DESIRABLE EFFORT FAIRNESS AND OPTIMALITY TRADE-OFFS IN STRATEGIC LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Strategic learning studies how decision rules interact with agents who may strategically change their inputs/features to achieve better outcomes. In standard settings, models assume that the decision-maker's sole scope is to learn a classifier that maximizes an objective (e.g., accuracy) assuming that agents will best respond. However, real decision-making systems' goals do not always align *exclusively* with producing good predictions. They may need to consider the downstream effects of inducing certain incentives, which translates into certain features being regarded as more *desirable* to change for the decision maker. Not only that, but the principal may also need to incentivize desirable feature changes equally across heterogeneous agents. *How much does this constrained optimization (i.e., maximize the principal's objective, while minimizing the disparity in terms of incentivizing desirable effort) cost the principal?* We propose a unified model of principal-agent interaction that captures this trade-off under three additional components: (1) causal dependencies between features, such that changes in one feature affect others; (2) heterogeneous manipulation costs across the agent population; and (3) peer learning, through which agents infer the principal's algorithm. We provide theoretical guarantees on the principal's optimality loss constrained to a particular desirability fairness tolerance for multiple broad classes of fairness measures. Finally, we demonstrate through experiments on real datasets an explicit tradeoff between maximizing accuracy and fairness in desirability effort.

1 INTRODUCTION

Most automated decision-making systems–whose task is to assign a human agent a numeric score or a classification based on which a *decision* on the individual is made–must contend with the fact that, given sufficient information and ability, agents may strategically alter their features to improve their assigned outcomes, thus jeopardizing the effectiveness of the originally proposed decision-making policies; e.g., a bank's loan-approval algorithm may incentivize applicants to take out additional credit cards even when this does not improve true creditworthiness. Similar human adaptations to automated decision-making rules have been observed in healthcare (Chang et al., 2024; Efthymiou et al., 2025) and recommendation systems (Haupt et al., 2023; Fedorova et al., 2025).

In the classic strategic learning literature (e.g., (Hardt et al., 2015; Dong et al., 2018; Chen et al., 2020)), this is modeled as a Stackelberg game: the principal publicly commits to an accuracy-maximizing algorithm, anticipating that agents will best respond by submitting altered features that maximize their utility about predicted outcomes. In its most classical form, the game is the following: the principal commits to a scoring rule $\mathbf{w} \in \mathbb{R}^d$, the agents observe $\mathbf{w}$ along with their true feature-score pair $(\mathbf{x}, y)$ (where $\mathbf{x} \in \mathbb{R}^d$ and $y \in [0, 1]$), and they report feature $\mathbf{x}'$ (where $\mathbf{x}'$ is their "best-response", i.e., it maximizes their underlying utility function for obtaining a better outcome $\mathbf{w}^\top \mathbf{x}'$). The principal's goal is to identify the optimal $\mathbf{w}$ such that the loss between predictions on the altered features $\mathbf{w}^{\star\top} \mathbf{x}'$ and the true scores y is minimized.

This classical model–albeit great at highlighting *some* of the complexities that arise from the agents' best-responding behavior–fails to capture several of the intricacies of the real-world settings they wish to model, as discussed next. Take for example a content recommendation platform (e.g., YouTube) that algorithmically determines how much to promote a video based on various video features such as topic, title, explicitness etc. Content creators want their videos highly promoted and

can change their videos (to an extent) so that they are more likely to be promoted. This is a nuanced interaction between creators and the platform. On a creator's side, there are several issues that complicate her video changes beyond the perfect best response (i.e., the perfect video edit for maximum promotability): **(1)** She may only know about the algorithm what she/friends can test or learn; **(2)** Her video edits influence each other (e.g., changing topic may also affect how explicit the video is); and **(3)** Some changes really *do* change how popular a video should be, i.e., they are not purely gaming of the recommender system! As for the platform, independent of promotion accuracy, creating incentives to alter certain video features might be (un)desirable. For example, clickbait titles may genuinely make clicking the video more appealing, but becoming known as a platform inundated with clickbait titles (if creators are incentivized to use them) will hurt the platform's reputation to advertisers and users alike. In the same vein, to be neutral in the eyes of advertisers, a platform may further be concerned with *which creators* are more/less incentivized to clickbait.

Altogether, the aforementioned example illustrates several complexities for the modeling of the *agents* and the *principal* in algorithmic decision making systems. For the agents: **(A1)** they may not know the principal's algorithm fully when choosing their best responses; **(A2)** changes in certain features can causally trigger changes in other features; **(A3)** not all feature alterations are "bad"– instead, some represent genuine improvement. And for the principal: **(P1)** independent of their predictive power, some feature changes may be more/less *desirable* for external stakeholders to whom the principal is accountable; and so **(P2)** the principal may need to create equitable incentives to improve desirable features even for heterogeneous agents. Putting the complex considerations **(A1) – (A3)** and **(P1) – (P2)** together and focusing on the principal's perspective, we ask:

How much of his objective value (e.g., optimal accuracy) does the principal trade off to give heterogeneous agents equitable incentives with respect to desirable features?

1.1 Our Contributions

Our model. In Section 2, we present a game-theoretic model capturing the principal-agent interaction with stakeholder input, under properties **(P1), (P2)** for the principal and **(A1)–(A3)** for the agents. We assume that the agent population is comprised by 2 heterogeneous groups. An external stakeholder selects which features are *desirable* for change and the definition of the desirable effort *discrepancy function*; as its name suggests, this is the function measuring how disparate the 2 groups are in terms of desirable effort that they exert when altering their features. The principal in turn chooses a tolerance β capturing the extent to which he is willing to constrain his scoring rule to achieve more equitable incentives according to the discrepancy function set by the stakeholder. He then deploys the optimal (according to either accuracy or social welfare) such constrained scoring rule. Agents, belonging to one of 2 groups, best respond to a peer-learned estimate of the principal's rule by adding "exogenous effort" to their features. An agent's final altered feature set is determined using a causal graph that captures how exogenous feature changes affect each other. While some of these complications have been modeled in prior work, we are the first to study them all together while formally reasoning about the principal's tradeoffs of optimality vs desirable effort discrepancy.

Our goal is to reason about the principal's tradeoffs without restricting ourselves to a single discrepancy function; after all, different stakeholders can have vastly different ways of quantifying discrepancy in terms of desirable effort exerted across different groups. For that, our analysis is split into two parts based on broad families of discrepancy functions, outlined below.

Optimality loss guarantees under convex constraints. In Section 3, we provide theoretical guarantees on the maximum optimality loss (as a function of tolerance) a principal suffers in equilibrium given that the stakeholders' function of desirable effort discrepancy is convex and satisfies certain natural regularity properties; we also provide examples of discrepancy measures that satisfy said properties. In particular, natural vector comparison functions such as sum of squared or absolute value differences are within the classes of functions captured.

Optimality loss guarantees under nonconvex constraints. What if the stakeholders measure desirable effort discrepancy in a way that is asymmetric (i.e., one group may be more desirably incentivized, but not the other way around)? Such a function may be nonconvex and thus yield nonconvex constraints on the principal's problem. In Section 4, we provide the theoretical guarantees for the maximum optimality loss (as a function of tolerance) a principal suffers in equilibrium given the

measure of desirable effort discrepancy selected by the stakeholders is *nonconvex* and satisfies various properties. We also provide natural nonconvex discrepancy measures a stakeholder may be interested in, particularly when one agent group is already privileged over another.

Experiments. In Section 5, we run experiments on the ADULT dataset to map out the desirable effort fairness and optimality trade-off a principal experiences for different tolerance selections under convex discrepancy function examples from Section 3. Because we compute the optimal rule for every β, we can analyze the impact of group disparities on the equilibria. In particular, the results highlight that in cases where groups' disparity aligns with desirability, constraining policies to induce equitable desirable incentives forces the principal to lose more accuracy at a given β-fair equilibrium and increases the maximum tolerance at which the principal is no longer at lower optimal accuracy.

1.2 Summary of related works

Our work is related to streams of literature: *strategic learning* and *algorithmic recourse*. We give a brief overview, and defer a more thorough discussion to Appendix A.1.1.

Strategic learning/classification models (see Podimata (2025) for a review) consider a principal/learner robustness problem, in which the principal wishes to construct an optimal (in terms of accuracy) algorithm under the assumption that agents will have knowledge of this algorithm and "game" their features to earn a better score/classification (Hardt et al., 2015; Dong et al., 2018; Chen et al., 2020; Ahmadi et al., 2021; Trachtenberg & Rosenfeld, 2025; Rosenfeld & Rosenfeld, 2024; Podimata, 2025). However, further work (including ours) complicates this idea by rejecting the assumption that all feature alterations are gaming and studies what this means for various learner and agent perspectives (Efthymiou et al., 2025; Bechavod et al., 2022; Alhanouti & Naghizadeh, 2025; Miller et al., 2020; Shavit et al., 2020; Bechavod et al., 2021; Kleinberg & Raghavan, 2019; Harris et al., 2021; Alon et al., 2020; Haghtalab et al., 2021; Tsirtsis & Gomez-Rodriguez, 2020; Horowitz & Rosenfeld, 2023) Additionally, existing research also considers the added complication that agents may not have full information about the learner's policy and what this means for optimal algorithms, accuracy, agents' feature alterations, and social welfare (Bechavod et al., 2022; Avasarala et al., 2025; Braverman & Garg, 2020; Ghalme et al., 2021; Cohen et al., 2025; Ahmadi et al., 2023; Ebrahimi et al., 2025; Efthymiou et al., 2025) as we do for its impact on equity in induced desirable effort. Finally, there exists a connection between our focus on constraining only to policies that create *fair* incentives and strategic learning analysis that considers other types of fairness induced by the learner's optimal policies (Estornell et al., 2023; Milli et al., 2019; Diana et al., 2025; Alhanouti & Naghizadeh, 2024; Keswani & Celis, 2023; Levanon & Rosenfeld, 2021; Alhanouti & Naghizadeh, 2025). Outside of strategic classification/learning, in algorithmic recourse, rather than a learner who induces feature alterations through the incentives of his algorithm, researchers study how a learner may provide explanations or recommended actions to agents who receive unfavorable scores (von Kügelgen et al., 2020; Ehyaei et al., 2023; Karimi et al., 2021; Perello et al., 2025; Gupta et al., 2019), see Karimi et al. (2022) for a review.

2 Model

2.1 Notation

We use $\mathbb{I}_{\dots}$ as an indicator function s.t. $\mathbb{I}_{\dots} = 1$ if subscript is satisfied and $\mathbb{I}_{\dots} = 0$ otherwise. Matrices are capital (i.e., $C \in \mathbb{R}^{d\times d}$), vectors are lower-case and bolded (i.e., $\mathbf{w} \in \mathbb{R}^d$), and one-dimensional variables are lower-case (i.e., $y \in \mathbb{R}$). $C_{j,i}$ corresponds to the element in the jth row and ith column of a matrix C. We use $\ker(C)$ to denote the *kernel* of a matrix C, i.e., all $\mathbf{w}$ s.t. $C\mathbf{w} = 0$. We use $H(M)$ to denote the Hoffman constant of matrix $M \in \mathbb{R}^{k\times d}$ (see Appendix A.2.1 for details). $\lambda_d(M)$ and $\sigma_d(M)$ are the d-th largest eigenvalues and singular values of M respectively. For a vector, $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{w}_+$ denotes a vector $\in \mathbb{R}^d$, such that $(\mathbf{w}_+)_i = \mathbb{I}_{\mathbf{w}_i\geq 0}\mathbf{w}_i$. We use capital calligraphic letters for sets (e.g., $\mathcal{W}(\cdot,\cdot)$). We use $\mathcal{B}(\rho)$ to denote the Euclidean ball of radius ρ, i.e., $\{\mathbf{w} \in \mathbb{R}^d : \langle \mathbf{w}, \mathbf{w}\rangle \leq \rho\}$. We use $\langle\cdot,\cdot\rangle_G$ and $\|\cdot\|_G$ to be generalized inner products and norms w.r.t to some positive definite matrix G. Thus, $\langle\cdot,\cdot\rangle_G = \|\cdot\|_G^2 = \langle\cdot, G\cdot\rangle$. For a function, f, $\partial f(\mathbf{x})$ denotes its set of subgradients at $\mathbf{x}$. A table of notation can be in Appendix A.2.

Protocol 1 Principal-Agent Interaction with Stakeholder Input

Nature selects $\mathcal{G}$, $\mathbf{w}^\star$, A_1, and A_2. ▷ causal graph, ground truth rule, and cost matrices
Stakeholder selects Π_D and $\Delta(\mathbf{w})$. ▷ desirability scores and discrepancy measure
Principal chooses β. ▷ desirability discrepancy tolerance
Principal deploys an optimal (based on $\texttt{OBJ}(\cdot,\cdot)$) $\mathbf{w}$ subject to $\Delta(\mathbf{w}) \leq \beta$.
Agents draw initial features $\mathbf{x} \sim \mathcal{D}_g$.
Agents alter and reveal features: $\mathbf{x}' \in \arg\max(\texttt{Score}(\mathbf{x}', g) - \texttt{Cost}(\mathbf{x}_e, g))$.

2.2 Model Summary

We begin by summarizing our model's components; in the following subsections, we analyze each of the moving pieces in detail.

We focus on a Stackelberg game between a principal and an agent population comprised by 2 subpopulations, with different distributions over the feature space and movement costs encoded as cost matrices A_1, A_2. An agent (she) belongs to group $g \in \{1, 2\}$ and has initial features $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ drawn from distribution $\mathcal{D}_g$ over the feature subspace $\mathcal{S}_g \subseteq \mathcal{X}$. Let $\Pi_g \in \mathbb{R}^{d\times d}$ be the orthogonal projection matrix into a group subspace. Let $\mathbf{w}^* \in \mathbb{R}^d$ denote the ground truth linear scoring rule; i.e., $\mathbb{E}[y|\mathbf{x}] = \langle \mathbf{w}^*, \mathbf{x}\rangle$ is the expected (true) quality of an agent with features $\mathbf{x} \in \mathcal{X}$. Similarly to (Bechavod et al., 2022), $\mathbf{w}^\star$ may be optimal for *prediction* accuracy, but it may not be optimal for satisfying all of the other principal concerns that we are analyzing.

Altered features. In response to information she has learned about $\mathbf{w}$ from her group, g, (see Section 2.3 for details) an agent alters her feature vector to $\mathbf{x}'$ in order to maximize her anticipated score, $\texttt{Score}(\mathbf{x}', g)$, minus the cost of alteration, $\texttt{Cost}(\mathbf{x}_e, g)$. Following (Efthymiou et al., 2025), we assume that altering one feature may causally affect others thus, $\mathbf{x}' := C\mathbf{x}_e + \mathbf{x}$ where $C \in \mathbb{R}^{d\times d}$ is a "contribution matrix" of a causal graph, $\mathcal{G}$, representing relationships between features and $\mathbf{x}_e \in \mathbb{R}^d$ is the actual (exogenous) effort an agent added to each feature.

External stakeholder & principal's goals. An external stakeholder assigns a "desirability score" to each feature, $\texttt{des}(i) > 0$ where $i \in [d]$; let $\Pi_D := \text{Diag}(\{\texttt{des}(i)\}_{i\in[d]})$ denote the desirability score matrix. Roughly, a higher desirability score for a feature i means that the stakeholder wants to *incentivize* the agent to exert effort and improve i. Additionally, we assume that the external stakeholder selects a function $\Delta(\mathbf{w}) \in [0, 1]$, that measures the group discrepancy in desirable effort incentivized by $\mathbf{w}$. To remain trustworthy to the external stakeholder, the principal chooses a desirability discrepancy tolerance, β, and must find the optimal policy (also referred to as rule) $\mathbf{w}$ (according to their own objective function $\texttt{OBJ}(\mathbf{w}; \mathbf{w}^\star)$); see Section 2.4 for details.

