Robust Strategic Classification under Decision-Dependent Cost Uncertainty

Sura Alhanouti
ISE Department
The Ohio State University and
Jordan U. of Science & Technology
alhanouti.1@osu.edu

Güzin Bayraksan
ISE Department
The Ohio State University
bayraksan.1@osu.edu

Parinaz Naghizadeh
ECE Deparment
UC San Diego
parinaz@ucsd.edu

Abstract

Humans have been found to manipulate their inputs to algorithmic decision systems to receive favorable outcomes. This has motivated a line of work on "strategic classification," wherein algorithmic decision rules are selected to prevent undesirable strategic responses. Prior works typically assume that the cost of such strategic behavior is fixed and independent of the classifier's decision. In practice, however, manipulation costs depend on past decisions: today's algorithmic decisions influence tomorrow's costs of strategic response. To capture this dependency, we propose to formulate the problem of strategic classification as a two-stage robust optimization problem with a decision-dependent uncertainty set. We formalize this problem, develop approximations and reformulations to solve it, and numerically illustrate our algorithm's ability to mitigate gaming of the algorithmic system.

1 Introduction

Machine learning algorithms are increasingly deployed in decision making systems, including for making lending, hiring, recidivism, and school admission decisions. A key issue arising in these systems is *strategic behavior* by humans, who can modify their inputs to the algorithm (whether through genuine effort or by misrepresenting their features) to gain favorable outcomes [14, 16, 39, 46, 51]. This has led to a literature on strategic machine learning [21, 28, 31, 37, 38], which studies the design of algorithms that can prevent "gaming" of the algorithmic system.

Much of the existing works in this area have assumed that both the decision maker and the agents have full information about the system and decision boundaries [27, 34, 48, 56]. Recent work has started deviating from these assumptions, by considering agents with incomplete or biased perceptions of the classifier [10, 12, 22] or classifiers uncertain about agent responses [1, 41, 44]. (We review additional related work in Appendix B). A type of uncertainty that arises in practice, but remains unaddressed by this existing literature, is that classifier decisions can *endogenously* shape the environment. Specifically, the choices made by a classifier influence the future costs that the agents face when attempting to strategically respond to the algorithmic system in the future. For instance, test-optional policies adopted by colleges during the COVID-19 pandemic reduced demand and prices for test preparation in the following years, while simultaneously shifting spending toward alternative costly signals such as essay coaching, and extracurricular activities [8, 30, 32, 54]. Thus, a school's decision today shape next year's applicants' response costs. Yet, the exact impact that this year's decisions will have on subsequent years' costs is not fully known at the time of decision-making.

To capture such situations, we propose to formulate the classifier choice as a two-stage robust optimization problem (TSRO) with *decision-dependent and uncertain costs*. In the first stage, the classifier anticipates that its choices will influence future response costs by the agents; we model this through a decision-dependent uncertainty set for the second-stage problem in our proposed

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

formulation. We then develop approximations and reformulations to make the proposed problem amenable to existing algorithms for solving TSROs, specifically the "Benders and Constraint and Column Generation (C&CG)" algorithm of [55]. Through numerical experiments, we illustrate how our dependency-aware classifier strategically sacrifices some first-stage performance to gain second-stage robustness, reducing overall loss and limiting manipulations more effectively than a dependency-unaware baseline.

2 Problem Setting and Preliminaries

We study the problem of strategic binary classification, where a firm makes accept/reject decisions on agents with observable features $x \in \mathcal{X} \subseteq \mathbb{R}^d$ and hidden true labels $y \in \mathcal{Y} = \{\pm 1\}$. Let $(X, Y) \sim P_{XY}$ denote the joint distribution over the support $\mathcal{X} \times \mathcal{Y}$, with (x, y) a realization of (X, Y). The problem unfolds over two stages, where a long-lived firm interacts with short-lived agents (different populations in each stage). In the first stage, the firm selects a linear classifier $\text{sign}(\beta^T x)$, where $\beta \in \mathcal{B}$, and agents may strategically modify their features to gain acceptance without altering their true label. Formally, the applicant can strategically alter their initial feature x to

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

where $u \geq 0$ is the utility from a positive classification, and $c(x, \hat{x})$ is the cost of changing x to $\hat{x}$. An agent responds strategically only if the modification flips the outcome to positive and $c(x, \hat{x}) \leq u$.

