Fair Robust Strategic Classification under Decision-Dependent Cost Uncertainty

Sura Alhanouti ALHANOUTI.1@OSU.EDU; SZALHANOTY14@ENG.JUST.EDU.JO
The Ohio State University, Ohio, USA
Jordan University of Science and Technology, Irbid, Jordan

Güzin Bayraksan BAYRAKSAN.1@OSU.EDU
The Ohio State University, Ohio, USA

Parinaz Naghizadeh PARINAZ@UCSD.EDU
University of California, San Diego, USA

Abstract

Humans are increasingly finding ways to strategically respond to algorithmic decision systems, raising concerns about the robustness and fairness of AI systems in critical contexts. Existing works on fair strategic classification, which aim to address these concerns, have largely focused on *static and known* cost environments. We instead propose a framework for *endogenously evolving and uncertain* cost environments, where strategic costs evolve over time in response to past decisions. We model the firm's fair classifier design problem as a two-stage robust optimization problem with decision-dependent uncertainty and endogenously evolving costs under the demographic parity (DP)-fairness constraint. We develop an analytically supported reformulation of the fairness constraint, enabling the solution of the resulting classifier design problem. Our analysis highlights trade-offs between robustness and fairness across stages and demographic groups, notably an asymmetric robustness that preserves stronger protections from disadvantaged groups.

1. Introduction

With the growing use of machine learning algorithms in high-stakes decision-making settings, such as lending, hiring, and resource allocation, individuals have begun to strategically modify their behaviors or inputs to algorithmic systems to improve their outcomes [14, 22, 24, 28, 30, 32]. This raises the need to design algorithms that are robust to such strategic "gaming" behavior; left unaddressed, this gaming will interfere with accurate predictions [e.g., 1, 7, 15, 18, 23].

At the same time, this re-design of algorithmic systems to account for humans' strategic responses has significant implications for the *algorithmic fairness* of the resulting models. Prior works suggest that strategic behavior in classification can induce or exacerbate group disparities when groups face different strategic response costs [10, 15, 23]. Moreover, any fairness interventions need to account for strategic responses [2, 12, 17, 20, 35], including by considering their long-term implications [35, 38]. (We review additional related work in Appendix A.)

In this work, we are similarly interested in designing classifiers that are robust to humans' strategic responses, while at the same time satisfying the fairness notion of *demographic parity*. Our main distinction is to account for the *endogenously-evolving and uncertain nature of the costs of strategic*

© S. Alhanouti, G. Bayraksan & P. Naghizadeh.

response. Much of the existing literature on strategic classification assumes that strategic response costs are *static* and *known*, an assumption that may not hold in practice. Particularly, the classifier's current decisions can shape the strategic response costs that agents will incur in the future. As an example, the increased shift towards test-optional admission policies during the COVID-19 pandemic caused a temporary decline in demand for standardized testing in succeeding years. Simultaneously, many families shifted their spending toward costly essay consultants and extracurricular enhancement services [8, 9, 11, 25, 27, 31]. Therefore, the schools' policy in COVID years shaped the application costs faced by future applicants, even though, crucially, the magnitude of those downstream effects was uncertain at the time of decision-making. Such choices of admission policies not only shaped applicants' costs, but may have also shifted inequalities; for instance, [26] found that students with a higher socio-economic status list more extracurricular activities than their peers, suggesting that the gap in achievements is primarily driven by unequal access to resources and opportunities rather than differences in individual initiative. Hence, as costs evolve endogenously and heterogeneously across different demographic groups, new disparities may emerge over time.

Motivated by these challenges, we develop a fair temporal strategic classification framework under endogenous and uncertain agent response costs. Building on recent work in [3, 4], we formulate the problem as a two-stage robust optimization (TSRO) model with decision-dependent uncertainty, and introduce the Demographic Parity (DP) fairness constraint in this framework. Our main technical contribution is to develop an analytically supported reformulation of the fairness constraint, which enables us to solve the resulting two-stage robust classifier design problem efficiently. Specifically, using an equivalent surrogate-score representation of the post-strategic acceptance rule, we establish a fairness-controlled distortion bound under a ramp surrogate loss and show that achieving ϵ-fairness requires stricter training-time constraints. We then develop a tractable solution approach integrating the convex-concave procedure with Benders column-and-constraint generation.

We also conduct numerical experiments using our developed framework and solution approach. We find that our fair dependency-aware classifier relaxes robustness to uncertainty for the advantaged group while maintaining stricter robustness for the disadvantaged group, which results in fewer undesired strategic responses by the disadvantaged group in both stages. Furthermore, compared to an unfair, dependency-aware baseline, it incurs a larger loss in the second stage to achieve fairness. This further underscores the trade-off between fairness and robustness to uncertainty.

2. Problem Setting

We consider a strategic binary classification setting in which a firm makes accept or reject decisions on agents with an observable *feature vector* $x \in \mathcal{X} \subseteq \mathbb{R}^d$, a binary *sensitive attribute* (e.g., sex, race) $s \in \mathcal{S} = \{a, b\}$, and an unobserved *qualification state* $y \in \mathcal{Y} = \{\pm 1\}$. Let $(X, Y, S) \sim P_{XYS}$ denote the joint distribution over $\mathcal{X} \times \mathcal{Y} \times \mathcal{S}$, with (x, y, s) representing a realization. Further, let p_s be the proportion of group s in the population. There are two time steps; at each, a new batch of agents arrive. For each step i, the firm selects a linear decision rule $\operatorname{sign}((\beta_s^i)^\top x)$ for each demographic group s, where $\beta_s^i \in \mathcal{B}$ represents the classifier coefficients for group s. We denote the stacked vector of group-specific classifier parameters at step i by $\beta^i := [\beta_a^i, \beta_b^i]^\top \in \mathbb{R}^{2d}$.

An agent from group s responds to any classifier β_s by adjusting its features x to

$$\hat{x}(\beta_s; x) := \arg\max_{\hat{x} \in \mathcal{X}} \left[\mathbb{1}(\beta_s^\top \hat{x} \geq 0)\, u - c_s(x, \hat{x}) \right], \tag{1}$$

where $u \geq 0$ is the utility of a positive classification and $c_s(x, \hat{x})$ is the cost of changing x to $\hat{x}$. Movement occurs only if the change yields a positive outcome and $c_s(x, \hat{x}) \leq u$. Similar to [3, 4, 29], we consider a general form for the cost function for group s, given by $c_s(x, \hat{x}) = \phi(\|\Sigma_s^{1/2}(\hat{x} - x)\|)$, where $\Sigma_s \succ 0$ is a positive definite *cost matrix* which uniquely parametrizes the cost function and encodes the geometry of the cost landscape for each group s, $\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a non-decreasing function, and $\|\cdot\|$ is the standard p-norm ($p \geq 1$).

We assume that the firm knows the exact strategic response cost function $c_s(x, \hat{x})$ in the first time step, parameterized by a homogeneous fixed Σ_0 across groups. The firm also anticipates that its first-step classifier influences the second-step costs; however, the exact magnitude of this impact is uncertain in the first step, and revealed only in the second stage. Specifically, we let the second step's cost matrix for each group s be a random function $\Sigma(\omega_s) := Q\big(\omega_s^{-2}\big) \cdot \Sigma_0$, where the random vector $\omega_s \in \Omega(\beta_s) \subseteq \mathbb{R}_+^d$, drawn from the uncertainty set $\Omega(\beta_s)$, encodes how the first-step classifier β_s affects the second-step cost, and $Q(\cdot) \in \mathbb{R}^{d \times d}$ is a matrix-valued transformation mapping the element-wise scaling factors $(\omega_{j,s}^{-2})$ to a cost-scaling matrix.

To summarize, let $\mathbf{\Sigma} := (\Sigma_s)_{s \in S}$ denote the profile of group-specific cost matrices. In the first step, this profile is given by the known shared matrix $\mathbf{\Sigma}_0 := (\Sigma_0)_{s \in S}$, while in the second step it is given by the uncertain, decision-dependent profile $\mathbf{\Sigma}(\omega) := (\Sigma(\omega_s))_{s \in S}$.

3. The Firm's Optimization Problem

We model the firm's fair decision process as a constrained two-stage robust optimization problem with decision-dependent uncertainty. We detail this problem in this section.

The learning objective. Given that efficient optimization of the typical learning objective of (expected) 0–1 loss is hindered by two issues, namely, non-convexity from the sign function and discontinuities in $\hat{x}(\beta_s; x)$, we utilize the *cost-aware strategic hinge loss* across groups in both stages as the learning objective, as proposed by [19, 29]; this incorporates agents' best responses through a cost-dependent margin adjustment. Specifically, we consider the empirical strategic hinge risk

$$\widehat{R}_{\mathbf{\Sigma}}(\beta) := \frac{1}{N} \sum_{s \in S} \sum_{i=1}^{N_s} \max\{0, 1 - y_i\big(\beta_s^\top x_i + u_* \|\beta_s\|_{*,\Sigma_s}\big)\}, \tag{2}$$

where $N = \sum_{s \in S} N_s$, u_* is the largest value satisfying $\phi(u_*) \leq u$, and $\|\beta_s\|_{*,\Sigma_s} := \sup_{\|v\|_{\Sigma_s}=1} \beta_s^\top v = \|\Sigma_s^{-1/2}\beta_s\|_*$ is the Σ_s*-transformed dual norm of* β_s, with $\|\cdot\|_*$ denoting the dual norm.

Constrained classifier design. We focus on Demographic Parity (DP) as the fairness criterion, which requires that selection (acceptance) rates be similar across demographic groups. Specifically, based on the induced post-strategic feature $\hat{X}(\beta_s)$, we define the population demographic parity (DP) gap of a classifier β as $\Delta_F(\beta) := \mathbb{P}\big(\beta_a^\top \hat{X}(\beta_a) \geq 0 \mid S = a\big) - \mathbb{P}\big(\beta_b^\top \hat{X}(\beta_b) \geq 0 \mid S = b\big)$. This quantity captures the disparity in acceptance probabilities between groups, which, crucially, should be evaluated under the *post-strategic* distribution of features. Accordingly, we define the population ϵ-fair feasible set as $\mathcal{B}_F(\epsilon) := \big\{\beta \in \mathbb{R}^{2d} : |\Delta_F(\beta)| \leq \epsilon\big\}$. We further introduce an optional operational feasible set capturing structural or policy restrictions on the classifier parameters: $\mathcal{B}_o := \big\{\beta \in \mathbb{R}^{2d} : \mathbf{A}\beta \geq \mathbf{b}\big\}$; see Appendix B for further details on this constraint set.

Combining fairness and operational constraints, the first-stage feasible set is $\mathcal{B} := \mathcal{B}_o \cap \mathcal{B}_F(\epsilon)$. Given training data, the empirical fairness set $\widehat{\mathcal{B}}_F(\epsilon)$ is obtained by replacing population proba-

bilities with their empirical counterparts. The resulting first-stage empirical feasible set is $\widehat{\mathcal{B}} := \mathcal{B}_o \cap \widehat{\mathcal{B}}_F(\epsilon)$. Similarly, after uncertainty realization ω, the second-stage feasible and empirical feasible sets are $\mathcal{B}'(\beta, \omega) := \mathcal{B}'_o(\beta, \omega) \cap \mathcal{B}'_F(\omega, \epsilon)$, and $\widehat{\mathcal{B}}'(\beta, \omega) := \mathcal{B}'_o(\beta, \omega) \cap \widehat{\mathcal{B}}'_F(\omega, \epsilon)$, respectively.

Modeling the uncertainty set. Our model of the *decision-dependent uncertainty set* follows that of [3, 4], extended to the multi-group setting. Define the stacked uncertainty vector $\omega := [\omega_a, \omega_b]^\top \in \mathbb{R}^{2d}_+$. The corresponding joint decision-dependent uncertainty set is defined as

$$\omega \in \Omega(\beta) := \left\{\omega : \mathbf{F}^{\text{blk}}(\beta)\omega \leq \mathbf{h}^{\text{blk}} + \mathbf{G}^{\text{blk}}\beta\right\}, \tag{3}$$

where $\mathbf{F}^{\text{blk}}(\beta)$ be the block-diagonal matrix with blocks $\mathbf{F}(\beta_s)$, and $\mathbf{h}^{\text{blk}}$ and $\mathbf{G}^{\text{blk}}$ analogously. We provide additional details about the choice of this set and its parameters in Appendix C.

