Y'_j(0)|X'_j = x]$ is identifiable from this dataset. A set of sufficient conditions for this would be if ignorability was satisfied, $\{Y'_j(0), Y'_j(1)\} \perp\!\!\!\perp T'_j | X'_j$, and $P(T'_j = 0|X_j = x) > 0 \; \forall x \in \mathcal{X}$, and SUTVA was satisfied, $Y'_j = Y'_j(T'_j)$ (encompassing both consistency and non-interference).

b) The relationship between the untreated potential outcome $Y'_j(0)$ and the covariates X'_j is the same as in the treatment population between $Y_i(0)$ and X_i. That is, $E[Y'_j(0)|X'_j = x] = E[Y_i(0)|X_i = x]$.

c) The support of X'_j covers the support of X_i. Formally, the distribution of X_i is absolutely continuous with respect to the distribution of X'_j.

2. **Untreated patients in the treatment population.** In the main paper, we have assumed that the principal only observed outcomes Y_i for treated units with $T_i^\pi = 1$. If the principal also observed a fraction of outcomes for untreated units, with $T_i^\pi = 0$, then these may also be incorporated into the estimate for $\hat{\mu}_0(x)$. Sufficient conditions for identifying $\mu_0(x)$ from a dataset of the agent's untreated units are:

   a) No information asymmetry: the agent's treatment policy π only depends on X_i. Under this assumption, we have no confounding: $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i^\pi | X_i$.

   b) Positivity: $P(T_i^\pi = 0|X_i = x) > 0$ for all $x \in \mathcal{X}$, and Y_i is observed for a nonzero proportion of untreated units for all x.

   The positivity assumption is particularly tricky. In the principal-agent game, there is no guarantee that the agent's best response policy π would satisfy positivity. Positivity may be enforced by restricting the treatment rule class Π to only include treatment rules where $\pi(x) < 1$, but this may not always be possible or ethical from a policy standpoint. In practice, it may be possible to create a composite dataset that combines data from both sources. That way, data from untreated patients in the treatment population could supplement auxiliary untreated data to produce a better estimator $\hat{\mu}_0(x)$ than using auxiliary untreated data alone.

Necessity of estimating the untreated potential outcome

Here we address the necessity of including an estimate of the untreated potential outcome in the reward function w. We show that if w is completely unable to differentiate between different distributions over the untreated potential outcome, perhaps through incorporating some estimate of some function of $Y_i(0)$ or through other constraints (e.g. co-monotonicity of $\mu_1(x)$ and $\mu_0(x)$), then regret is necessarily unbounded.

Formally, let $\mathcal{D}$ denote a distribution over $X_i, Y_i(1), Y_i(0)$. Let $\mathcal{D}'$ denote an alternate distribution such that $E_{\mathcal{D}'}[Y_i(0)|X_i = x] = \mu'_0(x)$, and the joint distribution $X_i, Y_i(1)$ under $\mathcal{D}'$ remains unchanged. Let Y_i continue to denote the observed outcome under treatment assignments $Y_i = Y_i(T_i^\pi)$, and note that the distribution of Y_i changes under $\mathcal{D}'$ as well. We show that if the reward function w cannot differentiate between worlds $\mathcal{D}$ and $\mathcal{D}'$, then regret will necessarily be unbounded.

Assumption 6 (Non-degenerate w)**.** Assume that w does not always induce a degenerate best response: that is, there exists a distribution $\mathcal{D}$ such that $\arg\max_{\pi \in \Pi} E_{\mathcal{D}}[w(\mathbf{X}, \mathbf{T}^\pi, \mathbf{Y})]$

contains π^w with $E_{\mathcal{D}}[\pi^w(X_i)] > 0$ (in other words, there exists a non-measure-zero set $\overline{\mathcal{X}}$ with $\pi^w(\overline{x}) > 0 \; \forall \overline{x} \in \overline{\mathcal{X}}$).

Theorem 13. *Suppose w satisfies Assumption 6. Let $\mathcal{S}$ be the set of all pairs of distributions $\mathcal{D}, \mathcal{D}'$ such that the distributions of $X_i, Y_i(1)$ remain unchanged between $\mathcal{D}$ and $\mathcal{D}'$. If for all pairs of distributions $\mathcal{D}, \mathcal{D}' \in \mathcal{S}$, $E_{\mathcal{D}}[w(\mathbf{X}, \mathbf{T}^\pi, \mathbf{Y})] = E_{\mathcal{D}'}[w(\mathbf{X}, \mathbf{T}^\pi, \mathbf{Y})]$ for all π, then regret is unbounded.*

Proof. Let $\pi' \in \arg\max_{\pi \in \Pi} E_{\mathcal{D}'}[w(\mathbf{X}, \mathbf{T}^\pi, \mathbf{Y})]$, with $\pi'(\overline{x}) = \delta > 0$ for all $\overline{x} \in \overline{\mathcal{X}}$, with $\overline{\mathcal{X}}$ being a set of measure $\rho > 0$ (which exists by Assumption 6). Let $\mu_1(\overline{x}) = E[Y_i(1)|X = \overline{x}] = \alpha$ for all $\overline{x} \in \overline{\mathcal{X}}$. Let $\mathcal{D}$ be distribution such that $\mu_0(\overline{x}) = E_{\mathcal{D}}[Y_i(0)|X = \overline{x}] = \beta$ for all $\overline{x} \in \overline{\mathcal{X}}$. Since $E_{\mathcal{D}}[w(\mathbf{X}, \mathbf{T}^\pi, \mathbf{Y})] = E_{\mathcal{D}'}[w(\mathbf{X}, \mathbf{T}^\pi, \mathbf{Y})]$ for all π, we have $\pi' \in \arg\max_{\pi \in \Pi} E_{\mathcal{D}}[w(\mathbf{X}, \mathbf{T}^\pi, \mathbf{Y})]$ as a candidate best response to w under distribution $\mathcal{D}$.

Denote the welfare maximizing policy for distribution $\mathcal{D}$ as

$$\pi^* \in \arg\max_{\pi \in \Pi} E_{\mathcal{D}}[\pi(X_i)(\mu_1(X_i) - \mu_0(X_i))].$$

Choose $\beta > \alpha$. Then the set of welfare maximizing policies contains a π^* with $\pi^*(\overline{x}) = 0$ for all $\overline{x} \in \overline{\mathcal{X}}$.

The regret is then

$$R(\pi') = E_{\mathcal{D}}[(\pi^*(X_i) - \pi'(X_i))(\mu_1(X_i) - \mu_0(X_i))] \geq -\delta(\alpha - \beta)\rho.$$

Fixing $\delta > 0, \rho > 0, \alpha < \infty$ and choosing β to be arbitrarily high results in arbitrarily high regret. □

D.2 Proofs From Section 5.4 on Reward Function Comparisons

This section provides proofs and formal statements for the best responses and regrets corresponding to each reward function in Section 5.4.

Notation

Let $\int f(Z) dP_Z$ denote integration of the function $f(Z)$ with respect to the probability measure for the random variable Z. Let the bold variable $\mathbf{Z}$ denote the vector of i.i.d. random variables $\{Z_i\}_{i=1}^n$. Let $\mathbb{1}(\cdot)$ denote an indicator function with

$$\mathbb{1}(x \in S) = \begin{cases} 1 & \text{if } x \in S \\ 0 & \text{otherwise.} \end{cases}$$

Let π^w denote the agent's best response for the reward function w:

$$\pi^w = \arg\max_{\pi \in \Pi} E[w(\mathbf{X}, \mathbf{T}^\pi, \mathbf{Y})].$$

Additional reward function

A simple extension from w_{ATO} would be to measure the *total treated outcome (TO)*:

Reward Function 5 (TO):

$$w_{\text{TO}}(\mathbf{X}, \mathbf{T}^{\pi}, \mathbf{Y}) = \sum_{i=1}^{n} Y_i T_i^{\pi}. \tag{D.1}$$

The unconstrained best response is $\pi^{w_{\text{TO}}}(x) = \mathbb{1}(\mu_1(x) > 0)$. Like w_{ATO}, this incurs unbounded regret; however, this rule treats more people due to treating all individuals for whom the treated outcome is positive. Thus, there is no longer the "creaming" Muller [2019] issue of maximizing the *average* effect at the expense of the *total* effect.

Formal statements and proofs for agent best responses

We give formal statements and proofs for the agent best responses to the different reward functions stated in Section 5.4.

Proposition 12 (ATO Best Response)**.** Suppose Π is the set of all functions $\pi : \mathcal{X} \to [0, 1]$, and suppose X is a discrete random variable supported on $\mathcal{X}$. Then the agent's best response to w_{ATO} is:

$$\pi^{w_{\text{ATO}}}(x) = \begin{cases} 1 & \text{if } x \in \arg\max_x \mu_1(x) \text{ and } \mu_1(x) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Proof. By the tower property,

$$E[w_{\text{ATO}}(\mathbf{X}, \mathbf{T}^{\pi}, \mathbf{Y})] = E[E[Y_i|T_i^{\pi} = 1, X_i]|T_i^{\pi} = 1] = \int E[Y_i|T_i^{\pi} = 1, X_i] dP_{X_i|T_i^{\pi}=1}.$$

Since T_i^{π} only depends on X_i by construction, we have the *strong ignorability* property that $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i^{\pi} | X_i$. Therefore,

$$\int E[Y_i|T_i^{\pi} = 1, X_i] dP_{X_i|T_i^{\pi}=1} = \int \mu_1(X_i) dP_{X_i|T_i^{\pi}=1}.$$

For discrete X_i, this is equal to

$$\sum_{x \in \mathcal{X}} \mu_1(x) P(X_i = x | T_i^{\pi} = 1) = \sum_{x \in \mathcal{X}} \mu_1(x) \frac{\pi(x) P(X_i = x)}{\sum_{z \in \mathcal{X}} \pi(z) P(X_i = z)}.$$

The π function that maximizes this is exactly that given in Proposition 12. □

As a tool to prove further results, we give the following lemma for the optimal treatment rule.

Lemma 26 (Optimal Treatment Rule)**.** If the agent's treatment rule π depends only on X_i, then the treatment rule π^* that maximizes $V(\pi)$ is

$$\pi^*(x) = \begin{cases} 1 & \text{if } \tau(x) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Proof.

$$\begin{aligned} V(\pi) &= E[Y_i(T_i^\pi) - Y_i(0)] \\ &= E[Y_i(T_i^\pi) - Y_i(0) | T_i^\pi = 1] P(T_i^\pi = 1) + E[Y_i(T_i^\pi) - Y_i(0) | T_i^\pi = 0] P(T_i^\pi = 0) \\ &= E[Y_i(1) - Y_i(0) | T_i^\pi = 1] P(T_i^\pi = 1). \end{aligned}$$

By the tower property,

$$E[Y_i(1) - Y_i(0) | T_i^\pi = 1] = E[E[Y_i(1) - Y_i(0) | T_i^\pi = 1, X_i] | T_i^\pi = 1].$$

Since we have ignorability in X_i,

$$E[Y_i(1) - Y_i(0) | T_i^\pi = 1, X_i] = E[Y_i(1) - Y_i(0) | X_i] = \tau(X_i).$$

Thus,

$$V(\pi) = E[\tau(X_i) | T_i^\pi = 1] P(T_i^\pi = 1) = \int \tau(X_i) P(T_i^\pi = 1) dP_{X_i | T_i^\pi = 1}.$$

By Bayes' theorem,

$$\int \tau(X_i) P(T_i^\pi = 1) dP_{X_i | T_i^\pi = 1} = \int \tau(X_i) P(T_i^\pi = 1 | X_i) dP_{X_i} = \int \tau(X_i) \pi(X_i) dP_{X_i}.$$

The π function that maximizes this is exactly as given in Lemma 26. □

Proposition 13 (ATT Best Response)**.** Suppose Π is the set of all functions $\pi : \mathcal{X} \to [0, 1]$, and suppose X is a discrete random variable supported on $\mathcal{X}$. Then the agent's best response to w_{ATT} is:

$$\pi^{w_{\text{ATT}}}(x) = \begin{cases} 1 & \text{if } x \in \arg\max_x \tau(x) \text{ and } \tau(x) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Proof.

$$\begin{aligned} E[w_{\text{ATT}}(\mathbf{X}, \mathbf{T}^\pi, \mathbf{Y})] &= E[Y_i - \hat{\mu}_0(X_i) | T_i^\pi = 1] \\ &= E[Y_i | T_i^\pi = 1] - E[E[\hat{\mu}_0(X_i) | T_i^\pi = 1, X_i] | T_i^\pi = 1]. \end{aligned}$$

Since $E[\hat{\mu}_0(x)] = \mu_0(x)$ for all $x \in \mathcal{X}$, we have $E[\hat{\mu}_0(X_i) | T_i^\pi = 1, X_i] = \mu_0(X_i)$. Thus,

$$E[w_{\text{ATT}}(\mathbf{X}, \mathbf{T}^\pi, \mathbf{Y})] = E[Y_i | T_i^\pi = 1] - E[\mu_0(X_i) | T_i^\pi = 1].$$

Since we have ignorability in X_i,

$$E[Y_i|T_i^\pi = 1] = E[E[Y_i|T_i^\pi = 1, X_i]|T_i^\pi = 1] = E[E[Y_i(1)|X_i]|T_i^\pi = 1] = E[\mu_1(X_i)|T_i^\pi = 1].$$

Combining these,

$$E[w_{\text{ATT}}(\mathbf{X}, \mathbf{T}^\pi, \mathbf{Y})] = E[\mu_1(X_i) - \mu_0(X_i)|T_i^\pi = 1] = E[\tau(X_i)|T_i^\pi = 1].$$

For discrete X_i, this is equal to

$$\sum_{x\in\mathcal{X}} \tau(x)P(X_i = x|T_i^\pi = 1) = \sum_{x\in\mathcal{X}} \tau(x)\frac{\pi(x)P(X_i = x)}{\sum_{z\in\mathcal{X}} \pi(z)P(X_i = z)}.$$

The π function that maximizes this is exactly that given in Proposition 13. □

Proposition 14 (TO Best Response)**.** Suppose Π is the set of all functions $\pi : \mathcal{X} \to [0, 1]$. Then the agent's best response to w_{TO} is:

$$\pi^{w_{\text{TO}}}(x) = \begin{cases} 1 & \text{if } \mu_1(x) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Proof.

$$E[w_{\text{TO}}(\mathbf{X}, \mathbf{T}^\pi, \mathbf{Y})] = nE[Y_i|T_i^\pi = 1]P(T_i^\pi = 1).$$

Since we have ignorability in X_i, we have $E[Y_i|T_i^\pi = 1] = E[\mu_1(X_i)|T_i^\pi = 1]$ (shown in more detail in the above proofs). Thus,

$$E[w_{\text{TO}}(\mathbf{X}, \mathbf{T}^\pi, \mathbf{Y})] = nE[\mu_1(X_i)|T_i^\pi = 1]P(T_i^\pi = 1) = n\int \mu_1(X_i)P(T_i^\pi = 1)dP_{X_i|T_i^\pi=1}.$$

By Bayes' theorem,

$$n\int \mu_1(X_i)P(T_i^\pi = 1)dP_{X_i|T_i^\pi=1} = n\int \mu_1(X_i)P(T_i^\pi = 1|X_i)dP_{X_i} = n\int \mu_1(X_i)\pi(X_i)dP_{X_i}.$$

The π function that maximizes this is exactly that given in Proposition 14. □

Proposition 15 (TT Best Response)**.** Suppose Π is the set of all functions $\pi : \mathcal{X} \to [0, 1]$. Then the agent's best response to w_{TT} is:

$$\pi^{w_{\text{TT}}}(x) = \begin{cases} 1 & \text{if } \tau(x) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Proof.

$$E[w_{\text{TT}}(\mathbf{X}, \mathbf{T}^\pi, \mathbf{Y})] = nE[Y_i - \hat{\mu}_0(X_i)|T_i^\pi = 1]P(T_i^\pi = 1).$$

Since we have ignorability in X_i, we have $E[Y_i|T_i^\pi = 1] = E[\mu_1(X_i)|T_i^\pi = 1]$. (shown in more detail in the above proofs). Since $E[\hat{\mu}_0(x)] = \mu_0(x)$ for all $x \in \mathcal{X}$, we have $E[\hat{\mu}_0(X_i)|T_i^\pi = 1] = E[\mu_0(X_i)|T_i^\pi = 1]$. Thus,

$$E[w_{\text{TT}}(\mathbf{X}, \mathbf{T}^\pi, \mathbf{Y})] = nE[\tau(X_i)|T_i^\pi = 1]P(T_i^\pi = 1) = nV(\pi).$$

Applying Lemma 26, the π function that maximizes this is exactly that given in Proposition 15. □

Regret proofs

We give proofs for the regret bounds stated for each reward function in Section 5.4.