2.3 Agents

In our model, we assume that the agents do *not* have full information about the scoring rule $\mathbf{w}$. Instead, they engage in *peer learning* as modeled in (Bechavod et al., 2022); each agent sees their features $\mathbf{x}_{g,i}$ (chosen by nature) and policy outcomes, $\hat{y}_{g,i} = \langle \mathbf{x}_{g,i}, \mathbf{w}\rangle$, of N_g random (nonstrategic) agents from her own group g and does empirical risk minimization (ERM) to get $\mathbf{w}_{\texttt{EST}}$, the estimated policy; note that this model captures agents that are risk-averse, fully rational, and have no other information about $\mathbf{w}$ except the peer dataset. Formally, $\mathbf{w}_{\texttt{EST}}(g)$ is the solution to the following:

$$\begin{aligned} \min_{\tilde{\mathbf{w}}\in W} \quad & \langle \tilde{\mathbf{w}}, \tilde{\mathbf{w}}\rangle \\ \text{subject to} \quad & W = \{\mathbf{w} : \mathbf{w} = \arg\min_{\mathbf{w}'} \sum_{i\in N_g} (\mathbf{x}_{g,i}^\top \mathbf{w}' - \hat{y}_{g,i})^2\} \end{aligned} \tag{1}$$

In anticipation of the principal's policy $\mathbf{w}$, an agent modifies her features from $\mathbf{x}$ to $\mathbf{x}'$ by exerting exogenous effort, $\mathbf{x}_e \in \mathbb{R}^d$, that is added (after a linear transformation) to her original features. The final $\mathbf{x}'$ is a *best response*, meaning it is the maximizer of her utility function: $\mathcal{U}(\mathbf{x}, \mathbf{x}', g) := \texttt{Score}(\mathbf{x}', g) - \texttt{Cost}(\mathbf{x}_e, g)$ where $\texttt{Score}(\mathbf{x}', g) := \langle \mathbf{w}_{\texttt{EST}}, \mathbf{x}'\rangle$ is the score she believes she will have after modification and $\texttt{Cost}(\mathbf{x}_e, g) := \frac{1}{2}\mathbf{x}_e^\top A_g \mathbf{x}_e$ is the cost she must incur for her exogenous effort, $\mathbf{x}_e$ determined by a known positive definite (PD) cost matrix A_g. Formally, we write $\mathbf{x}'(\mathbf{x}; \mathbf{w}, g)$ to

denote the best-response of an agent from group g with original features $\mathbf{x}$; we drop the dependence on $\mathbf{x}, g$ whenever clear from context.

Relationship between $\mathbf{x}'$ and $\mathbf{x}_e$. Effort on one feature may induce changes in another. We refer to the initial effort as "exogenous", to distinguish it from the "spillover" effects across features. These spillovers are modeled using a weighted causal graph, where nodes represent features. The contribution matrix C captures the weighted flow of effort across features. (Note that C is the transpose of the matrix used in Efthymiou et al. (2025), who adopt a similar causal-flow perspective.)

Definition 2.1 (Contribution matrix) *Given a weighted directed acyclic graph (DAG) $\mathcal{G} = ([d], \mathcal{A}, \omega)$ where $[d]$ is the set of feature-nodes, $\mathcal{A}$ is the set of edges indicating causality between features, and ω is a weight function on the edges, the contribution matrix C is such that:*

$$C_{ii} = 1 \quad \forall i \in [d]$$
$$C_{ij} = \sum_{p \in \mathcal{P}_{ij}} \omega(p) \quad \forall i, j \in [d], i \neq j \tag{2}$$

where $\mathcal{P}_{ij}$ is the set of all direct paths from node i to node j on $\mathcal{G}$ and $\omega(p)$ for $p \in \mathcal{P}_{ij}$ is the sum of the weights along path p, i.e., : $\omega(p) = \sum_{a \in p} \omega(a)$. Note that $\ker(C) = \emptyset$ (Lemma A.1).

Hence, although an agent's exogenous effort is characterized by $\mathbf{x}_e$, the causal interactions between features result in $\mathbf{x}' = \mathbf{x} + C\mathbf{x}_e$, which is ultimately what the principal observes. Putting everything together, we can find the closed form for $\mathbf{x}_e$; see Appendix A.2.2 for the proof.

Proposition 2.1 *Given $\mathbf{w}$, the effort of an agent from group g is $\mathbf{x}_e^{(g)}(\mathbf{w}) = A_g^{-1} C^\top \Pi_g \mathbf{w}$.*

2.4 The Principal's Problem

We compare the Stackelberg equilibrium (SE) of Protocol 1 for a selected β (Eq. (3)) to the Stackelberg equilibrium were there no stakeholder (Eq. (4)). We refer to Equations (3) and (4) as the *fairness-constrained* and *unconstrained* principal problem respectively.

$$\max_{\mathbf{w} \in \mathcal{B}(1)} \texttt{OBJ}(\mathbf{w}; \mathbf{w}^\star), \quad \text{subject to} \quad \Delta(\mathbf{w}) \leq \beta \tag{3}$$

$$\max_{\mathbf{w} \in \mathcal{B}(1)} \texttt{OBJ}(\mathbf{w}; \mathbf{w}^\star) \tag{4}$$

Importantly, these problems have the same objectives and only differ by *feasible region*! The feasible region of the fairness-constrained problem is $\mathcal{B}(1) \cap \{\mathbf{w} \in \mathbb{R}^d : \Delta(\mathbf{w}) \leq \beta\}$ while the other is only $\mathcal{B}(1)$. Notice that because we analyze equilibrium, optimizations are done as if $\mathbf{w}^\star$ is known. If $\mathbf{w}^\star$ is unknown, then one can follow (Bechavod et al., 2022) (Appendix A) and obtain the same results (up to a small error term). To simplify exposition, we stick with the known $\mathbf{w}^\star$ assumption. The principal may choose between two different objectives: Accuracy, `ACC`, or Social Welfare, `SW`. Recall that $\Delta(\mathbf{w})$ is a function capturing the discrepancy (between agent groups) in terms of desirable effort incentivized by policy $\mathbf{w}$; the function $\Delta(\cdot)$ is chosen by the external stakeholder.

Formally, the accuracy and social welfare objectives are defined as follows:

$$\texttt{ACC}(\mathbf{w}; \mathbf{w}^\star) := -\sum_{g \in [2]} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_g} \left[\left(\langle \mathbf{w}^\star, \mathbf{x}'(\mathbf{x}; \mathbf{w}, g) \rangle - \langle \mathbf{w}, \mathbf{x}'(\mathbf{x}; \mathbf{w}, g) \rangle \right)^2 \right]$$
$$\texttt{SW}(\mathbf{w}; \mathbf{w}^\star) := \sum_{g \in [2]} \mathbb{E}_{x \sim \mathcal{D}_g} [\langle \mathbf{x}'(\mathbf{x}; g), \mathbf{w}^\star \rangle]$$

Remark 2.1 *We define accuracy to be the negative of squared loss in order to discuss both problems as a maximization. Our results can viewed equivalently as for a minimization of squared loss.*

2.4.1 Desirability Fairness Constraint

The function $\Delta(\mathbf{w})$–chosen by the external stakeholder–represents the group-discrepancy in incentivized desirable effort induced by $\mathbf{w}$. After choosing tolerance β, the principal constrains his problem such that no deployed rule induces desirable effort incentives whose discrepancy across agents, as measured by $\Delta(\mathbf{w})$, is greater than β. We call the set of $\mathbf{w}$ that satisfy this constraint, β-fair.

Table 1: Upper bounds on opt loss in β-fair SE under Properties 3.1–3.3. Propositions yielding bounds are in Appendix A.3.1. $\tilde{\mathbf{w}} := (CA_1^{-1}C^\top\Pi_1 + CA_2^{-1}C^\top\Pi_2)^\top\mathbf{w}^\star$, $\mathbf{w}' = \mathbb{I}_{\mathbf{w}^\star\in\mathcal{B}(1)}\mathbf{w}^\star + \mathbb{I}_{\mathbf{w}^\star\notin\mathcal{B}(1)}\mathbf{w}^\star/\|\mathbf{w}^\star\|$, $s_{\min} = \sqrt{\beta/\lambda_1(Q)}$ and $r = (\|\mathbf{w}^\star\| - 1)_+$

Prop. 3.1	Prop. 3.2	Prop. 3.3	Accuracy	SW
			$4\big(\|\mathbf{w}^\star\|_2 + 1\big)$	$2\|\tilde{\mathbf{w}}\|_2$
✓			$\big[H(M)\,\|(M\mathbf{w}' - \beta\mathbf{1})_+\|_2\big]^2$	$2\|\tilde{\mathbf{w}}\|_2$
	✓		$4\big(\|\mathbf{w}^\star\|_2 + 1\big)$	$\sqrt{2}$
	✓	✓	$(2r + 1 - s_{\min})(r + 1 - s_{\min})$	$\|\tilde{\mathbf{w}}\|_2 - \sqrt{\beta}\,\|\tilde{\mathbf{w}}\|_{Q^{-1}}$

Definition 2.2 ($\mathcal{W}(\beta;\Delta)$, the set of β-fair rules) $\mathcal{W}(\beta;\Delta) := \{\mathbf{w}\in\mathbb{R}^d : \Delta(\mathbf{w}) \le \beta\}$.

Our theoretical guarantees assume little about $\Delta(\cdot)$ beyond its satisfaction of broad properties, so we briefly provide intuition for natural structures of $\Delta(\cdot)$ that will appear throughout the paper. Recall that $\mathbf{x}_e^{(g)}(\mathbf{w})$ is the exogenous effort vector across features a group-g-agent exerts in best response to what she knows about a policy, $\mathbf{w}$, via peer learning. Thus, $\Pi_D\mathbf{x}_e^{(g)}(\mathbf{w})$ is a *desirability-weighted* effort vector for the agent. Therefore, a natural measure of the desirability discrepancy of a rule is $\Delta(\mathbf{w}) = \texttt{Dist}(\Pi_D\mathbf{x}_e^{(1)}(\mathbf{w}), \Pi_D\mathbf{x}_e^{(2)}(\mathbf{w}))$ where `Dist` is some vector comparison function.

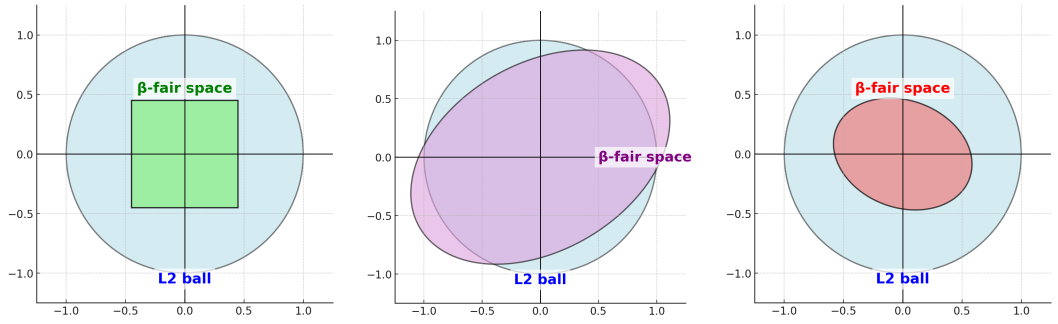
3 Convex Fairness Constraints

In this section, we present optimality loss bounds for the principal's β-fair SE when the discrepancy functions, $\Delta(\cdot)$, are convex in $\mathbf{w}$. We focus on $\Delta(\cdot)$ such that the β-fair rules, $\mathcal{W}(\beta;\Delta)$, form an ellipsoidal or polyhedral feasible region for the fairness-constrained principal problem (Eq. (3)); the formal definitions follow next. Common vector comparison functions (e.g., sum of absolute or squared value differences) create such regions.

Property 3.1 (Constrained feasible region is polyhedral (Fig 1a)) *$\mathcal{W}(\beta;\Delta)$ is such that $\mathcal{B}(1)\cap\mathcal{W}(\beta;\Delta) = \{\mathbf{w} : \mathbf{w}\in\mathbb{R}^d, M\mathbf{w}\le\beta\mathbf{1}\}$ for some $M\in\mathbb{R}^{k\times d}$ and $\mathbf{1}\in\mathbb{R}^k$ where $k\in\mathbb{N}$.*

Property 3.2 (β-fair space is ellipsoidal (Fig 1b)) *$\mathcal{W}(\beta;\Delta) = \{\mathbf{w} : \mathbf{w}\in\mathbb{R}^d, \mathbf{w}^\top Q\mathbf{w}\le\beta\}$ for some $Q\succ 0$.*

Property 3.3 (Constrained feasible region is ellipsoidal (Fig 1c)) *$\mathcal{W}(\beta;\Delta)$ is such that $\mathcal{B}(1)\cap\mathcal{W}(\beta;\Delta) = \{\mathbf{w} : \mathbf{w}\in\mathbb{R}^d, \mathbf{w}^\top Q\mathbf{w}\le\beta\}$ for some $Q\succ 0$.*



(a) Feasible region of the fairness constrained problem (Equation 3) is polyhedral (Property 3.1)

(b) β-fair space is ellipsoidal (Property 3.2)

(c) Feasible region of the fairness constrained problem (Equation 3) is ellipsoidal (Property 3.3)

Figure 1: Examples of β-fair spaces in 2 dimensions that satisfy Properties 3.1 (feasible region polyhedral), 3.2 (β-fair space ellipsoidal), and/or 3.3 (feasible region ellipsoidal)

Table 1 provides upper bounds on accuracy and welfare loss of an algorithmic decision making system in β-fair SE as a function of setting parameters, alteration incentive desirability, and the definition of desirable incentive fairness that stakeholders care about. For a principal, given the stakeholders' discrepancy function satisfies any of the properties, if he has knowledge/estimates of the system parameters (C, A_g, Π_g, and $\mathbf{w}^\star$), he has an estimate of his worst-case loss in the system's equilibrium. Note, several bounds are in terms of his discrepancy tolerance. Meaning, he may use these to get (worst case) trade-offs he will suffer for selecting different βs when this system reaches equilibrium! To concretely illustrate these guarantees, we present numerical examples A.1 and A.2.

3.1 Examples of property-satisfying constraints

Properties 3.1–3.3 are not niche. Polyhedral/ellipsoidal spaces form when the discrepancy function, Δ, compares $\Pi_D \mathbf{x}_e^{(1)}(\mathbf{w})$ and $\Pi_D \mathbf{x}_e^{(2)}(\mathbf{w})$ (desirability-weighted effort vectors) using a sum of either the absolute or squared values of the difference. Consider Examples 3.1 and 3.2.