The two-stage robust optimization problem. We now present the firm's fairness-constrained, two-stage robust optimization (TSRO) problem under decision-dependent uncertainty. Formally, at the first time step, the firm solves the following nested min–max–min program:

$$\min_{\beta \in \widehat{\mathcal{B}}} \left(\widehat{R}_{\mathbf{\Sigma}_0}(\beta) + \max_{\omega \in \Omega(\beta)} \min_{\beta' \in \widehat{\mathcal{B}}'(\beta,\omega)} \widehat{R}_{\mathbf{\Sigma}(\omega)}(\beta')\right), \tag{4}$$

where $\widehat{R}_{\mathbf{\Sigma}_0}(\beta)$ and $\widehat{R}_{\mathbf{\Sigma}(\omega)}(\beta')$ denote the empirical strategic hinge risks defined in (2), evaluated with cost matrices $\mathbf{\Sigma}_0$, $\mathbf{\Sigma}(\omega)$, respectively; $\Omega(\beta)$ is the joint decision-dependent uncertainty set introduced in (3); and $\widehat{\mathcal{B}}$ and $\widehat{\mathcal{B}}'(\beta, \xi)$ denote the first- and second-stage feasible sets. In the first step, the firm selects group-specific classifiers, subject to both fairness and operational constraints, to minimize its overall risk from first-stage agents and across both groups, while not only anticipating the first-stage agents' strategic behavior, but also looking forward to the need to uphold robustness and fairness under the worst-case uncertainty in second-stage costs that is induced by the firm's own first-stage decisions. Once the first-stage concludes, and the second-stage manipulation cost is realized, the firm will be selecting the second-stage classifier accordingly, anticipating the second-stage agents' strategic behavior under the *realized* endogenous costs.

This framework admits a Stackelberg-type leader-follower interpretation, where the firm anticipates strategic responses when selecting classifiers; however, note that we assume agents are short-lived. Therefore, unlike repeated or dynamic games, agents' equilibrium behavior across multiple rounds is outside the scope of the current framework.

4. Fairness Constraint Reformulation

Solving the problem in (4) poses several computational challenges. The challenges arising from the objective function are discussed in [3, 4] and are omitted in interest of space. Here, we focus on the additional challenges introduced by the demographic parity fairness constraints: group acceptance rates depend on post-strategic features in (1), which induce an implicit nonconvex response mapping without a closed-form expression. Moreover, the indicator-based acceptance rates render the fairness constraints discontinuous and nonsmooth. We address these challenges in this section.

Surrogate post-strategic score. Recall that, following the idea proposed in [19, 29], we replaced the true post-strategic score $\beta_s^\top \hat{x}_i(\beta_s; x_i)$ with a *surrogate* score $\beta_s^\top x_i + u_* \|\beta_s\|_{*,\Sigma_s}$. Our key result regarding its implications on the DP fairness constraints is stated below.

Lemma 1 *Under the strategic response model in* (1)*, for any cost matrix* Σ_s*, classifier* $\beta_s \in \mathcal{B}$*, and instance* x_i*, the acceptance decision based on the true post-strategic feature is equivalent to that based on the surrogate score. That is,* $\mathbb{1}\{\beta_s^\top \hat{x}_i(\beta_s; x_i) \geq 0\} = \mathbb{1}\{\beta_s^\top x_i + u_* \|\beta_s\|_{*,\Sigma_s} \geq 0\}$.

In words, the lemma states that the true post-strategic score and the surrogate score are sign-equivalent (see proof in Appendix F.1). We define the surrogate score as $\widetilde{SC}(\beta_s, x_i) := \beta_s^\top x_i + u_* \|\beta_s\|_{*,\Sigma_s}$. Using this score, we define the population DP gap under surrogate scores as $\Delta_{\widetilde{F}}(\beta) := \mathbb{P}\big(\widetilde{SC}(\beta_a, X) \geq 0 \mid S = a\big) - \mathbb{P}\big(\widetilde{SC}(\beta_b, X) \geq 0 \mid S = b\big)$. As a consequence of Lemma 1, surrogate scores induce identical acceptance decisions at both the population and empirical levels, leading to the following corollary. (See Appendix H.1 for numerical validation.)

Corollary 2 *For the strategic response model in* (1)*,* $|\Delta_{\widetilde{F}}(\beta)| - |\Delta_F(\beta)| = 0$.

Motivated by this equivalence, we can work with surrogate scores in place of the true post-strategic scores in the subsequent reformulations of the fairness constraints.

Ramp surrogate for acceptance indicators. We next address the discontinuity of the acceptance indicator by replacing it with a ramp-hinge surrogate defined as $g_r(\widetilde{SC}(\beta_s, x)) := \min\{1, \max\{0, (\widetilde{SC}(\beta_s, x) + \tau)/(2\tau)\}\}$, where $\tau > 0$ controls the width of the margin around the decision boundary. As $\tau \to 0$, this surrogate increasingly approximates the 0-1 indicator function. We define the *ramp-DP gap* using the surrogate scores as

$$\Delta_{\widetilde{F},g_r}(\beta) := \mathbb{E}\big[g_r(\widetilde{SC}(\beta_a, x)) \mid S = a\big] - \mathbb{E}\big[g_r(\widetilde{SC}(\beta_b, x)) \mid S = b\big],$$

and, similarly, its corresponding empirical counterpart $\widehat{\Delta}_{\widetilde{F},g_r}(\beta)$. The following Lemma establishes a bound on how much the population fairness gap assessed under the indicator function differs from the fairness gap assessed under our proposed ramp-hinge relaxation of the indicator function.

Lemma 3 *Let* $g_r : \mathbb{R} \to [0, 1]$ *be a ramp surrogate with parameter* $\tau > 0$*. Then, for any decision vector* β*, the DP fairness distortion satisfies* $\big|\Delta_{\widetilde{F}}(\beta) - \Delta_{\widetilde{F},g_r}(\beta)\big| \leq \frac{1}{2}\sum_{s\in\{a,b\}} \alpha_s(\tau, \beta_s)$*, where* $\alpha_s(\tau, \beta_s) := \mathbb{P}(|\beta_s^T X + u_* \|\beta_s\|_{*,\Sigma_s}| \leq \tau)$ *is the* τ-margin band-mass *of group* s *under* β_s.

Lemma 3 shows that the fairness distortion induced by the ramp-hinge surrogate is governed by the choice of the margin parameter τ and by the probability mass near the decision boundary (see Appendix F.2 for proof, and Appendix H.2 for numerical validation). The following corollary uses Lemma 3 to specify fairness guarantees obtainable under the ramp-hinge loss.

Corollary 4 *Let* $g_r : \mathbb{R} \to [0, 1]$ *be the ramp surrogate with parameter* $\tau > 0$*. Define the total margin band-mass* $\alpha(\tau, \beta) := \frac{1}{2}\sum_{s\in\{a,b\}} \alpha_s(\tau, \beta_s)$*. Then for any* $\epsilon \geq 0$*,* $|\Delta_{\widetilde{F},g_r}(\beta)| \leq \epsilon \quad \Rightarrow |\Delta_{\widetilde{F}}(\beta)| \leq \epsilon + \alpha(\tau, \beta)$.

In words, if we want an ϵ delivered fairness gap, we need to impose a *more strict* fairness requirement when training under the ramp-hinge relaxation (proof in Appendix F.3). Specifically, in order to guarantee an ϵ fairness gap, we define the optimal τ as the maximum value such that $\alpha(\tau, \beta) \leq \bar{\alpha}$, where $\bar{\alpha}$ represents the maximum allowable distortion. This choice effectively bounds the gap between the assessed and true fairness metrics for a fixed classifier. Then, for a given parameter τ of the ramp-hinge loss, if we want a fairness gap of ϵ, we should impose a fairness

gap of $\epsilon - \alpha(\tau, \beta)$ during training. The parameter τ can then be adjusted relative to ϵ, to prevent $\epsilon - \alpha(\tau, \beta)$ from becoming too small, thus limiting the fair feasible set and compromising optimality. (Additional empirical fairness and risk bounds are provided in Appendix D.)

We are now ready to solve the firm's fair robust classifier design problem in (4). Specifically, to solve (4) using the C&CG algorithm with the DP-ramp-hinge constraint, we use the Difference-of-Convex (DC) structure of the ramp function, enabling linearization of the anticipated second-stage fairness constraints. We then apply the convex-concave procedure (CCP) to iteratively enforce fairness in the nonlinear first-stage problem. (See Appendix E for detailed discussion.)

We note that our theoretical analysis establishes fairness guarantees for the first-stage classifier choice through the proposed reformulation. For the *anticipated* second-stage problem embedded in the TSRO, we employ a *fixed* ramp-surrogate fairness reformulation. This is a relaxation as the distortion induced by a fixed surrogate parameter may exceed the prescribed bound, and therefore the corresponding true second-stage fairness guarantee may not hold. That said, we note that this concerns only the *anticipated* second-stage problem within the TSRO rather than the *realized* deterministic second-stage problem; the realized classifier for the second-stage is selected after the realization of uncertainty and therefore satisfies fairness guarantees. That said, extending the fairness guarantee of the anticipated second-stage (potentially through an *adaptive* ramp-surrogate fairness reformulation) could improve solution quality and remains open. An adaptive selection of τ is discussed in Appendix E.1; incorporating it into the C&CG framework remains for future work, as varying τ alters the second-stage feasible set and the corresponding dual formulation used to derive the feasibility and optimality cuts.

5. Numerical Experiments

We now numerically evaluate the performance of our proposed solution approach. We conduct experiments on a synthetic dataset. See Appendix G for dataset, experiment setup, and algorithmic implementation details. Note that the reported second-stage results correspond to the realized values for the second-stage optimization *after* the uncertainty is observed, rather than the anticipated second-stage problem that is solved while embedded within the TSRO.

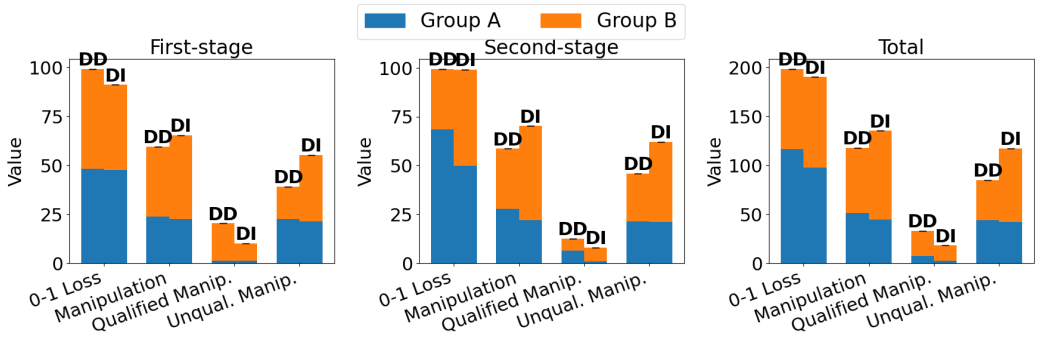


Figure 1: Performance comparison: fair DD-classifier vs. DI-classifier.

We first compare our fair decision-dependent (F-DD) classifier against a baseline fair decision-independent (F-DI) classifier that ignores the dependence of strategic response costs on decisions.