Proposition 2 (ATO Regret)**.** *If the conditional mean untreated potential outcomes $\mu_0(x)$ are unbounded, then the regret for the reward function w_{ATO} may be arbitrarily large.*

Proof. We construct an family of distributions for which the regret is unbounded. Let $\mathcal{X} = \{0, 1\}$ with $P(X_i = 1) = p$, and let $\mu_1(0) = 0$, $\mu_1(1) = 1$. Let $\mu_0(0) = \alpha$, and $\mu_0(1) = \beta$. Suppose Π is the set of all functions $\pi : \mathcal{X} \to [0, 1]$.

For w_{ATO}, the agent's best response is $\pi^{w_{\text{ATO}}}(0) = 0$ and $\pi^{w_{\text{ATO}}}(1) = 1$.

We illustrate two failure modes for the reward function w_{ATO}. First, suppose $\alpha < 0$, and $\beta = 0$. Then the regret is given by

$$R(\pi^w) = \max_{\pi\in\Pi} V(\pi) - V(\pi^w) = -\alpha(1-p) + p - p = -\alpha(1-p).$$

This is unbounded for unbounded α. Intuitively, this illustrates an example where the agent ignores the higher treatment effect of the patients with a lower treated outcome, such as sicker patients with higher mortality probability but more benefit from surgery. Since the agent's best response does not account for the patient's untreated potential outcome, the agent thus ignores the fact that sicker patients would otherwise have very poor outcomes without treatment. This matches the findings from Dranove et al. [2003].

A second failure mode would be if $\alpha = -1$, and $\beta > 1$. Then the regret is given by

$$R(\pi^w) = (1-p) - p(\beta - 1) = 1 - p\beta.$$

This is also unbounded for unbounded β. This illustrates an example where the agent treats the patients with a higher treated outcome, even though treatment actually harms those patients, such as healthier patients who might incur more risks or side effects from treatment. □

Proposition 3 (ATT Regret)**.** *Suppose Π is the set of all functions $\pi : \mathcal{X} \to [0, 1]$. Then the regret for the reward function w_{ATT} is upper bounded by*

$$R(\pi^{w_{ATT}}) \leq \max_{\pi\in\Pi} V(\pi).$$

Proof. Let $\pi^* = \arg\max_{\pi\in\Pi} V(\pi)$. Then

$$\pi^*(x) = \begin{cases} 1 & \text{if } \tau(x) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

$$\begin{aligned} R(\pi^{w_{\text{ATT}}}) &= V(\pi^*) - V(\pi^{w_{\text{ATT}}}) \\ &= E[\pi^*(X_i)\tau(X_i)] - E[\pi^{w_{\text{ATT}}}(X_i)\tau(X_i)] \end{aligned}$$

$$= E[\tau(X_i)\,\mathbb{1}(\tau(X_i) > 0)] - E[\tau(X_i)\,\mathbb{1}(\tau(X_i) > 0 \cap X_i \in \arg\max_{x\in\mathcal{X}} \tau(x))]$$
$$= E[\tau(X_i)\,\mathbb{1}(\tau(X_i) > 0 \cap X_i \notin \arg\max_{x\in\mathcal{X}} \tau(x))]$$

□

Proposition 4 (TT Regret)**.** *If $E[\hat{\mu}_0(X)] = \mu_0(X)$, then the regret from applying the reward function w_{TT} is $R(\pi^{w_{TT}}) = 0$.*

Proof. As shown in Proposition 15,

$$E[w_{\mathrm{TT}}(\mathbf{X}, \mathbf{T}^{\pi}, \mathbf{Y})] = nE[\tau(X_i)|T_i^{\pi} = 1]P(T_i^{\pi} = 1) = nV(\pi).$$

Therefore, the agent's best response $\pi^{w_{\mathrm{TT}}}$ also maximizes $V(\pi)$ over the same feasible set, Π. □

D.3 Proofs From Section 5.5 on Ranking

We give expanded notation and proofs for the results in Section 5.5 on modifications of w_{TT} to satisfy ranking desiderata.

Detailed notation for ranking with multiple agents

We expand the notation for ranking with multiple agents, as applied in Section 5.5. Suppose there are K agents, where each agent k observes its own sample of n_k patients with covariates $\mathbf{X}^{(k)} = \{X_i^{(k)}\}_{i=1}^{n_k}$ drawn i.i.d. from distribution $P_{X^{(k)}}$ with support $\mathcal{X}$. Let $Y_i^{(k)}(t)$ denote the potential outcomes when agent k treats the patient with treatment t. Let $\mu_t^{(k)}(x) = E[Y_i^{(k)}(t)|X_i^{(k)} = x]$, and $\tau^{(k)}(x) = E[Y_i^{(k)}(1) - Y_i^{(k)}(0)|X_i^{(k)} = x]$.

For rankings to be meaningful, we assume that if the same patient with covariate value $x \in \mathcal{X}$ were to be treated by either agent j or agent k, their potential outcomes would follow each agent's respective conditional potential outcome distributions for covariate value x. Furthermore, the conditional potential untreated outcome has the same distribution for all k: for each $x \in \mathcal{X}$, the distributions $P_{Y^{(k)}(0)|X^{(k)}=x}$ are identical for all $k \in \{1, ..., K\}$. Let $\mu_0(x)$ denote the shared mean conditional untreated potential outcome, with $\mu_0^{(k)}(x) = \mu_0(x)$ for all k. In short, one provider not treating a patient is equivalent to another provider not treating the same patient.

Suppose agent k chooses treatment policy π_k, thus producing realized treatments denoted $\mathbf{T}^{\pi_k} = \{T_i^{\pi_k}\}_{i=1}^{n_k}$ and outcomes $Y_i^{(k)} = Y_i^{(k)}(T_i^{\pi_k})$, $\mathbf{Y}^{(k)} = \{Y_i^{(k)}\}_{i=1}^{n_k}$.

Suppose the reward function w is used to rank these K agents in the following way: the principal publishes score functions $w_k : \mathcal{X}^{n_k} \times \{0,1\}^{n_k} \times \mathbb{R}^{n_k} \to \mathbb{R}$, and each agent k gets score $w_k(\mathbf{X}^{(k)}, \mathbf{T}^{\pi_k}, \mathbf{Y}^{(k)})$ after choosing their treatment policy π_k. The agents are then ranked from highest to lowest score function values.

We assume that each agent seeks to maximize their individual ranking, and make the simplifying assumption that the agents act independently: that is, each agent does not consider the potential actions of other agents when choosing their actions. This may be realistic in a setting with a large number of hospitals serving more-or-less independent populations, though more complex competitive multi-agent models may make for interesting future extensions.

Proofs

We give proofs for the proposition and theorems from Section 5.5.

Proposition 5. *If*

$$w_k(\mathbf{X}^{(k)}, \mathbf{T}^{\pi_k}, \mathbf{Y}^{(k)}) = \sum_{i=1}^{n_k} (Y_i^{(k)} - \hat{\mu}_0(X_i^{(k)}))T_i^{\pi_k},$$

then both ranking properties in Definitions 3 and 4 will be violated.

Proof. Suppose for agents j and k, Π_j and Π_k are both unconstrained. For w_k as defined above,

$$\max_{\pi_j \in \Pi_j} E[w_j(\mathbf{X}^{(j)}, \mathbf{T}^{\pi_j}, \mathbf{Y}^{(k)})] = n_j E[\tau^{(j)}(X^{(j)})\,\mathbb{1}(\tau^{(j)}(X^{(j)}) > 0)],$$

$$\max_{\pi_k \in \Pi_k} E[w_k(\mathbf{X}^{(k)}, \mathbf{T}^{\pi_k}, \mathbf{Y}^{(k)})] = n_k E[\tau^{(k)}(X^{(k)})\,\mathbb{1}(\tau^{(k)}(X^{(k)}) > 0)].$$

For Definition 3, suppose $\tau^{(j)}(x) \geq \tau^{(k)}(x)$ for all $x \in \mathcal{X}$, and $X^{(j)}$ and $X^{(k)}$ are identically distributed. Then,

$$E[\tau^{(j)}(X^{(j)})\,\mathbb{1}(\tau^{(j)}(X^{(j)}) > 0)] \geq E[\tau^{(k)}(X^{(k)})\,\mathbb{1}(\tau^{(k)}(X^{(k)}) > 0)].$$

Let n_j, n_k be a pair such that

$$n_j < n_k \frac{E[\tau^{(k)}(X^{(k)})\,\mathbb{1}(\tau^{(k)}(X^{(k)}) > 0)]}{E[\tau^{(j)}(X^{(j)})\,\mathbb{1}(\tau^{(j)}(X^{(j)}) > 0)]}.$$

This immediately results in Definition 3 being violated.

Definition 4 is also violated, since $\tau^{(j)}(x) \geq \tau^{(k)}(x)$ for all $x \in \mathcal{X}$ implies that $E[\tau^{(j)}(X_0)] \geq E[\tau^{(k)}(X_0)]$ for any reference population P_{X_0}. However, for the above n_j, n_k, we've shown that

$$\max_{\pi_j \in \Pi_j} E[w_j(\mathbf{X}^{(j)}, \mathbf{T}^{\pi_j}, \mathbf{Y}^{(k)})] < \max_{\pi_k \in \Pi_k} E[w_k(\mathbf{X}^{(k)}, \mathbf{T}^{\pi_k}, \mathbf{Y}^{(k)})].$$

□

Theorem 10 (Incentive Alignment). *Suppose $E[\hat{\mu}_0(x)] = \mu_0(x)$, and suppose Π is the set of all functions $\pi : \mathcal{X} \to [0, 1]$. For any function $g : \mathcal{X} \to \mathbb{R}^+$, w_{TT}^g yields an agent best response with zero regret.*

Proof. By the tower property,

$$\begin{aligned} E[w^g_{\mathrm{TT}}(\mathbf{Y}, \mathbf{T}^\pi, \mathbf{X})] &= nP(T_i^\pi = 1)E[E[(Y_i - \hat{\mu}_0(X_i))g(X_i)|T_i^\pi = 1, X_i]|T_i^\pi = 1] \\ &= nP(T_i^\pi = 1)E[g(X_i)E[Y_i - \hat{\mu}_0(X_i)|T_i^\pi = 1, X_i]|T_i^\pi = 1]. \end{aligned}$$

Since $E[\hat{\mu}_0(x)] = \mu_0(x)$ and we have ignorability in X_i,

$$E[Y_i - \hat{\mu}_0(X_i)|T_i^\pi = 1, X_i] = \tau(X_i).$$

This is shown in more detail in the proof for Proposition 15.

Combining this with the above, we have

$$E[w^g_{\mathrm{TT}}(\mathbf{Y}, \mathbf{T}^\pi, \mathbf{X})] = nP(T_i^\pi = 1)E[g(X_i)\tau(X_i)|T_i^\pi = 1].$$

Applying Bayes' theorem as in Lemma 26,

$$\begin{aligned} P(T_i^\pi = 1)E[g(X_i)\tau(X_i)|T_i^\pi = 1] &= \int g(X_i)\tau(X_i)P(T_i^\pi = 1)dP_{X_i|T_i^\pi=1} \\ &= \int g(X_i)\tau(X_i)P(T_i^\pi = 1|X_i)dP_{X_i} \\ &= \int g(X_i)\tau(X_i)\pi(X_i)dP_{X_i}. \end{aligned}$$

Therefore,

$$E[w^g_{\mathrm{TT}}(\mathbf{Y}, \mathbf{T}^\pi, \mathbf{X})] = nE[g(X_i)\tau(X_i)\pi(X_i)].$$

Since $g(X_i) > 0$, the treatment rule π that maximizes this is the same as π^* from Lemma 26. Therefore, the regret is zero. □

Theorem 11 (Ranking Desiderata Satisfied)**.** *Let $P_{X^{(k)}}$ be absolutely continuous with respect to P_{X_0}, and let $g_k = \frac{1}{n_k}\frac{dP_{X_0}}{dP_{X^{(k)}}}$ be the normalized Radon–Nikodym derivative of the reference distribution P_{X_0} with respect to agent k's covariate distribution $P_{X^{(k)}}$. Then*

$$w_k(\mathbf{X}^{(k)}, \mathbf{T}^{\pi_k}, \mathbf{Y}^{(k)}) = \sum_{i=1}^{n_k}(Y_i^{(k)} - \hat{\mu}_0(X_i^{(k)}))T_i^{\pi_k}g_k(X^{(k)}) \tag{D.2}$$

satisfies both ranking properties in Definitions 3 and 4 as long as Π_k is unconstrained and treatment effects are nonnegative, $\tau^{(k)}(x) \geq 0$, for all $k \in \{1, ..., K\}$.

Proof. As shown in the proof of Theorem 10,

$$E[w_k(\mathbf{X}^{(k)}, \mathbf{T}^{\pi_k}, \mathbf{Y}^{(k)})] = n_kE[\tau^{(k)}(X_i^{(k)})\pi_k(X_i^{(k)})g_k(X_i^{(k)})].$$

With $g_k = \frac{1}{n_k}\frac{dP_{X_0}}{dP_{X^{(k)}}}$, we have

$$n_kE[\tau^{(k)}(X_i^{(k)})\pi_k(X_i^{(k)})g_k(X_i^{(k)})] = E[\tau^{(k)}(X_0)\pi_k(X_0)].$$

Let Π_j, Π_k both be unconstrained. Then

$$\max_{\pi_j \in \Pi_j} E[w_j(\mathbf{X}^{(j)}, \mathbf{T}^{\pi_j}, \mathbf{Y}^{(k)})] = E[\tau^{(j)}(X_0)\,\mathbb{1}(\tau^{(j)}(X_0) > 0)],$$

$$\max_{\pi_k \in \Pi_k} E[w_k(\mathbf{X}^{(k)}, \mathbf{T}^{\pi_k}, \mathbf{Y}^{(k)})] = E[\tau^{(k)}(X_0)\,\mathbb{1}(\tau^{(k)}(X_0) > 0)].$$

Definition 3 is immediately satisfied, since $\tau^{(j)}(x) \geq \tau^{(k)}(x)$ for all $x \in \mathcal{X}$ implies $E[\tau^{(k)}(X_0)\,\mathbb{1}(\tau^{(k)}(X_0) > 0)] \geq E[\tau^{(k)}(X_0)\,\mathbb{1}(\tau^{(k)}(X_0) > 0)]$.

If $\tau^{(k)}(x) \geq 0$ for all $x \in \mathcal{X}$ and all $k \in \{1, ..., K\}$, then Definition 4 is satisfied, since

$$E[\tau^{(j)}(X_0)] \geq E[\tau^{(k)}(X_0)] \implies E[\tau^{(j)}(X_0)\,\mathbb{1}(\tau^{(j)}(X_0) > 0)] \geq E[\tau^{(k)}(X_0)\,\mathbb{1}(\tau^{(k)}(X_0) > 0)].$$

□

D.4 Proofs and Additional Results for Section 5.6 on Information Asymmetry

In this section, we give proofs and additional regret bound results for Section 5.6 on information asymmetry. First, we prove the regret bounds in Section 5.6 when $\hat{\mu}_0(x)$ is estimated from auxiliary data where the mean untreated potential outcome conditioned on X, $\mu_0(x)$, is identifiable.

Second, we prove similar results when $\hat{\mu}_0(x)$ is estimated from the untreated units in the treatment population, where $T^\pi = 0$. Under information asymmetry, the mean untreated potential outcome conditioned on X, $\mu_0(X)$, is no longer identifiable. Therefore, $\hat{\mu}_0(x)$ will be subject to confounding bias. Still, if we apply a similar but stronger assumption than Assumption 2, we can get similar regret bounds to Theorem 12.

Detailed notation for information asymmetry

We model information asymmetry in our principal agent game as follows: suppose the agent observes additional covariates per patient $U_i \in \mathcal{U}$, and selects a treatment rule $\pi : \mathcal{X}, \mathcal{U} \to [0, 1]$ from a feasible set of treatment rules Π, with $\pi(X, U) = P(T_i^\pi = 1 | X, U)$. Suppose the principal still observes only $\{X_i, T_i^\pi, Y_i\}_{i=1}^n$, and chooses a reward function $w : \mathcal{X}^n \times \{0, 1\}^n \times \mathbb{R}^n \to \mathbb{R}$ with which to reward the agent. Notably, the principal's reward function w cannot depend on U. Let $\mu_t(X, U) = E[Y_i(t)|X, U]$ and $\tau(X, U) = E[Y_i(1) - Y_i(0)|X, U]$.

The utility is still defined as in Section 5.3, and with the additional U variable can be rewritten as

$$V(\pi) = E[\pi(X_i, U_i)\tau(X_i, U_i)].$$

The regret is also still defined as in Section 5.3.