Example 3.1 (Sum of absolute value differences) *Let the space of β-fair policies be defined as:* $\mathcal{W}(\beta;\Delta) := \{\mathbf{w} : \mathbf{w} \in \mathbb{R}^d, \Delta(\mathbf{w}) \leq \beta\}$, *where* $\Delta(\mathbf{w}) := \sum_{i\in[d]} \left|(\Pi_D \mathbf{x}_e^{(1)}(\mathbf{w}) - \Pi_D \mathbf{x}_e^{(2)}(\mathbf{w}))_i\right|$

Example 3.2 (Sum of squared differences) *Let the space of β-fair policies be defined as:* $\mathcal{W}(\beta;\Delta) := \{\mathbf{w} : \mathbf{w} \in \mathbb{R}^d, \Delta(\mathbf{w}) \leq \beta\}$ *where* $\Delta(\mathbf{w}) := \sum_{i\in[d]} (\Pi_D \mathbf{x}_e^{(1)}(\mathbf{w}) - \Pi_D \mathbf{x}_e^{(2)}(\mathbf{w}))_i)^2$

In Example 3.1, the feasible region for the fairness constrained problem forms a polyhedron (i.e., Property 3.1). Likewise, in Example 3.2, the set of β-fair rules form an ellipsoid (i.e., Property 3.2). Technically, both of these require mild regularity conditions (Appendices A.3.2 and A.3.3): the kernel of a product of setting parameters must be empty. However, should not be hard to satisfy as randomness in matrix entries (e.g., that which is induced by noise in estimating parameters) generally induces non-singularity with high probability.

4 Non-Convex Fairness Constraints

What if the stakeholder wants a non-convex desirability discrepancy function, Δ? This may happen if the stakeholder compares desirability effort vectors, $\Pi_D \mathbf{x}_e^{(1)}(\mathbf{w})$ and $\Pi_D \mathbf{x}_e^{(2)}(\mathbf{w})$, asymmetrically, which is natural if one group is already privileged. We present optimality loss bounds for β-fair Stackelberg equilibrium (SE) when the space of β-fair policies, $\mathcal{W}(\beta;\Delta)$, belongs to class $\mathcal{F}$:

Definition 4.1 ($\mathcal{F}$, a class of nonconvex fairness constraints) $\mathcal{W}(\beta;\Delta) \in \mathcal{F}$ *if the following is true for some $Q \in \mathbb{R}^{d\times d}$ and $Q \succ 0$:*

- $\Delta(\mathbf{w}) = \langle \mathbf{w}, \mathbf{w}\rangle_Q - f(\mathbf{w})$ *where* $f : \mathbb{R}^d \to \mathbb{R}$ *is nonnegative in* $\mathbf{w}$
- $\beta \leq \lambda_d(Q)$

$\mathcal{F}$ is a class of β-fair spaces such that Δ is a generalized inner product minus a positive function. Thus, Δ is (generally) nonconvex. $\mathcal{F}$ ensures a simple, nonempty ellipsoidal restriction of the β-fair space, allowing us to get optimality loss bounds using Property 3.3 from Section 3. While nonnegative assumptions on f may seem strong, recall that Δ measures the desirability discrepancy of effort vectors. As we will see in Example 4.1, a function defined by norms, which is a natural measurement of the size of $\Pi_D \mathbf{x}_e^{(g)}(\mathbf{w})$, the desirability effort vector, fits this definition.

4.1 Optimality loss bounds

We can easily restrict any $\mathcal{W}(\beta;\Delta) \in \mathcal{F}$ to a nonempty ellipsoid, $\mathcal{E}(\beta) = \{\mathbf{w} \in \mathbb{R}^d : \mathbf{w}^\top Q \mathbf{w} \leq \beta\}$, inside the feasible region of the principal's fairness constrained problem (Appendix A.4.1). Invoking Table 1 bounds for internal ellipsoids (Property 3.3) we have:

$$\text{Accuracy loss} \leq (2r + 1 - s_{\min})(r + 1 - s_{\min}) \quad \text{and} \quad \text{SW loss} \leq \|\tilde{\mathbf{w}}\|_2 - \sqrt{\beta}\|\tilde{\mathbf{w}}\|_{Q^{-1}} \tag{5}$$

Equation 5 is an upper bound to accuracy and welfare loss of an algorithmic decision making system in β-fair SE. Given the stakeholders' discrepancy function satisfies (1) in Definition 4.1, for any tolerance that satisfies (2), the principal has an estimate of worst-case optimality loss in the system's equilibrium (when he has knowledge/estimates of system parameters). Because bounds are in terms of the tolerance, he thus has a (worst-case) trade-off he will suffer for β selection when this system reaches equilibrium!

4.2 Example of such a nonconvex restriction

In Section 3, we present discrepancy functions based on sum of square or absolute value differences (Examples 3.1, 3.2). They share two traits: (1) Desirability unfairness is constrained *symmetrically*, (2) Difference in desirable effort is calculated *by feature* then summed. Although (1) seems natural, a group g might be already privileged, making it only worthwhile to intervene if group g' is poorly incentivized. Additionally, (2) ensures convexity, but may be too granular if a stakeholder cares about *overall* desirability. The Δ of Example 4.1 represents an alternative to both of these traits.

Example 4.1 (Asymmetric desirability fairness) *Let the space of β-fair rules be defined as:* $\mathcal{W}(\beta;\Delta) := \{\mathbf{w} : \mathbf{w} \in \mathbb{R}^d, \Delta(\mathbf{w}) \leq \beta\}, \quad \Delta(\mathbf{w}) := \|\Pi_D \mathbf{x}_e^{(g)}(\mathbf{w})\|_2^2 - \|\Pi_D \mathbf{x}_e^{(g')}(\mathbf{w})\|_2^2$

In Example 4.1, $\Delta(\mathbf{w}) > 0$ when group g is more desirably incentivized. Thus, upper bounding by β means that the principal's rules cannot better incentivize group g than g' by more than β, but better incentivizing group g' is fine. This may be useful if group g is already externally privileged. When Δ is defined as described in Example 4.1, we have that $\mathcal{W}(\beta;\Delta) \in \mathcal{F}$ meaning bounds of Equation 5 hold! Technically, this requires mild regularity conditions formally presented in Appendices A.4.2, but they are not hard to satisfy. What are these conditions intuitively? Equation 5 will hold as long as the following is true: (1) the feature subspace of the privileged group, g, spans the whole space (2) principal's tolerance toward group g being more desirably incentivized is not too big.

5 Experimental Evaluation

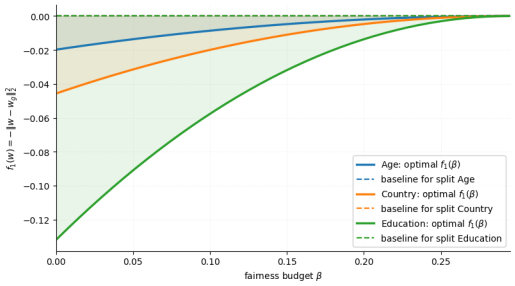
In Sections 3 and 4, we presented upper bounds on principal trade-offs. However, it is unclear how much group disparities (i.e., along subspace, Π_g, and cost, A_g) change the impact of β-desirability fairness. On the `ADULT` dataset, we analyze how the same constraint impacts the principal's SE optimal value were groups variously disparate. We see that when disparity aligns with desirability, imposing β-fairness particularly hits the principal's optimal accuracy harder than when groups are disparate randomly. Interestingly, social welfare is less sensitive disparity/desirability correlation.

5.1 Experimental Setup

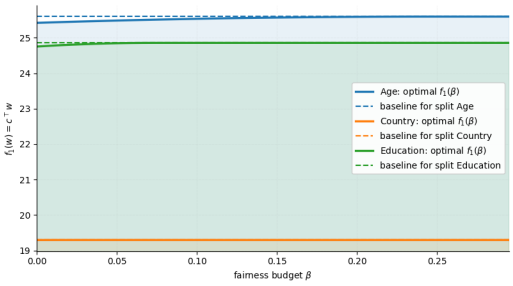
We use the `ADULT` dataset and study 3 agent groupings. By: age, country, and education level. These are used to form 3 sets of Π_1 and Π_2. We then compute the β-fair equilibrium for various cost matrices, A_1, A_2, and desirable feature sets, Π_D. Further details are in Appendix A.5.1

5.2 Results

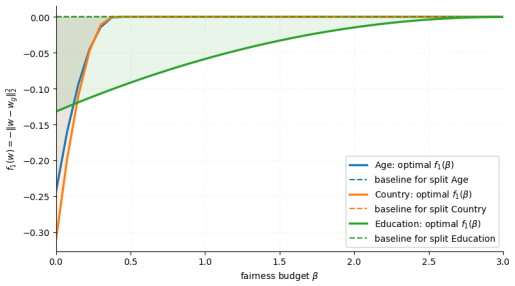
Figure 2 plots of the optimal value at various β-fair equilibria using the discrepancy function of Example 3.1, henceforth called the ℓ_1-fairness constraint, under different groupings and cost disparities. In general, as the tolerance, β, increases, the fairness constraint relaxes and the optimal value (accuracy, social welfare) increases. In Figure 2a, the *Education* split, which separates agents into well and less educated groups, is consistently the most constrained, its curve starts at the lower point and improves most slowly. In this case, the desirable attributes (education, workclass, occupation) are very aligned with this group disparity, so the discrepancy function, Δ, penalizes movements along the most predictive directions. Interestingly, from Figure 2b, we see that Social Welfare is not very sensitive to disparities in Π_g as for all groups, the constrained optimal value approaches the unconstrained very quickly. We then introduce group disparity by cost ($A_1 \neq A_2$). In Figure 2c, we see that optimality loss for the *Education* split retains the same shape, though the maximum β until recovery has increased. Meanwhile, both *Country* and *Age* have significant optimality loss at the tightest fairness constraints, though recover 0 accuracy much faster than *Education*. This behavior



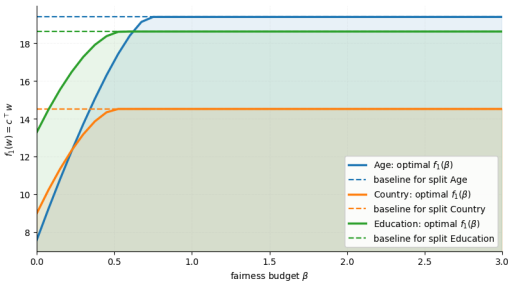
(a) *Accuracy comparison*: **Uniform costs** $A_1 = A_2$; **desirable**: occupation, workclass, education, under ℓ_1-fairness constraint



(b) *Social-welfare comparison*: **Uniform costs** $A_1 = A_2$; **desirable**: occupation, workclass, education, under ℓ_1-fairness constraint



(c) *Accuracy comparison*: **Non-uniform costs** $A_1 = I, A_2 = 2A_1$; **desirable**: occupation, workclass, education, under ℓ_1-fairness constraint



(d) *Social-welfare comparison*: **Non-uniform costs** $A_1 = I, A_2 = 2A_1$; **desirable**: occupation, workclass, education, under ℓ_1-fairness constraint

Figure 2: Optimal value in various β-fair equilibria.

follows from the conclusions under uniform costs. The *Education* split is actually correlated with desirable characteristics and thus fairness constraints continue to have long lasting effects even at high β. Now that disparity has increased, optimality in β-fair equilibria for *Country* and *Age* suffers at the harshest constraints, but as they are not closely aligned with desirability, it does not impact looser constraints. In Figure 2d we continue to see a similar effect. As Social welfare is not so sensitive to disparities in Π_g (Fig 2b), creating similar cost disparities on all groups similar invokes optimality loss for all groups, but there is no relative difference. In the appendix (A.5.2) we include experiements under ℓ_2-fairness constraint (see Example 3.2 for the definition of this constraint) and random cost matrices.

6 Discussion

We formalize and study the problem of guaranteeing that strategic principals induce *equitable* incentives across heterogeneous agents. We do so by analyzing the trade-offs a principal may take in an algorithmic decision making system that must fairly incentivize heterogeneous agents toward changes that have external value outside of the chosen outcome. Theoretically, we provide guarantees on the principal's maximum loss in the system's Stackelberg Equilibrium were he to commit to providing fair incentives. In an empirical study, we map the optimality loss a fairly-incentivizing principal suffers in the β-fair equilibrium of a real setting. We see that for an accuracy-maximizing principal, fair incentivization "hurts more" when agents are disparate in a way that is aligned with the alterations that are externally important to incentive. There are a couple natural avenues for future work. (1) In order to compute their best response, agents should have access to the causal graph, this may be replaced with looser assumptions such as estimates or even group-disparate misspecified beliefs. (2) While we focus on equilibria, they may be non-trivial to reach, thus it would be interesting to consider the [online] learning perspective.

REFERENCES

Saba Ahmadi, Hedyeh Beyhaghi, Avrim Blum, and Keziah Naggita. The strategic perceptron. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, EC '21, pp. 6–25, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385541. doi: 10.1145/3465456.3467629. URL https://doi.org/10.1145/3465456.3467629.

Saba Ahmadi, Avrim Blum, and Kunhe Yang. Fundamental bounds on online strategic classification. In *Proceedings of the 24th ACM Conference on Economics and Computation*, EC '23, pp. 22–58, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701047. doi: 10.1145/3580507.3597818. URL https://doi.org/10.1145/3580507.3597818.

Sura Alhanouti and Parinaz Naghizadeh. Could anticipating gaming incentivize improvement in (fair) strategic classification? In *2024 IEEE 63rd Conference on Decision and Control (CDC)*, pp. 6028–6035, 2024. doi: 10.1109/CDC56724.2024.10886604.

Sura Alhanouti and Parinaz Naghizadeh. Anticipating gaming to incentivize improvement: Guiding agents in (fair) strategic classification, 2025. URL https://arxiv.org/abs/2505.05594.

Tal Alon, Magdalen Dobson, Ariel Procaccia, Inbal Talgam-Cohen, and Jamie Tucker-Foltz. Multiagent Evaluation Mechanisms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(02):1774–1781, April 2020. doi: 10.1609/aaai.v34i02.5543. URL https://ojs.aaai.org/index.php/AAAI/article/view/5543.

Srikanth Avasarala, Serena Wang, and Juba Ziani. The disparate effects of partial information in bayesian strategic learning, 2025. URL https://arxiv.org/abs/2506.00627.

Yahav Bechavod, Katrina Ligett, Steven Wu, and Juba Ziani. Gaming helps! learning from strategic interactions in natural dynamics. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1234–1242. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/bechavod21a.html.

Yahav Bechavod, Chara Podimata, Steven Wu, and Juba Ziani. Information discrepancy in strategic learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1691–1715. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/bechavod22a.html.

Mark Braverman and Sumegha Garg. The Role of Randomness and Noise in Strategic Classification. In Aaron Roth (ed.), *1st Symposium on Foundations of Responsible Computing (FORC 2020)*, volume 156 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 9:1–9:20, Dagstuhl, Germany, 2020. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-142-9. doi: 10.4230/LIPIcs.FORC.2020.9. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.FORC.2020.9.

Trenton Chang, Lindsay Warrenburg, Sae-Hwan Park, Ravi Parikh, Maggie Makar, and Jenna Wiens. Who's gaming the system? a causally-motivated approach for detecting strategic adaptation. *Advances in Neural Information Processing Systems*, 37:42311–42348, 2024.

Yiling Chen, Yang Liu, and Chara Podimata. Learning strategy-aware linear classifiers. *Advances in Neural Information Processing Systems*, 33:15265–15276, 2020.

Silvia Chiappa and Thomas P. S. Gillam. Path-specific counterfactual fairness, 2018. URL https://arxiv.org/abs/1802.08139.

Lee Cohen, Saeed Sharifi-Malvajerdi, Kevin Stangl, Ali Vakilian, and Juba Ziani. Bayesian strategic classification. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2025. Curran Associates Inc. ISBN 9798331314385.

Emily Diana, Saeed Sharifi-Malvajerdi, and Ali Vakilian. Minimax group fairness in strategic classification. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 753–772, 2025. doi: 10.1109/SaTML64287.2025.00047.

Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 55–70, 2018.