Figure 1 shows that F-DD exhibits a higher overall true loss (198.53 vs. 190.11), but achieves lower total manipulation (118.29 vs. 135.34). This reduction is driven by a shift in who manipulates: more manipulation by qualified agents (33.04 vs. 18.07). This pattern holds across both groups, but is more pronounced for the disadvantaged group (B). At the same time, we observe less manipulation by unqualified agents overall (84.89 vs. 117.27), with the reduction being particularly evident within the disadvantaged group. This behavior arises from a sacrifice in robustness to decision-dependent uncertainty in manipulation costs, particularly for group A, as illustrated in Figure 2. The F-DD classifier hold stronger robustness for group B, but is relaxed for group A. In fact, compared to the F-DI classifier, slightly more unqualified agents from group A manipulate under the F-DD classifier (44.04 vs. 42.46). Overall, the F-DD classifier is more robust to manipulation by unqualified agents across both groups, but less robust to manipulation by qualified agents.

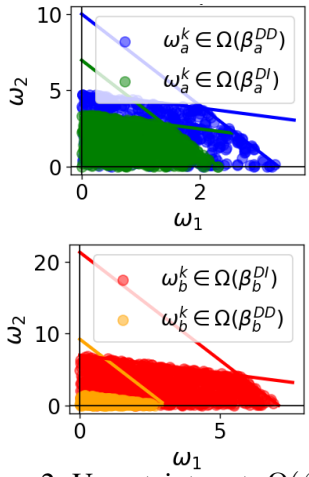


Figure 2: Uncertainty sets $\Omega(\beta_s^{\mathrm{DD}})$ vs $\Omega(\beta_s^{\mathrm{DI}})$.

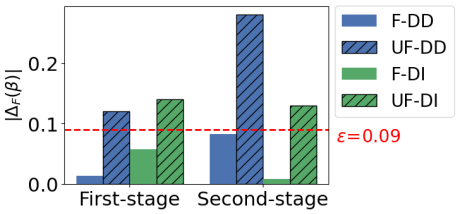


Figure 3: Fairness gap: DD vs. DI.

Finally, Figure 3 shows that the true demographic parity gap of the F-DD classifier satisfies the fairness constraint at both stages (target was $\leq \epsilon = 0.09$). In the first stage, F-DD achieves tighter fairness than F-DI, at the cost of higher loss, mainly driven by group B, which contributes more manipulation and thus higher loss, including more manipulation of qualified agents. However, it exhibits looser fairness in the second stage than F-DI, while both have similar loss. Under F-DD, loss is primarily incurred by the advantaged group (A), and the classifier remains robust to manipulation, particularly by unqualified agents in the disadvantaged group. Overall, incorporating fairness with decision-dependent costs induces a trade-off: higher loss in exchange for improved robustness, fairness, and control of manipulation, especially in the disadvantaged group. Lastly, compared to unfair dependency-independent (UF-DI) classifiers, unfair dependency-dependent (UF-DD) classifiers exhibit lower fairness, particularly in the second stage, indicating a trade-off of fairness for improved robustness to uncertainty and strategic manipulation. For completeness, we compare unfair DD [4] and unfair DI for both groups in Appendix G.5. We also compare the performance of fair and unfair dependency-aware classifiers in Appendix G.4.

Camera-Ready Edits Summary

The camera-ready version incorporates revisions based on the reviewers' feedback. The main changes are summarized below.

- **Theoretical guarantees and technical contributions.** We clarify the scope of the theoretical guarantees by distinguishing the first-stage fairness guarantee from the anticipated second-stage problem within the two-stage robust optimization framework. In particular, we clarify that the anticipated second-stage problem employs a fixed-surrogate fairness formulation, for which the corresponding true fairness guarantee is not established, and distinguish this from the realized deterministic second-stage problem. We also better distinguish the technical contributions of this work from components adopted from prior robust strategic classification frameworks.
- **Game-theoretic contextualization.** We clarify the Stackelberg-type leader–follower interpretation of the proposed framework, discuss its relation to robust bilevel optimization, and explicitly state that repeated strategic adaptation and multi-round equilibrium analysis are outside the scope of this work.
- **Algorithmic implementation details.** We added implementation details for the proposed CCP–Benders C&CG algorithm, including stopping criteria, iteration counts, Benders cuts, and runtime, in a new appendix.
- **Related work and references.** We updated the introduction with more focused citations and additional references supporting the motivating examples.

References

[1] Saba Ahmadi, Kunhe Yang, and Hanrui Zhang. Strategic littlestone dimension: Improved bounds on online strategic classification. *Advances in Neural Information Processing Systems*, 37:101696–101724, 2024.

[2] Sura Alhanouti and Parinaz Naghizadeh. Anticipating gaming to incentivize improvement: Guiding agents in (fair) strategic classification. *arXiv preprint arXiv:2505.05594*, 2025.

[3] Sura Alhanouti, Guzin Bayraksan, and Parinaz Naghizadeh. Robust strategic classification under decision-dependent cost uncertainty. In *NeurIPS 2025 Workshop MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making*, 2025.

[4] Sura Alhanouti, Guzin Bayraksan, and Parinaz Naghizadeh. Robust strategic classification under decision-dependent cost uncertainty. In *International Conference on Machine Learning*, 2026.

[5] Yahav Bechavod, Chara Podimata, Steven Wu, and Juba Ziani. Information discrepancy in strategic learning. In *International Conference on Machine Learning*, 2022.

[6] Mark Braverman and Sumegha Garg. The role of randomness and noise in strategic classification. In *1st Symposium on Foundations of Responsible Computing, FORC 2020*, 2020.

[7] Lee Cohen, Saeed Sharifi-Malvajerdi, Kevin Stangl, Ali Vakilian, and Juba Ziani. Bayesian strategic classification. *Advances in Neural Information Processing Systems*, 37:111649–111678, 2025.

[8] David Deming. The worst way to do college admissions, 2024. URL https://www.theatlantic.com/ideas/archive/2024/03/standarized-testing-requirements-act-sat/677667/. Accessed: 2025-07-30.

[9] Halle Edwards. How much do ap tests cost?, 2022. URL https://blog.prepscholar.com/how-much-do-ap-tests-cost. Accessed: 2025-07-30.

[10] Valia Efthymiou, Ekaterina Fedorova, and Chara Podimata. Desirable effort fairness and optimality trade-offs in strategic learning. *arXiv preprint arXiv:2510.19098*, 2025.

[11] Val Ellicott. Standardized testing: Should higher education drop admission exams permanently? *CQ Researcher*, 2022. doi: 10.4135/cqresrre20220506.

[12] Andrew Estornell, Sanmay Das, Yang Liu, and Yevgeniy Vorobeychik. Unfairness despite awareness: Group-fair classification with strategic agents. In *AAMAS 2022 Workshop on Learning with Strategic Agents*, 2022.

[13] 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*, pages 389–399, 2023.

[14] Haleluya Hadero. The internet is filled with fake reviews. here are some ways to spot them. The Associated Press, December 2024. URL https://apnews.com/article/fake-online-reviews-generative-ai-40f5000346b1894a778434ba295a0496. Accessed: Jac 24, 2026.

[15] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 259–268, 2019.

[16] Xiaolin Huang, Lei Shi, and Johan AK Suykens. Ramp loss linear programming support vector machine. *The Journal of Machine Learning Research*, 15(1):2185–2211, 2014.

[17] 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*, pages 1–11, 2023.

[18] Tosca Lechner, Ruth Urner, and Shai Ben-David. Strategic classification with unknown user manipulations. In *International Conference on Machine Learning*, pages 18714–18732. PMLR, 2023.

[19] Sagi Levanon and Nir Rosenfeld. Generalized strategic classification and the case of aligned incentives. In *International Conference on Machine Learning*, 2022.

[20] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.

[21] Lydia T Liu, Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer Chayes. The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 381–391, 2020.

[22] Christopher Maag. The hustlers who make $6,000 a month by gaming citi bikes. The New York Times, September 2024. URL https://www.nytimes.com/2024/09/19/nyregion/citi-bike-scam-nyc.html. Online; accessed August 31, 2025.

[23] 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*, pages 230–239, 2019.

[24] Marieke Möhlmann and Lior Zalmanson. Hands on the wheel: Navigating algorithmic management and uber drivers'. In *Proceedings of the International Conference on Information Systems*, 2017.

[25] Navagant. Test prep industry report. Technical report, Navagant, 2024. URL https://navagant.com/wp-content/uploads/2024/08/Test-Prep-Industry-Report_vF.pdf.

[26] Julie J Park, Brian Heseung Kim, Nancy Wong, Jia Zheng, Stephanie Breen, Pearl Lo, Dominique J Baker, Kelly Rosinger, Mike Hoa Nguyen, and OiYan A Poon. Inequality beyond standardized tests: Trends in extracurricular activity reporting in college applications across race and class. *American Educational Research Journal*, 62(2):336–377, 2025.

[27] Anna Petrosino. Scores in, policies out: Schools reconsider stances on standardized testing, 2025. URL https://www.inklingsnews.com/b/2025/05/09/scores-in-policies-out-schools-reconsider-stances-on-standardized-testing. Accessed: 2025-07-30.

[28] Priyanjana Pramanik. What happens when job candidates face ai instead of humans? Medical Research News, Jun 2025. URL https://www.news-medical.net/news/20250625/What-happens-when-job-candidates-face-AI-instead-of-humans.aspx. Online; accessed Jan 24, 2026.

[29] Elan Rosenfeld and Nir Rosenfeld. One-shot strategic classification under unknown costs. In *Proceedings of the 41st International Conference on Machine Learning*, pages 42719–42741, 2024.

[30] Paul Solman. How uber drivers game the app and force surge pricing. PBS NewsHour, August 2017. URL https://www.pbs.org/newshour/economy/uber-drivers-game-app-force-surge-pricing. Available online.

[31] Laura Spitalniak. Wealthier students, those at private schools list more extracurriculars on college applications, 2023. URL https://www.k12dive.com/news/application-advantage-extracurriculars-wealthy-white-asianstudents/648171. Accessed: 2025-07-30.

[32] Ashley Stahl. 5 resume hacks to pass ATS. Forbes, 2022. URL https://www.forbes.com/sites/ashleystahl/2022/12/12/5-resume-hacks-to-pass-ats.

[33] Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 526–536, 2021.

[34] Bo Zeng and Wei Wang. Two-stage robust optimization with decision dependent uncertainty. *arXiv preprint arXiv:2203.16484*, 2022.

[35] Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, and Cheng Zhang. How do fair decisions fare in long-term qualification? In *Advances in Neural Information Processing Systems*, 2020.

[36] Xueru Zhang, Mohammad Mahdi Khalili, Kun Jin, Parinaz Naghizadeh, and Mingyan Liu. Fairness interventions as (dis) incentives for strategic manipulation. In *International Conference on Machine Learning*, pages 26239–26264. PMLR, 2022.

[37] Zhiqun Zuo and Mohammad Mahdi Khalili. Individual fairness in strategic classification. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[38] Zhiqun Zuo, Tian Xie, Xuwei Tan, Xueru Zhang, and Mohammad_Mahdi Khalili. Lookahead counterfactual fairness. *Transactions on machine learning research*, 2024.

[39] Zhiqun Zuo, Ding Zhu, and Mohammad Mahdi Khalili. Post-processing for fair regression via explainable svd. *arXiv preprint arXiv:2504.03093*, 2025.

Appendix A. Related Work

Fairness in machine learning algorithms with strategic behavior has gained increasing attention, especially as we find that real-world agents actively respond to algorithms to gain more favorable outcomes. The majority of the existing literature assumes full information and transparency between the classifier and agents. The key complexity in incorporating fairness in strategic classification settings is that the constraint changes the agents' responses, which leads to shifting the data distribution that the fairness metric meant to handle. Prior works have addressed this challenge through constrained optimization [2, 13, 35–37], randomized classification [6, 37], and post-processing algorithms [39]. However, these works operate under the assumption of full information about agents' costs, and do not consider the uncertainty or endogenous evolution of costs.

The prevailing assumption of full information has been challenged in the context of fair classification; in particular, [33] study classification with group-dependent noisy labels and show that ignoring such uncertainty can simultaneously harm both fairness and accuracy. However, this work does not account for strategic settings.