Proofs for regret bounds with unbiased counterfactual estimate

Suppose the principal estimates the mean conditional untreated potential outcome from auxiliary data $\{X'_j, T'_j, Y'_j\}_{j=1}^m$ drawn i.i.d. from auxiliary dataset $\mathcal{Q}$, denoted $\hat{\mu}_0^{\mathcal{Q}}(x)$. Suppose we have a "best case scenario" where the relationship between X'_j and $Y'_j(0)$ is the same in the auxiliary data as the relationship between X_i and $Y_i(0)$ in the treatment population, and the mean untreated potential outcome conditional on X'_j is identifiable from $\mathcal{Q}$, such that

$$E[\hat{\mu}_0^{\mathcal{Q}}(x)] = E[Y'_j(0)|X'_j = x] = E[Y_i(0)|X_i = x].$$

Outside of this ideal setting, any distribution shift or problems with identifiability in the $\mathcal{Q}$ dataset would increase the regret. Proposition 16 below is an intermediate result that does not actually rely on identifiability of $\mu_0(x)$, and may provide a good starting point for future analyses of distribution shift or non-identifiability. To simplify the proofs below, we first give Lemma 27.

Lemma 27. Under information asymmetry,

$$E[w_{\mathrm{TT}}(\mathbf{X}, \mathbf{T}^\pi, \mathbf{Y})] = n(E[\pi(X_i, U_i)\mu_1(X_i, U_i)] - E[\pi(X_i, U_i)\hat{\mu}_0(X_i)]).$$

Proof. We have previously shown that

$$\begin{aligned} E[w_{\mathrm{TT}}(\mathbf{X}, \mathbf{T}^\pi, \mathbf{Y})] &= nP(T_i^\pi = 1)E[Y_i - \hat{\mu}_0(X_i)|T_i^\pi = 1] \\ &= n(P(T_i^\pi = 1)E[Y_i|T_i^\pi = 1] - P(T_i^\pi = 1)E[\hat{\mu}_0(X_i)|T_i^\pi = 1]). \end{aligned}$$

Considering the first term, since we have ignorability in X_i and U_i,

$$\begin{aligned} P(T_i^\pi = 1)E[Y_i|T_i^\pi = 1] &= P(T_i^\pi = 1)E[E[Y_i|X_i, U_i, T_i^\pi = 1]|T_i^\pi = 1] \\ &= P(T_i^\pi = 1)E[\mu_1(X_i, U_i)|T_i^\pi = 1] \\ &= P(T_i^\pi = 1)\int \mu_1(X_i, U_i)dP_{X_i,U_i|T_i^\pi=1} \\ &= \int \mu_1(X_i, U_i)P(T_i^\pi = 1|X_i, U_i)dP_{X_i,U_i} \\ &= E[\pi(X_i, U_i)\mu_1(X_i, U_i)]. \end{aligned}$$

For the second term, we again apply Bayes' theorem:

$$\begin{aligned} P(T_i^\pi = 1)E[\hat{\mu}_0(X_i)|T_i^\pi = 1] &= P(T_i^\pi = 1)\int \hat{\mu}_0(X_i)dP_{X_i,U_i|T_i^\pi=1} \\ &= \int \hat{\mu}_0(X_i)P(T_i^\pi = 1|X_i, U_i)dP_{X_i,U_i} \\ &= E[\pi(X_i, U_i)\hat{\mu}_0(X_i)]. \end{aligned}$$

□

Proposition 16. Suppose the principal applies the reward function w_{TT} with an estimate $\hat{\mu}_0^{\mathcal{Q}}(X)$. Then the regret is bounded by the average bias in the conditional untreated potential outcome estimate.

$$R(\pi^w) \leq 2E[|\hat{\mu}_0^{\mathcal{Q}}(X_i) - \mu_0(X_i, U_i)|]. \tag{D.3}$$

Proof. Let $\hat{V}(\pi) = \frac{1}{n}E[w_{\mathrm{TT}}(\mathbf{X}, \mathbf{T}^{\pi}, \mathbf{Y})]$. Then $\pi^{w_{\mathrm{TT}}}$ maximizes $\hat{V}(\pi)$ as well.

$$\begin{aligned}
V(\pi^*) - V(\pi^{w_{\mathrm{TT}}}) &\leq V(\pi^*) - V(\pi^{w_{\mathrm{TT}}}) + \hat{V}(\pi^{w_{\mathrm{TT}}}) - \hat{V}(\pi^*) \\
&\leq |V(\pi^*) - \hat{V}(\pi^*)| + |\hat{V}(\pi^{w_{\mathrm{TT}}}) - V(\pi^{w_{\mathrm{TT}}})| \\
&\leq 2\max_{\pi\in\Pi} |\hat{V}(\pi) - V(\pi)| \\
&= 2\max_{\pi\in\Pi} |E[\pi(X_i, U_i)(\hat{\mu}_0^{\mathcal{Q}}(X_i) - \mu_0(X_i, U_i))]|,
\end{aligned}$$

where the last line follows from Lemma 27. By Jensen's inequality,

$$\begin{aligned}
\max_{\pi\in\Pi} |E[\pi(X, U)(\hat{\mu}_0^{\mathcal{Q}}(X_i) - \mu_0(X_i, U_i))]| &\leq \max_{\pi\in\Pi} E[\pi(X_i, U_i)|\hat{\mu}_0^{\mathcal{Q}}(X_i) - \mu_0(X_i, U_i)|] \\
&\leq E[|\hat{\mu}_0^{\mathcal{Q}}(X_i) - \mu_0(X_i, U_i)|].
\end{aligned}$$

□

Theorem 12 (Regret With Information Asymmetry)**.** *If the principal applies the reward function from w_{TT} with an unbiased estimate $\hat{\mu}_0(X)$ where $E[\hat{\mu}_0(x)] = \mu_0(x)$, then under Assumption 2, the regret is upper bounded as*

$$R(\pi^{w_{TT}}) \leq 2\gamma_{marg}.$$

Proof. This follows similarly to the proof of Proposition 16:

$$\begin{aligned}
V(\pi^*) - V(\pi^{w_{\mathrm{TT}}}) &\leq V(\pi^*) - V(\pi^{w_{\mathrm{TT}}}) + \hat{V}(\pi^{w_{\mathrm{TT}}}) - \hat{V}(\pi^*) \\
&\leq |V(\pi^*) - \hat{V}(\pi^*)| + |\hat{V}(\pi^{w_{\mathrm{TT}}}) - V(\pi^{w_{\mathrm{TT}}})| \\
&\leq 2\max_{\pi\in\Pi} |\hat{V}(\pi) - V(\pi)| \\
&= 2\max_{\pi\in\Pi} |E[\pi(X_i, U_i)(\hat{\mu}_0(X_i) - \mu_0(X_i, U_i))]|
\end{aligned}$$

By the tower property,

$$\begin{aligned}
E[\pi(X, U)\hat{\mu}_0(X)] &= E[E[\hat{\mu}_0(X_i)\pi(X_i, U_i)|X_i, U_i]] \\
&= E[\pi(X_i, U_i)E[\hat{\mu}_0(X_i)|X_i, U_i]] \\
&= E[\pi(X_i, U_i)\mu_0(X_i)].
\end{aligned}$$

Therefore,

$$E[\pi(X_i, U_i)(\hat{\mu}_0(X_i) - \mu_0(X_i, U_i))] = E[\pi(X_i, U_i)(\mu_0(X_i) - \mu_0(X_i, U_i))].$$

By Jensen's inequality,

$$\begin{aligned}\max_{\pi\in\Pi} |E[\pi(X_i,U_i)(\mu_0(X_i)-\mu_0(X_i,U_i))]| &\leq \max_{\pi\in\Pi} E[\pi(X_i,U_i)|\mu_0(X_i)-\mu_0(X_i,U_i)|] \\ &\leq E[|\mu_0(X_i)-\mu_0(X_i,U_i)|].\end{aligned}$$

The regret bound then follows directly from Assumption 2 which says that

$$E[|\mu_0(X_i)-\mu_0(X_i,U_i)|] \leq \gamma_{\text{marg}}.$$

□

Proposition 6. *For all $\varepsilon > 0$, there exists a distribution of $X_i, U_i, Y_i(0), Y_i(1)$ wherein $R(\pi^{w_{TT}}) \geq \gamma_{marg} - \varepsilon$.*

Proof. We construct a family of distributions of $X_i, U_i, Y_i(0), Y_i(1)$ that achieves this regret bound. Let $U \in \{0,1\}$, with $P(U=1)=\frac{1}{2}$. Suppose X is entirely uncorrelated with $Y_i(0), Y_i(1)$, such that $E[Y(t)|X] = E[Y(t)] = \mu_t$. For $\alpha > 0, \beta > 0$, let

$$\mu_1(x,u) = \begin{cases} 0 & \text{if } u=1 \\ \beta & \text{if } u=0 \end{cases}, \quad \mu_0(x,u) = \begin{cases} -\alpha & \text{if } u=1 \\ \alpha & \text{if } u=0 \end{cases},$$

which also means that $\mu_0(x) = \mu_0 = 0$ for all x. Suppose Π includes all functions $\pi : \mathcal{X}, \mathcal{U} \to [0,1]$. Then by Lemma 27, and assuming $\hat{\mu}_0(x)$ is unbiased, we have

$$E[w_{\text{TT}}(\mathbf{X}, \mathbf{T}^\pi, \mathbf{Y})] = n(E[\pi(X_i,U_i)(\mu_1(X_i,U_i)-\mu_0(X_i))].$$

The resulting policy $\pi^{w_{\text{TT}}}$ that maximizes this is

$$\pi^{w_{\text{TT}}}(x,u) = \begin{cases} 0 & \text{if } u=1 \\ 1 & \text{if } u=0 \end{cases}.$$

Suppose $\beta < \alpha$. Then the optimal policy is

$$\pi^*(x,u) = \begin{cases} 1 & \text{if } u=1 \\ 0 & \text{if } u=0 \end{cases}.$$

with utility $V(\pi^*) = \frac{\alpha}{2}$. The regret $R(\pi^{w_{\text{TT}}})$ is then

$$R(\pi^{w_{\text{TT}}}) = \frac{\alpha}{2} - \frac{\beta-\alpha}{2} = \alpha - \frac{\beta}{2}.$$

Note that $\gamma_{\text{marg}} = \alpha$. For any ε, choosing $\beta = \varepsilon$ gets $R(\pi^{w_{\text{TT}}}) = \gamma_{\text{marg}} - \frac{\varepsilon}{2}$, satisfying the bound in Proposition 6. □

Additional regret bounds when estimating the counterfactual using agent data

Suppose the principal estimates $\hat{\mu}_0(x)$ from the untreated data from the agent's treatment population, i.e. those individuals for whom $T_i^\pi = 0$. As discussed in Sections 5.6 and D.1, under full information symmetry, the mean conditional untreated potential outcome is identifiable as long as the agent's treatment rule class Π maintains positivity, $\pi(x) > 0$ for all $x \in \mathcal{X}$. Under information asymmetry, positivity is no longer sufficient for identifiability, as the agent's additional information makes T_i^π depend on U_i, and thus ignorability in X_i is no longer satisfied.

Still, we can analyze what happens when the principal constructs a counterfactual estimator from the agent's untreated outcomes. This estimator depends on the agent's treatment rule π, and unlike in Section D.1, information asymmetry means that the agent's treatment rule π affects this estimator as well. Let $\hat{\mu}_0^\pi(X) = E[Y_i|T^\pi = 0, X]$ denote the principal's estimate of the mean untreated potential outcome from the agent's data.

As with the auxiliary data, if we assume a bound on how much U_i affects the untreated potential outcome given X_i, we can still bound the regret if the principal were to apply the reward function w_{TT} using $\hat{\mu}_0^\pi(X)$. However, the required assumption is a bit stronger.

Assumption 7. The maximum effect of the unobserved attributes U on the conditional untreated potential outcome is bounded on average. Define the maximum difference in the untreated potential outcome for a given x, u to be

$$\Delta(x,u) = \max_{\tilde{u} \in \mathcal{U}} |\mu_0(x,u) - \mu_0(x,\tilde{u})|.$$

The average difference is bounded as:

$$E[\Delta(X,U)] \leq \gamma_{\max}.$$

Assumption 7 is stronger than Assumption 2 in the sense that Assumption 2 is not sufficient to bound the regret when the principal estimates the mean conditional untreated potential outcome from the agent's data using $\hat{\mu}_0^\pi(X)$. Furthermore, the bound in Assumption 7 implies that Assumption 2 is satisfied with $\gamma_{\text{marg}} \leq \gamma_{\max}$. Intuitively, Assumption 7 is not sufficient to bound the regret when the principal uses $\hat{\mu}_0^\pi(X)$ because the agent can choose π to allocate treatment to single values of $\tilde{u} \in \mathcal{U}$, such that $\hat{\mu}_0^\pi(x)$ ends up matching a single value $\mu_0(x, \tilde{u})$.

More generally, Assumption 7 is closer to bounds from senstivity analysis on expected outcome functions under unobserved confounding Kennedy [2022]. While many existing sensitivity analyses bound the effect of unobserved confounding on treatment in prior data Yadlowsky et al. [2022], in this case the agent's simultaneous treatment selection with data collection makes it less reasonable to bound the dependence of the treatment on U.

Theorem 14. *Suppose for all $\pi \in \Pi$, $\pi(x,u) > 0$ for all $x \in \mathcal{X}$, $u \in \mathcal{U}$. If the principal applies the reward function w_{TT} using an estimate $\hat{\mu}_0^\pi(x)$, then under Assumption 7, the regret is upper bounded as*

$$R(\pi^w) \leq 2\gamma_{max}.$$

Proof. This also follows similarly to the proof of Proposition 16:

$$\begin{aligned}
V(\pi^*) - V(\pi^{w_{\mathrm{TT}}}) &\leq V(\pi^*) - V(\pi^{w_{\mathrm{TT}}}) + \hat{V}(\pi^{w_{\mathrm{TT}}}) - \hat{V}(\pi^*) \\
&\leq |V(\pi^*) - \hat{V}(\pi^*)| + |\hat{V}(\pi^{w_{\mathrm{TT}}}) - V(\pi^{w_{\mathrm{TT}}})| \\
&\leq 2\max_{\pi\in\Pi} |\hat{V}(\pi) - V(\pi)| \\
&= 2\max_{\pi\in\Pi} |E[\pi(X_i,U_i)(\hat{\mu}_0^\pi(X_i) - \mu_0(X_i,U_i))]|
\end{aligned}$$

By Jensen's inequality,

$$\begin{aligned}
\max_{\pi\in\Pi} |E[\pi(X,U)(\hat{\mu}_0^\pi(X) - \mu_0(X,U))]| &\leq \max_{\pi\in\Pi} E[\pi(X,U)|\hat{\mu}_0^\pi(X) - \mu_0(X,U)|] \\
&\leq \max_{\pi\in\Pi} E[|\hat{\mu}_0^\pi(X) - \mu_0(X,U)|]
\end{aligned}$$

Define

$$\Delta^\Pi(x,u) = \max_{\pi\in\Pi} |\hat{\mu}_0^\pi(x) - \mu_0(x,u)|.$$

Since for all $\pi \in \Pi$, $\pi(x,u) > 0$ for all $x \in \mathcal{X}$, $u \in \mathcal{U}$, we can apply Lemma 28 below, which says that $\hat{\mu}_0^\pi(x) \in [\min_u \mu_0(x,u), \max_u \mu_0(x,u)]$. Therefore, for all x, u,

$$\Delta^\Pi(x,u) \leq \Delta(x,u).$$

Putting this together,

$$\max_{\pi\in\Pi} E[|\hat{\mu}_0^\pi(X) - \mu_0(X,U)|] \leq E[\max_{\pi\in\Pi} |\hat{\mu}_0^\pi(X) - \mu_0(X,U)|] \leq E[\Delta^\Pi(X,U)] \leq E[\Delta(X,U)].$$

Therefore, under assumption 7, $R(\pi^w) \leq 2\gamma_{\max}$. □

Lemma 28. Suppose that for all $\pi \in \Pi$, $P_{U|X=x,T^\pi=0}$ is a well defined probability distribution. Let $\hat{\mu}_0^\pi(x) = E[Y|X=x, T^\pi = 0]$. Then for all π, $\hat{\mu}_0^\pi(x) \in [\min_u \mu_0(x,u), \max_u \mu_0(x,u)]$.