We consider cost functions of the form $c(x, \hat{x}) = \phi(\|\hat{x} - x\|_\Sigma)$, where $\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a non-decreasing function and $\|\hat{x} - x\|_\Sigma := \|\Sigma^{\frac{1}{2}}(\hat{x} - x)\|$ where $\Sigma \in \mathbb{R}^{d \times d}$ is a positive definite (PD) *cost matrix* ($\Sigma \succ 0$) which uniquely parametrizes the cost function, and $\|\cdot\|$ is the standard p-norm. This cost function was initially proposed by [41], and includes the previously studied ℓ_p-norm costs [10, 13, 52], such as the widely used ℓ_2-norm [12, 20, 27], as a special case. In matrix Σ, each diagonal element σ_i sets the relative cost of changing feature x_i, while off-diagonal terms capture correlations in feature-change costs. In the first stage, we assume the firm knows the cost function $c(x, \hat{x})$ via Σ_0 and uses it to anticipate agents' responses when selecting β. This stage, in isolation, is a classic strategic classification setup [1, 10, 21] but with a more general cost model.

Our model differs from existing ones through its two-stage structure and their coupling: we assume the firm's first-stage decision influences the second-stage cost of strategic responses. Importantly, while the firm anticipates this dependence, the exact impact is unknown at the time of the first-stage decision. Formally, the second-stage cost of changing x to $\hat{x}$ is given by

$$c(x, \hat{x}) = \phi\left(\|\hat{x} - x\|_{\Sigma(\omega)}\right) . \tag{2}$$

That is, the second stage has a cost matrix $\Sigma(\omega) \in \mathbb{R}^{d \times d}$, defined as a function of a random vector $\omega \in \mathbb{R}^d_+$ whose components capture how the cost depends on the first-stage decision β.

Specifically, we introduce *a decision-dependent uncertainty set* $\Omega(\beta)$ from which the random vector ω is drawn. During the first stage, the firm's uncertainty about the second stage is parametrized by this set, containing all possible realizations of the cost-driving vector ω that could arise given the first-stage decision β. We formally detail the firm's model of this uncertainty set in Section 3. At the beginning of the second-stage, a realization ω is drawn from this set, which in turn realizes the cost matrix for the second-stage according to

$$\Sigma(\omega) := Q\big(g(\omega)\big) \cdot \Sigma_0, \tag{3}$$

where $\Sigma_0 \in \mathbb{R}^{d \times d}$ is the first-stage cost matrix, and $Q(g(\cdot)) \in \mathbb{R}^{d \times d}$ is a matrix-valued transformation that maps the element-wise scaling factors $g(\omega_j)$ into a full cost-scaling matrix. Each ω_i scales the cost of feature x_i and may also affect other features via the linear transformation $Q(g(\cdot))$. To keep the structure tractable, $Q(g(\omega))$ is restricted to be component-wise linear, with bounded and strictly positive scaling functions $g(\cdot)$ ensuring $\Sigma(\omega) \succ 0$, and $g(\omega)^{-1/2}$ is assumed to be linear in ω. For instance, in the special case when $Q(g(\omega))$ and Σ_0 are both diagonal, and $g(\omega) = \omega^2$ or $g(\omega) = \omega^{-2}$, the firm faces linearly scaled second-stage costs $\omega_i \sigma_i$ or $\frac{\sigma_i}{\omega_i}$, respectively.

In summary, the dependence of the second-stage cost on the first-stage decision β can be expressed as the following chain of mappings: $\beta \mapsto \Omega(\beta) \ni \omega \mapsto \Sigma(\omega) \mapsto \|\cdot\|_{\Sigma(\omega)} \mapsto c(x, \hat{x})$. Note that once the second stage begins, a linear classifier $\text{sign}(\beta'^T x)$ is chosen from some set $\mathcal{B}'$, *after* the realization of the decision-dependent costs (i.e., second-stage agents' costs becomes known to the firm at the second stage). It is only the choice of the first-stage classifier β that must both limit gaming in the present and anticipate its uncertain influence on future costs. We provide a motivating example for our proposed model in Appendix A.1.

3 The Firm's Optimization Problem

To model the two-stage decision-making problem outlined in Section 2, we formulate the firm's problem as a two-stage robust optimization (TSRO) with a decision-dependent uncertainty set.