Although recent work [3, 4] studies endogenously evolving manipulation costs, it considers only unfair settings at the population level and does not analyze the impact across demographic groups. In contrast, we incorporate fairness under temporal shifts in the environment induced by first-stage classifier decisions by jointly minimizing current loss and anticipated worst-case future loss while enforcing fairness across both stages. This yields robustness to strategic adaptations and preserves fairness in the post-response distribution.

We also note that another differentiator between our proposed model and the existing models of fair strategic classification is in the modeling of the cost functions underlying agents' responses. Most works use either an action-based cost [2, 21, 36] or a distance-based cost [13, 37], where feature dimensions are treated as equally weighted, failing to account for the nuanced difficulties of altering different attributes. In contrast, recent work has introduced cost matrices to capture heterogeneous costs of altering different attributes as well as the correlations between them [5, 17, 29]. Both Bechavod et al. [5] and Keswani and Celis [17] incorporate such cost matrices within quadratic cost functions: the former studies information design and welfare implications, while the latter addresses the "social burden gap" through a constrained optimization framework. Our work departs from these approaches in both objective and formulation. Specifically, rather than focusing on welfare maximization or recourse equality, we study fairness through a constrained (two-stage) optimization problem that directly minimizes misclassification loss. Building on the norm induced by the cost matrix introduced by Rosenfeld and Rosenfeld [29], we analyze fairness interventions.

Our key methodological contribution is a ramp-surrogate reformulation of the demographic parity constraint that makes the resulting optimization problem tractable. This reformulation enables integration of the convex-concave procedure (CCP) with Benders column-and-constraint generation (C&CG) to solve the resulting fair TSRO problem under decision-dependent uncertainty.

Appendix B. Operational Feasible Set: Detailed Description

The matrix $\mathbf{A}$ may be block-diagonal, imposing structural constraints independently within each group, or may include cross-group terms encoding additional modeling or operational restrictions. For example, in the context of university admissions, an institution may restrict the weight assigned to GPA, limit interaction terms between GPA and essay scores, or require certain coefficients to

be equal across demographic groups. Such requirements can be naturally encoded through the constraint matrix $\mathbf{A}$ and vector $\mathbf{b}$.

We denote the first-stage operational constraints by $(\mathbf{A}_1, \mathbf{b}_1)$ and the second-stage constraints by $(\mathbf{A}_2, \mathbf{b}_2)$. In the second stage, $\mathbf{b}_2$ may depend on the first-stage decision β and realized uncertainty ω, enabling the feasible set to capture adaptive policy adjustments, such as tightening or relaxing bounds on feature weights in response to observed behavioral shifts. For example, in school admissions, shifts in applicant behavior following first-stage policy changes may necessitate bounds on certain feature weights, which can be encoded through $(\mathbf{A}_2, \mathbf{b}_2)$.

Appendix C. Interpreting the Decision-Dependent Uncertainty (DDU) Set Parameter

For the group-dependent decision-dependent uncertainty (DDU) set defined as

$$\omega_s \in \Omega(\beta_s) := \left\{ \omega \in \mathbb{R}^d_+ : \mathbf{F}(\beta_s)\, \omega \leq \mathbf{h} + \mathbf{G}\beta_s \right\}.$$

Here, $\mathbf{h}$ serves as a baseline constraint (e.g., capturing average budgets or behavioral limits). The matrices $\mathbf{F}(\beta_s)$ and $\mathbf{G}$ capture how β_s influences the structure and bounds of the uncertainty set, respectively

If $\mathbf{F}$ is a fixed (β_s-independent) matrix, the inequalities constrain each cost component independently. In contrast, allowing $\mathbf{F}$ to depend on β_s means that the classifier can alter how different cost components interact. In practice, each component of the matrices can be set using domain expertise or estimated from historical data on strategic behavior; see Appendix C for more discussion This formulation allows the firm's initial decision for each group to influence both the magnitude and direction of the uncertainty governing second-stage costs.

In practice, each component of the matrices $\mathbf{F}$,$\mathbf{h}$, and $\mathbf{G}$ encode economically meaningful constraints and can be specified using domain expertise, data-driven estimation, or both. The vector $\mathbf{h}$ determines the baseline upper bounds on the components of ω, which determine the realization of the second-stage cost $\Sigma(\omega)$ via $g(\omega)$. For instance, in a simple illustrative case where $\Sigma(\omega) = \mathrm{diag}(2\omega_1, 5\omega_2)$ and $\omega_i \leq h_i$, choosing $h_i = 1$ could encode expected decreases in preparation costs based on expert assessment or historical data. More generally, $\mathbf{h}$ can be set using domain expertise (e.g., anticipated shifts in coaching or tutoring markets) or estimated from data on past behavioral responses.

The matrix $\mathbf{G}$ determines how the classifier's coefficients influence the feasible magnitude of manipulation incentives. Positive diagonal entries G_{ii} indicate that increasing weight on feature i expands the admissible range of the corresponding ω_i, while negative values imply the opposite. Off-diagonal terms encode cross-feature effects—for example, how emphasizing one criterion may indirectly relax or tighten incentives to adjust another. These relationships can likewise be informed by domain experts (e.g., admissions officers' beliefs about how applicants reallocate effort) or estimated empirically from behavioral data linking past policy changes to observed manipulation patterns. In this way, $\mathbf{G}$ provides a structured mechanism for capturing decision–uncertainty interactions.

The matrix $\mathbf{F}$ imposes structural relationships among the components of ω. Off-diagonal entries represent how increases in one type of preparation constrain or interact with other forms of preparation. For example, additional GPA-focused tutoring may limit resources available for extracurricular coaching, or vice versa. These couplings can be estimated from historical correlations

in preparation behavior or specified by experts familiar with how various investments substitute for or complement each other.

Appendix D. Empirical and Risk Guarantees

Empirical fairness gap bounds. We next establish that the above population-level fairness gap bounds when adopting our proposed ramp-hinge relaxation can be extended to the statistical learning setting. Specifically, the following lemma shows that enforcing surrogate ramp-fairness constraints on finite samples leads to population fairness guarantees.

Lemma 5 (Population-empirical distortion for surrogate ramp-fairness) *Let g_r be the ramp function, which is L-Lipschitz (with $L = 1/(2\tau)$). Assume that $\|\beta_s\|_{*,\Sigma_s} \leq B$ for all $\beta \in \mathcal{B}$ and that $\|X\| \leq R$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the draw of the sample from both groups,*

$$\sup_{\beta\in\mathcal{B}} \left|\Delta_{\widetilde{F},g_r}(\beta) - \widehat{\Delta}_{\widetilde{F},g_r}(\beta)\right| \leq \eta, \qquad \eta := \sum_{s\in\{a,b\}} \frac{2LB(R+u_*) + \sqrt{\ln(4/\delta)/2}}{\sqrt{n_s}}.$$

The proof is in Appendix F.4. This lemma quantifies how well the empirical ramp-fairness gap estimates its population counterpart, with the bound highlighting that generalization improves with larger group-specific sample sizes. Consequently, we obtain the following generalization guarantee on the surrogate ramp-fairness constraints.

Corollary 6 (Generalization of surrogate ramp-fairness constraints) *Under the conditions of Lemma 5, if a classifier $\hat{\beta}$ satisfies the empirical constraint $|\widehat{\Delta}_{\widetilde{F},g_r}(\hat{\beta})| \leq \varepsilon$, then with probability at least $1-\delta$,*

$$|\Delta_{\widetilde{F},g_r}(\hat{\beta})| \leq \varepsilon + \eta.$$

The proof is in Appendix F.5. In words, the corollary states that a classifier trained to satisfy surrogate ramp-fairness constraints on finite samples will remain approximately fair in expectation, with an explicit trade-off between sample size and fairness tolerance.

Bounds on risk evaluation. Lastly, we establish a population-level comparison of the objective function (risk) obtained from surrogate hinge-ramp relaxations vs. the true indicator-based fairness constraints. The following result shows how optimizing the strategic hinge risk under a surrogate ramp-fair constraint relates to hinge-risk optimization under a suitably relaxed indicator fairness constraint, and how this relationship induces a bound on the corresponding population 0-1 risk.

Lemma 7 (Population risk certification under surrogate ramp-fairness) *Let $R_{\mathbf{\Sigma}}(\beta)$ be the population strategic hinge risk corresponding to the empirical risk $\widehat{R}_{\mathbf{\Sigma}}(\beta)$ defined in (2), and $R^{0-1}_{\mathbf{\Sigma}}(\beta)$ denote the associated population 0-1 risk. Let $\beta^*_{\widetilde{F}}(\epsilon)$ and $\beta^*_{\widetilde{F},g_r}(\epsilon)$ denote any minimizer of $R_{\mathbf{\Sigma}}(\cdot)$ with DP-fairness target ϵ, under the surrogate scores and the ramp-hinge surrogate with parameter $\tau > 0$, respectively. Then*

$$R^{0-1}_{\mathbf{\Sigma}}\left(\beta^*_{\widetilde{F}}(\epsilon)\right) \leq R_{\mathbf{\Sigma}}\left(\beta^*_{\widetilde{F}}(\epsilon)\right) \leq R_{\mathbf{\Sigma}}\left(\beta^*_{\widetilde{F},g_r}(\epsilon - \alpha(\tau))\right).$$

The first inequality reflects the standard dominance of the hinge loss over the 0-1 loss, while the second inequality shows that the hinge risk achieved under the surrogate ramp-DP fairness constraint at the more conservative level $\epsilon - \alpha(\tau)$ upper bounds the hinge risk attainable under the relaxed indicator at fairness level ϵ. This follows from the fact that $\beta^*_{\widetilde{F}}(\epsilon)$ is a minimizer of $R_{\mathbf{\Sigma}}(\cdot)$, and further that a more conservative fairness target $\epsilon - \alpha(\tau)$ restrict the search space compared to a fairness constraint with target level ϵ, under which $\beta^*_{\widetilde{F}}(\epsilon)$ is obtained.

Together with the preceding results, this establishes an end-to-end guarantee linking tractable surrogate ramp-DP fair optimization to population-level fairness and classification performance.

Appendix E. Tractable reformulations

We have analytically established that the use of the surrogate score, a ramp-hinge approximation, and a stricter fairness gap, allows us to obtain any desired fairness guarantee that would be attainable had we used the exact scores and the indicator function with the fairness constraints.

First, for issues in the second-stage objective, we employ same reformulation and linearization in [4]. Here, we present the second-stage problem after removing the uncertainty from the objective function and linearization.

$$\min_{\beta'^+,\beta'^-,s_2,t_2^q,t_2^\omega,z} \frac{1}{N}\sum_{i=1}^{N} s_{2,i} \tag{5}$$

subject to

$$s_{2,i} \geq 1 - y_i\left((\beta'^+ - \beta'^-)^\top x_i + u_* M z\right), \forall i \in [N], \tag{(5).a}$$

$$-t_2^q \leq \beta_j'^+ - \beta_j'^- \leq t_2^q, \quad \forall j \in [d], \tag{(5).b}$$

$$-t_2^\omega \leq \textstyle\sum_{r=1}^{d} [\Sigma(\omega)^{-\frac{1}{2}}]_{jr} \leq t_2^\omega, \quad \forall j \in [d], \tag{(5).c}$$

$$\mathbf{B}_2(\beta'^+ - \beta'^-) \geq \mathbf{d} - \mathbf{B}_1(\beta^+ - \beta^-) - \mathbf{E}\omega, \tag{(5).d}$$

$$t_2^q \leq t_{2,\max}^q, \quad t_2^\omega \leq t_{2,\max}^\omega, \tag{(5).e}$$

$$z \leq t_{2,\max}^\omega t_2^q, \quad z \leq t_{2,\max}^q t_2^\omega, \tag{(5).f}$$

$$z \geq t_{2,\max}^q t_2^\omega + t_{2,\max}^\omega t_2^q - t_{2,\max}^\omega t_{2,\max}^q, \tag{(5).g}$$

$$\beta_j'^+, \beta_j'^-, s_{2,i}, t_2^q, t_2^\omega, z \geq 0, \ \forall i \in [N], \forall j \in [d]. \tag{(5).h}$$

Despite its generalization guarantees, the ramp-based fairness constraint still leads to a nonconvex optimization problem due to the ramp function $g_r(\cdot)$, which causes computational challenges for solving our TSRO problem. To address this, we exploit the Difference-of-Convex (DC) structure of the ramp function, which facilitates linearization of the fairness constraints in the second stage. We first introduce a linearization of the fairness constraints specifically for the second-stage problem. We then turn to the first-stage problem.