Proof. We decompose $\hat{\mu}_0^\pi(x)$ as

$$\hat{\mu}_0^\pi(x) = E[Y|X=x, T^\pi = 0] = \int \mu_0(x,u) dP_{U|X=x,T^\pi=0}$$

For any π, $\int \mu_0(x,u) dP_{U|X=x,T^\pi=0} = \int \mu_0(x,u) dW(u)$ for some distribution W; therefore,

$$\left\{ \int \mu_0(x,u) dP_{U|X=x,T^\pi=0} : \pi \in \Pi \right\}$$

$$\subseteq \left\{ \int \mu_0(x,u)dW(u) : W \text{ is a distribution over } U, \int dW(u) = 1 \right\}.$$

Both the maximizer and the minimizer of the smaller set are contained in the larger set. Specifically,

$$\max_\pi \hat{\mu}_0^\pi(x) = \max_\pi \int \mu_0(x,u)dP_{U|X=x,T^\pi=0} \leq \max_{W:\int dW(u)=1} \int \mu_0(x,u)dW(u) \leq \max_u \mu_0(x,u)$$

$$\min_\pi \hat{\mu}_0^\pi(x) = \min_\pi \int \mu_0(x,u)dP_{U|X=x,T^\pi=0} \geq \min_{W:\int dW(u)=1} \int \mu_0(x,u)dW(u) \geq \min_u \mu_0(x,u)$$

Therefore, for all π, $\hat{\mu}_0^\pi(x) \in [\min_u \mu_0(x,u), \max_u \mu_0(x,u)]$. □

Overall, we have discussed two plausible data collection settings for the principal to estimate $\mu_0(x)$ to implement the reward function w_{TT}. In the first setting using auxiliary data where $\mu_0(x)$ is identifiable, the regret is bounded by the gap between the conditional effects of U on $Y_i(0)$ and the marginal effect of only X on $Y_i(0)$. In the second setting using untreated units from the agent's treatment population, $\mu_0(x)$ is not identifiable, and the regret can be bounded under a stronger assumption on the sensitivity of $\mu_0(X,U)$ to the agent's private information U_i.

D.5 Additional Experiment Details and Results

We give additional experiment details and results here.

Implementation

All experiment code is available at `https://github.com/serenalwang/counterfactual_metrics`. All models were trained using the linear regression and logistic regression packages from scikit-learn [Pedregosa et al., 2011]. All categorical features were one-hot encoded.

Additional dataset details

We provide additional setup and training details for each dataset.

Horse Colic dataset

For the Horse Colic dataset from UCI Dua and Graff [2017], we let X consist of all features observed prior to surgery, which includes 13 categorical features and 7 numerical features. All features used are listed in Figure D.2. We removed all examples in which the horse was euthanized, and used the remaining "outcome" variable as Y, where we set $Y_i = 1$ if the horse lived, and $Y_i = -1$ if the horse died. We only used the main `horse-colic.data` dataset, and did not use the "test" dataset included in the UCI directory.

To simulate $\mu_t(x)$, we assume that the outcome distribution takes the parametric form,

$$P(Y(t) = 1|X = x) = \sigma(\beta_0 + \beta_1^\top x + \beta_2 t + \beta_3^\top xt)$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the standard logistic function.

We trained a logistic regression model of Y on X, T, and the interaction term XT to estimate parameters $\hat{\beta}$, and used the resulting estimate to compute $\mu_t(x) = \sigma(\hat{\beta}_0 + \hat{\beta}_1^\top x + \hat{\beta}_2 t + \hat{\beta}_3^\top xt)$. For the fitted model, the AUC was 0.9924, and the accuracy was 0.6824. When computing agent best responses and regret, we take this function $\mu_t(x)$ to be given as synthetic mean conditional potential outcomes. Note that the clinical validity of $\mu_t(x)$ as actual potential outcomes cannot be verified from the dataset alone. There may be error in both our unconfoundedness assumption with respect to X, and in the logistic parametric specification of the outcome model.

In Table 5.1, the "demographic information" for the Horse Colic dataset consists of only the *age* feature.

International Stroke Trial dataset

For the International Stroke Trial dataset International Stroke Trial Collaborative Group [1997], we let X include all clinical data prior to treatment, which includes all "Randomisation data" except for dates and times. This includes 17 categorical features and 3 numerical features. All features used are listed in Figure 5.1. For the outcome variable Y, we apply the negative of the scalarized composite outcome score from Kallus and Zhou [2018]. Specifically,

$$\begin{aligned} Y = &-2\,\mathbb{1}(\text{death}) - \mathbb{1}(\text{recurrent stroke}) - 0.5\,\mathbb{1}(\text{pulmonary embolism or intracranial bleeding}) \\ &- 0.5\,\mathbb{1}(\text{other side effects}) + 2\,\mathbb{1}(\text{full recovery at 6 months}) + \mathbb{1}(\text{discharge within 14 days}). \end{aligned}$$

This results in $Y \in [-4, 3]$.

To simulate $\mu_t(x)$, we assume the conditional mean of the potential outcome distribution has the linear parametric form,

$$E[Y(t)|X = x] = \beta_0 + \beta_1^\top x + \beta_2 t + \beta_3^\top xt.$$

We trained a linear regression model of Y on X, T, and the interaction term XT to estimate parameters $\hat{\beta}$, and used the resulting estimate to compute $\mu_t(x) = \hat{\beta}_0 + \hat{\beta}_1^\top x + \hat{\beta}_2 t + \hat{\beta}_3^\top xt$. The fitted RMSE was 1.34 and the R^2 was 0.26. As with the Horse Colic dataset, when computing agent best responses and regret, we take this function $\mu_t(x)$ to be given as synthetic mean conditional potential outcomes. In this dataset the unconfoundedness assumption should hold as this data came from a randomized control trial. However, there may still be error in the linear parametric specification of the outcome model.

In Table 5.1, the "demographic information" for the Stroke Trial dataset consists of both the *age* feature and the *sex* feature.

Effect of better estimates of $\mu_t(x)$

Better estimates of $\mu_t(x)$ may come from doubly robust estimators. However, the contribution of this work is not to improve estimation methods for $\mu_t(x)$, nor do the proposed reward policies rely on the quality or variance of $\mu_t(x)$ estimators. The quality of $\mu_t(x)$ only affects the how well the regrets reported in experiments might approximate real regrets. This is because in experiments, we estimate $\mu_t(x)$ to simulate potential outcomes for two purposes: *(i)* simulating the agent's perception of potential outcomes and thus yielding the agent's best responses; and *(ii)* calculating an approximate regret. Better estimates of $\mu_t(x)$ would improve the alignment with both of these simulations with reality.

Our theory only assumes that the principal has an estimator $\hat{\mu}_0(x)$ which is unbiased with respect to the agent's preceived $\mu_0(x)$ function which the agent uses to calculate their best response. For the experiments, this unbiasedness assumption is true, as we assume both principal and agent have access to the same synthetic $\mu_t(x)$ values. We do not directly simulate the principal's calculation of $\hat{\mu}_0(x)$ from a subset of the data, since this would not affect the actions of the expectation maximizing agent.

The main limitation of our experiments is that in the absence of true counterfactual outcomes, they rely on parametric estimates of μ_t which we can't guarantee are well specified. An ideal dataset for evaluating true welfare impacts would have a structure that identifies P_{X_i} and $\mu_t(x)$, where $\mu_t(x)$ ideally matches the agent's *perceived* mean conditional potential outcomes. While we can't guarantee that our estimated imputed values for $\mu_t(x)$ match real providers, our experiments provide a structure by which such experiments could be run in the future if regulatory agencies have internal access to more ideal data, perhaps even through surveying providers themselves.

Calculating $\gamma_{\mathbf{marg}}$

When simulating information asymmetry, we compute empirical estimates of γ_{marg} over the data. For each individual feature in Figures 5.1 and D.2, let X represent the individual feature, and let U represent the set of all other features. We use the full regression result as $\mu_0(x, u)$. To calculate $\mu_0(x)$ where x represents the individual feature, we take the empirical conditional mean over the dataset:

$$\mu_0(x) = \frac{\sum_{i=1}^{n} \mu_0(X_i, U_i)\, \mathbb{1}(X_i = x)}{\mathbb{1}(X_i = x)}.$$

Then, γ_{marg} is calculated empirically over the dataset as

$$\gamma_{\text{marg}} = \frac{1}{n} \sum_{i=1}^{n} |\mu_0(X_i) - \mu_0(X_i, U_i)|.$$

We take these empirical estimates as given in the absence of a closed form for the joint distribution of all features. More sophisticated Bayesian distribution estimation or smoothing may produce different γ_{marg} estimates.

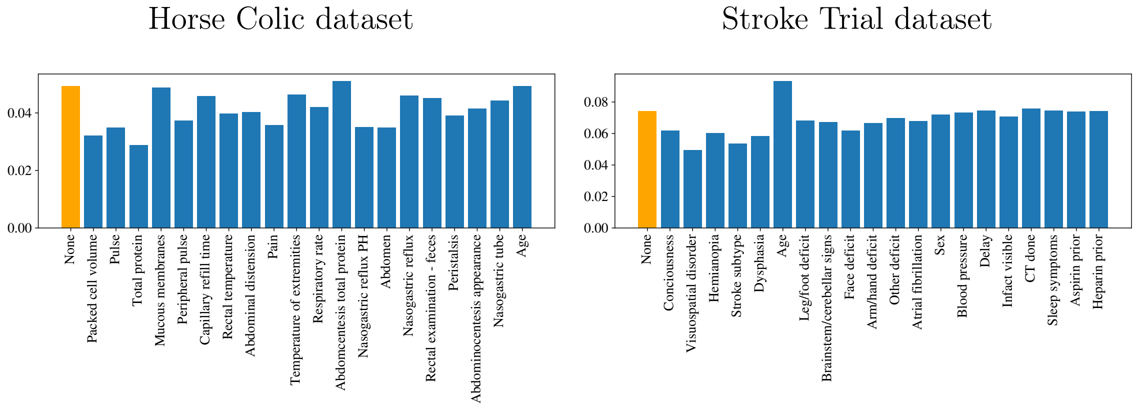
Additional results

Table D.1 shows the utility and regret comparisons for different reward functions, including an additional comparison to Reward Function 5 in the Appendix, w_{TO}.

Figure D.1 shows the regrets when the principal knows only each individual feature. Interestingly, use of several of these individual features leads to worse regret than if the principal knew no features and just estimated the marginal expected untreated potential outcome $E[Y(0)]$. This confirms that the regret amplification effect can occur for features other than the demographic information in Table 5.1.

Table D.1: Utility and regret comparisons for different reward functions. For each reward function w, we report utility $V(\pi^w)$, regret $R(\pi^w)$, and the realized treatment rate $P(T^{\pi^w} = 1)$.

	Horse Colic dataset			Stroke Trial dataset		
Reward function	Utility	Regret	Treat rate	Utility	Regret	Treat rate
w_{ATO}	0.00000	0.1470	0.1922	0.00004	0.0251	0.0001
w_{ATT}	0.00784	0.1391	0.0039	0.00013	0.0250	0.0001
w_{TO}	0.08599	0.0610	0.6706	-0.08278	0.1079	0.6689
w_{TT}	0.14695	0.000	0.2431	0.02518	0.000	0.1872
w_{TT} (no info)	0.09761	0.0493	0.6274	-0.04888	0.0741	0.4829
w_{TT} (demographic info)	0.09761	0.0493	0.6274	-0.06392	0.0891	0.5041


Horse Colic dataset

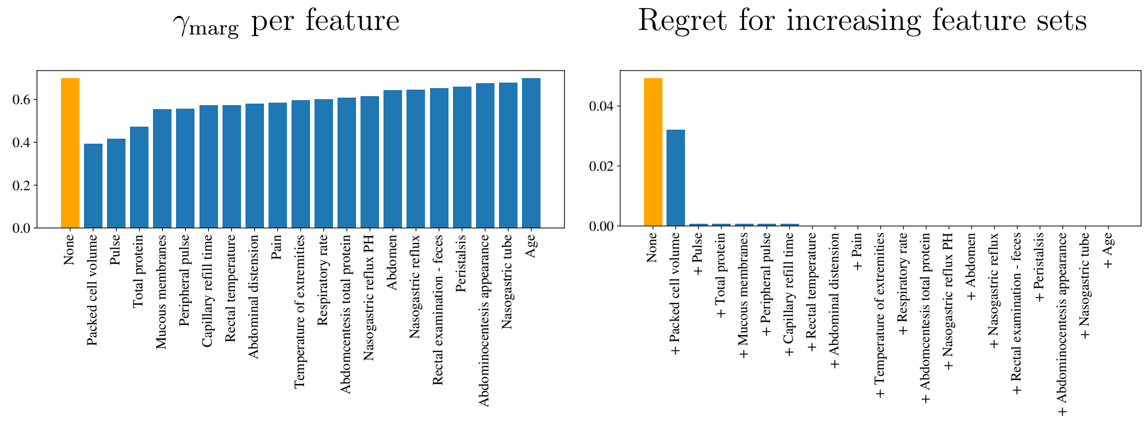
Stroke Trial dataset

Figure D.1: Regret if the principal only knows the labeled feature. Features are sorted in order of their individual γ_{marg}.

Figure D.2 is the equivalent of Figure 5.1 for the Horse Colic dataset. For the Horse Colic dataset, the regret drops to close to zero if the principal knows only a few features. This is likely due to the fact that there is little variation in other feature values in the dataset conditioned *packed cell volume* and *pulse*, exacerbated by the fact that the dataset is so small. The value of γ_{marg} for these two features combined is small compared to the individual γ_{marg} values in Figure D.2 (*left*), at 0.035.

This may change for larger datasets, where it may be more likely to observe more samples with the same *packed cell volume* and *pulse* values alongside more variation in other feature values. If in practice other features did not vary much given *packed cell volume* and *pulse*, then the regret shape in Figure D.2 (*right*) would hold. More data would be needed to verify the joint distribution of X, U to confirm this.

The Stroke Trial dataset does not exhibit a similar effect since its three numerical features take significantly fewer possible values relative to the size of the dataset.



γ_{marg} per feature

Regret for increasing feature sets

Figure D.2: Regret under fine-grained information asymmetry on the Horse Colic dataset. The *left* plot shows γ_{marg} values if the principal only knows each individual feature. The *right* plot shows regret as the principal accumulates features from the left (most important).

D.6 Additional Discussion of Modeling Assumptions

We do not explicitly model the agent's costs, and all budget constraints are contained in the agent's feasible set of treatment rules Π. This may be viewed as considering the agent's incentives when there is no cost to treatment, thus decoupling our analysis of the agent's incentives to do well on healthcare report cards from the broader treatment pricing market. In practice, the compensation hospitals and doctors receive for treatment could yield a positive adjustment or sometimes a negative adjustment to their utility per treated unit. Analysis of how other external incentives pair with the quality measure incentives would be interesting future work, but for now we focus on the incentives induced solely by the accountability metrics.

Whether the information asymmetry modeled here is common or severe in the medical space has been debated and may change over time. Dranove et al. [2003] remark that “providers may be able to improve their ranking by selecting patients on the basis of characteristics that are unobservable to the analysts but predictive of good outcomes.” However, medical treatment protocols are also generally heavily codified, with online clinical decision support tools becoming increasingly widely used 30 years of UpToDate: The evolution of clinical decision support and the future of evidence-based medicine. Still, if the regulatory agency does not have full access to patents’ medical records, then information asymmetry will arise.

Appendix E

Deferred Proofs for Chapter 6

E.1 Proofs from Section 6.3

Here we give full proofs from Section 6.3.

Proof of Proposition 7

To prove Proposition 7, we first reorganize the difference between the agent's concealed and revealed utilities:

$$V_{\text{con}}(p^*) - V_{\text{rev}}(p_1^*, p_1^*) = \theta \Delta V_1 - (1-\theta)\Delta V_0$$

where

$$\Delta V_0 \coloneqq V_0(p_0^*) - V_0(p^*); \quad \Delta V_1 \coloneqq V_1(p^*) - V_1(p_1^*).$$

We begin with Lemma 29 below which upper bounds ΔV_0.

Lemma 29. For any concave and continuously differentiable CDF F_0, $\Delta V_0 \leq p_0^* - \overline{p}$ for any $\overline{p} < p_0^*$.

Proof. $\Delta V_0 \leq p_0^* - \overline{p}$ if

$$\frac{V_0(p_0^*) - V_0(\overline{p})}{p_0^* - \overline{p}} \leq 1.$$

We upper bound this difference by differentiating V_0:

$$V_0'(p) = \frac{d}{dp}\int_0^p (p-c) f_0(c)dc = F_0(p)$$

Since F_0 is a concave and continuously differentiable CDF, by the mean value theorem,

$$\frac{V_0(p_0^*) - V_0(\overline{p})}{p_0^* - \overline{p}} \leq \sup_p V_0'(p) = \sup_p F_0(p) \leq 1.$$

□

We now leverage Lemma 29 to prove the full proposition.