The learning objective. Recall that agents may respond to classifier β by modifying features from x to $\hat{x}(\beta)$ (as shown in (1)). Define the change vector $\delta_\Sigma(x;\beta) := \hat{x}(\beta) - x$, where the cost function is parameterized by matrix Σ. If no action is taken, $\delta_\Sigma(x;\beta)$ is simply a zero vector.

The firm's objective is to select a β that minimizes the expected 0–1 loss under strategic responses. For each strategic agent (x, y) with cost matrix Σ, the incurred 0–1 loss is: $\ell_\Sigma(\beta^\top x, y) := \mathbb{1}\{\text{sign}(\beta^\top(x + \delta_\Sigma(x;\beta))) \neq y\}$. Although well-defined, optimizing this expected 0-1 loss faces two challenges: the non-convexity introduced by the sign function, and discontinuities in the agent movement $\delta_\Sigma(x;\beta)$. The first is standard and addressed via surrogate losses (e.g., hinge loss), while the second is more subtle and has been handled by prior work [34, 41] through a cost-aware strategic hinge loss:

$$\ell_{\Sigma,\text{s-hinge}}(\beta^\top x, y) := \max\{0, 1 - y(\beta^\top x + u_*\|\beta\|_{*,\Sigma}\} \,, \tag{4}$$

where u_* is the largest value satisfying $\phi(u_*) \leq u$, and $\|\beta\|_{*,\Sigma} := \sup_{\|v\|_\Sigma=1} \beta^\top v = \|\Sigma^{-\frac{1}{2}}\beta\|_*$ is the Σ-*transformed dual norm* of β, with $\|\cdot\|_*$ denoting the dual norm of the adopted p-norm. Rosenfeld and Rosenfeld [41] formally show that this cost-aware strategic hinge loss is a valid proxy for the strategic 0–1 loss: it always upper-bounds the true risk, and minimizing it yields uniform generalization guarantees with error $O(\frac{1}{\sqrt{n}})$, ensuring fidelity to the true loss. We likewise adopt the cost-aware strategic hinge loss as the firm's learning function in both stages, setting $\Sigma = \Sigma_0$ in the first stage and $\Sigma = \Sigma(\omega)$ in the second. Specifically, we will assume that the firm selects β to (in part) minimize the following empirical risk evaluated on the N training data points:

$$R_\Sigma(\beta) := \frac{1}{N}\sum_{i=1}^{N} \max\left\{0, 1 - y_i\left(\beta^\top x_i + u_*\|\beta\|_{*,\Sigma}\right)\right\}. \tag{5}$$

Modeling the uncertainty set. We propose to model a *decision-dependent uncertainty* set as

$$\Omega(\beta) = \{\omega \in \mathbb{R}^d_+ : \mathbf{F}(\beta)\omega \leq \mathbf{h} + \mathbf{G}\beta\}, \tag{6}$$

where ω represents the cost-driving vector (uncertain variable), and β denotes the classifier's first-stage decision vector. Here, $\mathbf{h}$ serves as a baseline constraint (e.g., capturing average budgets) while the matrices $\mathbf{F}(\beta)$ and $\mathbf{G}$ capture how the decision β influences the structure and bounds of the uncertainty set, respectively. Note that if $\mathbf{F}$ is a fixed (β-independent) matrix, the inequalities constrain each cost component independently. In contrast, allowing $\mathbf{F}$ to depend on β means that the classifier can alter how different cost components interact.

We provide a motivating example for this proposed model of decision-dependent uncertainty sets in Appendix A.2. From a technical viewpoint, these polyhedral sets also allow us to reformulate the two-stage robust optimization problem efficiently, using linear programs, as detailed later.

The two-stage robust optimization problem. We can now state the firm's two-stage robust optimization (TSRO) problem with decision-dependent uncertainty (DDU):

$$\min_{\beta\in\mathcal{B}} \left(R_{\Sigma_0}(\beta) + \max_{\omega\in\Omega(\beta)} \min_{\beta'\in\mathcal{B}'(\beta,\omega)} R_{\Sigma(\omega)}(\beta') \right) \,, \tag{7}$$

where $R_\Sigma(\beta)$ is the empirical risk from (5), and $\Omega(\beta)$ is the DDU set (6). In words, the firm first chooses β to minimize the cost-aware hinge loss under Σ_0, while accounting for the worst-case second-stage costs $\Sigma(\omega)$ and anticipating that the adjusted classifier β' will minimize loss once ω is realized. Note that classifier choices may also be constrained, *if desired*. Specifically, we let $\mathcal{B} = \{\beta \in \mathbb{R}^d : \mathbf{A}\beta \geq \mathbf{b}\}$ and $\mathcal{B}'(\beta,\omega) = \{\beta' \in \mathbb{R}^d : \mathbf{B}_2\beta' \geq \mathbf{d} - \mathbf{B}_1\beta - \mathbf{E}\omega\}$, which allow for feature-importance restrictions, including allowable adjustments once ω is realized.