Raman Ebrahimi, Kristen Vaccaro, and Parinaz Naghizadeh. The double-edged sword of behavioral responses in strategic classification: Theory and user studies. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pp. 868–886, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714825. doi: 10.1145/3715275.3732056. URL https://doi.org/10.1145/3715275.3732056.

Valia Efthymiou, Chara Podimata, Diptangshu Sen, and Juba Ziani. Incentivizing desirable effort profiles in strategic classification: The role of causality and uncertainty, 2025. URL https://arxiv.org/abs/2502.06749.

Ahmad-Reza Ehyaei, Amir-Hossein Karimi, Bernhard Schoelkopf, and Setareh Maghsudi. Robustness implies fairness in causal algorithmic recourse. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pp. 984–1001, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594057. URL https://doi.org/10.1145/3593013.3594057.

Andrew Estornell, Sanmay Das, Yang Liu, and Yevgeniy Vorobeychik. Group-fair classification with strategic agents. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pp. 389–399, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594006. URL https://doi.org/10.1145/3593013.3594006.

Ekaterina Fedorova, Madeline Kitch, and Chara Podimata. User altruism in recommendation systems. *arXiv preprint arXiv:2506.04525*, 2025.

Ganesh Ghalme, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen, and Nir Rosenfeld. Strategic classification in the dark. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3672–3681. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/ghalme21a.html.

Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubramanian. Equalizing recourse across groups, 2019. URL https://arxiv.org/abs/1909.03166.

Nika Haghtalab, Nicole Immorlica, Brendan Lucier, and Jack Z. Wang. Maximizing welfare with incentive-aware evaluation mechanisms. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021. ISBN 9780999241165.

Moritz Hardt, Nimrod Megiddo, Christos H. Papadimitriou, and Mary Wootters. Strategic classification. *CoRR*, abs/1506.06980, 2015. URL http://arxiv.org/abs/1506.06980.

Keegan Harris, Hoda Heidari, and Zhiwei Steven Wu. Stateful strategic regression. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Andreas Haupt, Dylan Hadfield-Menell, and Chara Podimata. Recommending to strategic users. *arXiv preprint arXiv:2302.06559*, 2023.

Alan J. Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49(4):263–265, 1952.

Guy Horowitz and Nir Rosenfeld. Causal strategic classification: a tale of two shifts. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 259–268, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287597. URL https://doi.org/10.1145/3287560.3287597.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 353–362, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445899. URL https://doi.org/10.1145/3442188.3445899.

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Comput. Surv.*, 55(5), December 2022. ISSN 0360-0300. doi: 10.1145/3527848. URL https://doi.org/10.1145/3527848.

Vijay Keswani and L. Elisa Celis. Addressing strategic manipulation disparities in fair classification. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400703812. doi: 10.1145/3617694.3623252. URL https://doi.org/10.1145/3617694.3623252.

Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation*, EC '19, pp. 825–844, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367929. doi: 10.1145/3328526.3329584. URL https://doi.org/10.1145/3328526.3329584.

Sagi Levanon and Nir Rosenfeld. Strategic classification made practical. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6243–6253. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/levanon21a.html.

John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Smitha Milli, John Miller, Anca D. Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 230–239, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287576. URL https://doi.org/10.1145/3287560.3287576.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Nicholas Perello, Cyrus Cousins, Yair Zick, and Przemyslaw Grabowicz. Discrimination induced by algorithmic recourse objectives. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pp. 1653–1663, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714825. doi: 10.1145/3715275.3732110. URL https://doi.org/10.1145/3715275.3732110.

Javier Peña, Juan C. Vera, and Luis F. Zuluaga. New characterizations of Hoffman constants for systems of linear constraints. *Mathematical Programming*, 187(1):79–109, May 2021. ISSN 1436-4646. doi: 10.1007/s10107-020-01473-6. URL https://doi.org/10.1007/s10107-020-01473-6.

Chara Podimata. Incentive-aware machine learning; robustness, fairness, improvement & causality. *arXiv preprint arXiv:2505.05211*, 2025.

Elan Rosenfeld and Nir Rosenfeld. One-shot strategic classification under unknown costs. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Yonadav Shavit, Benjamin L. Edelman, and Brian Axelrod. Causal strategic linear regression. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Benyamin Trachtenberg and Nir Rosenfeld. Strategic classification with non-linear classifiers, 2025. URL https://arxiv.org/abs/2505.23443.

Stratis Tsirtsis and Manuel Gomez-Rodriguez. Decisions, counterfactual explanations and strategic behavior. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Julius von Kügelgen, Umang Bhatt, Amir-Hossein Karimi, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. On the fairness of causal algorithmic recourse. *CoRR*, abs/2010.06529, 2020. URL https://arxiv.org/abs/2010.06529.

A Supplemental Material

A.1 Supplemental material for Section 1

A.1.1 Supplemental material for Section 1.2

Strategic learning/classification models (see Podimata (2025) for a review) consider a learner (or principal) robustness problem, in which the learner must construct an optimal (according to accuracy or loss) algorithm under the assumption that agents will have knowledge of this algorithm and "game" their features to earn a better score/classification subject to some (potentially unknown to the learner) costs (Hardt et al., 2015; Dong et al., 2018; Chen et al., 2020; Ahmadi et al., 2021; Trachtenberg & Rosenfeld, 2025; Rosenfeld & Rosenfeld, 2024). However, as discussed by Miller et al. (2020), who create a formal *causal* framework in which to study strategic learning, agents' feature alterations are not always gaming. To this end, further work (including ours) complicates this idea by rejecting the assumption that all feature alterations are gaming. Several works consider different optimal algorithms for the learner in a causal setting. Horowitz & Rosenfeld (2023) present maximally accurate algorithms for learner under a setting in which agents' features may be causal, non-causal, or unobserved. Shavit et al. (2020) find a learner algorithm that, under similar causality assumptions, maximizes agent outcomes and show that if all causal features are observed, estimates of ground truth model parameters are improved. Bechavod et al. (2021) show that in an online setting with "meaningful" features, the learner can use strategic behavior to purposefully uncover those features that lead to real improvements and thus purposefully incentivize them. Others, directly create learner rules or algorithms to incentivize good feature alterations. Kleinberg & Raghavan (2019) present algorithms for the learner which allow him to incentivize chosen "good" effort profiles. While Harris et al. (2021) consider a similar learner problem of incentivizing effort, but in a setting where the agent and principal interact multiple times and the agent's effort accumulates to form different states. Another perspective is mechanism design under strategic behavior (Alon et al., 2020; Haghtalab et al., 2021),which aims to promote beneficial effort—and, in some cases, provide counterfactual explanations that guide individuals on how to improve outcomes (Tsirtsis & Gomez-Rodriguez, 2020). While, like all of these, our work also allows agents to genuinely change their features rather than game, our approach differs in that we take an equilibrium analysis perspective i.e., we study the learner's optimal value rather than provide specific algorithms, and we study a learner who must provide *fair* incentives to improvement across agents. Most relevant to our model's approach toward improvement/causality is Efthymiou et al. (2025), who study agents' desirable effort profiles when they must determine how to make alterations to their features given features causally affect each other and potentially partial information on both the causal graph and the learner's rule. Though we similarly model causal flow between features and analyze effort desirability, our focus and results are on inducing *fair* desirable incentives from the *principal's perspective*.

Models also consider the added complication that agents may not have full information about the learner's policy. One such work that also studies improvements (rather than gaming) is Bechavod et al. (2022). They find that in cases where agents must learn about the learner's rule from peers, under some types of group disparity, the learner's optimal policy may induce deterioration (i.e. feature manipulations that cause agents to have worse ground truth outcomes). Our work uses their model of peer learning but adds causal flow between features to study how the use of desirable (exogenous) effort constraints on the learner's policy impact his optimal value. Additional relevant models of agent's incomplete information include (Ebrahimi et al., 2025) where they explore how agent's biases/misconception affect their ability to best-respond, (Ghalme et al., 2021) where similarly with us they consider a scenario where a classifier is not publicly available, and (Avasarala et al., 2025) where agents have access only to a noisy signal of the rule. In addition, there are works that examine strategic classification under the lens of Bayesian theory (see (Cohen et al., 2025)) where agents hold priors over a class of possible classification rules. Another related line of research studies the deployment of random classifiers, which is a natural resource of incomplte infomation and studies its effect on the classification quality as well as agent's ability to best-respond (Braverman & Garg, 2020; Ahmadi et al., 2023).

Finally, there exists a natural connection between our focus on constraining only to policies that induce *fair* incentives and strategic learning analysis that considers other types of fairness induced by the learner's optimal policies. Most related, (Alhanouti & Naghizadeh, 2024) find that in a strate-

gic classification setting, imposing fairness constraints that equalize true positive rates or equalize acceptance rates actually hurts incentive equity between advantaged and disadvantaged agents. Our work attempts to address the issue they document in directly constraining (desirable) incentive discrepancy between agent groups. Other works include Estornell et al. (2023) and Ahmadi et al. (2021) who show that traditionally fair classifiers experience no longer achieve these goals when considered in a strategic setting, Milli et al. (2019) and Hu et al. (2019) who analyze the disparate effects of an optimal strategic learning rule on heterogeneous agents. And Diana et al. (2025), which deals with fairness from an algorithmic point of view, present learner algorithms that provide optimal rules subject to minimax fairness constraints.

Outside of strategic classification/learning, in algorithmic recourse, rather than a learner who induces feature alterations through the incentives of his algorithm, researchers study how a learner may provide explanations or recommended actions to agents who receive unfavorable scores (von Kügelgen et al., 2020; Ehyaei et al., 2023; Karimi et al., 2021; Perello et al., 2025; Gupta et al., 2019), see Karimi et al. (2022) for a review. While these settings similarly consider agents who may change their features, our model (and strategic classification as a whole) takes a more mechanism design approach in that the learner indirectly creates "recourse" for agents by incentivizing changes exclusively through deploying the algorithm or policy rather than direct recommendations to agents.

A.2 Supplemental material for Section 2

Table 2: Notation Table

Symbol	Meaning
d	dimension of features
C	Contribution matrix
A_g	group g cost matrix
Π_g	group g projection matrix
Π_D	diagonal desirability score matrix
$\texttt{des}(i)$	feature i desirability score
g	group
$\mathcal{D}_g$	group g distribution of features
$\mathcal{S}_g$	group g feature subspace
$\mathbf{x}$	initial feature
$\mathbf{x}'$	altered feature
$\mathbf{x}_e$	exogenous effort
$\mathbf{w}$	principal's rule
$\mathbf{w}^\star$	ground truth rule
$\texttt{Score}(\mathbf{x}', g)$	group g agent's estimated score
$\texttt{Cost}(\mathbf{x}_e; g)$	agent cost for exerting effort $\mathbf{x}_e$
$\Delta(\mathbf{w})$	discrepancy in desirable effort incentivized by $\mathbf{w}$
$\mathcal{G}$	causal graph
$\mathcal{A}$	set of edges
ω	edge weights
$\mathcal{P}_{ij}$	set of all direct paths from node i to node j
$\omega(p)$	sum of weights on path $p \in \mathcal{P}_{ij}$
β	principal's discrepancy tolerance
$\mathcal{U}(\mathbf{x}, \mathbf{x}', g)$	group g agent's utility as a function alteration
$\mathcal{B}(1)$	euclidean ball with radius 1
$\texttt{ACC}(\mathbf{w}, \mathbf{w}^\star)$	accuracy of rule $\mathbf{w}$
$\texttt{SW}(\mathbf{w}, \mathbf{w}^\star)$	social welfare of rule $\mathbf{w}$
$\mathcal{W}(\beta; \Delta)$	set of β-fair rules
$\Pi_D \mathbf{x}_e^{(g)}(\mathbf{w})$	desirability-weighted effort vector for group g agent
$H(M)$	Hoffman constant of matrix M
$\mathcal{E}(\tilde{\beta})$	ellipsoidal restriction of a $\mathcal{W}(\beta; \Delta) \in \mathcal{F}$

A.2.1 SUPPLEMENTAL MATERIAL FOR SECTION 2.1

Hoffman constant. We use $H(M)$ to denote the Hoffman constant of matrix $M \in \mathbb{R}^{k\times d}$, i.e., a constant such that $\forall \mathbf{b} \in \mathcal{M} + \mathbb{R}^k_{\geq 0}$ and $\forall \mathbf{z} \in \mathbb{R}^d$ it is true that

$$\texttt{Dist}(\mathbf{z}, P_A(\mathbf{b})) \leq H(M)\|(M\mathbf{z} - \mathbf{b})_+\|_2$$

where $\texttt{Dist}(\mathbf{z}, P_A(\mathbf{b})) := \min\{\|\mathbf{z} - \mathbf{x}\|_2 : \mathbf{x} \in P_A(\mathbf{b})$. $\mathcal{M} := \{M\mathbf{w} : \mathbf{w} \in \mathbb{R}^d\}$ and $P_A(\mathbf{b}) := \{\mathbf{w} \in \mathbb{R}^d : M\mathbf{w} \leq \mathbf{b}\}$.

In particular, this is a Hoffman constant for $p = 2$ (i.e., $\texttt{Dist}$ is the $l2$ norm). This is an equivalent definition to the one used by Peña et al. (2021).

A.2.2 SUPPLEMENTAL MATERIAL FOR SECTION 2.3

Lemma A.1 (Contribution matrix of a DAG is invertible and kernel zero) *Let C be the contribution matrix of a DAG. Then $\ker(C) = \emptyset$. Equivalently, C is invertible.*

Proof of Lemma A.1. Recall that for some exogenous effort $\mathbf{x}_e \in \mathbb{R}^d$, we have the post-causality effort $\mathbf{x} := C\mathbf{x}_e$. We shall prove $\ker(C) = \emptyset$ by contradiction. Suppose $\ker(C) \neq \emptyset$, then it must be the case that there exists $\mathbf{x}_e$ where $\mathbf{x}_e \neq \mathbf{0}$, s.t. $C\mathbf{x}_e = \mathbf{0}$ and thus this exogenous effort "cancels itself out". Let $\mathbf{x}_e$ be a nonzero vector in $\ker(C)$ and define $\mathcal{I} := \{i \in [d] : x_{e_i} \neq 0\}$. Because $C\mathbf{x}_e = \mathbf{0}$, $\forall i \in \mathcal{I}$, node i in the causal graph must have at least one in-degree from some $j \in \mathcal{I}$ or else there is no way that $\mathbf{c_i}^\top \mathbf{x}_e = 0$. To see this, recall that by construction, a row $\mathbf{c_i}$, of C is made up of paths *into* node i and $C_{i,i} = 1$. Consider the subgraph, $\tilde{\mathcal{G}}$, represented by the collection of nodes in $\mathcal{I}$ and the edges between them. Each of these nodes have at least one in-degree from another in the subgraph. Thus, no node in the finite directed subgraph has 0 in-degree. Clearly this means there must be cycle because if we consider traversing the graph from any vertex, we must eventually repeat a vertex as they are finite and all have an in-degree. However, this poses a contradiction to our assumption that C comes from a DAG. Therefore, it must be the case that $\ker(C) = \emptyset$. □

Proof of Proposition 2.1. From Lemma 3.1 of Bechavod et al. (2022), agents' estimate of $\mathbf{w}_{est}$ can be solved in closed form as a function of $\Pi_g, \mathbf{w}$: $\mathbf{w}_{est}(g) = \Pi_g \mathbf{w}$. Thus:

$$\begin{aligned}\mathcal{U}(\mathbf{x}, \mathbf{x}'; g) :&= \langle \Pi_g \mathbf{w}, \mathbf{x}' \rangle - \frac{1}{2}||\sqrt{A_g}(\mathbf{x}_e)||^2 \\ &= \langle \Pi_g \mathbf{w}, \mathbf{x} + C\mathbf{x}_e \rangle - \frac{1}{2}||\sqrt{A_g}(\mathbf{x}_e)||^2\end{aligned}$$