Proposition 7 (Sufficient concealment condition with zero-cost type)**.** *Suppose F_0 is a concave and continuously differentiable CDF. Suppose $C|X=1$ takes value 0 with probability 1. Suppose the ratio $\frac{F_0(p)}{f_0(p)}$ is strictly monotone increasing for $p>0$. Then $V_{con}(p^*) > V_{rev}(p_0^*, p_1^*)$ if*

$$\theta > (1-\theta)\frac{1}{\eta((1-\theta)p_0^*)} - \frac{1}{\eta_0\left(p_0^*\right)}$$

where $\eta(p) = \frac{p(1-\theta)f_0(p)}{(1-\theta)F_0(p)+\theta}$ and $\eta_0(p) = \frac{pf_0(p)}{F_0(p)}$ are the respective price elasticities for task completion quantity for the mixture distribution C and the conditional distribution $C|X=0$.

Proof. We consider the extreme case where $C|X=1$ has value 0 with probability 1. For this distribution of $C|X=1$, we show that Equation (6.3) implies that

$$\theta\Delta V_1 > (1-\theta)\Delta V_0. \tag{E.1}$$

First, for any nonzero $C|X=0$, we have that $p^* < p_0^*$. Thus, Lemma 29 gives that

$$(1-\theta)\Delta V_0 \le (1-\theta)(p_0^* - p^*).$$

Next, we further upper bound this by showing that for any F_0 that satisfies Equation (6.3),

$$(1-\theta)(p_0^* - p^*) < \theta p^*. \tag{E.2}$$

Equation (E.1) then follows from the fact that $\Delta V_1 = p^*$ when $C|X=1$ is always zero.

We now prove that equation (E.2) holds under equation (6.3). First, note that

$$(1-\theta)(p_0^* - p^*) < \theta p^* \iff (1-\theta)p_0^* < p^*.$$

Since F_0 is concave and continuously differentiable, p^* satisfies the following first-order condition:

$$p^* + \frac{(1-\theta)F_0(p^*)+\theta}{(1-\theta)f_0(p^*)} = b. \tag{E.3}$$

Since $\frac{F_0(p)}{f_0(p)}$ is strictly monotone increasing for $p>0$ and $F_0(p)$ is concave, $p + \frac{(1-\theta)F_0(p)+\theta}{(1-\theta)f_0(p)}$ is also strictly monotone increasing for $p>0$. Therefore, $(1-\theta)p_0^* < p^*$ if and only if

$$(1-\theta)p_0^* + \frac{(1-\theta)F_0((1-\theta)p_0^*)+\theta}{(1-\theta)f_0((1-\theta)p_0^*)} < b.$$

We also have that p_0^* satisfies the first-order condition

$$p_0^* + \frac{F_0(p_0^*)}{f_0(p_0^*)} = b.$$

Therefore, $(1-\theta)p_0^* < p^*$ if and only if

$$(1-\theta)p_0^* + \frac{(1-\theta)F_0((1-\theta)p_0^*)+\theta}{(1-\theta)f_0((1-\theta)p_0^*)} < p_0^* + \frac{F_0(p_0^*)}{f_0(p_0^*)},$$

which is equivalent to the condition in equation (6.3).

□

Proofs for Propositions 8 and 9

Here we give proofs for Propositions 8 and 9. These analyses parallel those of Aguirre et al. [2010] for the effects of price discrimination on total welfare.

We first prove Lemma 16, which follows from Assumption 3 and 4.

Lemma 16. *Under Assumptions 3 and 4, $V_{const}(r)$ is strictly quasi-convex for $r \in [0, p_0^* - p_1^*]$. That is, if there exists $\hat{r} \in [0, p_0^* - p_1^*]$ such that $V_{const}(\hat{r}) = 0$, then $V''_{const}(\hat{r}) > 0$.*

Proof. The constraint in equation (6.4) is binding when $r \in [0, p_0^* - p_1^*]$. Therefore, the optimization problem in equation (6.4) can be rewritten as

$$\max_{p_1} \ \Pi_1(p_1) + \Pi_0(p_1 + r),$$

yielding a first-order condition that $\Pi'_1(p_1) + \Pi'_0(p_1 + r) = 0$. Further differentiating this first-order condition, as done by Aguirre et al. [2010], yields that

$$p'_1(r) = \frac{-\Pi''_0(p_0(r))}{\Pi''_0(p_0(r)) + \Pi''_1(p_1(r))}.$$

A similar method shows that

$$p'_0(r) = \frac{\Pi''_1(p_1(r))}{\Pi''_0(p_0(r)) + \Pi''_1(p_1(r))}.$$

Thus, we have that

$$\begin{aligned} V'_{\text{const}}(r) &= (1-\theta)V_0(p_0(r))p'_0(r) + \theta V_1(p_1(r))p'_1(r) \\ &= \left(\frac{-\Pi''_1(p_1(r)\Pi''_0(p_0(r)))}{\Pi''_0(p_0(r)) + \Pi''_1(p_1(r))}\right)\left(\frac{\theta V'_1(p_1(r))}{\Pi''_1(p_1(r))} - \frac{(1-\theta)V'_0(p_0(r))}{\Pi''_0(p_0(r))}\right) \\ &= \left(\frac{-\Pi''_1(p_1(r)\Pi''_0(p_0(r)))}{\Pi''_0(p_0(r)) + \Pi''_1(p_1(r))}\right)(\theta w_1(p_1(r)) - (1-\theta)w_0(p_0(r))) \end{aligned} \tag{E.4}$$

where $w_x(p) := \frac{V'_x(p)}{\Pi''_x(p)}$.

Taking the second derivative, we have that

$$\begin{aligned} V''_{\text{const}}(r) = &\left(\frac{-\Pi''_1(p_1(r)\Pi''_0(p_0(r)))}{\Pi''_0(p_0(r)) + \Pi''_1(p_1(r))}\right)(\theta w'_1(p_1(r))p'_1(r) - (1-\theta)w'_0(p_0(r))p'_0(r)) \\ &+ (\theta w_1(p_1(r)) - (1-\theta)w_0(p_0(r)))\frac{\partial}{\partial r}\left(\frac{-\Pi''_1(p_1(r)\Pi''_0(p_0(r)))}{\Pi''_0(p_0(r)) + \Pi''_1(p_1(r))}\right). \end{aligned}$$

The first term $\left(\frac{-\Pi''_1(p_1(r)\Pi''_0(p_0(r)))}{\Pi''_0(p_0(r)) + \Pi''_1(p_1(r))}\right)$ is positive by strict concavity given by Assumption 3.

If $V'_{\text{const}}(\hat{r}) = 0$, then $\theta w_1(p_1(\hat{r})) - (1-\theta)w_0(p_0(\hat{r})) = 0$. By the DRC, $w'_1(p_1(\hat{r}))p'_1(\hat{r}) > 0$ since $w'_1(p_1(\hat{r})) < 0$ and $p'_1(\hat{r}) < 0$. Similarly, $w'_0(p_0(\hat{r}))p'_0(\hat{r}) < 0$. Therefore, $V''_{\text{const}}(\hat{r}) > 0$.

□

Given Lemma 16, we now prove Propositions 8 and 9 by signing the derivative $V'_{\text{const}}(r)$ for extreme values of r.

Proposition 8 (Sufficient concealment condition under DRC)**.** *Under Assumptions 3 and 4, $V_{con}(p^*) > V_{rev}(p_0^*, p_1^*)$ if*

$$\frac{(1-\theta)(b-p_0^*)}{2-\sigma_0(p_0^*)} < \frac{\theta(b-p_1^*)}{2-\sigma_1(p_1^*)},$$

where $\sigma_x(p) = \frac{F_x(p)f'_x(p)}{f_x^2(p)}$ is the curvature of the inverse of the task completion quantity function $F_x(p)$.

Proof. If $V_{\text{const}}(r)$ is strictly monotone decreasing in r, then $V_{\text{con}}(p^*) > V_{\text{rev}}(p_0^*, p_1^*)$. Since $V_{\text{const}}(r)$ is strictly quasi-convex, a sufficient condition for $V_{\text{const}}(r)$ to be strictly monotone decreasing is $V'_{\text{const}}(p_0^* - p_1^*) < 0$.

By equation (E.4), we have that $V'_{\text{const}}(p_0^* - p_1^*) < 0$ if $\theta w_1(p_1^*) - (1-\theta)w_0(p_0^*) < 0$. By the first-order condition that

$$b - p_x^* = \frac{F_x(p_x^*)}{f_x(p_x^*)}, \tag{E.5}$$

we have that

$$w_x(p_x^*) = \frac{F_x(p_x^*)}{\Pi''_x(p_x^*)} = \frac{b-p_x^*}{\Pi''_x(p_x^*)/f_x(p_x^*)}$$

Note that

$$\Pi''_x(p) = -2f_x(p) + f'_x(p)(b-p)$$

Therefore, also applying the first-order condition from equation (E.5), we have

$$\Pi''_x(p_x^*)/f_x(p_x^*) = -2 + \frac{F_x(p)f'_x(p)}{f_x^2(p)} = -2 + \sigma_x(p).$$

□

A similar argument yields Proposition 9.

Proposition 9 (Sufficient revelation condition under DRC)**.** *Under Assumptions 3 and 4, $V_{con}(p^*) < V_{rev}(p_0^*, p_1^*)$ if*

$$\frac{2 + L(p^*)\alpha_1(p^*)}{2 + L(p^*)\alpha_0(p^*)} > \frac{\theta F_1(p^*)/f_1(p^*)}{(1-\theta)F_0(p^*)/f_0(p^*)},$$

where $L(p) = \frac{b-p}{p}$ is the Lerner index, and $\alpha_x(p) = \frac{-pf'_x(p)}{f_x(p)}$ is the curvature of the task completion quantity function $F_x(p)$.

Proof. If $V_{\text{const}}(r)$ is strictly monotone increasing in r, then $V_{\text{con}}(p^*) < V_{\text{rev}}(p_0^*, p_1^*)$. Since $V_{\text{const}}(r)$ is strictly quasi-convex, a sufficient condition for $V_{\text{const}}(r)$ to be strictly monotone increasing is $V'_{\text{const}}(0) > 0$.

By equation (E.4), we have that $V'_{\text{const}}(0) > 0$ if $\theta w_1(p^*) - (1-\theta)w_0(p^*) > 0$.

$$w_x(p^*) = \frac{F_x(p^*)/f_x(p^*)}{-2 + (b - p^*)(f'_x(p^*)/f_x(p^*))} = \frac{F_x(p^*)/f_x(p^*)}{-2 - L(p^*)\alpha_x(p^*)}$$

Therefore, $\theta w_1(p^*) - (1-\theta)w_0(p^*) > 0$ if

$$\frac{\theta F_1(p^*)/f_1(p^*)}{2 + L(p^*)\alpha_1(p^*)} < \frac{(1-\theta)F_0(p^*)/f_0(p^*)}{2 + L(p^*)\alpha_0(p^*)}$$

□

Comparison of decreasing ratio condition to increasing ratio condition

The results in Section 6.3 depend on the decreasing ratio condition (DRC) given in Assumption 4. This is analogous to the "increasing ratio condition (IRC)" from Aguirre et al. [2010], which says that the ratio $\frac{W'_x(p)}{\Pi''_x(p)}$ is increasing in p. We now discuss in more detail the relationship between the DRC and the IRC, including sufficient conditions under which the DRC holds. In introducing the IRC, Aguirre et al. [2010] describe a "very large set of demand functions" for which the IRC holds. Aguirre et al. [2010] give sufficient conditions for the IRC to hold in Appendix B from their paper, which includes linear functions and exponential and constant elasticity functions.

For all of the sufficient conditions that Aguirre et al. [2010] proposes for the IRC, these are also sufficient conditions for the DRC if paired with the additional condition that $\frac{F_x(p)}{f_x(p)}$ is increasing in p for $x \in \{0,1\}$. For example, a specific sufficient condition for the DRC, which also implies the IRC, is the following: Let $\sigma(p) = \frac{F(p)f'(p)}{f(p)^2}$. If $\sigma(p) \leq 1$, and $\alpha(p) = -\frac{pf'(p)}{f(p)}$ is non-decreasing and positive in p, then the DRC holds. The IRC would also hold. A similar analogy can be made for all other conditions given in Appendix B of Aguirre et al. [2010].

Proofs from Section 6.3

Lemmas 17 and 18 show that the principal always benefits from more information being revealed. Assumption 5 further implies that the principal strictly benefits from revelation.

Lemma 17 (Principal prefers revelation)**.** *Revealing X never decreases the value of the principal: $\Pi_{rev}(\rho^*) \geq \Pi_{con}(p^*)$, where $\rho^* \in \arg\max_\rho \Pi_{rev}(\rho)$. Revealing X strictly increases the value of the principal only if X and C are not independent.*

Proof. Assuming that the principal's feasible set of payments does not change between markets, the solution $\hat{\rho}(x) = p^*$ is in the feasible set of the principal's optimization problem with information revealed. Therefore,

$$\max_\rho \Pi_{\text{rev}}(\rho) \geq \Pi_{\text{rev}}(\hat{\rho}) = \Pi_{\text{con}}(p^*).$$

If X and C are independent, we have $F_x = F$ for all $x \in \mathcal{X}$, so

$$\max_\rho \Pi_{\text{rev}}(\rho) = \max_\rho \mathbb{E}[F(\rho(X))(b-\rho(X))]$$

By Jensen's inequality, we have

$$\max_\rho \mathbb{E}[F(\rho(X))(b-\rho(X))] \leq \mathbb{E}[\max_\rho F(\rho(X))(b-\rho(X))] = \mathbb{E}[F(p^*)(b-p^*)] = \Pi_{\text{con}}(p^*).$$

Therefore, if X and C are independent, then $\Pi_{\text{rev}}(\rho^*) \leq \Pi_{\text{con}}(p^*)$. □

Lemma 18 (Principal strictly benefits from revelation)**.** *Let F_0 and F_1 be continuously differentiable CDFs. If the MLRP holds (Assumption 5), then the principal strictly benefits when X is revealed: $\Pi_{rev}(p_0^*, p_1^*) > \Pi_{con}(p^*)$.*

Proof. Since F_0, F_1 are continuously differentiable, the first-order necessary conditions hold for the optimal payments p_0^*, p_1^* in equation (E.5). By these conditions, $p_0^* = p_1^*$ only if there exists a value p such that

$$p + \frac{F_0(p)}{f_0(p)} = p + \frac{F_1(p)}{f_1(p)} = b.$$

Such a value p cannot exist if $\frac{F_0(p)}{f_0(p)} \neq \frac{F_1(p)}{f_1(p)}$ for all p. The MLRP implies that $\frac{F_0(p)}{f_0(p)} < \frac{F_1(p)}{f_1(p)}$ for all p; therefore, $p_0^* \neq p_1^*$, and maximum value for Π_{rev} is strictly greater than the maximum value for Π_{con}. □

Proofs from Section 6.3

Lemma 19. *Total welfare increases under revelation ($W_{rev}(p_0^*, p_1^*) - W_{con}(p^*)$) only if task completion quantity increases under revelation,*

$$F(p^*) < (1-\theta)F_0(p_0^*) + \theta F_1(p_1^*).$$

Proof.

$$\begin{aligned}
W_{\text{rev}}(p_0^*, p_1^*) - W_{\text{con}}(p^*) &= (1-\theta)\int_0^{p_0^*}(b-c)f_0(c)dc + \theta\int_0^{p_1^*}(b-c)f_1(c)dc - \int_0^{p^*}(b-c)f(c)dc \\
&= (1-\theta)\int_0^{p_0^*}(b-c)f_0(c)dc + \theta\int_0^{p_1^*}(b-c)f_1(c)dc \\
&\quad - (1-\theta)\int_0^{p^*}(b-c)f_0(c)dc - \theta\int_0^{p^*}(b-c)f_1(c)dc \\
&= (1-\theta)\int_{p^*}^{p_0^*}(b-c)f_0(c)dc - \theta\int_{p_1^*}^{p^*}(b-c)f_1(c)dc
\end{aligned}$$

Upper bounding this:

$$(1-\theta)\int_{p^*}^{p_0^*}(b-c)f_0(c)dc-\theta\int_{p_1^*}^{p^*}(b-c)f_1(c)dc \leq (1-\theta)\int_{p^*}^{p_0^*}(b-p^*)f_0(c)dc-\theta\int_{p_1^*}^{p^*}(b-p^*)f_1(c)dc$$

$$\implies W_{\text{rev}}(p_0^*,p_1^*)-W_{\text{con}}(p^*) \leq (b-p^*)\left((1-\theta)\int_{p^*}^{p_0^*}f_0(c)dc-\theta\int_{p_1^*}^{p^*}f_1(c)dc\right).$$

The right hand term is exactly the difference in output:

$$\begin{aligned}\left((1-\theta)\int_{p^*}^{p_0^*}f_0(c)dc-\theta\int_{p_1^*}^{p^*}f_1(c)dc\right) &= (1-\theta)(F_0(p_0^*)-F_0(p^*))-\theta(F_1(p^*)-F_1(p_1^*))\\ &= (1-\theta)F_0(p_0^*)+\theta F_1(p_1^*)-((1-\theta)F_0(p^*)+\theta F_1(p^*))\\ &= (1-\theta)F_0(p_0^*)+\theta F_1(p_1^*)-F(p^*)\end{aligned}$$

$$\implies W_{\text{rev}}(p_0^*,p_1^*)-W_{\text{con}}(p^*) \leq (b-p^*)\left((1-\theta)F_0(p_0^*)+\theta F_1(p_1^*)-F(p^*)\right)$$

Since $(b-p^*)>0$, if $(1-\theta)F_0(p_0^*)+\theta F_1(p_1^*)-F(p^*)\leq 0$, then $W_{\text{rev}}(p_0^*,p_1^*)-W_{\text{con}}(p^*)\leq 0$. □

E.2 Proofs from Section 6.4

Here we give proofs for results for the garbling model presented in Section 6.4.