4 Reformulating and Solving the Firm's Optimization Problem

We next propose an algorithm for solving the firm's TSRO with DDU in (7). Existing methods for solving TSRO problems [7, 40, 55, 58] cannot be applied directly due to structural differences in

our objective and uncertainty sets. We address this by approximating and reformulating the original tri-level nonlinear model into a linear reformulation, enabling the use of the *Benders and C&CG* algorithm from [55]. Specifically, the "Benders C&CG" algorithm [55] works by formulating a master problem that is iteratively refined with cuts from worst-case uncertainty realizations. This requires converting the tri-level optimization problem into a single-level problem, typically by dualizing the second-stage to yield a bilevel problem, and then replacing the lower-level problems with KKT conditions. In [55], this procedure works as the second stage problem has a linear and decision-independent objective function and decision-dependent constraints. However, our second-stage problem in (7) does not meet these requirements, particularly as the uncertainty ω enters directly into the objective via the dual-norm term $\|\beta'\|_{*,\Sigma(\omega)}$. As a result, the dual feasible region varies with each ω, so dual extreme points cannot be reused across iterations, which prevents systematic Benders cut generation in methods such as C&CG.

We begin by addressing this by first approaximating the Σ-induced dual norm, leveraging the norm equivalence theorem and ℓ_∞'s submultiplicativity to bound the dual norm, as follows.

Lemma 1. *There exists a constant $M > 0$ such that, $\|\beta\|_{*,\Sigma(\omega)} \leq M\|\Sigma(\omega)^{-\frac{1}{2}}\|_\infty\|\beta\|_\infty$.*

We next linearize the max operator in (5) using slack variables, and then apply the McCormick envelope method to handle the nonlinear product term arising from the application of Lemma 1, ultimately obtaining a linear tri-level reformulation amenable to "Benders C&CG". The detailed derivation can be found in Appendix C. With the reformulation in place, we can apply "Benders C&CG," which relies on the bilevel and single-level reformulations to construct the master problem, together with the subproblem formulations and algorithmic steps detailed in Appendices D–F.

5 Numerical Experiments

We evaluate our two-stage robust strategic classification framework with DDU costs on synthetic data. More details on the experiment setup are in Appendix G. In the first stage, our decision-dependent (DD) classifier is obtained as detailed in Section 4, whereas the decision-independent (DI) baseline ignores cost dependence and minimizes the loss in (4). In the second stage, both our DD classifier and the benchmark DI classifier aim to minimize the loss in (4) for the realized second-stage costs, but the second-stage costs are different due to the difference in the two methods' first-stage classifiers.

Table 1 summarizes the results. The DD classifier incurs slightly higher first-stage loss (25.36 vs. 22.80) but achieves far lower second-stage loss (5.22 vs. 43.92). Overall, the DD approach reduces both total loss (30.58 vs. 66.72) and manipulations (26.66 vs. 66.46). This reduction occurs because the DD classifier anticipates second-stage manipulation costs, making manipulation more difficult. As shown in Figure 1, the uncertainty set $\Omega(\beta^{\text{DI}})$ admits larger ω values than $\Omega(\beta^{\text{DD}})$, implying cheaper manipulation under the DI classifier. In contrast, $\Omega(\beta^{\text{DD}})$ restricts $\omega < 1$, so it both limits the reduction in second-stage costs and ensures manipulation remains consistently expensive. Moreover, the DD classifier accepts fewer unqualified manipulators, highlighting how anticipating DD costs improves robustness at (a limited) expense of first-stage accuracy. Additional details are provided in Appendix G.