This function should be concave (sum of 3 concave functions: a constant plus a linear term minus a norm) and hence:

$$\begin{aligned}\nabla \mathcal{U}(\mathbf{x}, \mathbf{x}'; g) = C^\top \Pi_g \mathbf{w} - A_g \mathbf{x}_e = 0 \iff \\ \mathbf{x}_e = A_g^{-1} C^\top \Pi_g \mathbf{w}\end{aligned}$$

Therefore the best-response is:

$$\mathbf{x}'(\mathbf{x}; g) = \mathbf{x} + CA_g^{-1}C^\top \Pi_g \mathbf{w} \tag{6}$$

□

A.2.3 SUPPLEMENTAL MATERIAL FOR SECTION 2.4

Lemma A.2 (Equivalent Accuracy Objective) *An accuracy-maximizing principal can solve either Problem 3 or 4 using the following objective to find the optimal $\mathbf{w}_{\texttt{ACC}}$ policy in equilibrium*

$$\max_{\mathbf{w}\in\mathbb{R}^d} \quad -\|\mathbf{w}^\star - \mathbf{w}\|_2^2 \tag{7}$$

Proof of Lemma A.2. Using the solution from 2.1

$$
\begin{aligned}
-\text{ACC} &= \sum_{g\in[2]} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_g}\left[\left(\mathbf{w}^{\star\top}\hat{\mathbf{x}}(\mathbf{x};g) - \mathbf{w}^{\top}\hat{\mathbf{x}}(\mathbf{x};g)\right)^2\right] \\
&= \sum_{g\in[2]} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_g}\left[(\mathbf{w}^{\star\top}\hat{\mathbf{x}}(\mathbf{x};g))^2 + (\mathbf{w}^{\top}\hat{\mathbf{x}}(\mathbf{x};g))^2 - 2(\mathbf{w}^{\star\top}\hat{\mathbf{x}}(\mathbf{x};g))(\mathbf{w}^{\top}\hat{\mathbf{x}}(\mathbf{x};g))\right] \\
&= \langle \mathbf{w}^\star, \mathbf{w}^\star\rangle + \langle \mathbf{w}, \mathbf{w}\rangle - 2\langle \mathbf{w}^\star, \mathbf{w}\rangle \\
&= \langle \mathbf{w}^\star - \mathbf{w}, \mathbf{w}^\star - \mathbf{w}\rangle \\
&= \|\mathbf{w}^\star - \mathbf{w}\|_2^2
\end{aligned}
$$

□

Lemma A.3 (Equivalent SW Objective) *A social-welfare-maximizing principal can solve either Problem 3 or 4 using the following objective to find the optimal* $\mathbf{w}_{\text{SW}}$ *policy in equilibrium*

$$\max_{\mathbf{w}\in\mathbb{R}^d} \quad \langle (CA_1^{-1}C^\top\Pi_1 + CA_2^{-1}C^\top\Pi_2)^\top\mathbf{w}^\star, \mathbf{w}\rangle \tag{8}$$

Proof of Lemma A.3. Recall that $\text{SW} := \sum_{i\in[2]} \mathbb{E}_{x\sim\mathcal{D}_g}[\langle\hat{\mathbf{x}}(\mathbf{x};g), \mathbf{w}^\star\rangle]$. Using the $\hat{\mathbf{x}}$ solution from 2.1, we see that this is equivalent to the following

$$
\begin{aligned}
\text{SW} &= \sum_{g\in[2]} \mathbb{E}_{x\sim\mathcal{D}_g}[\langle\hat{\mathbf{x}}(\mathbf{x};g), \mathbf{w}^\star\rangle] \\
&= \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_1}\left[\langle\mathbf{x} + \Delta_1(\mathbf{w}), \mathbf{w}^\star\rangle\right] + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_2}\left[\langle\mathbf{x} + \Delta_2(\mathbf{w}), \mathbf{w}^\star\rangle\right] \\
&= \langle\Delta_1(\mathbf{w}), \mathbf{w}^\star\rangle + \langle\Delta_2(\mathbf{w}), \mathbf{w}^\star\rangle + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_1}\left[\langle\mathbf{x}, \mathbf{w}^\star\rangle\right] + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_2}\left[\langle\mathbf{x}, \mathbf{w}^\star\rangle\right] \\
&= \langle CA_1^{-1}C^\top\Pi_1\mathbf{w}, \mathbf{w}^\star\rangle + \langle CA_2^{-1}C^\top\Pi_2\mathbf{w}, \mathbf{w}^\star\rangle + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_1}\left[\langle\mathbf{x}, \mathbf{w}^\star\rangle\right] + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_2}\left[\langle\mathbf{x}, \mathbf{w}^\star\rangle\right] \\
&= \langle \mathbf{w}^\star, CA_1^{-1}C^\top\Pi_1\mathbf{w}\rangle + \langle \mathbf{w}^\star, CA_2^{-1}C^\top\Pi_2\mathbf{w}\rangle + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_1}\left[\langle\mathbf{x}, \mathbf{w}^\star\rangle\right] + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_2}\left[\langle\mathbf{x}, \mathbf{w}^\star\rangle\right] \\
&= \mathbf{w}^{\star\top}CA_1^{-1}C^\top\Pi_1\mathbf{w} + \mathbf{w}^{\star\top}CA_2^{-1}C^\top\Pi_2\mathbf{w} + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_1}\left[\langle\mathbf{x}, \mathbf{w}^\star\rangle\right] + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_2}\left[\langle\mathbf{x}, \mathbf{w}^\star\rangle\right] \\
&= \langle (CA_1^{-1}C^\top\Pi_1)^\top\mathbf{w}^\star, \mathbf{w}\rangle + \langle (CA_2^{-1}C^\top\Pi_2)^\top\mathbf{w}^\star, \mathbf{w}\rangle + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_1}\left[\langle\mathbf{x}, \mathbf{w}^\star\rangle\right] + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_2}\left[\langle\mathbf{x}, \mathbf{w}^\star\rangle\right] \\
&= \langle (CA_1^{-1}C^\top\Pi_1 + CA_2^{-1}C^\top\Pi_2)^\top\mathbf{w}^\star, \mathbf{w}\rangle + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_1}\left[\langle\mathbf{x}, \mathbf{w}^\star\rangle\right] + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_2}\left[\langle\mathbf{x}, \mathbf{w}^\star\rangle\right]
\end{aligned}
$$

The two expectation terms are constants with respect to $\mathbf{w}$, so they can be ignored when finding an optimal solution. Since the learner wants to maximize the linear objective, this is equivalent to minimizing the negative. □

A.3 Supplemental material for Section 3

A.3.1 Supplemental material of Table 1

We will now derive the bounds that correspond to each combination in the table.

First, notice that Accuracy and Social Welfare objectives are very nice. We can use this to loosely bound the optimality loss even when the only property satisfied by the fairness space is convexity.

Proposition A.1 (Optimality loss bounds with no fairness space properties) *For the Accuracy objective, the optimality loss between the unconstrained and fairness-constrained equilibrium is upper-bounded:*

$$|f(\mathbf{w}_c^\star) - f(\mathbf{w}_u^\star)| \leq 4(\|\mathbf{w}^\star\| + 1) \tag{9}$$

For the Social Welfare objective, the optimality loss between the unconstrained and fairness-constrained equilibrium is upper-bounded:

$$|f(\mathbf{w}_c^\star) - f(\mathbf{w}_u^\star)| \leq 2L_{\text{SW}} \tag{10}$$

Where $L_{\text{SW}} := \|(CA_1^{-1}C^\top\Pi_1 + CA_2^{-1}C^\top\Pi_2)^\top\mathbf{w}^\star\|_2$

Proof. Notice that for an L-lipschitz objective function:

$$\begin{aligned}
|f(\mathbf{w}_c^\star) - f(\mathbf{w}_u^\star)| &\leq L\|\mathbf{w}_u^\star - \mathbf{w}_c^\star\|_2 \quad \forall \mathbf{w}_c^\star, \mathbf{w}_u^\star \in \mathcal{B}(1) && \text{(def of } L\text{-lipschitz in ball)} \\
&\leq L \sup_{\mathbf{w} \in \mathcal{W}(\beta) \cap \mathcal{B}(1)} \|\mathbf{w}_u^\star - \mathbf{w}\|_2 && \text{(def of constrained problem)} \\
&\leq 2L && \text{(diameter of } \mathcal{B}(1))
\end{aligned}$$

Now we shall just prove that the accuracy and social welfare objectives are $2(\|\mathbf{w}^\star\| + 1)$- and L_{SW}-lipschitz on the euclidean ball respectively.

Accuracy objective is simply the squared $l2$ distance between $\mathbf{w}^\star$ (Lemma A.2) and its projection onto the feasible region.

$$\begin{aligned}
|f(\mathbf{w}) - f(\mathbf{w}')| &= |-\|\mathbf{w}^\star - \mathbf{w}\|_2^2 + \|\mathbf{w}^\star - \mathbf{w}'\|_2^2| \\
&= |(\|\mathbf{w}^\star - \mathbf{w}'\|_2 + \|\mathbf{w}^\star - \mathbf{w}\|_2)(\|\mathbf{w}^\star - \mathbf{w}'\|_2 - \|\mathbf{w}^\star - \mathbf{w}\|_2)| \\
&= (\|\mathbf{w}^\star - \mathbf{w}'\|_2 + \|\mathbf{w}^\star - \mathbf{w}\|_2)|\|\mathbf{w}^\star - \mathbf{w}'\|_2 - \|\mathbf{w}^\star - \mathbf{w}\|_2| \\
&\leq (\|\mathbf{w}^\star - \mathbf{w}'\|_2 + \|\mathbf{w}^\star - \mathbf{w}\|_2)\|\mathbf{w} - \mathbf{w}'\| && \text{triangle ineq} \\
&\leq 2(\|\mathbf{w}^\star\| + 1)\|\mathbf{w} - \mathbf{w}'\| \quad \forall \mathbf{w}, \mathbf{w}' \in \mathcal{B}(1)
\end{aligned}$$

Recall that Social Welfare is a linear objective (Lemma A.3). Specifically: $\langle \mathbf{c}, \mathbf{w} \rangle$ where $\mathbf{c} := (CA_1^{-1}C^\top \Pi_1 + CA_2^{-1}C^\top \Pi_2)^\top \mathbf{w}^\star$

$$\begin{aligned}
|f(\mathbf{w}) - f(\mathbf{w}')| &= |\langle \mathbf{c}, \mathbf{w} - \mathbf{w}' \rangle| \\
&\leq \|\mathbf{c}\|_2 \|\mathbf{w} - \mathbf{w}'\|_2 \quad \forall \mathbf{w}, \mathbf{w}' \qquad \text{Cauchy-schwarz}
\end{aligned}$$

□

Proposition A.2 (Ellipsoidal (Property 3.2) social welfare loss) *For fairness spaces satisfying Property 3.2, when the learner's objective, f, is social welfare, optimality loss is upper bounded.*

$$|f(\mathbf{w}_u^\star) - f(\mathbf{w}_c^\star)| \leq \sqrt{2}$$

Proof. Recall that Social Welfare is a linear objective (Lemma A.3). Specifically: $\langle \mathbf{c}, \mathbf{w} \rangle$ where $\mathbf{c} := (CA_1^{-1}C^\top \Pi_1 + CA_2^{-1}C^\top \Pi_2)^\top \mathbf{w}^\star$

Using KKT in the unconstrained problem we get that the optimum w_u^* lies in the direction of c.

1. If $\mathcal{W}(\beta) \supset \mathcal{B}_2(1)$ then the constrained optimum coincides with the unconstrained (it lies in the L2-ball), and the angle is zero.
2. If $\mathcal{W}(\beta) \subset \mathcal{B}_2(1)$ then in the constrained optimization problem the ball constraint is not active and therefore by KKT we infer that the optimimum lies in the direction of $Q^{-1}c$ and in particular : $w_c^* = \frac{Q^{-1}c}{c^\top Qc}$. Now observe that

   $$\cos(x_u^*, x_c^*) = \cos(c, A^{-1}c) = \frac{c^\top A^{-1}c}{\|c\|\|A^{-1}c\|} > 0$$

   since we know that $A \succ 0$ (by definition of ellipsoid) and therefore A^{-1} exists and it is also a PD.
3. If neither of the above is happening, then the KKT conditions tell us that $2\lambda_1 w_c^* + 2\lambda_2 Q w_c^* = c$ where $\lambda_1, \lambda_2 \geq 0$ are the Langrange multipliers. Clearly, if $\lambda_1, \lambda_2 > 0$* , $A = \lambda_1 I + \lambda_2 Q$ must be invertible as the sum of two PD matrices is a PD matrix. Hence $w_c^* = \frac{1}{2}A^{-1}c$ does not follow the direction of either c or Q^{-1}c but a linear combination of those. In any case, A^{-1} is a PD matrix itself and therefore:

   $$\cos(x_u^*, x_c^*) = \frac{c^\top A^{-1}c}{\|c\|\|A^{-1}c\|} > 0, \quad \forall c \neq 0$$

   * If $\lambda_1 = 0, \lambda_2 > 0$ then $w_c^* = \frac{1}{2\lambda_2}Q^{-1}c$ and since $Q \succ 0$ from assumption we will also have $c^\top Q^{-1}c \succ 0 \; \forall c \in \mathbb{R}^n \implies \cos(x_u^*, x_c^*) > 0$.
   If $\lambda_1 > 0, \lambda_2 = 0$ then w_c^* and w_u^* lie in the same direction and the angle is 0.