Proofs from Section 6.4

Lemma 20 (Monotonic price changes)**.** *Suppose $\gamma=\theta=\frac{1}{2}$. Suppose F_0, F_1 are continuously differentiable CDFs, the principal's utility is strictly concave (Assumption 3), and the MLRP holds (Assumption 5). Then $p_0'(\varepsilon)>0$ and $p_1'(\varepsilon)<0$ for all $\varepsilon\in[0,1]$.*

Proof. $p_0(\varepsilon), p_1(\varepsilon)$ must satisfy first-order necessary conditions for optimality:

$$p_0(\varepsilon)+\frac{\frac{1+\varepsilon}{2}F_0(p_0(\varepsilon))+\frac{1-\varepsilon}{2}F_1(p_0(\varepsilon))}{\frac{1+\varepsilon}{2}f_0(p_0(\varepsilon))+\frac{1-\varepsilon}{2}f_1(p_0(\varepsilon))}=b;\quad p_1(\varepsilon)+\frac{\frac{1+\varepsilon}{2}F_1(p_1(\varepsilon))+\frac{1-\varepsilon}{2}F_0(p_1(\varepsilon))}{\frac{1+\varepsilon}{2}f_1(p_1(\varepsilon))+\frac{1-\varepsilon}{2}f_0(p_1(\varepsilon))}=b$$

Differentiating these first-order conditions, we have:

$$p_0'(\varepsilon)=\frac{\Pi_0'(p_0(\varepsilon))-\Pi_1'(p_0(\varepsilon))}{-2\Pi_{Y=0}''(p_0(\varepsilon))};\quad p_1'(\varepsilon)=\frac{\Pi_1'(p_1(\varepsilon))-\Pi_0'(p_1(\varepsilon))}{-2\Pi_{Y=1}''(p_1(\varepsilon))},\tag{E.6}$$

where

$$\Pi_{Y=0}(p)=\frac{1+\varepsilon}{2}\Pi_0(p)+\frac{1-\varepsilon}{2}\Pi_1(p);\quad \Pi_{Y=1}(p)=\frac{1+\varepsilon}{2}\Pi_1(p)+\frac{1-\varepsilon}{2}\Pi_0(p).$$

Strict concavity from Assumption 3 makes both denominators of $p_x'(\varepsilon)$ positive.

The MLRP also implies that $p_0(\varepsilon)<p_0^*$ and $p_1(\varepsilon)>p_1^*$ for any ε. Therefore, by strict concavity of $\Pi_x(p)$, we have $\Pi_0'(p_0(\varepsilon))>0$, and $\Pi_1'(p_0(\varepsilon))<0$, implying that $p_0'(\varepsilon)>0$. Similarly, $\Pi_1'(p_1(\varepsilon))<0$, and $\Pi_0'(p_1(\varepsilon))>0$, implying that $p_1'(\varepsilon)<0$. □

Proofs from Section 6.4

We first expand $V_{\text{garb}}(\varepsilon)$ for $\gamma = \theta = \frac{1}{2}$.

$$V_{\text{garb}}(\varepsilon) = \frac{1}{2}\left(\frac{1+\varepsilon}{2}V_0(p_0(\varepsilon)) + \frac{1-\varepsilon}{2}V_1(p_0(\varepsilon)) + \frac{1+\varepsilon}{2}V_1(p_1(\varepsilon)) + \frac{1-\varepsilon}{2}V_0(p_1(\varepsilon))\right). \quad \text{(E.7)}$$

To prove Proposition 10, we first give Lemma 30 to handle the anchored zero-cost agent.

Lemma 30. Suppose $C|X = 1$ takes value 0 with probability 1. Suppose f_0 is bounded: $f_0(p) < B$ for all p in the support. Then for any fixed value $b > 0$, there exists $\delta > 0$ such that for any $\varepsilon > \delta$, $p_1(\varepsilon) = 0$.

Proof. Define $h(p, \varepsilon) = p + \frac{1}{f_0(p)}\frac{1+\varepsilon}{1-\varepsilon} + \frac{F_0(p)}{f_0(p)}$. By choosing δ close to 1, we can make the term $\frac{1+\delta}{1-\delta}$ arbitrarily large, and consequently $\frac{1}{f_0(p_1)}\frac{1+\delta}{1-\delta}$ arbitrarily large, since $f_0(p) > 0$. Then for any b, we choose δ close enough to 1 such that $\frac{1}{B}\frac{1+\delta}{1-\delta} > b$. □

Proposition 10 (Sufficient garbling condition with zero-cost type)**.** *Suppose $\gamma = \theta = \frac{1}{2}$. Suppose $C|X = 1$ takes value 0 with probability 1. Suppose F_0 is continuously differentiable and $f_0(c)$ is bounded. $V_{garb}(\varepsilon)$ is maximized at $\varepsilon^* < 1$ if*

$$\frac{v - p_0^*}{2 - \sigma_0(p_0^*)} < g_0(p_0^*),$$

where $g_0(p) = \int_0^p (1 - F_0(c))dc$ is the restricted mean cost of task completion, and $\sigma_0(p) = \frac{F_0(p)f_0'(p)}{f_0(p)^2}$ is the curvature of the inverse quantity function.

Proof. We show that the condition in equation (6.11) implies that $V'_{\text{garb}}(1) < 0$.

For $C|X = 1$ taking value 0 with probability 1, we have $V_1(p) = p$. Substituting this into equation (E.7),

$$V_{\text{garb}}(\varepsilon) = \frac{1}{2}\left(\frac{1+\varepsilon}{2}V_0(p_0(\varepsilon)) + \frac{1-\varepsilon}{2}p_0(\varepsilon) + \frac{1+\varepsilon}{2}p_1(\varepsilon) + \frac{1-\varepsilon}{2}V_0(p_1(\varepsilon))\right).$$

Differentiating this, we have

$$\begin{aligned} 2V'_{\text{garb}}(\varepsilon) =& p_0'(\varepsilon)\left(\frac{1+\varepsilon}{2}F_0(p_0(\varepsilon)) + \frac{1-\varepsilon}{2}\right) + p_1'(\varepsilon)\left(\frac{1+\varepsilon}{2} + \frac{1-\varepsilon}{2}F_0(p_1(\varepsilon))\right) \\ &+ \frac{1}{2}\left((V_0(p_0(\varepsilon)) - V_0(p_1(\varepsilon))) - (p_0(\varepsilon) - p_1(\varepsilon))\right) \end{aligned}$$

Evaluating this derivative at $\varepsilon = 1$, Lemma 30 implies that $p_1(\varepsilon) = 0$ and $p_1'(1) = 0$.

$$2V'_{\text{garb}}(1) = p_0'(1)F_0(p_0^*) + \frac{1}{2}(V_0(p_0^*) - p_0^*).$$

$$V_0(p_0^*) - p_0^* = E[(p_0^* - C)\,\mathbb{1}(C < p_0^*)|X = 0] - p_0^* = -\frac{1}{2}g_0(p_0^*).$$

$$\implies 4V'_{\text{garb}}(1) = -g_0(p_0^*) + 2F_0(p_0^*)p'_0(1)$$

Simplifying $2F_0(p_0^*)p'_0(1)$:

$$\begin{aligned}
2F_0(p_0^*)p'_0(1) &= \frac{F_0(p_0^*)f_0(p_0^*)}{2f_0(p_0^*)^2 - F_0(p_0^*)f'_0(p_0^*)} \\
&= \frac{\frac{F_0(p_0^*)}{f_0(p_0^*)}}{2 - \frac{F_0(p_0^*)f'_0(p_0^*)}{f_0(p_0^*)^2}} \\
&= \frac{v - p_0^*}{2 - \sigma_0(p_0^*)}
\end{aligned}$$

Therefore,

$$V'_{\text{garb}}(1) < 0 \iff -g_0(p_0^*) + \frac{v - p_0^*}{2 - \sigma_0(p_0^*)} < 0.$$

□

Proposition 11 (Sufficient garbling condition)**.** *Suppose $\gamma = \theta = \frac{1}{2}$. Suppose F_0, F_1 are continuously differentiable. $V_{garb}(\varepsilon)$ is maximized at $\varepsilon^* < 1$ if*

$$-\Pi'_1(p_0^*)\left(\frac{b - p_0^*}{2 - \sigma_0(p_0^*)}\right) - \Pi'_0(p_1^*)\left(\frac{b - p_1^*}{2 - \sigma_1(p_1^*)}\right) < \Delta(p_0^*, p_1^*).$$

Proof. Differentiating with respect to ε, we have

$$\begin{aligned}
2V'_{\text{garb}}(\varepsilon) =& p'_0(\varepsilon)\left(\frac{1+\varepsilon}{2}F_0(p_0(\varepsilon)) + \frac{1-\varepsilon}{2}F_1(p_0(\varepsilon))\right) + p'_1(\varepsilon)\left(\frac{1+\varepsilon}{2}F_1(p_1(\varepsilon)) + \frac{1-\varepsilon}{2}F_0(p_1(\varepsilon))\right) \\
&+ \frac{1}{2}\left((V_0(p_0(\varepsilon)) - V_0(p_1(\varepsilon))) - (V_1(p_0(\varepsilon)) - V_1(p_1(\varepsilon)))\right)
\end{aligned}$$

Substituting in the price derivatives from equation (E.6) and the agent utility dominance identity from Definition 5,

$$\begin{aligned}
2V'_{\text{garb}}(\varepsilon) = &\left(\frac{\Pi'_0(p_0(\varepsilon)) - \Pi'_1(p_0(\varepsilon))}{-2\Pi''_{Y=0}(p_0(\varepsilon))}\right)\left(\frac{1+\varepsilon}{2}F_0(p_0(\varepsilon)) + \frac{1-\varepsilon}{2}F_1(p_0(\varepsilon))\right) \\
&+ \left(\frac{\Pi'_1(p_1(\varepsilon)) - \Pi'_0(p_1(\varepsilon))}{-2\Pi''_{Y=1}(p_1(\varepsilon))}\right)\left(\frac{1+\varepsilon}{2}F_1(p_1(\varepsilon)) + \frac{1-\varepsilon}{2}F_0(p_1(\varepsilon))\right) \\
&+ \frac{1}{2}\left(-\Delta(p_0(\varepsilon), p_1(\varepsilon))\right)
\end{aligned}$$

Let

$$z_0^\varepsilon(p) = \frac{V'_{Y=0}(p)}{\Pi''_{Y=0}(p)} = \frac{\frac{1+\varepsilon}{2}F_0(p) + \frac{1-\varepsilon}{2}F_1(p)}{\Pi''_{Y=0}(p)},$$
$$z_1^\varepsilon(p) = \frac{V'_{Y=1}(p)}{\Pi''_{Y=1}(p)} = \frac{\frac{1+\varepsilon}{2}F_1(p) + \frac{1-\varepsilon}{2}F_0(p)}{\Pi''_{Y=1}(p)}.$$

Then

$$\begin{aligned}4V'_{\text{garb}}(\varepsilon) = &\left(\Pi'_0(p_0(\varepsilon)) - \Pi'_1(p_0(\varepsilon))\right)\left(-z_0^\varepsilon(p_0(\varepsilon))\right) + \left(\Pi'_1(p_1(\varepsilon)) - \Pi'_0(p_1(\varepsilon))\right)\left(-z_1^\varepsilon(p_1(\varepsilon))\right)\\ &- \Delta(p_0(\varepsilon), p_1(\varepsilon))\end{aligned}$$

Evaluating this at $\varepsilon = 1$, we have

$$\begin{aligned}4V'_{\text{garb}}(1) = &\left(\Pi'_0(p_0^*) - \Pi'_1(p_0^*)\right)\left(-z_0(p_0^*)\right) + \left(\Pi'_1(p_1^*) - \Pi'_0(p_1^*)\right)\left(-z_1(p_1^*)\right)\\ &- \Delta(p_0^*, p_1^*)\end{aligned}$$

where

$$z_x(p) = \frac{V'_x(p)}{\Pi''_x(p)}$$

Note that $V'_x(p_x^*) = W'_x(p_x^*)$, so at p^*, this is identical to the z function from Aguirre et al. [2010]. By first-order optimality conditions,

$$-z_x(p_x^*) = \frac{b - p_x^*}{2 - \sigma_x(p_x^*)}$$

Therefore, $V'_{\text{garb}}(1) < 0$ if

$$(-\Pi'_1(p_0^*))\left(\frac{b - p_0^*}{2 - \sigma_0(p_0^*)}\right) + (-\Pi'_0(p_1^*))\left(\frac{b - p_1^*}{2 - \sigma_1(p_1^*)}\right) < \Delta(p_0^*, p_1^*).$$

□

Proofs from Section 6.4

Lemma 21. *$\Pi_{con}(p^*) \leq \Pi_{garb}(p_0(\varepsilon), p_1(\varepsilon)) \leq \Pi_{rev}(p_0^*, p_1^*)$ for all ε.*

Proof. For $\varepsilon \in \{0, 1\}$, the inequalities clearly hold. Fix $\varepsilon \in (0, 1)$. For the lower bound,

$$\Pi_{\text{garb}}(p_0(\varepsilon), p_1(\varepsilon)) \geq \Pi_{\text{garb}}(p^*, p^*) = \Pi_{\text{con}}(p^*).$$

For the upper bound, since the noise ξ is independent of X and C,

$$\Pi_{\text{garb}}(p_0, p_1) = \phi(\theta, \gamma)\Pi_{\text{rev}}(p_0, p_1) + \psi(\theta, \gamma)\Pi_{\text{rev}}(p_1, p_0) \leq \Pi_{\text{rev}}(p_0^*, p_1^*)$$

where ϕ, ψ are some positive functions of θ, γ. □

Lemma 22 (Strict principal improvement)**.** *Suppose $\gamma = \theta = \frac{1}{2}$. Suppose the principal's utility is strictly concave (Assumption 3). If the MLRP holds (Assumption 5), then $\Pi'_{garb}(\varepsilon) > 0$ for all $\varepsilon \in [0,1]$.*

Proof.

$$\begin{aligned}\Pi_{\text{garb}}(\varepsilon) = \frac{1}{2}(b - p_0(\varepsilon))\left(\frac{1+\varepsilon}{2}F_0(p_0(\varepsilon)) + \frac{1-\varepsilon}{2}F_1(p_0(\varepsilon))\right)\\ + \frac{1}{2}(b - p_1(\varepsilon))\left(\frac{1+\varepsilon}{2}F_1(p_1(\varepsilon)) + \frac{1-\varepsilon}{2}F_0(p_1(\varepsilon))\right)\end{aligned}$$

Differentiating with respect to ε:

$$\begin{aligned}4\Pi'_{\text{garb}}(\varepsilon) =& F_0(p_0(\varepsilon))(b - p_0(\varepsilon)) - F_0(p_1(\varepsilon))(b - p_1(\varepsilon))\\ &+ F_1(p_1(\varepsilon))(b - p_1(\varepsilon)) - F_1(p_0(\varepsilon))(b - p_0(\varepsilon))\end{aligned}$$

By the MLRP and strict concavity of $\Pi_0(p), \Pi_1(p)$, we have that $p_1(\varepsilon) < p_0(\varepsilon) < p_0^*$. Strict concavity of $\Pi_0(p)$ and the optimality of p_0^* for Π_1 then implies that

$$F_0(p_1(\varepsilon))(v - p_1(\varepsilon)) < F_0(p_0(\varepsilon))(v - p_0(\varepsilon)).$$

Similarly, by the MLRP and strict concavity of $\Pi_0(p), \Pi_1(p)$, we have that $p_1^* < p_1(\varepsilon) < p_0(\varepsilon)$. Strict concavity of $\Pi_1(p)$ and the optimality of p_1^* for Π_1 implies that

$$F_1(p_0(\varepsilon))(v - p_0(\varepsilon)) < F_1(p_1(\varepsilon))(v - p_1(\varepsilon)).$$