DC decomposition of the ramp function. Let $r_{i,s} = \frac{\widetilde{SC}(\beta_s, x_i)+\tau}{2\tau}$. We express the ramp function using a difference-of-convex decomposition [16]

$$g_r(r_{i,s}) = (r_{i,s})_+ - (r_{i,s} - 1)_+,$$

where $(z)_+ := \max\{0, z\}$. Using this representation, the ramp-DP fairness gap can be expressed as a difference of convex functions,

$$\Delta_{\tilde{F},g_r}(\beta) = P(\beta) - N(\beta), \tag{7}$$

where

$$P(\beta) = \Big(\mathbb{E}[(r_a)_+] + \mathbb{E}[(r_b - 1)_+]\Big), \tag{8}$$

$$N(\beta) = \Big(\mathbb{E}[(r_a - 1)_+] + \mathbb{E}[(r_b)_+]\Big). \tag{9}$$

Both $P(\beta)$ and $N(\beta)$ are convex functions of β, which yields a DC representation of the fairness constraint.

Fair second-stage problem linearization. To impose the ramp-based fairness constraint in the second stage, we handle each max operator via epigraph variables $e_{1,i,s}, e_{2,i,s} \in \mathbb{R}_+$, for $i \in [N_s]$ and $s \in \{a, b\}$, subject to

$$e_{1,i,s} \geq r_{i,s}, \quad e_{2,i,s} \geq r_{i,s} - 1.$$

Therefore, $P(\beta) = \frac{1}{N_a}\sum_{i=1}^{N_a} e_{1,i,a} + \frac{1}{N_b}\sum_{i=1}^{N_b} e_{2,i,b}$, and $N(\beta) = \frac{1}{N_a}\sum_{i=1}^{N_a} e_{2,i,a} + \frac{1}{N_b}\sum_{i=1}^{N_b} e_{1,i,b}$. Let $e_{1,s} = (e_{1,1,s}, \ldots, e_{1,N_s,s}) \in \mathbb{R}_+^{N_s}$, and $e_{2,s} = (e_{2,1,s}, \ldots, e_{2,N_s,s}) \in \mathbb{R}_+^{N_s}$ denote the vectors of these new variables at the second-stage for $s \in \{a, b\}$. We further penalize the ramp variables in the objective, which encourages these epigraph constraints to be tight at optimality.

Specifically, the second-stage objective in (5) becomes

$$\frac{1}{N}\sum_{s\in\{a,b\}}\sum_{i=1}^{N_s} s_{2,i,s} + \lambda_F \sum_{s\in\{a,b\}}\sum_{i=1}^{N_s}\big(e_{1,i,s} + e_{2,i,s}\big).$$

The following approximated and linearized fairness constraints are added to the second-stage formulation, in addition to ((5).a)–((5).h). Unless stated otherwise, all constraints below are imposed and should hold for each $s \in \{a, b\}$:

$$e_{1,i,s} \geq \frac{(\beta_s^+ - \beta_s^-)^\top x_i + u_* M z_s + \tau}{2\tau}, \quad \forall i \in [N_s], \tag{(5).i}$$

$$e_{2,i,s} \geq \frac{(\beta_s^+ - \beta_s^-)^\top x_i + u_* M z_s + \tau}{2\tau} - 1, \forall i \in [N_s], \tag{(5).j}$$

$$\Big(\frac{1}{N_a}\sum_{i=1}^{N_a} e_{1,i,a} + \frac{1}{N_b}\sum_{i=1}^{N_b} e_{2,i,b}\Big) - \Big(\frac{1}{N_a}\sum_{i=1}^{N_a} e_{2,i,a} + \frac{1}{N_b}\sum_{i=1}^{N_b} e_{1,i,b}\Big) \leq \epsilon. \tag{(5).k}$$

$$\Big(\frac{1}{N_a}\sum_{i=1}^{N_a} e_{2,i,a} + \frac{1}{N_b}\sum_{i=1}^{N_b} e_{1,i,b}\Big) - \Big(\frac{1}{N_a}\sum_{i=1}^{N_a} e_{1,i,a} + \frac{1}{N_b}\sum_{i=1}^{N_b} e_{2,i,b}\Big) \leq \epsilon, \tag{(5).l}$$

$$e_{1,i,s}, e_{2,i,s} \geq 0, \qquad \forall i \in [N_s]. \tag{(5).m}$$

Fair first-stage problem formulation. The first stage problem remains nonlinear, unlike [4]. We therefore apply the Convex-Concave Procedure (CCP) to iteratively enforce fairness via linearized cuts. Specifically, at each iteration, CCP linearizes the concave component of the fairness constraint, generating convex cuts that progressively refine the feasible region of the master problem.

Since the fairness constraint is given by $|\Delta_{\tilde{F},g_r}(\beta)| \leq \epsilon$, the resulting nonlinear first-stage problem is

$$\min_{\beta} \frac{1}{N} \sum_{s\in\{a,b\}} \sum_{i=1}^{N_s} \Big(1 - y_i\Big(\beta_s^\top x_i + u_* \|\beta_s\|_{*,\Sigma_0}\Big)\Big)_+ \tag{11}$$

subject to

$$P(\beta) - N(\beta) \leq \epsilon \tag{((11).a)}$$
$$-P(\beta) + N(\beta) \leq \epsilon \tag{((11).b)}$$

E.1. Final reformulation of the firm's problem.

The final, tri-level reformulation of our original problem (4) under decision-dependent uncertainty is

$$\min_{\beta\in\mathcal{S}_1} \left(\widehat{R}_{\mathbf{\Sigma}_0}(\beta) + \max_{\omega\in\Omega(\beta)} \min_{v_2\in\mathcal{S}_{t2}(\omega)} \sum_{s\in\{a,b\}} \sum_{i=1}^{N_s} \frac{1}{N} s_{2,i,s} + \lambda_F\big(e_{1,i,s} + e_{2,i,s}\big) \right). \tag{13}$$

Here, the second-stage decision vector is $v_2 := \Big(\beta_s'^{+}, \beta_s'^{-}, s_{2,s}, t_{2,s}^{\omega_s}, t_{2,s}^{q}, z_s, e_{1,s}, e_{2,s}\Big)_{s\in\{a,b\}}$, and the feasible sets for the first- and second-stage problems are given by

$$\mathcal{S}_1 = \Big\{\beta \in \widehat{\mathcal{B}} \mid ((11).a) - ((11).b)\Big\}, \text{ and}$$

$$\mathcal{S}_{t2}(\omega) = \Big\{\beta_s'^{+}, \beta_s'^{-}, s_{2,s}, z_s, t_{2,s}^{q}, t_{2,s}^{\omega_s}, e_{1,s}, e_{2,s} \;\Big|\; ((5).a) - ((5).m)\Big\}.$$

To solve the reformulated problem, we adopt a C&CG-based approach, following the Benders C&CG algorithm of Zeng and Wang [34]. In addition, we integrate the CCP into this framework. The master problem is addressed by leveraging the DC formulation of the fairness constraint and applying a cut-based CCP to locally approximate the original constraint, as discussed below.

At each iteration, the master problem solution β^k is used to evaluate the fairness constraint. If $|\Delta_{\tilde{F},g_r}(\beta^k)| \leq \epsilon$, the solution satisfies the constraint. Otherwise, we identify the concave component to linearize. For instance, if $\Delta_{\tilde{F},g_r}(\beta^k) > \epsilon$, we linearize $N(\beta)$ and impose the cut

$$P(\beta^k) - \Big(N(\beta^k) + \nabla N(\beta^k)^\top(\beta - \beta^k)\Big) \leq \epsilon.$$

Otherwise, we linearize $P(\beta)$ and impose the cut

$$N(\beta^k) - \Big(P(\beta^k) + \nabla P(\beta^k)^\top(\beta - \beta^k)\Big) \leq \epsilon.$$

The CCP-based enforcement mechanism complements the Benders C&CG procedure: while Benders C&CG generates cuts to account for worst-case realizations of the decision-dependent uncertainty set, CCP generates cuts to enforce fairness in the first-stage problem. As a result, the master problem is refined simultaneously along both robustness and fairness dimensions. Finally, based on Corollary 4, to guarantee an ϵ-fairness gap, we select at each iteration an appropriate

ramp parameter τ to control the discrepancy between the surrogate and indicator functions, subject to a prescribed distortion level $\bar{\alpha}$. CCP is then applied to enforce the resulting tightened fairness constraint with $\hat{\epsilon} = \epsilon - \alpha(\tau, \beta^k)$.

When the manipulation cost is uncertain, similar to the evolving second stage setting, one could adapt τ at each first-stage candidate β^k and realized uncertainty set $\Omega(\beta^k)$ by solving

$$\tau^* \in \arg\max_{\tau} \quad \text{s.t.} \sup_{\mathbf{\Sigma}(\omega^k),\, \beta',\, \omega^k \in \Omega(\beta^k)} \alpha(\tau, \beta') \leq \bar{\alpha}.$$

However, this approach remains under investigation. In particular, it requires incorporating optimality and feasibility cuts for the second-stage problem under varying choices of τ. In Section 5, we conduct numerical experiments with τ fixed in the second-stage ramp surrogate fairness constraint. We acknowledge that this constitutes a limitation and leave a more comprehensive treatment for future work, as fixing τ at a prescribed value does not necessarily guarantee satisfaction of the true ϵ–fairness gap.

Appendix F. Proofs Section 4

F.1. Proof Lemma 1

Proof According to Rosenfeld and Rosenfeld [29], the maximum score attainable under the movement budget $c(x_i, \hat{x}_i(\beta_s; x_i)) \leq u$ is $\beta_s^\top x_i + u_* \|\beta_s\|_{*,\Sigma_s}$. Hence, $u_* \|\beta_s\|_{*,\Sigma_s}$ is the maximal possible increase in the agent's score due to strategic behavior under (β_s, Σ_s).

We show that

$$\mathbb{1}\{\beta_s^\top \hat{x}_i(\beta_s; x_i) \geq 0\} = \mathbb{1}\{\beta_s^\top x_i + u_* \|\beta_s\|_{*,\Sigma_s} \geq 0\}$$

If $\beta_s^\top x_i \geq 0$, then $\hat{x}_i(\beta_s; x_i) = x_i$ is feasible. Therefore,

$$\mathbb{1}\{\beta_s^\top \hat{x}_i(\beta_s; x_i) \geq 0\} = 1.$$

Moreover,

$$\beta_s^\top x_i + u_* \|\beta_s\|_{*,\Sigma_s} \geq \beta_s^\top x_i \geq 0,$$

so $\mathbb{1}\{\beta_s^\top x_i + u_* \|\beta_s\|_{*,\Sigma_s} \geq 0\} = 1$. Thus, the equality holds in this case.

For $\beta_s^\top x_i < 0$, we distinguish two subcases. First, if the agent moves, this happens if and only if there exists a feasible $\hat{x}_i(\beta_s; x_i)$ such that

$$c(x_i, \hat{x}_i) < u \text{ and } \beta_s^\top \hat{x}_i(\beta_s; x_i) \geq 0.$$

In this case, $\mathbb{1}\{\beta_s^\top \hat{x}_i(\beta_s; x_i) \geq 0\} = 1$. Since the maximal feasible score change is $u_* \|\beta_s\|_{*,\Sigma_s}$, we have

$$u_* \|\beta_s\|_{*,\Sigma_s} \geq \beta_s^\top (\hat{x}_i(\beta_s; x_i) - x_i) \to \beta_s^\top x_i + u_* \|\beta_s\|_{*,\Sigma_s} \geq \beta_s^\top \hat{x}_i(\beta_s; x_i) \geq 0.$$

Therefore, $\mathbb{1}\{\beta_s^\top x_i + u_* \|\beta_s\|_{*,\Sigma_s} \geq 0\} = 1$.