Using the generalized pythagorean:

$$\|x_c^* - x_u^*\| \leq \|x_c^*\|^2 + \|x_u^*\|^2 - 2\cos(x_u^*, x_c^*) \leq \|x_c^*\|^2 + \|x_u^*\|^2$$

Since both x_c^* and x_u^* belong in the ball we get:

$$\|x_c^* - x_u^*\| \leq \|x_c^*\|^2 + \|x_u^*\|^2 \leq 1 + 1 = \sqrt{2}$$

□

Proposition A.3 (Internal polyhedron (Property 3.1) accuracy loss) *For fairness spaces satisfying Property 3.1, when the learner's objective, f, is accuracy, optimality loss is upper bounded.*

$$\textit{if } \mathbf{w}^\star \in \mathcal{B}(1) : |f(\mathbf{w}_u^\star) - f(\mathbf{w}_c^\star)| \leq [H(M)\|(M\mathbf{w}^\star - \beta\mathbf{1})_+\|_2]^2$$

$$\textit{if } \mathbf{w}^\star \notin \mathcal{B}(1) : |f(\mathbf{w}_u^\star) - f(\mathbf{w}_c^\star)| \leq \left[H(M)\|(M\frac{\mathbf{w}^\star}{\|\mathbf{w}^\star\|_2} - \beta\mathbf{1})_+\|_2\right]^2$$

Where $M \in \mathbb{R}^{k\times d}$ and $\mathbf{1} \in \mathbb{R}^k$ define the polyhedral representation of the fairness space. That is, the feasible region, $\mathcal{B}(1) \cap \mathcal{W}(\beta) = \{\mathbf{w} \in \mathbb{R}^d : M\mathbf{w} \leq \mathbf{b}\}$

Proof. First, recall that accuracy optimization is simply euclidean projection onto the respective feasible region (Lemma A.2). Therefore, $\mathbf{w}_u^*$ and $\mathbf{w}_c^*$ are projections of $\mathbf{w}^*$ onto $\mathcal{B}(1)$ and $\mathcal{B}(1) \cap \mathcal{W}(\beta)$ respectively. In order to prove Proposition A.3, we will leverage that $\mathbf{w}_c^*$ is closer to $\mathbf{w}^\star$ than the projection of $\mathbf{w}_u^*$ would be onto $\mathcal{W}(\beta) \cap \mathcal{B}(1)$. Let $\mathbf{z} := P_{\mathcal{W}(\beta)\cap\mathcal{B}(1)}(\mathbf{w}_u^*)$ be this projection. Clearly, we have:

$$\begin{aligned} \|\mathbf{w}_c^\star - \mathbf{w}^\star\|_2^2 &\leq \|\mathbf{z} - \mathbf{w}^\star\|_2^2 && (\mathbf{w}_c^\star \text{ is optimal}) \\ &\leq \|\mathbf{w}_u^\star - \mathbf{w}^\star\|_2^2 + \|\mathbf{z} - \mathbf{w}_u^\star\|_2^2 && (\text{triangle ineq}) \end{aligned}$$

Using Lemma A.2 this implies that

$$|f(\mathbf{w}_u^\star) - f(\mathbf{w}_c^\star)| = \|\mathbf{w}_c^\star - \mathbf{w}^\star\|_2^2 - \|\mathbf{w}_u^\star - \mathbf{w}^\star\|_2^2 \leq \|\mathbf{z} - \mathbf{w}_u^\star\|_2^2$$

Note that clearly $\|\mathbf{w}_c^\star - \mathbf{w}^\star\|_2 \geq \|\mathbf{w}_u^\star - \mathbf{w}^\star\|_2$. So now, we must simply upper bound $\|\mathbf{z} - \mathbf{w}_u^\star\|_2$. Using a Hoffman bound (Hoffman (1952)), we have that $\exists \mathbf{w}_0 \in \mathcal{W}(\beta) \cap \mathcal{B}(1), H(M) > 0$ such that,

$$[H(M)\|(M\mathbf{w}^\star - \beta\mathbf{1})_+\|_2]^2 \geq \|\mathbf{w}_u^\star - \mathbf{w}_0\|_2^2 \geq \|\mathbf{w}_u^* - \mathbf{z}\|_2^2$$

Of course, we want this bound in terms of $\mathbf{w}^\star$ not $\mathbf{w}_u^\star$, but this is simple because since $\mathbf{w}_u^\star$ is the projection onto $\mathcal{B}(1)$, we have a closed form in terms of $\mathbf{w}^\star$. In particular, if $\mathbf{w}^\star \in \mathcal{B}(1)$, then $\mathbf{w}^\star = \mathbf{w}_u^\star$. Otherwise, we normalize it by the l-2 norm: $\mathbf{w}^\star/\|\mathbf{w}^\star\|_2 = \mathbf{w}_u^\star$ □

Lemma A.4 *Consider the following convex optimization problem, where $Q \in \mathbb{R}^{d\times d}$ and $Q \succ 0$. and $\tilde{\mathbf{w}} \neq \mathbf{0}$*

$$\begin{aligned} \textit{minimize}_{w\in\mathbb{R}^d} \quad & \langle \tilde{\mathbf{w}}, \mathbf{w} \rangle \\ \textit{subject to} \quad & \mathbf{w}^\top Q\mathbf{w} - \beta \leq 0 \end{aligned} \tag{11}$$

The optimal solution is

$$\hat{\mathbf{w}}_{ellipsoid}^* = \frac{-\sqrt{\beta}Q^{-1}\tilde{\mathbf{w}}}{\|\sqrt{Q^{-1}}\tilde{\mathbf{w}}\|_2}$$

Proof. By Slater's condition, we have that the KKT conditions are necessary and sufficient for optimality. Therefore, any $\mathbf{w}$ satisfying them must be optimal. We shall proceed by solving the KKT conditions.

$$\begin{aligned} \text{minimize}_{w\in\mathbb{R}^d} \quad & \langle \tilde{\mathbf{w}}, \mathbf{w} \rangle \\ \text{subject to} \quad & \mathbf{w}^\top Q\mathbf{w} - \beta \leq 0 \end{aligned} \tag{12}$$

The KKT conditions state:

$$\begin{aligned} -\tilde{\mathbf{w}} &= 2\lambda Q\mathbf{w} \\ \lambda &\geq 0 \\ \lambda(\mathbf{w}^\top Q\mathbf{w} - \beta) &= 0 \\ \mathbf{w}^\top Q\mathbf{w} &\leq \beta \end{aligned}$$

$\lambda \neq 0$ because if it were, then $\tilde{\mathbf{w}} = \mathbf{0}$, which would be a contradiction. Therefore it must be the case that $\lambda > 0$.

Using the first KKT condition we have that $\mathbf{w} = \frac{-Q^{-1}\tilde{\mathbf{w}}}{2\lambda}$. Q is PD, therefore it is also invertible. Because $\lambda > 0$, it must be the case that $\mathbf{w}^\top Q\mathbf{w} = \beta$ (3rd KKT condition). Thus we have:

$$\mathbf{w}^\top\mathbf{w} = \left[\frac{-Q^{-1}\tilde{\mathbf{w}}}{2\lambda}\right]^\top Q\frac{-Q^{-1}\tilde{\mathbf{w}}}{2\lambda} = \frac{1}{4\lambda^2}\tilde{\mathbf{w}}^\top Q^{-1}\tilde{\mathbf{w}} = \beta$$

Notice that if Q is PD, it is symmetric. Solving for λ, we see that $\lambda^\star = \frac{1}{2\sqrt{\beta}}\|\sqrt{Q^{-1}}\tilde{\mathbf{w}}\|_2$. Substituting this into $\mathbf{w} = \frac{-Q^{-1}\tilde{\mathbf{w}}}{2\lambda}$ we have

$$\hat{\mathbf{w}}^*_{ellipsoid} = \frac{-\sqrt{\beta}Q^{-1}\tilde{\mathbf{w}}}{\|\sqrt{Q^{-1}}\tilde{\mathbf{w}}\|_2}$$

□

Proposition A.4 (Internal ellipsoid (Property 3.3) SW loss) *Assume desirability fairness space, $\mathcal{W}(\beta)$ satisfies property 3.3. Then social welfare loss is exactly:*

$$SW(\mathbf{w}^\star_u) - SW(\mathbf{w}^\star_c) = \|(CA_1^{-1}C^\top\Pi_1 + CA_2^{-1}C^\top\Pi_2)^\top\mathbf{w}^\star\|_2 - \sqrt{\beta}\|(CA_1^{-1}C^\top\Pi_1 + CA_2^{-1}C^\top\Pi_2)^\top\mathbf{w}^\star\|_{Q^{-1}}$$

Proof. Recall that Social Welfare is a functionally linear objective (Lemma A.3). Specifically: $\langle\mathbf{c}, \mathbf{w}\rangle$ where $\mathbf{c} := (CA_1^{-1}C^\top\Pi_1 + CA_2^{-1}C^\top\Pi_2)^\top\mathbf{w}^\star$

Using Lemma A.4 we get the closed from for the solution for the unconstrained and fairness constrained problem, which we then plug back in for the optimal value. In each case, $\tilde{\mathbf{w}} := -(CA_1^{-1}C^\top\Pi_1 + CA_2^{-1}C^\top\Pi_2)^\top\mathbf{w}^\star$

Equation 4: $Q = I, \beta = 1$. This yields:

$$\mathbf{w}^\star_u := \frac{(CA_1^{-1}C^\top\Pi_1 + CA_2^{-1}C^\top\Pi_2)^\top\mathbf{w}^\star}{\|(CA_1^{-1}C^\top\Pi_1 + CA_2^{-1}C^\top\Pi_2)^\top\mathbf{w}^\star\|_2} \tag{13}$$

$$\begin{aligned} SW_u &= \langle(CA_1^{-1}C^\top\Pi_1 + CA_2^{-1}C^\top\Pi_2)^\top\mathbf{w}^\star, \frac{(CA_1^{-1}C^\top\Pi_1 + CA_2^{-1}C^\top\Pi_2)^\top\mathbf{w}^\star}{\|(CA_1^{-1}C^\top\Pi_1 + CA_2^{-1}C^\top\Pi_2)^\top\mathbf{w}^\star\|_2}\rangle \\ &= \|(CA_1^{-1}C^\top\Pi_1 + CA_2^{-1}C^\top\Pi_2)^\top\mathbf{w}^\star\|_2 \end{aligned}$$

Equation 3: $Q = Q, \beta = \beta$. This yields:

$$\hat{\mathbf{w}}^\star_c := \frac{\sqrt{\beta}Q^{-1}(CA_1^{-1}C^\top\Pi_1 + CA_2^{-1}C^\top\Pi_2)^\top\mathbf{w}^\star}{\|\sqrt{Q^{-1}}(CA_1^{-1}C^\top\Pi_1 + CA_2^{-1}C^\top\Pi_2)^\top\mathbf{w}^\star\|_2} \tag{14}$$

$$\begin{aligned} SW_c &= \langle(CA_1^{-1}C^\top\Pi_1 + CA_2^{-1}C^\top\Pi_2)^\top\mathbf{w}^\star, \frac{\sqrt{\beta}Q^{-1}(CA_1^{-1}C^\top\Pi_1 + CA_2^{-1}C^\top\Pi_2)^\top\mathbf{w}^\star}{\|\sqrt{Q^{-1}}(CA_1^{-1}C^\top\Pi_1 + CA_2^{-1}C^\top\Pi_2)^\top\mathbf{w}^\star\|_2}\rangle \\ &= \sqrt{\beta}\|(CA_1^{-1}C^\top\Pi_1 + CA_2^{-1}C^\top\Pi_2)^\top\mathbf{w}^\star\|_{Q^{-1}} \end{aligned}$$

Subtracting the two yields the result of the proposition. □

Proposition A.5 (Internal ellipsoid(Property 3.3) accuracy loss) *Assume property 3.3 is satisfied. Let $\mathbf{w}'$ be a maximizer on the problem:*

$$\begin{aligned} \max \; & -\|\mathbf{w} - \mathbf{w}^\star\|_2^2 \\ & \mathbf{w}^\top Q\mathbf{w} \leq \beta \end{aligned}$$

Then we can bound the accuracy loss between $\mathbf{w}'$ and unconstrainted (normalized on L2) policy $\mathbf{w}^\star_u$

$$|ACC(\mathbf{w}') - ACC(\mathbf{w}^*_u)| \leq (2r + 1 - s_{\min})(r + 1 - s_{\min})$$

where $\mathbf{w}^\star$ is the ground-truth policy, $s_{\min} = \sqrt{\frac{\beta}{\lambda_{\max}(Q)}}$ and $r = (\|\mathbf{w}^\star\| - 1)_+$

Observation 1*: When w^* lies in the L2 ball then $r = 0$ and the bound is $(1 - s_{\min})^2$.*
Observation 2*: When the ellipsoid tends to cover the L2, i.e $s_{\min} \to 1$ the bound becomes simply $2r^2$ (the optimal w_u^* and w' will coincide and the bound will only depend on the position of w^* with relation to the L2-ball).*

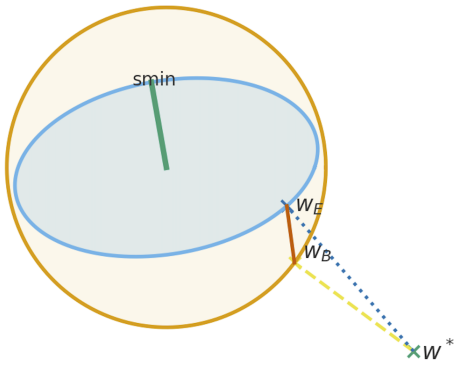


Figure 3: Useful visual: $s_{\min}$ is the minimum distance of any point in the boundary of the ellipsoid from its origin, w^* is the optimal while w_B is its projection onto the ball, w_E is the optimal over the intersection of the ellipsoid with the ball.

Proof. Define $\mathcal{E}(\beta) := \{\mathbf{w} \in \mathbb{R}^d : \mathbf{w}^\top Q\mathbf{w} \leq \beta\}$ to be the ellipsoid internal to the euclidean ball.

Let $s_{\min} = \sqrt{\frac{\beta}{\lambda_{\max}(Q)}}$ be the shortest radius from the origin to $\mathcal{E}(\beta)$'s boundary. Let $P_{\mathcal{E}(\beta)}(\mathbf{w})$ be the projection of $\mathbf{w}$ onto $\mathcal{E}(\beta)$ then $\|P_{\mathcal{E}(\beta)}(\mathbf{w}_u^\star) - \mathbf{w}_u^\star\| = 1 - s$ where s is the distance between the origin and the point projection point on the boundary of $\mathcal{E}(\beta)$. Clearly

$$s \geq s_{\min} \implies \|P_{\mathcal{E}(\beta)}(\mathbf{w}_u^\star) - \mathbf{w}_u^\star\| = 1 - s \leq 1 - s_{\min}$$

Thus,

$$\begin{aligned}
\|\mathbf{w}' - \mathbf{w}_u^\star\|_2 &= \|P_{\mathcal{E}(\beta)}(\mathbf{w}^\star) - \mathbf{w}_u^\star\|_2 \\
&\leq \|P_{\mathcal{E}(\beta)}(\mathbf{w}^\star) - P_{\mathcal{E}(\beta)}(\mathbf{w}_u^\star)\|_2 + \|P_{\mathcal{E}(\beta)}(\mathbf{w}_u^\star) - \mathbf{w}_u^\star\|_2 && \text{(triangle ineq)} \\
&\leq \|\mathbf{w}^\star - \mathbf{w}_u^\star\| + 1 - s_{\min}, && \text{(non-expansiveness of projections)} \\
&\leq (\|\mathbf{w}^\star\|_2 - 1)_+ + 1 - s_{\min}
\end{aligned}$$

where the last inequality comes from the fact that if $\mathbf{w}^\star$ is in the boundary of L_2 then distance is 0 and otherwise it is $\mathbf{w}_u^\star = \mathbf{w}^\star / \|\mathbf{w}^\star\|$.

Now we can write using Lemma A.2:

$$\begin{aligned}
|\text{Acc}(\mathbf{w}') - \text{Acc}(\mathbf{w}_u^\star)| &= |-\|\mathbf{w}' - \mathbf{w}^\star\|_2^2 + \|\mathbf{w}_u^\star - \mathbf{w}^\star\|_2^2| \\
&= |(\|\mathbf{w}' - \mathbf{w}^\star\|_2 + \|\mathbf{w}_u^\star - \mathbf{w}^\star\|_2)(|\mathbf{w}' - \mathbf{w}^\star\|_2 - \|\mathbf{w}_u^\star - \mathbf{w}^\star\|_2)| \\
&\leq (\|\mathbf{w}' - \mathbf{w}_u^\star\|_2 + \|\mathbf{w}_u^\star - \mathbf{w}^\star\|_2)(\|\mathbf{w}' - \mathbf{w}_u^\star\|_2) && \text{(triangle ineq)} \\
&\leq (r + 1 - s_{\min} + r)(r + 1 - s_{\min}) \\
&= (2r + 1 - s_{\min})(r + 1 - s_{\min})
\end{aligned}$$

where $r = (\|\mathbf{w}^\star\| - 1)_+$ □

To supplement Table 1 bounds, one may consider the following numerical examples.