□

Proofs from Section 6.4

Lemma 23 (Reducing garbling initially increases total welfare)**.** *Relative to full concealment with $\varepsilon = 0$, revealing Y with some garbled noise initially does not decrease total welfare: $W'_{garb}(0) \geq 0$. The inequality is strict if $\Pi'_{garb}(0) > 0$, which is true under strict concavity of $\Pi_0(p), \Pi_1(p)$ (Assumption 3) and the MLRP (Assumption 5).*

Proof.

$$W'_{\text{garb}}(\varepsilon) = V'_{\text{garb}}(\varepsilon) + \Pi'_{\text{garb}}(\varepsilon)$$

$V'_{\text{garb}}(0) = 0$, so $W'_{\text{garb}}(0) \geq 0$. The strict inequality comes from applying Lemma 22 that $\Pi'_{\text{garb}}(0) > 0$. □

Lemma 24 (Optimal garbling increases welfare over concealment)**.** *Let $\varepsilon^* \in \arg\max_{\varepsilon\in[0,1]} V_{garb}(\varepsilon)$. Then $W_{garb}(\varepsilon^*) \geq W_{garb}(0)$.*

Proof.

$$W_{\text{garb}}(\varepsilon^*) = V_{\text{garb}}(\varepsilon^*) + \Pi_{\text{garb}}(\varepsilon^*)$$

Lemma 21 implies $\Pi_{\text{garb}}(\varepsilon^*) \geq \Pi_{\text{garb}}(0)$. By optimality of ε^*, $V_{\text{garb}}(\varepsilon^*) \geq V_{\text{garb}}(0)$. □

Lemma 25 (More information increases total welfare relative to agent optimal garbling)**.** *$W'_{garb}(\varepsilon) \geq V'_{garb}(\varepsilon)$ for all $\varepsilon \in [0,1]$. The inequality is strict if $\Pi'_{garb}(0) > 0$, which is true under strict concavity of $\Pi_0(p), \Pi_1(p)$ (Assumption 3) and the MLRP (Assumption 5).*

Proof. Lemma 21 implies that $\Pi'_{\text{garb}}(\varepsilon) > 0$, which implies that $W'_{\text{garb}}(\varepsilon) \geq V'_{\text{garb}}(\varepsilon)$. The inequality is strict under the conditions of Lemma 22. □

Bibliography

30 years of UpToDate: The evolution of clinical decision support and the future of evidence-based medicine. 30 years of uptodate: The evolution of clinical decision support and the future of evidence-based medicine. *Wolters Kluwer*, 2022. URL https://www.wolterskluwer.com/en/expert-insights/30-years-of-uptodate-evolution-of-clinical-decision-support-future-of-evidence-based-medicine.

Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. Roles for computing in social change. In *Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020.

A. Agarwal, M. Dudik, and Z. S. Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning (ICML)*, 2019.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning (ICML)*, 2018.

Sina Aghaei, Mohammad Javad Azizi, and Phebe Vayanos. Learning optimal and fair decision trees for non-discriminative decision-making. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 1418–1426, 2019.

Inaki Aguirre, Simon Cowan, and John Vickers. Monopoly price discrimination and demand curvature. *American Economic Review*, 100(4):1601–1615, 2010.

Saba Ahmadi, Hedyeh Beyhaghi, Avrim Blum, and Keziah Naggita. On classification of strategic agents who can both game and improve. In *3rd Symposium on Foundations of Responsible Computing (FORC)*, 2022.

H. E. Allen, E. J. Latessa, and B. S. Ponder. *Corrections in America: An Introduction*. Pearson, 2015.

Tal Alon, Paul Dütting, Yingkai Li, and Inbal Talgam-Cohen. Bayesian analysis of linear contracts. *arXiv preprint arXiv:2211.06850*, 2022.

M. Anderson and S. Anderson. The status of machine ethics: A report from the AAAI Symposium. *Minds and Machines*, 17:1–10, 07 2007.

Rohan Anil, Gabriel Pereyra, Alexandre Tachard Passos, Robert Ormandi, George Dahl, and Geoffrey Hinton. Large scale distributed neural network training through online distillation. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/pdf?id=rkr1UDeC-.

N. P. Archer and S. Wang. Application of the back propagation neural network algorithm with monotonicity constraints for two-group classification problems. *Decision Sciences*, 24 (1):60–75, 1993.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

Kenneth J Arrow. Uncertainty and the welfare economics of medical care. *The American Economic Review*, 53(5):941–973, 1963.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89 (1):133–161, 2021.

Pranjal Awasthi, Matthäus Kleindessner, and Jamie Morgenstern. Equalized odds postprocessing under imperfect group information. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

M. Ayer, H. D. Brunk, G. M. Ewing, W. T. William, and E. Silverman. An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26:641–647, 1955.

O. Azar. What sustains social norms and how they evolve? the case of tipping. *Journal of Economic Behavior and Organization*, 54:49–64, 2004.

Deborah L. Bandalos. *Measurement Theory and Applications for the Social Sciences*. Guilford Publications, 2017.

Deborah L Bandalos. *Measurement Theory and Applications for the Social Sciences*. Guilford Publications, 2018.

R. E. Barlow, D. J. Bartholomew, and J. M. Bremner. *Statistical inference under order restrictions; the theory and application of isotonic regression*. Wiley, 1972.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Stephen Bates, Michael I. Jordan, Michael Sklar, and Jake A. Soloff. Principal-agent hypothesis testing. *arXiv preprint arXiv:2205.06812*, 2022.

T. L. Beauchamp and J. F. Childress. Principles of biomedical ethics. In *Oxford University Press*, 2001.

Yahav Bechavod, Katrina Ligett, Zhiwei Steven Wu, and Juba Ziani. Causal feature discovery through strategic modification. *arXiv preprint arXiv:2002.07024*, 3, 2020.

A. Ben-David. Automatic generation of symbolic multiattribute ordinal knowledge based DSS: methodology and applications. *Decision Sciences*, pages 1357–1372, 1992.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, October 2009.

Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Suman Bera, Deeparnab Chakrabarty, Nicolas Flores, and Maryam Negahbani. Fair algorithms for clustering. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4954–4965, 2019.

Dirk Bergemann and Stephen Morris. Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95, 2019.

Dirk Bergemann, Benjamin Brooks, and Stephen Morris. The limits of price discrimination. *American Economic Review*, 105(3):921–957, 2015.

R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth. A convex framework for fair regression. *FATML*, 2017a.

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017b.

Dimitris Bertsimas, David B Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011.

A. Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. H. Chi, and C. Goodrow. Fairness in recommendation through pairwise experiments. *KDD Applied Data Science Track*, 2019.

P. J. Bickel, E. A. Hammel, and J. W. O'Connell. Sex bias in graduate admissions: Data from Berkeley. *Science*, pages 398–404, 1975.

R. Binns. Fairness in machine learning: Lessons from political philosophy. *Conf. on Fairness, Accountability, and Transparency*, 2018.

David Blackwell. Comparison of experiments. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, volume 2, pages 93–103. University of California Press, 1951.

Raphael Boleslavsky and Christopher Cotton. Grading standards and education quality. *American Economic Journal: Microeconomics*, 7(2):248–279, 2015.

M. Bonakdarpour, S. Chatterjee, R. F. Barber, and J. Lafferty. Prediction rule reshaping. In *International Conference on Machine Learning (ICML)*, 2018.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Journal of Machine Learning Research (JMLR)*, 2018.

Lilah Burke. The death and life of an admissions algorithm. *Inside Higher Ed*, December, 2020.

Donald T Campbell. Assessing the impact of planned social change. *Evaluation and Program Planning*, 2(1):67–90, 1979.

K. Canini, A. Cotter, M. M. Fard, M. R. Gupta, and J. Pfeifer. Fast and flexible monotonic functions with ensembles of lattices. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019a.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 1567–1578, 2019b.

Gabriel Carroll. Robustness and linear contracts. *American Economic Review*, 105(2):536–63, 2015.

Lawrence P Casalino, David Gans, Rachel Weber, Meagan Cea, Amber Tuchovsky, Tara F Bishop, Yesenia Miranda, Brittany A Frankel, Kristina B Ziehler, Meghan M Wong, et al. Us physician practices spend more than $15.4 billion annually to report quality measures. *Health Affairs*, 35(3):401–406, 2016.

L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. Ranking with fairness constraints. In *ICALP*, 2018.

Keshigeyan Chandrasegaran, Ngoc-Trung Tran, Yunqing Zhao, and Ngai-Man Cheung. Revisiting label smoothing and knowledge distillation compatibility: What was missing?, 2022. URL https://arxiv.org/abs/2206.14532.

Robert S Chen, Brendan Lucier, Yaron Singer, and Vasilis Syrgkanis. Robust optimization for non-convex objectives. In *NeurIPS*, 2017.

D. Chetverikov, A. Santos, and A. M. Shaikh. The econometrics of shape restrictions. *Annual Review of Economics*, 2018.

S. Coate and G. C. Loury. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, 83:1220–1240, 1993.

Guy W. Cole and Sinead A. Williamson. Avoiding resentment via monotonic fairness. *arXiv preprint arXiv:1909.01251*, 2019.

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

A. Cotter, H. Jiang, and K. Sridharan. Two-player games for efficient non-convex constrained optimization. In *ALT*, 2019a.

A. Cotter, H. Jiang, and K. Sridharan. Two-player games for efficient non-convex constrained optimization. In *International Conference on Algorithmic Learning Theory (ALT)*, 2019b.

A. Cotter, H. Jiang, S. Wang, T. Narayan, M. R. Gupta, S. You, and K. Sridharan. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research (JMLR)*, 2019c.

Andrew Cotter, Heinrich Jiang, Serena Wang, Taman Narayan, Seungil You, Karthik Sridharan, and Maya R. Gupta. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research (JMLR)*, 20(172):1–59, 2019d.

Kimberle Crenshaw. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43:1241, 1990.

Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 715–724, 2021.

Jiequan Cui, Shu Liu, Zhuotao Tian, Zhisheng Zhong, and Jiaya Jia. Reslt: Residual learning for long-tailed recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.

Rachel Cummings, Katrina Ligett, Mallesh M Pai, and Aaron Roth. The strange case of privacy in equilibrium models. In *ACM Conference on Economics and Computation (EC)*. 2016.

Mark A. Davenport, Richard G. Baraniuk, and Clayton D. Scott. Tuning support vector machines for minimax and Neyman-Pearson classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44 (7):3366–3385, 2021.

Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

David Dranove, Daniel Kessler, Mark McClellan, and Mark Satterthwaite. Is more information better? the effects of "report cards" on health care providers. *Journal of Political Economy*, 111(3):555–588, 2003.

Mengnan Du, Subhabrata Mukherjee, Yu Cheng, Milad Shokouhi, Xia Hu, and Ahmed Hassan Awadallah. What do compressed large language models forget? robustness challenges in model compression, 2021.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *International Conference on Machine Learning (ICML)*, 2008.

Paul Dütting, Tim Roughgarden, and Inbal Talgam-Cohen. Simple versus optimal contracts. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 369–387, 2019.

Paul Dütting, Tim Roughgarden, and Inbal-Talgam Cohen. The complexity of contracts. In *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms*, pages 2688–2707. SIAM, Philadelphia, PA, 2020.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science (ITCS)*, pages 214–226. ACM, 2012.

Elad Eban, Mariano Schain, Alan Mackey, Ariel Gordon, Rif A. Saurous, and Gal Elidan. Scalable learning of non-decomposable objectives. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171:115–166, 2018.

E. Fehr and K. M. Schmidt. A theory of fairness, competition, and cooperation. In *The Quarterly Journal of Economics*, volume 114, pages 817–868, 1999.

Marion Fourcade and Kieran Healy. Classification situations: Life-chances in the neoliberal era. *Accounting, Organizations and Society*, 38(8):559–572, 2013.

Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Conference on Fairness, Accountability, and Transparency (FAccT)*, 2019.

Robert Gibbons, John Roberts, et al. *The Handbook of Organizational Economics*. Princeton University Press Princeton, NJ, 2013.

Talia B Gillis. False dreams of algorithmic fairness: The case of credit pricing. *Available at SSRN 3571266*, 2020.

Laurent G Glance, Caroline P Thirukumaran, Changyong Feng, Stewart J Lustik, and Andrew W Dick. Association between the physician quality score in the merit-based incentive payment system and hospital performance in hospital compare in the first year of the program. *JAMA Network Open*, 4(8):e2118449–e2118449, 2021.

N. Goel, M. Yaghini, and B. Faltings. Non-discriminatory machine learning through convex fairness criteria. *AAAI/ACM Conference on Artificial Intelligence, Ethics, Society*, 2018.

Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael Friedlander. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Andrew A Gonzalez and Amir A Ghaferi. Hospital safety scores: do grades really matter? *JAMA Surgery*, 149(5):413–414, 2014.

Charles AE Goodhart. Problems of monetary management: the uk experience. In *Monetary theory and practice*, pages 91–121. Springer, 1984.

Google AI Blog. Tensorflow lattice: Flexible, controlled and interpretable ml, 2020. URL https://blog.tensorflow.org/2020/02/tensorflow-lattice-flexible-controlled-and-interpretable-ML.html.

Vincent Grari, Boris Ruf, Sylvain Lamprier, and Marcin Detyniecki. Achieving fairness with decision trees: An adversarial approach. *Data Science and Engineering*, 5(2):99–110, 2020.

K. Gray and J. Graham. *Atlas of Moral Psychology.* Guilford Press, 2018.

P. Groeneboom and G. Jongbloed. *Nonparametric estimation under shape constraints.* Cambridge Press, New York, USA, 2014.

The GenIUSS Group. *Best Practices for Asking Questions to Identify Transgender and Other Gender Minority Respondents on Population-Based Surveys.* The Williams Institute, 2014.

R. Gruetzemacher. Rethinking AI strategy and policy as entangled super wicked problems. *AAAI/ACM Conference on Artificial Intelligence, Ethics, Society*, 2018.

Luke Guerdan, Amanda Coston, Kenneth Holstein, and Zhiwei Steven Wu. Counterfactual prediction under outcome measurement error. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1584–1598, 2023.

M. R. Gupta, A. Cotter, J. Pfeifer, K. Voevodski, K. Canini, A. Mangylov, W. Moczydlowski, and A. Van Esbroeck. Monotonic calibrated interpolated look-up tables. *Journal of Machine Learning Research*, 17(109):1–47, 2016. URL http://jmlr.org/papers/v17/15-243.html.

Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. Proxy fairness. *arXiv preprint arXiv:1806.11212*, 2018.

Nika Haghtalab, Nicole Immorlica, Brendan Lucier, and Jack Z Wang. Maximizing welfare with incentive-aware evaluation mechanisms. *arXiv preprint arXiv:2011.01956*, 2020.

J. Haidt. *The Righteous Mind.* Random House, New York, USA, 2013.

S. Hajian and J. Domino-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE Trans. on Knowledge and Data Engineering*, 25(7): 1445–1459, 2013.

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. *Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016a.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016b.

Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning (ICML)*, 2018.

T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman Hall, New York, 1990.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Yin-Yin He, Jianxin Wu, and Xiu-Shen Wei. Distilling virtual examples for long-tailed recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 235–244, 2021.

California Health and Human Services Agency. Cal hospital compare announces honor roll hospitals. *Press Release*, 2022. URL https://www.chhs.ca.gov/blog/2022/08/22/cal-hospital-compare-announces-honor-roll-hospitals/.

Miguel A Hernan and James M Robins. *Causal Inference: What If*. CRC Press, 2020.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015.

Keisuke Hirano and Jack R Porter. Asymptotics for statistical treatment rules. *Econometrica*, 77(5):1683–1701, 2009.

Bengt Holmström. Moral hazard and observability. *The Bell Journal of Economics*, pages 74–91, 1979.

Bengt Holmstrom and Paul Milgrom. Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design. *The Journal of Law, Economics, and Organization*, 7 (special_issue):24–52, 1991.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*, 2020.

Bell Hooks. *Yearning: Race, Gender, and Cultural Politics*. Routledge, 1992.

Keith Hoskin. The 'awful idea of accountability': inscribing people into the measurement of objects. *Accountability: Power, Ethos and the Technologies of Managing*, 265, 1996.

The White House. FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation that Protects Americans' Rights and Safety. *Statements and Releases*, 2023. URL https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/#:~:text=Public%20assessments%20of,at%20DEFCON%2031.

Peter Howitt and R Preston McAfee. Animal spirits. *The American Economic Review*, pages 493–507, 1992.

J. E. Hunter and F. L. Schmidt. Critical analysis of the statistical and ethical implications of various definitions of test bias. *Psychological Bulletin*, 1976.