Second, if the agent does not move, then there is no feasible $\hat{x}_i(\beta_s; x_i)$ that satisfies $c(x_i, \hat{x}_i) < u$ and $\beta_s^\top \hat{x}_i(\beta_s; x_i) \geq 0$. Hence, $\hat{x}_i(\beta_s; x_i) = x_i$ and

$$\mathbb{1}\{\beta_s^\top \hat{x}_i(\beta_s; x_i) \geq 0\} = 0.$$

We now argue by contradiction that

$$\beta_s^\top x_i + u_* \|\beta_s\|_{*,\Sigma_s} < 0.$$

Suppose instead that $\beta_s^\top x_i + u_* \|\beta_s\|_{*,\Sigma_s} \geq 0$. By the definition of the maximal attainable score under budget u, there would then exist a feasible $\hat{x}_i(\beta_s; x_i)$ achieving a score of at least zero. This contradicts the assumption that there were no feasible movements that yield a nonnegative score. Therefore,

$$\beta_s^\top x_i + u_* \|\beta_s\|_{*,\Sigma_s} < 0,$$

and $\mathbb{1}\{\beta_s^\top x_i + u_* \|\beta_s\|_{*,\Sigma_s} \geq 0\} = 0$. Combining all cases establishes the desired equality for every $x_i \in \mathbb{R}^d$.

■

F.2. Proof Lemma 3

Proof For the ramp hinge function $g_r(\widetilde{SC}(\beta_s, x)) = \min\{1, \max\{0, \frac{\widetilde{SC}(\beta_s,x)+\tau}{2\tau}\}\}$, for scores $\widetilde{SC}(\beta_s x) \geq \tau$ and $\widetilde{SC}(\beta_s x) \leq -\tau$, the

$$|\mathbb{1}(\widetilde{SC}(\beta_s x) \geq 0) - g_r(\widetilde{SC}(\beta_s x))| = 0.$$

However, for $\widetilde{SC}(\beta_s x) \in (-\tau, \tau)$ the difference is at most $\frac{1}{2}$ when $\widetilde{SC}(\beta_s x) = 0$, and $g_r(\widetilde{SC}(\beta_s x)) = \frac{1}{2}$. Hence, we can write

$$\left|\mathbb{1}(\widetilde{SC}(\beta_s x) \geq 0) - g_r(\widetilde{SC}(\beta_s x))\right| \leq \frac{1}{2}\mathbb{1}(|\widetilde{SC}(\beta_s x))| \leq \tau).$$

Therefore,

$$\begin{aligned}\mathbb{E}\Big[\Big|\mathbb{1}\big(\widetilde{SC}(\beta_s, x) \geq 0\big) - g_r\big(\widetilde{SC}(\beta_s, x)\big)\Big| \mid S = s\Big] &\leq \mathbb{E}\Big[\tfrac{1}{2}\,\mathbb{1}\big(\big|\widetilde{SC}(\beta_s, x)\big| \leq \tau\big) \mid S = s\Big] \\ &= \tfrac{1}{2}\,\mathbb{P}\big(\big|\widetilde{SC}(\beta_s, x)\big| \leq \tau \mid S = s\big). \qquad (14)\end{aligned}$$

Therefore,

$$\begin{aligned}\Delta_{\widetilde{F}}(\beta) &= \mathbb{P}(\widetilde{SC}(\beta_a, x) \geq 0 \mid S = a) - \mathbb{P}(\widetilde{SC}(\beta_b, x) \geq 0 \mid S = b). \\ &= \mathbb{P}(\widetilde{SC}(\beta_a, x) \geq 0 \mid S = a) - \mathbb{E}\left[g_r(\widetilde{SC}(\beta_a, x)) \mid S = a\right] \\ &\qquad + \mathbb{E}\left[g_r(\widetilde{SC}(\beta_b, x)) \mid S = b\right] - \mathbb{P}(\widetilde{SC}(\beta_b, x) \geq 0 \mid S = b) \\ &\qquad + \mathbb{E}\left[g_r(\widetilde{SC}(\beta_a, x)) \mid S = a\right] - \mathbb{E}\left[g_r(\widetilde{SC}(\beta_b, x)) \mid S = b\right]. \\ &\leq |\Delta_{\widetilde{F}, g_r}(\beta)| + \left|\mathbb{P}(\widetilde{SC}(\beta_a, x) \geq 0 \mid S = a) - \mathbb{E}\left[g_r(\widetilde{SC}(\beta_a, x)) \mid S = a\right]\right| \\ &\qquad + \left|\mathbb{P}(\widetilde{SC}(\beta_b, x) \geq 0 \mid S = b) - \mathbb{E}\left[g_r(\widetilde{SC}(\beta_b, x)) \mid S = b\right]\right|. \\ &\leq \Delta_{\widetilde{F}, g_r}(\beta) + \frac{1}{2}\sum_{S=\{a,b\}} \mathbb{P}(|\widetilde{SC}(\beta_s, x))| \leq \tau \mid S = s).\end{aligned}$$

Similarly,

$$\Delta_{\widetilde{F}}(\beta) \geq \Delta_{\widetilde{F},g_r}(\beta) - \frac{1}{2} \sum_{S=\{a,b\}} \mathbb{P}(|\widetilde{SC}(\beta_s, x))| \leq \tau \mid S = s).$$

Combining the above bounds, we obtain

$$\left|\Delta_{\widetilde{F}}(\beta) - \Delta_{\widetilde{F},g_r}(\beta)\right| \leq \frac{1}{2} \sum_{S=\{a,b\}} \mathbb{P}(|\widetilde{SC}(\beta_s, x))| \leq \tau \mid S = s).$$

■

F.3. Proof Corollary 4

Proof Fix any $\beta \in \mathbb{R}^{2d}$. By the ramp-indicator DP fairness distortion bound in Lemma 3, for $\alpha(\tau, \beta) = \frac{1}{2} \sum_{s \in \{a,b\}} \alpha_s(\tau, \beta_s)$, and the triangle inequality,

$$|\Delta_{\widetilde{F}}(\beta)| \leq |\Delta_{\widetilde{F},g_r}(\beta)| + |\Delta_{\widetilde{F}}(\beta) - \Delta_{\widetilde{F},g_r}(\beta)| \leq |\Delta_{\widetilde{F},g_r}(\beta)| + \alpha(\tau, \beta).$$

Therefore, if $|\Delta_{\widetilde{F},g_r}(\beta)| \leq \epsilon$, then $|\Delta_{\widetilde{F}}(\beta)| \leq \epsilon + \alpha(\tau, \beta)$. ■

F.4. Proof Lemma 5

Proof Fix a group s and consider the class $\mathcal{F}_s := \{x \mapsto g_r(\widetilde{SC}(\beta_s^\top x)) : \|\beta_s\|_{*,\Sigma} \leq B\}$, where $\widetilde{SC}(\beta_s^\top x) = \beta_s^\top x + u_* \|\beta_s\|_{*,\Sigma_s}$. Since g_r maps to $[0, 1]$, by the standard Rademacher generalization bound, with probability at least $1 - \delta$,

$$\sup_{f \in \mathcal{F}_s} \left|\mathbb{E}[f] - \widehat{\mathbb{E}}[f]\right| \leq 2\mathfrak{R}_{n_s}(\mathcal{F}_s) + \sqrt{\frac{\ln(2/\delta)}{2n_s}}.$$

Since the ramp-hinge surrogate g_r is L-Lipschitz with $L = 1/(2\tau)$, Talagrand's contraction lemma implies

$$\mathfrak{R}_{n_s}(\mathcal{F}_s) \leq L\, \mathfrak{R}_{n_s}(\{x \mapsto \beta_s^\top x + u_* \|\beta_s\|_{*,\Sigma_s} : \|\beta_s\|_{*,\Sigma_s} \leq B\}) \leq L \frac{B(R + u_*)}{\sqrt{n_s}},$$

where the last inequality is the standard Rademacher bound for linear classes under $\|x\| \leq R$. Combining the inequalities yields the first display. Summing over $s \in \{a, b\}$ and applying a union bound (e.g., using $\delta/2$ for each group) yields the corresponding second. ■

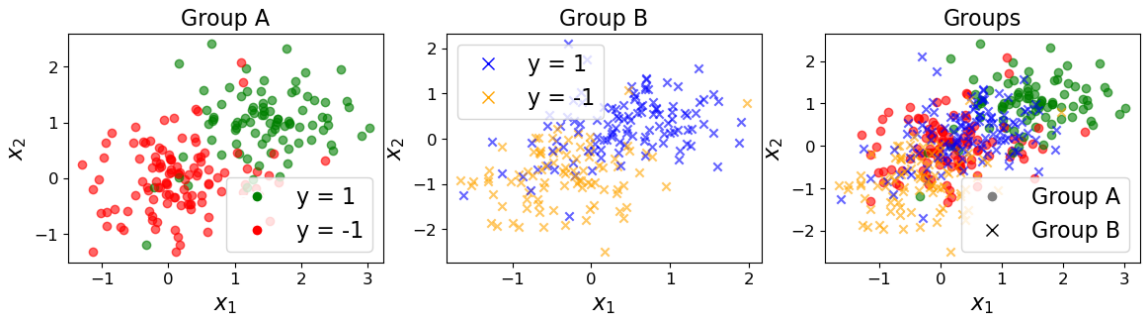


Figure 4: Two-dimensional synthetic dataset showing two demographic groups (a and b) with equal class priors. Scatter plot illustrates within-group separability and cross-group overlap, where Group a's negative class overlaps with Group b's positive class, increasing misclassification risk for Group b.

F.5. Proof Corollary 6

Proof The stated bound on $\sup_{\beta\in\mathcal{B}} |\Delta_{\widetilde{F},g_r}(\beta) - \widehat{\Delta}_{\widetilde{F},g_r}(\beta)|$ is precisely Lemma 6.4, yielding with probability at least $1-\delta_2$:

$$|\Delta_{\widetilde{F},g_r}(\beta)| \leq |\widehat{\Delta}_{\widetilde{F},g_r}(\beta)| + \eta \quad \text{for all } \beta \in \mathcal{B}.$$

Now fix any $\beta \in \widehat{\mathcal{B}}_{\widetilde{F},g_r}(\epsilon)$, so that $|\widehat{\Delta}_{\widetilde{F},g_r}(\beta)| \leq \epsilon$. Then the previous inequality implies $|\Delta_{\widetilde{F},g_r}(\beta)| \leq \epsilon + \eta$, hence $\beta \in \mathcal{B}_{\widetilde{F},g_r}(\epsilon+\eta)$. Therefore $\widehat{\mathcal{B}}_{\widetilde{F},g_r}(\epsilon) \subseteq \mathcal{B}_{\widetilde{F},g_r}(\epsilon+\eta)$. ■

Appendix G. Implementation Details, Data Generation, Experimental Details, and Additional Numerical Experiments

G.1. Synthetic Data Generation

We conduct experiments on a synthetic dataset with two demographic groups $S = \{a, b\}$, each containing $10,000$ samples in a two-dimensional feature space $x \in \mathbb{R}^2$. For each group, labels are drawn with group-specific base rates. We consider same qualification rate among groups, specifically, $\mathbb{P}(Y = 1 \,|\, S = a) = 0.5$, and $\mathbb{P}(Y = 1 \,|\, S = b) = 0.5$. Conditional on the label, features are sampled from a Gaussian distribution with shared covariance (standard deviation of 0.55, and correlation of 0.1), i.e., $x \mid (y = 1) \sim \mathcal{N}(\mu_{\text{pos}}, Cov), \quad x \mid (y = -1) \sim \mathcal{N}(\mu_{\text{neg}}, Cov)$. The class means are chosen to ensure within-group linearly separable structure while creating a targeted cross-group overlap: For Group a, the negative and positive class means are $\mu^a_{\text{neg}} = [0, 0], \quad \mu^a_{\text{pos}} = [1.5, 1.1]$. For Group b, the corresponding means are $\mu^b_{\text{pos}} = [0.55, 0.45], \quad \mu^b_{\text{neg}} = [-0.5, -0.9]$. This construction causes the negative class of Group a to overlap substantially with the positive class of Group b in feature space. Consequently, feature vectors with similar values may be associated with different labels depending on group membership. Under these settings, group b can be regarded as the disadvantaged group, as its positive instances are more likely to be confused with negative instances from Group a. See Figure 4.