Example A.1 (Bounded optimality loss given non-disparate costs) *Let agents have 2 features ($d = 2$). Cost is non-disparate and changing either feature requires unit cost, so $A_g = \mathbf{I}_2$. Features do not have any causal flow between one another, so $C = \mathbf{I}_2$. The first feature is desirable, while*

the 2nd feature is slightly less so, $\Pi_D : diag(1, 3/4)$. *Agents come from a feature distribution that results in* $\Pi_1 := diag(1, 0)$ *and* $\Pi_2 := diag(0, 1)$. *Finally* $\mathbf{w}^\star = (1/2, 1/2)$. *Using the fairness constraint of Example 3.1, how much social welfare or accuracy is lost in Stackelberg equilibrium?*

By Corollary A.2, if $\beta \leq 3/4$, *this satisfies Property 3.1. Thus using the polyhedral construction of Lemma A.5 and the bound in Table 1, accuracy loss is bounded:*

$$|f(\mathbf{w}_u^\star) - f(\mathbf{w}_c^\star)| \leq \left[H(\widehat{M})\right]^2 [(1/8 - \beta)_+]^2 + (7/8 - \beta)^2]$$

if we further have $1/8 \leq \beta \leq 3/4$*:*

$$|f(\mathbf{w}_u^\star) - f(\mathbf{w}_c^\star)| \leq \left[\frac{3H(\widehat{M})}{4}\right]^2$$

Where:

$$\widehat{M} = \begin{bmatrix} -1 & -3/4 \\ 1 & -3/4 \\ 1 & 3/4 \\ -1 & 3/4 \end{bmatrix}$$

Social welfare loss is also bounded:

$$|f(\mathbf{w}_u^\star) - f(\mathbf{w}_c^\star)| \leq 2\|(CA_1^{-1}C^\top\Pi_1 + CA_2^{-1}C^\top\Pi_2)^\top\mathbf{w}^\star\|_2 = \sqrt{1/2}$$

Example A.2 (Bounded accuracy loss given non-disparate feature distributions) *Let agents have 2 features* ($d = 2$). *Feature space is non-disparate and nonskewed, so* $\Pi_g = \mathbf{I}_2$. *Features do not have any causal flow between one another, so* $C = \mathbf{I}_2$. *The first feature is desirable, while the 2nd feature is slightly less so,* $\Pi_D : diag(1, 3/4)$. *Agents have disparate costs such that all change is easier for group 1 agents:* $\Pi_1 := diag(1/2, 1/2)$ *and* $\Pi_2 := diag(1, 1)$. *Finally* $\mathbf{w}^\star = (1/2, 1/2)$. *Using the fairness constraint of Example 3.1, how much social welfare or accuracy is lost in Stackelberg equilibrium?*

By Corollary A.3 if $\beta \leq 3/4$, *this satisfies Property 3.1. Thus using the polyhedral construction of Lemma A.5 and the bound in Table 1, accuracy loss is bounded:*

$$|f(\mathbf{w}_u^\star) - f(\mathbf{w}_c^\star)| \leq \left[H(\widehat{M})\right]^2 [(1/8 - \beta)_+]^2 + (7/8 - \beta)^2]$$

if we further have $1/8 \leq \beta \leq 3/4$*:*

$$|f(\mathbf{w}_u^\star) - f(\mathbf{w}_c^\star)| \leq \left[\frac{3H(\widehat{M})}{4}\right]^2$$

Where:

$$\widehat{M} = \begin{bmatrix} -1 & -3/4 \\ 1 & -3/4 \\ 1 & 3/4 \\ -1 & 3/4 \end{bmatrix}$$

Social welfare loss is also bounded:

$$|f(\mathbf{w}_u^\star) - f(\mathbf{w}_c^\star)| \leq 2\|(CA_1^{-1}C^\top\Pi_1 + CA_2^{-1}C^\top\Pi_2)^\top\mathbf{w}^\star\|_2 = \sqrt{1/2}$$

A.3.2 Supplemental material for Example 3.1

First we note some usually assumptions used in this Appendix section and Appendix A.3.3

Assumption A.1 (Unknown feature space is exclusive) $\ker(\Pi_1) \cap \ker(\Pi_2) = \emptyset$

Assumption A.2 (Estimated w is different) $\Pi_1\mathbf{w} \neq \Pi_2\mathbf{w} \quad \forall\mathbf{w} \neq \mathbf{0}$

Lemma A.5 ($\mathcal{W}(\beta)$ from Example 3.1 is a polyhedron) *$\mathcal{W}(\beta)$ of Example 3.1 can be rewritten as:*

$$\mathcal{W}(\beta) = \{\mathbf{w} : \mathbf{w} \in \mathbb{R}^d, \mathbf{a}^\top M\mathbf{w} \leq \beta \quad \forall \mathbf{a} \in \mathcal{A}\}$$

Where $M := \Pi_D A_1^{-1} C^\top \Pi_1 - \Pi_D A_2^{-1} C^\top \Pi_2$ and $\mathcal{A} := \{(a_1, \ldots, a_d) \in \mathbb{R}^d : a_i \in \{-1, 1\} \forall i \in [d]\}$

This is clearly a polyhedron of 2^d constraints each defined by $\mathbf{a}^\top M$

Proof.

$$\begin{aligned}\sum_{i\in[d]} |(\Pi_D \mathbf{x}_e^{(1)}(\mathbf{w}) - \Pi_D \mathbf{x}_e^{(2)}(\mathbf{w}))_i| &= \sum_{i\in[d]} |(\Pi_D A_1^{-1} C^\top \Pi_1 \mathbf{w} - \Pi_D A_2^{-1} C^\top \Pi_2 \mathbf{w})_i| \\ &= \sum_{i\in[d]} |((\Pi_D A_1^{-1} C^\top \Pi_1 - \Pi_D A_2^{-1} C^\top \Pi_2)\mathbf{w})_i| \\ &= \|M\mathbf{w}\|_1\end{aligned}$$

Where $M := \Pi_D A_1^{-1} C^\top \Pi_1 - \Pi_D A_2^{-1} C^\top \Pi_2$

Recall that $\{\mathbf{y} : \mathbf{y} \in \mathbb{R}^d, \|\mathbf{y}\|_1 \leq \beta\}$ describes a d-dimensional cube in $\mathbf{y}$ space, such shapes are clearly polyhedra. We will show that $\|M\mathbf{w}\|_1 \leq \beta$ creates the polyhedron described by the lemma.

We can describe $\{\mathbf{y} : \mathbf{y} \in \mathbb{R}^d, \|\mathbf{y}\|_1 \leq \beta\}$ equivalently as: $\{\mathbf{y} : \mathbf{y} \in \mathbb{R}^d, \mathbf{a}^\top \mathbf{y} \leq \beta \quad \forall \mathbf{a} \in \mathcal{A}\}$ where $\mathcal{A} := \{(a_1, \ldots, a_d) \in \mathbb{R}^d : a_i \in \{-1, 1\} \forall i \in [d]\}$ Note that $|\mathcal{A}| = 2^d$. Let $\mathbf{y} = M\mathbf{w}$ and this gives polyhedron of the lemma. □

Importantly, Property 3.1 requires that $\mathcal{B}(1) \cap \mathcal{W}(\beta)$ is a polyhedron! Therefore, for the optimality loss bound associated with this Property, we should ensure that $\mathcal{W}(\beta) \subseteq \mathcal{B}(1)$

Proposition A.6 (Necessary and sufficient conditions for Example 3.1 to satisfy property 3.1) *The fairness function described by Example 3.1 satisfies Property 3.1 if and only if M, where $M := \Pi_D A_1^{-1} C^\top \Pi_1 - \Pi_D A_2^{-1} C^\top \Pi_2$ is such that (1) $\ker(M) = \emptyset$ and (2) $\beta \leq \inf_{\|\mathbf{w}\|_2=1} \|M\mathbf{w}\|_1$*

Proof. First, notice that by Lemma A.5, the fairness space, $\mathcal{W}(\beta)$ is a polyhedron for any M and $\beta > 0$. Because $\mathcal{B}(1)$ is an ellipsoid, this means that for $\mathcal{W}(\beta) \cup \mathcal{B}(1)$ to be a polyhedron, $\mathcal{W}(\beta) \subseteq \mathcal{B}(1)$. So we must show that conditions (1) and (2) are necessary and sufficient to ensure that $\mathcal{W}(\beta) \subseteq \mathcal{B}(1)$.We will first prove that conditions (1) and (2) are necessary.

Notice that if $\ker(M) \neq 0$, then there exists some $\mathbf{w}_0 \in \ker(M)$ s.t. $\|\mathbf{w}\|_2 \geq 1$, but $\|M\mathbf{w}_0\|_1 = 0 < \beta$. That would mean $\mathcal{W}(\beta) \not\subseteq \mathcal{B}(1)$ Thus $\ker(M) = \emptyset$ must be necessary.

Let $\mu(M) := \inf_{\|\mathbf{w}\|_2=1} \|M\mathbf{w}\|_1$ Now notice that $\|M\mathbf{w}\|_1 = \|\mathbf{w}\|_2 \|M\frac{\mathbf{w}}{\|\mathbf{w}\|_2}\|_1 \geq \mu(M)\|\mathbf{w}\|_2 \quad \forall \mathbf{w}$ This implies $\frac{\|M\mathbf{w}\|_1}{\mu(M)} \geq \|\mathbf{w}\|_2$. Of course $\forall \mathbf{w} \in \mathcal{W}(\beta)$:

$$\frac{\beta}{\mu(M)} \geq \frac{\|M\mathbf{w}\|_1}{\mu(M)} \geq \|\mathbf{w}\|_2$$

From this, we see that in order for $\mathcal{W}(\beta) \subseteq \mathcal{B}(1)$, it is necessary that $\beta \leq \inf_{\|\mathbf{w}\|_2=1} \|M\mathbf{w}\|_1$.

Now we will show that conditions (1) and (2) are sufficient. We will do this by contradiction. Suppose that $\mathbf{w}_0 \in \mathcal{W}(\beta)$, but $\mathbf{w}_0 \notin \mathcal{B}(1)$ while both (1) and (2) hold. This means that $\|\mathbf{w}_0\|_2 > 1$ and $\|M\mathbf{w}_0\|_1 \leq \beta$. From condition (2), $\|\mathbf{w}_0\|_2 \|M\frac{\mathbf{w}_0}{\|\mathbf{w}_0\|_2}\|_1 = \|M\mathbf{w}_0\|_1 \leq \beta$. From condition (1) we know that $\mu(M) > 0$ and then from the definition of $\mu(M)$: $\|M\frac{\mathbf{w}_0}{\|\mathbf{w}_0\|_2}\|_1 \geq \mu(M)$. Thus it must be the case that $\|\mathbf{w}_0\|_2 \leq 1$. But this poses a contradiction! Thus we see that when conditions (1) and (2) hold, there cannot exist such a $\mathbf{w}$ where $\mathbf{w}_0 \in \mathcal{W}(\beta)$, but $\mathbf{w}_0 \notin \mathcal{B}(1)$, which means $\mathcal{W}(\beta) \subseteq \mathcal{B}(1)$. □

Corollary A.1 (Sufficient conditions for Example 3.1 to satisfy property 3.1) *The fairness function described by Example 3.1 satisfies Property 3.1 if M, where $M := \Pi_D A_1^{-1} C^\top \Pi_1 - \Pi_D A_2^{-1} C^\top \Pi_2$ is such that $\ker(M) = \emptyset$ and $\beta \leq \sigma_d(M)$*

Proof. We should show that $\beta \leq \sigma_d(M) \implies \beta \leq \inf_{\|\mathbf{w}\|_2=1} \|M\mathbf{w}\|_1$ This is simple. Note that rank of M must be d because it is a $d \times d$ matrix with an empty kernel. So the dth singular value is the smallest nonzero singular value. So we have $\inf_{\|\mathbf{w}\|_2=1} \|M\mathbf{w}\|_2 = \sigma_d(M)$ from the Min-Max theorem for singular values. And because $\forall \mathbf{y} \in \mathbb{R}^d$, $\|\mathbf{y}\|_2 \leq \|\mathbf{y}\|_2$, we have $\inf_{\|\mathbf{w}\|_2=1} \|M\mathbf{w}\|_2 = \sigma_d(M) \leq \inf_{\|\mathbf{w}\|_2=1} \|M\mathbf{w}\|_1$. Thus clearly $\beta \leq \sigma_d(M) \implies \beta \leq \inf_{\|\mathbf{w}\|_2=1} \|M\mathbf{w}\|_1$ □

These conditions, even in only the sufficient form are hard to interpret in the context of the setting specifically given that M is a function of several setting parameters. To make things a clearer, we can simplify to more specific settings in which groups have either cost or information discrepancy:

Corollary A.2 (Sufficient conditions for Example 3.1 to satisfy property 3.1 w/ no cost asymmetry) *Suppose that agents in our setting have the same cost to feature change (i.e. $A_1 = A_2 = A_g$). Then, the fairness function described by Example 3.1 satisfies Property 3.1 if Assumption A.1 is satisfied, $\Pi_1\mathbf{w} \neq \Pi_2\mathbf{w} \quad \forall \mathbf{w}, \mathbf{w} \neq \mathbf{0}$, and $\beta \leq \sigma_d(\Pi_d A_g^{-1} C^\top [\Pi_1 - \Pi_2])$.*

Proof.

$$\begin{aligned} M &= \Pi_D A_1^{-1} C^\top \Pi_1 - \Pi_D A_2^{-1} C^\top \Pi_2 \\ &= \Pi_D A_g^{-1} C^\top [\Pi_1 - \Pi_2] \end{aligned}$$

Clearly, the singular value condition is the same as the sufficient condition from A.1. Thus, all we must prove that Assumption A.1 and $\Pi_1\mathbf{w} \neq \Pi_2\mathbf{w} \quad \forall \mathbf{w}, \mathbf{w} \neq \mathbf{0} \implies \ker(M) = \emptyset$. Notice that that $\ker(\Pi_D A_g^{-1} C^\top) = \emptyset$ because Π_D and A_g are positive definite by setting assumptions and $\ker(C) = \emptyset$ by Lemma A.1. Thus all that matters is $\ker(\Pi_1 - \Pi_2)$. Clearly, as long as

1. $\forall \mathbf{w} \in \ker(\Pi_1)$, $\mathbf{w} \notin \ker(P_2)$ and $\forall \mathbf{w}' \in \ker(\Pi_2)$, $\mathbf{w}' \notin \ker(P_1)$
2. $\Pi_1\mathbf{w} \neq \Pi_2\mathbf{w} \quad \forall \mathbf{w}, \mathbf{w} \neq \mathbf{0}$

then $\ker(\Pi_D A_g^{-1} C^\top [\Pi_1 - \Pi_2])$ will be empty. □

Corollary A.3 (Sufficient conditions for Example 3.1 to satisfy property 3.1 w/ no info asymmetry) *Suppose that agents in our setting have the same information (i.e. $\Pi_1 = \Pi_2 = \Pi_g$). Then, the fairness function described by Example 3.1 satisfies Property 3.1 if $\ker(\Pi_g) = \emptyset$, $A_2 \succ A_1$ and $\beta \leq \sigma_d(\Pi_D[A_1^{-1} - A_2^{-1}]C^\top \Pi_g)$.*

Proof.

$$\begin{aligned} M &= \Pi_D A_1^{-1} C^\top \Pi_1 - \Pi_D A_2^{-1} C^\top \Pi_2 \\ &= \Pi_D [A_1^{-1} - A_2^{-1}] C^\top \Pi_g \end{aligned}$$