Ben Hutchinson and M. Mitchell. 50 years of test (un)fairness: Lessons for machine learning. In *Conference on Fairness, Accountability, and Transparency (FAccT)*, 2019.

Wenke Hwang, Jordan Derk, Michelle LaClair, and Harold Paz. Hospital patient safety grades may misrepresent hospital performance. *Journal of Hospital Medicine*, 9(2):111–115, 2014.

Physicians Advocacy Institute. 2022 merit-based incentive payment system (mips) scoring overview. *Medicare Quality Payment Program (QPP) Physician Education Initiative*, 2022. URL http://www.physiciansadvocacyinstitute.org/Portals/0/assets/docs/MIPS-Pathway/MIPS%20Scoring%20Overview.pdf.

International Stroke Trial Collaborative Group. The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19 435 patients with acute ischaemic stroke. *The Lancet*, 349(9065):1569–1581, 1997.

Ahmet Iscen, André Araujo, Boqing Gong, and Cordelia Schmid. Class-balanced distillation for long-tailed visual recognition. *arXiv preprint arXiv:2104.05279*, 2021.

K. V. Iserson and J. C. Moksop. Triage in medicine: Concept, history, types. *Annals of Emergency Medicine*, 2007.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. *arXiv preprint arXiv:1912.05511*, 2019.

Abigail Z Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 375–385, 2021.

A. Cecile J.W. Janssens. Validity of polygenic risk scores: are we measuring what we think we are? *Human molecular genetics*, 28(R2):R143–R150, 2019.

Kruskal JB. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29: 115–129, 1964.

William Stanley Jevons. *Investigations in currency and finance*. Macmillan and Company, 1884.

Ashish K Jha, E John Orav, Abigail B Ridgway, Jie Zheng, and Arnold M Epstein. Does the leapfrog program help identify high-quality hospitals? *The Joint Commission Journal on Quality and Patient Safety*, 34(6):318–325, 2008.

HJS Jones and L de Cossart. Risk scoring in surgical patients. *British Journal of Surgery*, 86 (2):149–157, 1999.

Haytham MA Kaafarani, Ann M Borzecki, Kamal MF Itani, Susan Loveland, Hillary J Mull, Kathleen Hickson, Sally MacDonald, Marlena Shin, and Amy K Rosen. Validity of selected patient safety indicators: opportunities and concerns. *Journal of the American College of Surgeons*, 212(6):924–934, 2011.

Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. *Advances in neural information processing systems*, 31, 2018.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020.

Emir Kamenica. Bayesian persuasion and information design. *Annual Review of Economics*, 11:249–272, 2019.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Maximilian Kasy and Rediet Abebe. Fairness, equality, and power in algorithmic decision-making. *ICML Workshop on Participatory Approaches to Machine Learning*, 2020.

KDD Cup. Predict funding requests that deserve an A+, 2014. URL https://www.kdd.org/kdd-cup/view/kdd-cup-2014.

M. Kearns, S. Neel, A. Roth, and S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning (ICML)*, 2018.

Edward H Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022.

Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, and Christos Thrampoulidis. Label-imbalanced and group-sensitive classification under overparameterization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)*, 8(4):1–23, 2020.

Niko Kolodny. Why equality of treatment and opportunity might matter. *Philosophical Studies*, 176:3357–3366, 2019.

J. Komiyama, A. Takeda, J. Honda, and H. Shimao. Nonconvex optimization for regression with fairness constraints. In *International Conference on Machine Learning (ICML)*, 2018.

Daniel Koretz. *The Testing Charade: Pretending to Make Schools Better*. The University of Chicago Press, 2017.

W. Kotlowski and R. Slowinski. Rule learning with monotonicity constraints. In *International Conference on Machine Learning (ICML)*, pages 537–544, 2009.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Ivar Krumpal. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality and Quantity*, 47:2025–2047, 2011.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

L. Bernstein. *Who Deserves a Liver?* Washington Post, September 2017. URL https://www.washingtonpost.com.

Natalie Lacireno-Paquet, Thomas T Holyoke, Michele Moser, and Jeffrey R Henig. Creaming versus cropping: Charter school enrollment practices in response to market incentives. *Educational Evaluation and Policy Analysis*, 24(2):145–158, 2002.

Jean-Jacques Laffont and David Martimort. *The Theory of Incentives*. Princeton University Press, 2009.

Jean-Jacques Laffont and Jean Tirole. Using cost observation to regulate firms. *Journal of Political Economy*, 94(3, Part 1):614–641, 1986.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H Chi. Fairness without demographics through adversarially reweighted learning. *arXiv preprint arXiv:2006.13114*, 2020.

Alexandre Lamy, Ziyuan Zhong, Aditya Krishna Menon, and Nakul Verma. Noise-tolerant fair classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Claire Lazar and Suhas Vijaykumar. A resolution in algorithmic fairness: Calibrated scores for fair classifications. *arXiv preprint arXiv:2002.07676*, 2020.

Sergio G Lazzarini, Sandro Cabral, Sergio Firpo, and Thomaz Teodorovicz. Why are counterfactual assessment methods not widespread in outcome-based contracts? a formal model approach. *Journal of Public Administration Research and Theory*, 32(3):509–523, 2022.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. CS 231N, 2015.

Abba Lerner. *The concept of monopoly and the measurement of monopoly power*. Springer, 1995.

Jiajin Li, Sen Huang, and Anthony Man-Cho So. A first-order algorithmic framework for Wasserstein distributionally robust logistic regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

M. Lichman. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning (ICML)*, 2018.

Lydia T Liu, Serena Wang, Tolani Britton, and Rediet Abebe. Reimagining the machine learning life cycle to improve educational outcomes of students. *Proceedings of the National Academy of Sciences*, 120(9):e2204781120, 2023.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

Robert E Lucas Jr. Econometric policy evaluation: A critique. In *Carnegie-Rochester conference series on public policy*, volume 1, pages 19–46. North-Holland, 1976.

Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *International Conference on Machine Learning (ICML)*. JMLR.org, 2020.

Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. Teacher's pet: understanding and mitigating biases in distillation. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=ph3AYXpwEb.

Charles F Manski. *Identification for Prediction and Decision*. Harvard University Press, 2008.

Steffen Mau. *The Metric Society: On the Quantification of the Social.* John Wiley & Sons, 2019.

Sean M McNee, John Riedl, and Joseph A Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on Human factors in computing systems*, pages 1097–1101, 2006.

Aditya K Menon, Ankit Singh Rawat, Sashank Reddi, Seungyeon Kim, and Sanjiv Kumar. A statistical perspective on distillation. In *International Conference on Machine Learning*, pages 7632–7642. PMLR, 2021a.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations (ICLR)*, 2021b.

Paul Milgrom and John Roberts. Relying on the information of interested parties. *The RAND Journal of Economics*, pages 18–32, 1986.

Paul R Milgrom. Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics*, pages 380–391, 1981.

Paul Robert Milgrom and John Roberts. *Economics, Organization, and Management.* Prentice Hall, 1992.

John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*, pages 6917–6926. PMLR, 2020.

J. H. Moor. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, pages 18–21, 2006.

Hussein Mozannar, Mesrob I Ohannessian, and Nathan Srebro. Fair learning with private demographic data. *arXiv preprint arXiv:2002.11651*, 2020.

Kenneth J Mukamal. The effects of smoking and drinking on cardiovascular disease and risk factors. *Alcohol Research & Health*, 29(3):199, 2006.

Jerry Z Muller. *The Tyranny of Metrics.* Princeton University Press, 2019.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems (NIPS)*, Red Hook, NY, USA, 2019. Curran Associates Inc.

Hongseok Namkoong and John Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Harikrishna Narasimhan and Aditya K Menon. Training over-parameterized models with non-decomposable objectives. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Harikrishna Narasimhan, Andrew Cotter, and Maya Gupta. Optimizing generalized rate metrics with three players. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019a.

Harikrishna Narasimhan, Andrew Cotter, and Maya R. Gupta. On making stochastic classifiers deterministic. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019b.

Harikrishna Narasimhan, Andrew Cotter, Yichen Zhou, Serena Wang, and Wenshuo Guo. Approximate heavily-constrained learning with lagrange multiplier models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Arvind Narayanan. Translation tutorial: 21 fairness definitions and their politics. In *Conference on Fairness, Accountability, and Transparency (FAccT)*, 2018.

Jersey Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10(1):1–51, 1923.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Michael Ostrovsky and Michael Schwarz. Information disclosure and unraveling in matching markets. *American Economic Journal: Microeconomics*, 2(2):34–63, 2010.

Art B. Owen. *Monte Carlo theory, methods and examples.* https://artowen.su.domains/mc/, 2013.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

J. Pearl, M. Glymour, and N. Jewell. *Causal Inference in Statistics: a Primer.* Wiley, 2016.

Judea Pearl. On a class of bias-amplifying variables that endanger effect estimates. *arXiv preprint arXiv:1203.3503*, 2012.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

A. Pérez-Suay, V. Laparra, G. Mateo-García, J. Muñoz-Marí, L. Gómez-Chova, and G. Camps-Valls. Fair kernel learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2017.

Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11557–11568, 2021.

Michael Power. *The audit explosion.* Number 7. Demos, 1994.

QCOR. 2023 OPO Interim Annual Public Aggregated Performance Report: Methodology for Donation and Transplant Measure Calculations. *Survey and Certification Quality, Certification and Oversight Reports (S&C QCOR)*, 2023. URL https://qcor.cms.gov/documents/2023%20OPO%20Interim%20Annual%20Public%20Aggregated%20Report.xlsx.

I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He. Data distillation: Towards omni-supervised learning. In *CVPR*, 2018.

E. Raff, J. Sylvester, and S. Mills. Fair forests: Regularized tree induction to minimize model bias. *AAAI/ACM Conference on Artificial Intelligence, Ethics, Society*, 2018.

J. Rawls. *A Theory of Justice.* Belknap Press, 1971.

Jennifer L Rodgers, Jarrod Jones, Samuel I Bolleddu, Sahit Vanthenapalli, Lydia E Rodgers, Kinjal Shah, Krishna Karia, and Siva K Panguluri. Cardiovascular risks associated with gender and aging. *Journal of cardiovascular development and disease*, 6(2):19, 2019.

Tina Rosenberg. To make hospitals less deadly, a dose of data. *The New York Times*, 2013. URL https://archive.nytimes.com/opinionator.blogs.nytimes.com/2013/12/04/to-make-hospitals-less-deadly-a-dose-of-data/.

W. D. Ross. *The Right and the Good.* Oxford Press Reprint of 1930 Original, Oxford, 2002.

Richard Rothstein. Holding accountability to account: How scholarship and experience in other fields inform exploration of performance incentives in education. working paper 2008-04. *National Center on Performance Incentives*, 2008.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association*, 75(371):591–593, 1980.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

Chris Russell, Matt J. Kusner, Joshua Loftus, and Ricardo Silva. When worlds collide: Integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Andrew M Ryan, James F Burgess Jr, Christopher P Tompkins, and Stanley S Wallack. The relationship between medicare's process of care quality measures and mortality. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, 46(3):274–290, 2009.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020a.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, 2020b.

Roshni Sahoo and Stefan Wager. Policy learning with competing agents. *arXiv preprint arXiv:2204.01884*, 2022.

C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort. When the algorithm itself is racist: diagnosing ethical harm in the basic components of software. *Intl. Journal of Communication*, 2016.

Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2):107–194, 2011.

Yonadav Shavit, Benjamin Edelman, and Brian Axelrod. Learning from strategic agents: Accuracy, improvement, and causality. In *International Conference on Machine Learning*. PMLR, 2020.

Zhiqiang Shen, Zechun Liu, Dejia Xu, Zitian Chen, Kwang-Ting Cheng, and Marios Savvides. Is label smoothing truly incompatible with knowledge distillation: An empirical study. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PObuuGVrGaZ.

J. Sill and Y. S. Abu-Mostafa. Monotonicity hints. In *Advances in Neural Information Processing Systems (NIPS)*, pages 634–640, 1997.

P. Singer. *Ethics in the Real World*. Princeton University Press, New York, 2016.

A. Singh and T. Joachims. Fairness of exposure in rankings. In *KDD*, 2018.

Maurice Sion. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.

Shawna N Smith, Heidi Reichert, Jessica Ameling, and Jennifer Meddings. Dissecting leapfrog: how well do leapfrog safe practices scores correlate with hospital compare ratings and penalties, and how much do they matter? *Medical Care*, 55(6):606, 2017.

Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Michael Spence. Job market signaling. In *Uncertainty in Economics*, pages 281–306. Elsevier, 1978.

Jann Spiess. Optimal estimation when researcher and social preferences are misaligned. Working paper, 2018. URL https://gsb-faculty.stanford.edu/jann-spiess/files/2021/01/alignedestimation.pdf.

Jörg Stoye. Minimax regret treatment choice with finite samples. *Journal of Econometrics*, 151(1):70–81, 2009.

Marilyn Strathern. 'Improving ratings': audit in the British university system. *European Review*, 5(3):305–321, 1997.

R. Sugden. Spontaneous order. In *Journal of Economic Perspectives*, 1989.

Lisa M Sullivan, Joseph M Massaro, and Ralph B D'Agostino Sr. Presentation of multivariate data for clinical use: The framingham study risk score functions. *Statistics in medicine*, 23 (10):1631–1660, 2004.

Aleksey Tetenov. An economic theory of statistical testing. Technical report, CeMMAP working paper, 2016.

J. P. Thiroux and K. W. Krasemann. *Ethics Theory and Practice*. Pearson, New York, 2017.

Hal R Varian. Price discrimination and social welfare. *The American Economic Review*, 75 (4):870–875, 1985.

Cédric Villani. *Optimal Transport, Old and New*, volume 338 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, 2009.

Rishi K Wadhera, Jose F Figueroa, Karen E Joynt Maddox, Lisa S Rosenbaum, Dhruv S Kazi, and Robert W Yeh. Quality measure development and associated spending by the centers for medicare & medicaid services. *JAMA*, 323(16):1614–1616, 2020.

Serena Wang and Maya R. Gupta. Deontological ethics by monotonicity shape constraints. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael Jordan. Robust optimization for fairness with noisy protected groups. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 5190–5203, 2020a.

Serena Wang, Stephen Bates, PM Aronow, and Michael I Jordan. On counterfactual metrics for social welfare: Incentives, ranking, and information asymmetry. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023a.

Serena Wang, Harikrishna Narasimhan, Yichen Zhou, Sara Hooker, Michal Lukasik, and Aditya Krishna Menon. Robust distillation for worst-class performance: on the interplay between teacher and student objectives. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 2237–2247. PMLR, 2023b.

Serena Wang, Michael I Jordan, Katrina Ligett, and R Preston McAfee. Information elicitation in agency games. *arXiv preprint arXiv:2402.14005*, 2024.

Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2020b.

Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

Georgia Wells, Jeff Horwitz, and Deepa Seetharaman. Facebook Knows Instagram is Toxic for Teen Girls, Company Documents Show. *The Wall Street Journal*, 14, 2021.

L. Wightman. LSAC national longitudinal bar passage study. *Law School Admission Council*, 1998.

Robert C Williamson, Elodie Vernet, and Mark D Reid. Composite multiclass losses. *Journal of Machine Learning Research*, 17:1–52, 2016.

Michael Woodford. Learning to believe in sunspots. *Econometrica: Journal of the Econometric Society*, pages 277–307, 1990.

Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory (COLT)*, pages 1920–1953, 2017.

Donald J Wright. Price discrimination with transportation costs and arbitrage. *Economics Letters*, 41(4):441–445, 1993.

Liuyu Xiang, Guiguang Ding, and Jungong Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *European Conference on Computer Vision*, pages 247–263. Springer, 2020.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020.

Canwen Xu, Wangchunshu Zhou, Tao Ge, Ke Xu, Julian McAuley, and Furu Wei. Beyond preserved accuracy: Evaluating loyalty and robustness of BERT compression. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

Steve Yadlowsky, Hongseok Namkoong, Sanjay Basu, John Duchi, and Lu Tian. Bounds on the conditional and average treatment effect with unobserved confounding factors. *The Annals of Statistics*, 50(5):2587–2615, 2022.

I-Cheng Yeh and Che hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36:2473–2480, 2009.

S. You, K. Canini, D. Ding, J. Pfeifer, and M. R. Gupta. Deep lattice networks and partial monotonic functions. *Advances in Neural Information Processing Systems (NIPS)*, 2017.

M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness constraints: A mechanism for fair classification. In *ICML Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2015.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research (JMLR)*, 20:1–42, 2019.

M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates. Fa*ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, 2017.

Shaoyu Zhang, Chen Chen, Xiyuan Hu, and Silong Peng. Balanced knowledge distillation for long-tailed learning. *arXiv preprint arXiv:2104.10510*, 2021.