G.2. Experimental setup

Individuals receiving negative outcomes (in either stage and from either group) may strategically manipulate their features to obtain a positive outcome, and receive utility $u = 1$. The manipulation cost is given by the ℓ_2-norm, with group-dependent cost function $c_s(x, \hat{x}) = \phi\big(\|\Sigma_s^{1/2}(\hat{x} - x)\|_2\big) = 0.5\|\Sigma_s^{1/2}(\hat{x} - x)\|_2$. We focus on a diagonal cost matrix and set the first-stage shared among group cost matrix to $\Sigma_0 = \mathrm{diag}(3, 6)$, so that manipulating x_2 is twice as costly as manipulating x_1. The second-stage cost matrix is given by $\Sigma(\omega_s) = \mathrm{diag}(g(\omega_{1,s}), g(\omega_{2,s})) \cdot \Sigma_0$, with $g(\omega_s) = \frac{1}{\omega_s^2}$. We assume a common decision-dependent uncertainty set across demographic groups. Specifically, for each group s, the uncertain cost-driving vector ω_s lies in $\Omega(\beta_s) = \{\omega_s \in \mathbb{R}_+^2 : \mathbf{F}\,\omega_s \leq \mathbf{h} + \mathbf{G}\beta_s\}$, with

$$\mathbf{F} = \begin{bmatrix} 0.9 & 0.3 \\ 0.3 & 0.6 \end{bmatrix}, \ \mathbf{h} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \ \mathbf{G} = \begin{bmatrix} 1.3 & 0.9 \\ 0.6 & 1.4 \end{bmatrix}.$$

Consequently, differences in uncertainty sets, and thus in second-stage costs, arise primarily from the group-dependent first-stage decision β_s.

As mentioned in the main paper, we set $\epsilon = 0.09$ and train the classifier under stricter fairness constraints to ensure that this target level is satisfied. This adjustment is tied to the maximum allowable distortion introduced by replacing the indicator acceptance function with the ramp-hinge approximation, denoted by $\bar{\alpha}$. We set $\bar{\alpha} = 0.05$. This distortion is controlled by the choice of the ramp parameter τ. To enforce a stricter fairness constraint we use $\hat{\epsilon} = \epsilon - \alpha(\tau^*, \beta)$, where $\tau^* = \arg\max_\tau \text{ s.t } \alpha(\tau, \beta^k) \leq \bar{\alpha}$. At each iteration of the master problem, given a candidate classifier β^k, we select the corresponding τ^*, while setting $\tau = 0.05$ for the second-stage problem.

For the second-stage problem, we also set $\bar{\alpha} = 0.05$, but fixing the ramp parameter $\tau = 0.05$ instead of optimizing over it. We select $\tau = 0.05$ as it is small enough to ensure that the distortion introduced by the ramp-hinge approximation remains limited, while not being so small as to make the approximation overly restrictive or numerically unstable. It remains an open problem and will be investigated.

For the (fair) baseline in the first stage, and for both models in the second stage, we optimize the strategic hinge loss in (2) with respect to $\mathbf{\Sigma} := (\Sigma_0)_{s \in S}$ and realized cost matrices profile $\mathbf{\Sigma}(\omega^k) := (\Sigma(\omega_s^k))_{s \in S}$, where $\omega_s^k \in \Omega(\beta_s)$. Performance is measured over 25 independent instances, each with 100 new test points from each group sampled from the 10,000-point dataset. Second-stage results are further averaged over 10 realizations of ω_s for both groups.

G.3. Algorithmic implementation details

For completeness, we summarize the implementation details of the proposed CCP–Benders C&CG algorithm.

Stopping criteria. The outer Benders C&CG procedure terminates when the optimality gap between the incumbent upper bound and master lower bound satisfies

$$UB - LB \leq \text{tol},$$

or when the maximum number of C&CG iterations is reached.

Within each master problem, the CCP procedure terminates when the surrogate fairness violation satisfies

$$|\Delta_{\widetilde{F}}(\beta)| \leq \hat{\epsilon} + \text{tol},$$

where $\hat{\epsilon} = \epsilon - \bar{\alpha}$, or when the maximum number of CCP iterations is reached.

Iteration counts. For the reported numerical experiments, the outer Benders C&CG algorithm converged in 12 iterations, corresponding to 12 Benders cuts. The number of CCP iterations required to solve the master problem varied across C&CG iterations, ranging from 0 to 101 iterations depending on the current master solution and fairness violation.

Runtime. On the reported instance, the total runtime of the proposed algorithm was 682.86 seconds.

G.4. Fair and Unfair Dependency Aware Classifier Numerical Comparision Experiments

Here, we compare the performance of fair and unfair dependency-aware classifiers. Figure 5 shows that the F-DD classifier incurs higher second-stage loss mostly from group A (middle panel) and greater total manipulation, leading to higher overall loss and manipulation (right panel). In contrast, the unfair dependency-aware (UF-DD) classifier incurs higher loss and manipulation in the first stage, reflecting a stronger emphasis on robustness to evolving manipulation costs, as illustrated in the middle panel, of much lower manipulation in the second stage.

This highlights a key trade-off: for the F-DD classifier, robustness is balanced against fairness, resulting in losses that reflect both objectives across stages. By comparison, the UF-DD classifier allocates its loss entirely toward robustness to uncertainty and evolving costs, without accounting for fairness.

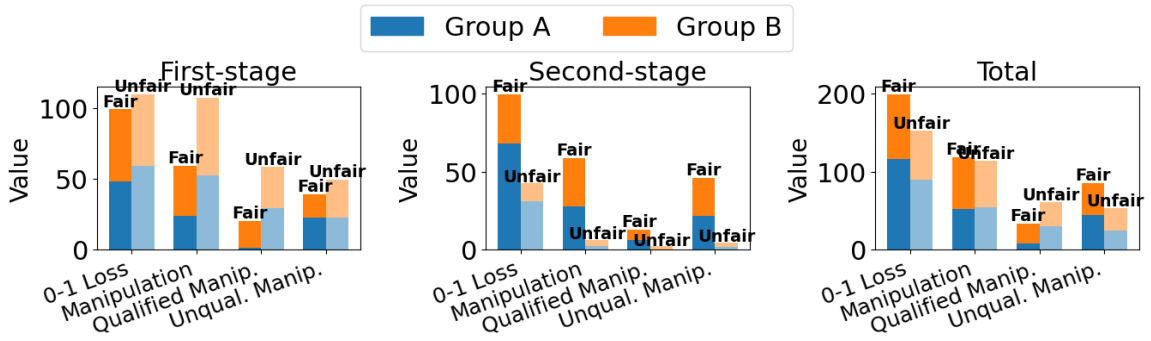


Figure 5: Performance comparison: fair vs. unfair DD-classifier.

G.5. Unfair Dependency-Aware and Dependency-Unaware Additional Experiments

For comparison, for an unfair classifier who account for the decision-dependent cost of manipulation and one that doesn't consider it. Consistent with the findings in [4], Figure 7 and Table 1 show that the DD classifier yields much lower second-stage loss (42.89 vs. 71.56) and manipulations (113.84 vs. 128.91) than the DI baseline. This advantage comes from anticipating how decisions affect manipulation costs, making manipulation harder ($\omega_{i,s} < 1, \forall i$) in both groups. Figure 6 visualizes this: our DD classifier's uncertainty sets $\Omega(\beta^{\text{DD}})$ is strictly smaller than those of the baseline DI classifier $\Omega(\beta^{\text{DI}})$ sets, and induce higher manipulation costs in the second stage, especially for group b.

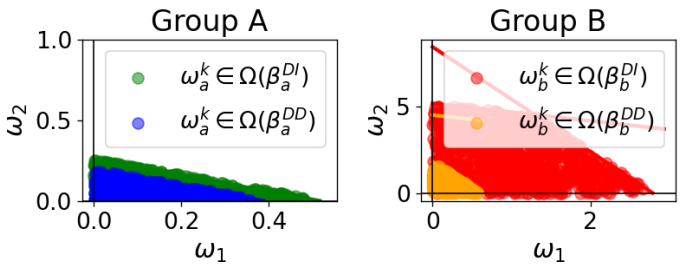


Figure 6: Uncertainty sets: decision-dependent-aware (DD) vs. -unaware (DI) without fairness considerations.

In the first stage, DD performs slightly worse (loss 109.52 vs. 84.48), but overall the dependency-aware model yields clear gains: total loss drops from 156.04 to 152.41, and manipulations fall from 128.91 to 113.84. These results confirm that modeling decision-dependent costs reduces both firm error and agents' induced strategic manipulations.

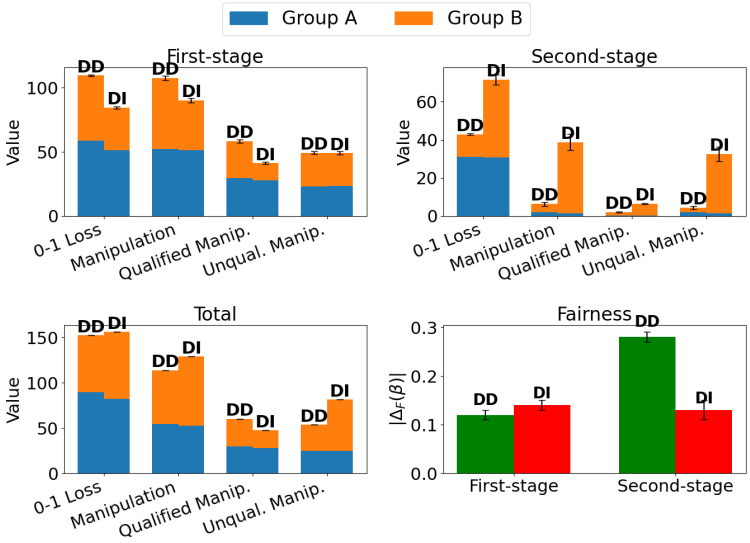


Figure 7: Performance comparison of unfair classifiers (DD vs. DI)

Appendix H. Numerical validation for Fair Problem

In this section, we consider a synthetic dataset where groups differ in their feature distribution parameters and qualification rates. We generate two groups $S \in \{a, b\}$ using Gaussian mixture models with group-specific class priors and shared covariance structure (See Appendix G.1). We provide numerical evidence supporting the theoretical results for the surrogate fairness-constrained model.