Clearly, the singular value condition is the same as the sufficient condition from A.1. Thus, all we must prove that $\ker(\Pi_g) = \emptyset$, $A_2 \succ A_1$ is sufficient to show that $\ker(M) = \emptyset$. Π_D and A_g are positive definite by setting assumptions and $\ker(C) = \emptyset$ by Lemma A.1. So we consider $\ker(A_1^{-1} - A_2^{-1})$ and $\ker(\Pi_g)$. Clearly if $\ker(A_1^{-1} - A_2^{-1}) = \emptyset$ and $\ker(\Pi_g) = \emptyset$ then $\ker(\Pi_D[A_1^{-1} - A_2^{-1}]C^\top \Pi_g) = \emptyset$. Note that:

$$A_2 \succ A_1 \implies A_1^{-1} \succ A_2^{-1} \implies A_1^{-1} - A_2^{-1} \succ 0 \implies \ker(A_1^{-1} - A_2^{-1}) = \emptyset$$

□

A.3.3 Supplemental material for Example 3.2

Proposition A.7 (Example 3.2 is (sometimes) an ellipsoid) *Example 3.2 represents and ellipsoid if and only if $M = \Pi_D(A_1^{-1} C^T P_1 - A_2^{-1} C^T P_2)$ is invertible.*

Proof. Expanding $\mathbf{x}_e^{(1)}, \mathbf{x}_e^{(2)}$ according to the closed-form solutions we identified in Proposition 2.1, we can think of Example 3.2 as $\|A\mathbf{w} - B\mathbf{w}\|_2^2 = \|(A-B)\mathbf{w}\|_2^2 \leq \beta$ where $A = \Pi_D(A_1^{-1} C^T P_1)$ and $B = \Pi_D(A_2^{-1} C^T P_2)$. Now can rewrite $\|(A-B)\mathbf{w}\|_2^2 = (A-B)^\top (A-B) \leq \beta$. This set represents an ellipsis when $M = (A-B)^\top (A-B) \succ 0$. Now we know $M \succeq 0$ always and $M \succ 0$ when $(A-B)$ is invertible.
Now we can study $(A-B) = \Pi_D(A_1^{-1} C^\top P_1 - A_2^{-1} C^\top P_2)$.
Hence $A - B$ is invertible iff $M = \Pi_D(A_1^{-1} C^\top P_1 - A_2^{-1} C^\top P_2)$ is invertible

□

Corollary A.4 (Sufficient conditions for Example 3.2 to satisfy property 3.2 w/ no cost asymmetry) *Suppose that agents in our setting have the same cost to feature change (i.e. $A_1 = A_2 = A_g$). Then, the fairness function described by Example 3.2 satisfies Property 3.2 if Assumptions A.1 and A.2 are satisfied.*

Proof.

$$\begin{aligned} M &= \Pi_D A_1^{-1} C^\top \Pi_1 - \Pi_D A_2^{-1} C^\top \Pi_2 \\ &= \Pi_D A_g^{-1} C^\top [\Pi_1 - \Pi_2] \end{aligned}$$

All we must prove that Assumptions A.1 and A.2 $\implies$ $\ker(M) = \emptyset$. Notice that that $\ker(\Pi_D A_g^{-1} C^\top) = \emptyset$ because Π_D and A_g are positive definite by setting assumptions and $\ker(C) = \emptyset$ by Lemma A.1. Thus all that matters is $\ker(\Pi_1 - \Pi_2)$. Clearly, as long as

1. $\forall \mathbf{w} \in \ker(\Pi_1)$, $\mathbf{w} \notin \ker(P_2)$ and $\forall \mathbf{w}' \in \ker(\Pi_2)$, $\mathbf{w}' \notin \ker(P_1)$
2. $\Pi_1 \mathbf{w} \neq \Pi_2 \mathbf{w} \quad \forall \mathbf{w}, \mathbf{w} \neq \mathbf{0}$

then $\ker(\Pi_D A_g^{-1} C^\top [\Pi_1 - \Pi_2])$ will be empty. □

Corollary A.5 (Sufficient conditions for Example 3.2 to satisfy property 3.2 w/ no info asymmetry) *Suppose that agents in our setting have the same information (i.e. $\Pi_1 = \Pi_2 = \Pi_g$). Then, the fairness function described by Example 3.2 satisfies Property 3.2 if $\ker(\Pi_g) = \emptyset$, $A_2 \succ A_1$.*

Proof.

$$\begin{aligned} M &= \Pi_D A_1^{-1} C^\top \Pi_1 - \Pi_D A_2^{-1} C^\top \Pi_2 \\ &= \Pi_D [A_1^{-1} - A_2^{-1}] C^\top \Pi_g \end{aligned}$$

Thus, all we must prove that $\ker(\Pi_g) = \emptyset$, $A_2 \succ A_1$ is sufficient to show that $\ker(M) = \emptyset$. Π_D and A_g are positive definite by setting assumptions and $\ker(C) = \emptyset$ by Lemma A.1. So we consider $\ker(A_1^{-1} - A_2^{-1})$ and $\ker(\Pi_g)$. Clearly if $\ker(A_1^{-1} - A_2^{-1}) = \emptyset$ and $\ker(\Pi_g) = \emptyset$ then $\ker(\Pi_D [A_1^{-1} - A_2^{-1}] C^\top \Pi_g) = \emptyset$. Note that:

$$A_2 \succ A_1 \implies A_1^{-1} \succ A_2^{-1} \implies A_1^{-1} - A_2^{-1} \succ 0 \implies \ker(A_1^{-1} - A_2^{-1}) = \emptyset$$

□

A.4 Supplemental material for Section 4

A.4.1 Supplemental material for Section 4.1

Proposition A.8 ($\mathcal{E}(\beta) \in \mathcal{W}(\beta; \Delta)$) *If $\mathcal{W}(\beta) \in \mathcal{F}$, then $\mathcal{E}(\beta) \subseteq \mathcal{W}(\beta) \cap \mathcal{B}(1)$ Where $\mathcal{E}(\beta) := \{\mathbf{w} : \mathbf{w} \in \mathbb{R}^d, \mathbf{w}^\top Q \mathbf{w} \leq \beta\}$*

Proof. Clearly if $\mathbf{w} \in \mathcal{E}(\beta)$ then we have: $\mathbf{w}^\top Q \mathbf{w} \leq \beta$. But by assumption $\mathcal{F}$ we have $\Delta(\mathbf{w}) \leq \mathbf{w}^\top Q \mathbf{w} \leq \beta$ thus $\mathcal{E}(\beta) \subseteq \mathcal{W}(\beta)$.

For the last part, note that (application of Löwner) a $\mathcal{E}(\beta) \in \mathcal{B}(1)$ iff $\beta \leq \lambda_d(Q)$ and this is true by definition of $\mathcal{F}$ □

A.4.2 Supplemental material for Section 4.2

Proposition A.9 ($\mathcal{W}(\beta) \in \mathcal{F}$ where $\mathcal{W}(\beta)$ defined by Definition 4.1) *$\mathcal{W}(\beta) \in \mathcal{F}$ if and only if $\ker(\Pi_g) = \emptyset$ and $\beta \leq \lambda_d(M^\top M)$ where $M := \Pi_D A_g^{-1} C^\top \Pi_g$*

Proof.

$$\begin{aligned} \Delta(\mathbf{w}) &= \|\Pi_D \mathbf{x}_e^{(g)}(\mathbf{w})\|_2^2 - \|\Pi_D \mathbf{x}_e^{(g')}(\mathbf{w})\|_2^2 \\ &= \langle \mathbf{w}, \mathbf{w} \rangle_{M_g^\top M_g} - \langle \mathbf{w}, \mathbf{w} \rangle_{M_{g'}^\top M_{g'}} \end{aligned}$$

Where $M_g := \Pi_D A_g^{-1} C^\top \Pi_g$. Point one of definition of $\mathcal{F}$ is that $M_g^\top M_g$ is PD. This is an iff with $\ker(M_g) = \emptyset$. The β conditions follows directly from the definition. □

A.5 SUPPLEMENTAL MATERIAL FOR SECTION 5

A.5.1 SUPPLEMENTAL MATERIAL FOR SECTION 5.1

We set the parameters $C, \Pi_1, \Pi_2, A_1, A_2, \Pi_D$, and $w^\star$ as follows.

Causal graph. We follow prior work on recourse/causal modeling for Adult von Kügelgen et al. (2020); Nabi & Shpitser (2018); Chiappa & Gillam (2018) (see their cited sources for the SCM) and instantiate an 8-node acyclic causal graph with nodes { sex,age,western,married,edu-num,workclass,occupation,hours}. Edge weights are sampled as non–negative values on the existing edges (respecting a topological order so the adjacency is strictly upper–triangular), yielding a weighted adjacency $A \in \mathbb{R}^{8\times 8}$ and the contribution matrix $C = \sum_{k=0}^{7} A^k$.

Groups and projectors. Following Bechavod et al. (2022), we form three sets of different 2 group splits. Groups sets are as follows:

- *Age* (≤ 35 vs. > 35),
- *Country* (western world vs. other)
- *Education* ($\geq$ high-school vs. $<$ high-school).

For each split $g \in \{1, 2\}$. We build a projection matrix $\Pi_g \in \mathbb{R}^{d\times d}$ by running SVD on the data points belonging to group g, taking the top k right singular vectors ($k = 5$), and setting $\Pi_g = V_{g,k}V_{g,k}^\top$.

Cost-matrices. For each group $g \in \{1, 2\}$ we sample a random matrix $G_g \in \mathbb{R}^{8\times 8}$ (with i.i.d. entries) and define $A_g = G_g^\top G_g + \rho I, \quad \rho > 0$. so that A_g is **random** yet **invertible** (symmetric positive definite).

Desirability. We choose *education*, *occupation*, and *workclass* as **desirable**, since these attributes are realistic to improve and likely to have downstream, external effects outside of income. Thus, to external entities (e.g. government bodies) they should be desirable to incentivize.

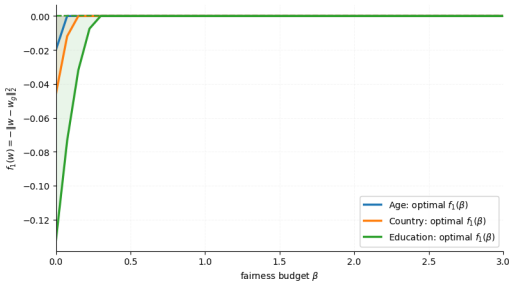
Group truth rule. We train a logistic-regression classifier on the ADULT dataset restricted to the eight variables that correspond to the nodes of our SCM (see above). Let $\tilde{\mathbf{w}} \in \mathbb{R}^8$ denote the learned coefficient vector. To make the comparisons, we project this vector onto $\tilde{w}$ the unit ℓ_2 ball, let w^* denote the projection. Our *accuracy loss* reports $\|\mathbf{w} - \mathbf{w}^\star\|_2^2$, and our *social welfare* uses the linear utility $u(\mathbf{w}) = \mathbf{c}^\top \mathbf{w}$ with $\mathbf{c}$ defined in Lemma A.3. Constrained learners are optimized under the ℓ_1-β-desirability fairness constraint $\|M_g\, \mathbf{w}\|_1 \leq \beta$ (see example 3.1). This normalization ensures the baseline is bounded for the social welfare problem while at the same time yields scale-invariant results.

A.5.2 SUPPLEMENTAL MATERIAL FOR SECTION 5.2

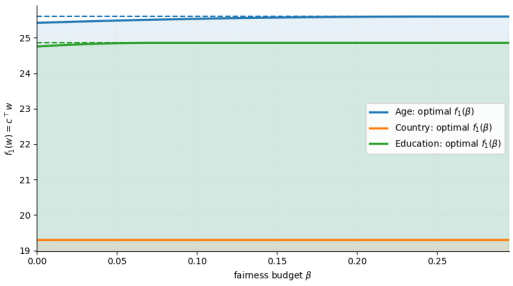
In this part we present the experiments using the ℓ_2-fairness desirability constraint (as defined in Example 3.2).

The results for ℓ_2 remain consistent with the results for ℓ_1 with the only noticeable difference being that the optimal accuracy is reached at a much faster rate (smaller value of β). This can be explained by the fact that $\|Mw\|_2 \leq \|Mw\|_1$ and therefore the ℓ_2 fairness constraint is more relaxed that the ℓ_1-fairness.

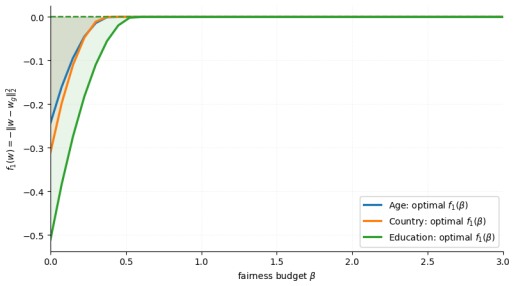
Interestingly, when we allow the cost matrices to be more complex than the unit (e.g random) then information disparities become more irrevelant. However, that's not suprising considering the fact that the fairness-desirability function values equally the desirability matrix and the projection matrices of the groups.



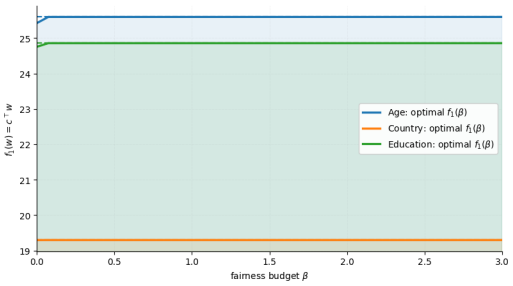
(a) *Accuracy comparison*: **Uniform costs** $A_1 = A_2$; **desirable**: occupation, workclass, education, under ℓ_2-fairness constraint



(b) *Social-welfare comparison*: **Uniform costs** $A_1 = A_2$; **desirable**: occupation, workclass, education, under ℓ_2-fairness constraint

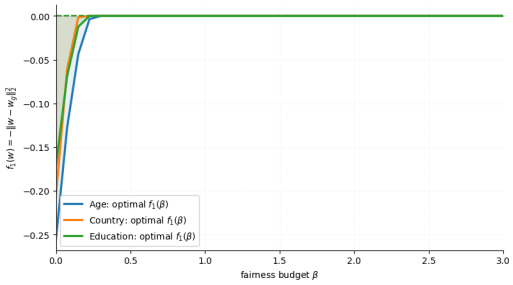


(c) *Accuracy comparison*: **Non-uniform costs** $A_1 = 2A_2 = 2I$; **desirable**: occupation, workclass, education, under ℓ_2-fairness constraint

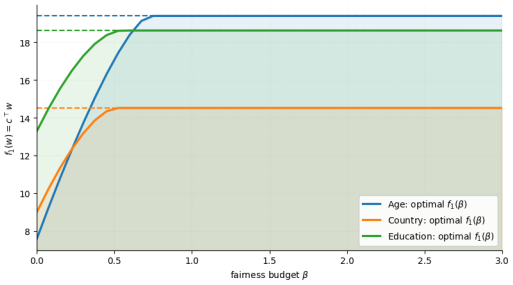


(d) *Social-welfare comparison*: **Non-uniform costs** $A_1 = 2A_2 = 2I$; **desirable**: occupation, workclass, education, under ℓ_2-fairness constraint

Figure 4: Optimal values for varying β for different group splits under ℓ_2**-fairness constraint**



(a) *Accuracy comparison*: **Non-uniform random costs** $A_1 = 2A_2$; **desirable**: occupation, workclass, education, under ℓ_1-fairness constraint



(b) *Social-welfare comparison*: **Non-uniform random costs** $A_1 = 2A_2$; **desirable**: occupation, workclass, education, under ℓ_1-fairness constraint

Figure 5: Optimal values for varying β for different group splits under **non-uniform random** cost matrices and the ℓ_1-fairness consntraint