Metric	G	First-stage DD	First-stage DI	Second-stage DD	Second-stage DI	Total DD	Total DI
0-1 Loss	A	58.88 ± 1.21	51.36 ± 1.21	31.04 ± 0.28	30.80 ± 0.28	89.92	82.16
	B	50.64 ± 1.17	33.12 ± 0.77	11.84 ± 0.24	40.76 ± 2.60	62.48	73.88
	T	109.52 ± 0.61	84.48 ± 0.96	42.89 ± 0.47	71.56 ± 2.69	**152.41**	156.04
Manip.	A	52.20 ± 1.24	51.16 ± 1.10	2.02 ± 0.39	1.49 ± 0.35	54.22	52.65
	B	55.32 ± 0.92	38.96 ± 0.92	4.31 ± 0.71	37.30 ± 4.11	59.63	76.26
	T	107.52 ± 1.71	90.12 ± 1.63	6.32 ± 1.03	38.79 ± 4.18	**113.84**	128.91
Qualified Manip.	A	29.40 ± 1.25	27.68 ± 1.03	0.09 ± 0.03	0.10 ± 0.04	29.49	27.78
	B	28.88 ± 0.77	13.48 ± 0.66	1.92 ± 0.29	6.22 ± 0.35	30.80	19.70
	T	58.28 ± 1.17	41.16 ± 0.96	2.01 ± 0.30	6.32 ± 0.38	**60.29**	47.48
Unqual. Manip.	A	22.80 ± 1.06	23.48 ± 0.89	1.92 ± 0.37	1.39 ± 0.32	24.72	24.87
	B	26.44 ± 0.83	25.48 ± 0.63	2.39 ± 0.44	31.08 ± 3.82	28.83	56.56
	T	49.24 ± 1.12	48.96 ± 1.23	4.31 ± 0.77	32.46 ± 3.88	**53.55**	81.42
$\lvert\Delta_F(\beta)\rvert$	–	0.12 ± 0.01	0.14 ± 0.01	0.28 ± 0.01	0.13 ± 0.02	0.40	0.27

Table 1: Comparison of unfair classifiers (DD vs. DI): average $\pm$ standard error across stages and overall, reported separately for groups A and B.

H.1. Surrogate Score: Identical Acceptance Decision

We numerically verify the theoretical equivalence between true strategic acceptance and its surrogate characterization in Lemma 1 ($\mathbb{1}\{\beta^\top \hat{x}_i(\beta) \geq 0\} = \mathbb{1}\{\beta^\top x_i + u_* \|\beta\|_{*,\Sigma} \geq 0\}$,), as well as, the zero fairness distortion in Corollary 2 ($|\Delta_{\widetilde{F}}(\beta)| - |\Delta_F(\beta)| = 0$.).

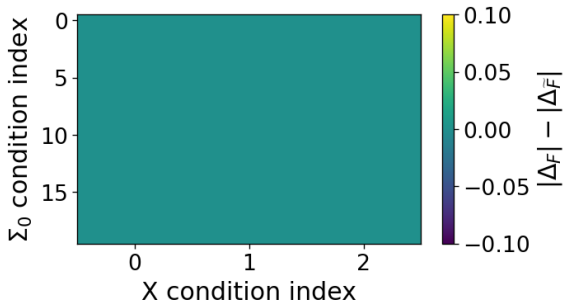


Figure 8: Heatmap of maximum demographic parity gap distortion across datasets and cost matrices Σ_0. Values are zero up to numerical precision, confirming Corollary 2.

To validate this identity, we generate synthetic datasets under multiple feature distributions and vary the baseline cost matrix Σ_0. For each experimental setting, we sweep the classifier direction β and compute both the true strategic acceptance indicator and its surrogate counterpart. We then measure the distortion, defined as the fraction of individuals for whom the two indicators disagree.

As shown in Figures 8, the distortion is identically zero up to numerical precision across all classifiers, feature distributions, and cost configurations. Figure 9 shows detailed results on the point-wise acceptance distortion from using the surrogate scores, which is zero for all different setups.

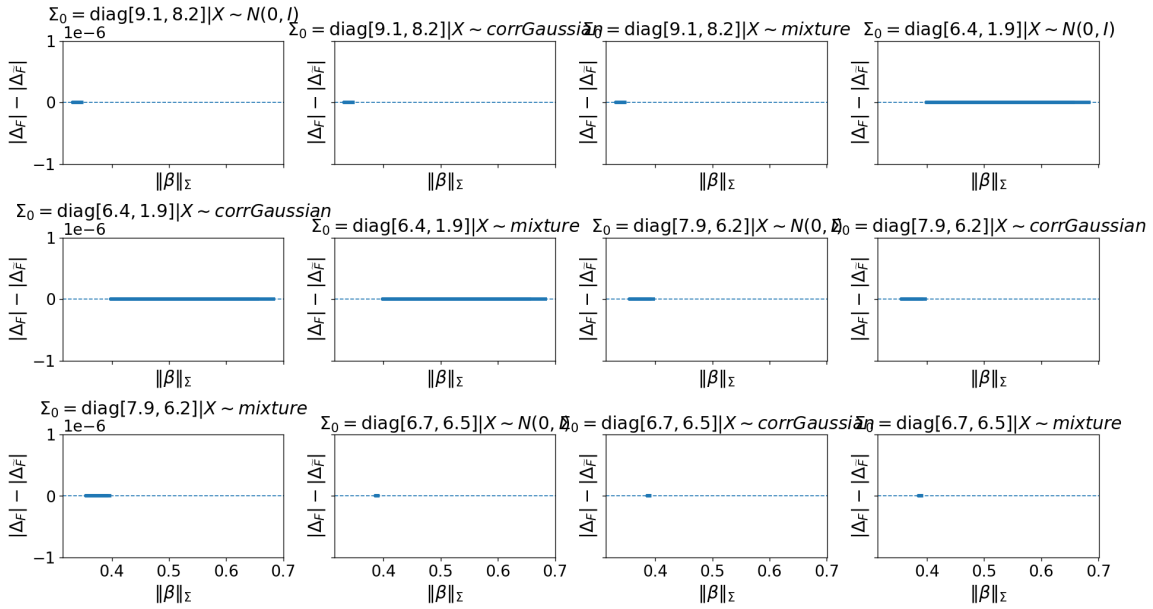


Figure 9: Demographic parity (DP) gap distortion between true strategic scores and surrogate scores. Distortion is zero across all datasets and cost matrices, confirming Corollary 2.

These results provide strong empirical confirmation that the surrogate rule exactly characterizes strategic acceptance outcomes.

H.2. Surrogate Ramp Hinge: Fairness distortion

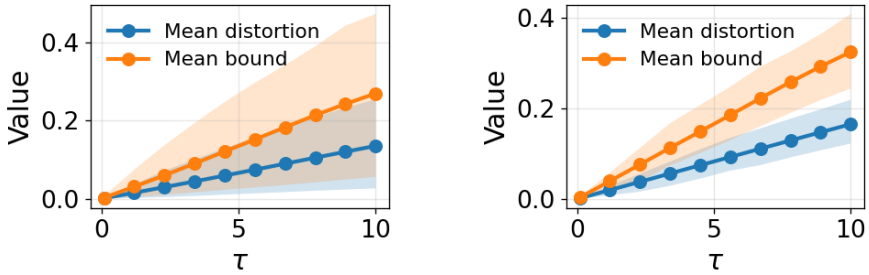
We first validate the point-wise-acceptance distortion bounds. We then study the fairness distortion indicated by the surrogate ramp-hinge function, which depends on the choice of τ, and the probability mass near the decision boundary (Lemma 3). Finally, we discuss the selection of τ and validation for Corollary 4.

Point-wise acceptance distortion We first run stress-test verification of $|\mathbb{1}(\widetilde{SC}(\beta_s, x) \geq 0) - g_r(\widetilde{SC}(\beta_s, x))| \leq \frac{1}{2}\mathbb{1}\{|\widetilde{SC}(\beta_s, x)| \leq \tau\}$ with $\widetilde{SC}(\beta_s, x) = \beta_s^\top x + \|\beta_s\|_{*,\Sigma_0}$. For each τ, 25 random $\beta_s \sim \mathrm{Unif}([-10, 10]^2)$ are sampled; we report averages across the 25 runs. We define the violation as

$$\mathrm{viol} = \Big(|\mathbb{1}(\widetilde{SC}(\beta_s, x) \geq 0) - g_r(\widetilde{SC}(\beta_s, x))|\Big) - \Big(\tfrac{1}{2}\mathbb{1}\{|\widetilde{SC}(\beta_s, x)| \leq \tau\}\Big).$$

Table 2 shows that the average point-wise distortion of acceptance is upper bounded by $\frac{1}{2}\mathbb{1}\{|\widetilde{SC}(\beta_s, x)| \leq \tau\}$ (In the table, Avg. #viol = 0 for all τ), which holds for both groups.

Fairness distortion induced by the surrogate ramp-hinge function Let $\alpha_s(\tau, \beta_s) = \widehat{\mathbb{P}}(|\widetilde{SC}(\beta_s, x)| \leq \tau \mid S = s)$, for a given $\Sigma_0 = \text{diag}[3, 6]$. Figure 10 shows the empirical average expected distortion $\mathbb{E}\Big[\Big|\mathbb{1}\big(\widetilde{SC}(\beta_s, x) \geq 0\big) - g_r\big(\widetilde{SC}(\beta_s, x)\big)\Big| \mid S = s\Big]$ (over 25 sample of (β_a, β_b)) is always upper bounded by the $\frac{1}{2}\alpha_s(\tau)$ at each τ (Lemma 3 holds). Moreover, as τ increases, the ramp-hinge deviates from the indicator on a larger subset of samples $(|\widetilde{SC}(\beta_s, x)| \leq \tau)$ in each group s, leading to higher empirical distortion; therefore, τ should be selected small enough to keep $\widehat{\mathbb{P}}(|\widetilde{SC}(\beta_s, x)| \leq \tau)$ for each group s controlled, while remaining large enough to provide numerical smoothness.



(a) Group A (b) Group B

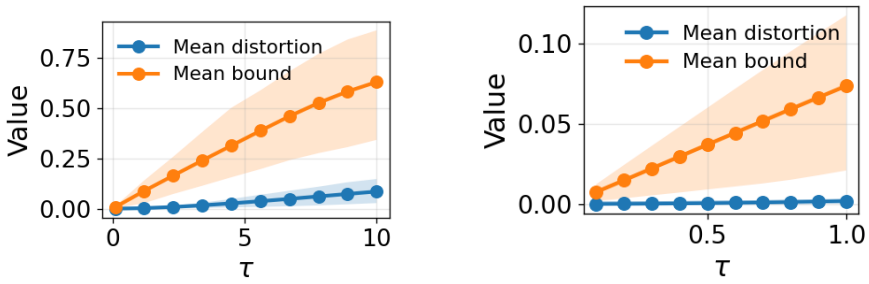
Figure 10: Mean empirical distortion vs. mean bound across β for different values of $\tau \in [0.1, 10]$. Panels (A) and (B) correspond to the advantaged and disadvantaged groups, respectively. Results confirm the second part of Lemma 3.

Let $\alpha(\tau, \beta) := \frac{1}{2}\sum_{s\in\{a,b\}} \alpha_s(\tau, \beta_s)$, Figure 11 shows the empirical average expected distortion $|\widehat{\Delta}_{\widetilde{F}}(\beta) - \widehat{\Delta}_{\widetilde{F},g_r}(\beta)|$ (over 25 sample of (β_a, β_b)) is always upper bounded by the $\alpha(\tau, \beta)$ at each τ. Moreover, as discussed above, as τ increases, the ramp-hinge deviates from the indicator on a larger subset of samples $(|\widetilde{SC}(\beta_s, x)| \leq \tau)$ in each group s, leading to higher empirical distortion; there-

τ	Avg. #viol	% any viol
0.1	0	0
1.2	0	0
2.3	0	0
3.4	0	0
4.5	0	0
5.6	0	0
6.7	0	0
7.8	0	0
8.9	0	0
10	0	0
Overall Avg	0	0

Table 2: Pointwise acceptance violations between ramp-hinge and indicator functions across values of the ramp parameter τ. No violations are observed for any τ.

fore, τ should be selected small enough to keep $\widehat{\mathbb{P}}(|\widetilde{SC}(\beta_s, x)| \leq \tau)$ for each group s controlled, while remaining large enough to provide numerical smoothness.



(a) Across $\tau \in [0.1, 10]$. (b) Across $\tau \in [0.01, 0.1]$.

Figure 11: Mean empirical distortion vs. mean bound across β for different values of τ. Panels (a) and (b) correspond to $\tau \in [0.1, 10]$ and $\tau \in [0.01, 0.1]$, respectively, confirming the second part of Lemma 3.