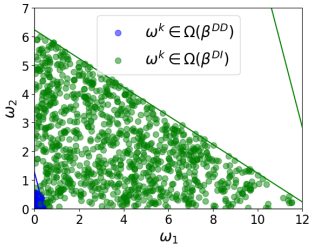

Figure 1: Second stage uncertainty sets $\Omega(\beta^{\text{DD}})$ vs $\Omega(\beta^{\text{DI}})$.

Metric	First-stage		Second-stage		Total	
	DD	DI	DD	DI	DD	DI
0-1 Loss	25.4 ± 1.01	22.8 ± 0.91	5.2 ± 0.21	43.9 ± 3.92	**30.6**	66.7
Manipulations	25.2 ± 1.02	22.8 ± 0.78	1.4 ± 0.45	43.7 ± 3.91	**26.7**	66.5
Qualified Manip.	2.8 ± 0.42	3.8 ± 0.35	0.6 ± 0.21	8.5 ± 2.17	**3.4**	12.3
Unqualified Manip.	22.4 ± 0.94	18.9 ± 0.69	0.8 ± 0.24	35.1 ± 3.58	**23.3**	54.0
Qualified Accepted default	46.3 ± 0.73	46.1 ± 0.77	46.8 ± 0.15	40.7 ± 2.18	**93.1**	86.8
Unqualified Accepted default	2.9 ± 0.37	3.9 ± 0.44	2.6 ± 0.11	8.8 ± 2.20	**5.5**	12.7

Table 1: Average $\pm$ standard error across stages and totals.

Acknowledgments and Disclosure of Funding. This work is supported in part by Cisco Research, and by The Jordan University of Science and Technology.

References

[1] Saba Ahmadi, Hedyeh Beyhaghi, Avrim Blum, and Keziah Naggita. The strategic perceptron. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 6–25, 2021.

[2] Miguel Alcobendas and Robert Zeithammer. Adjustment of bidding strategies after a switch to first-price rules. *Available at SSRN 4036006*, 2021.

[3] Sura Alhanouti and Parinaz Naghizadeh. Could anticipating gaming incentivize improvement in (fair) startegic classification. In *The IEEE Control and Decisions Conference (CDC)*, 2024.

[4] Yahav Bechavod, Katrina Ligett, Steven Wu, and Juba Ziani. Gaming helps! learning from strategic interactions in natural dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 1234–1242. PMLR, 2021.

[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] Yue Chen and Wei Wei. Robust generation dispatch with strategic renewable power curtailment and decision-dependent uncertainty. *IEEE Transactions on Power Systems*, 38(5):4640–4654, 2022.

[8] Allen Cheng. Sat and act tutoring is changing—here's why, 2022. URL https://blog.prepscholar.com/sat-act-tutoring-industry-changing. Accessed: 2025-07-30.

[9] Lee Cohen, Yishay Mansour, Shay Moran, and Han Shao. Learnability gaps of strategic classification. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1223–1259. PMLR, 2024.

[10] 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.

[11] 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*, pages 55–70, 2018.

[12] 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*, pages 868–886, 2025.

[13] Valia Efthymiou, Chara Podimata, Diptangshu Sen, and Juba Ziani. Incentivizing desirable effort profiles in strategic classification: The role of causality and uncertainty. *arXiv preprint arXiv:2502.06749*, 2025.

[14] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. First i "like" it, then i hide it: Folk theories of social feeds. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016.

[15] William Fitzsimmons and Robin Mamlet. On aps: Harvard and stanford admissions deans on advanced placement courses, 2009. URL https://mitadmissions.org/blogs/entry/on_aps_1/#:~:text=Harvard%3A. Accessed: 2025-08-14.

[16] Forbes. 5 resume hacks to pass ATS. https://www.forbes.com/sites/ashleystahl/2022/12/12/5-resume-hacks-to-pass-ats/?sh=3668530d4b2b, 2022. Accessed: March 15, 2024.

[17] Jack Geary and Henry Gouk. Strategic classification with randomised classifiers. *arXiv preprint arXiv:2502.01313*, 2025.

[18] Ganesh Ghalme, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen, and Nir Rosenfeld. Strategic classification in the dark. In *International Conference on Machine Learning*, pages 3672–3681. PMLR, 2021.

[19] Nico Grau and et al. The impact of college admissions policies on high school students' academic effort. *Economics of Education Review*, 2018.

[20] 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*, 2020.

[21] Nika Haghtalab, Nicole Immorlica, Brendan Lucier, and Jack Z. Wang. Maximizing welfare with incentive-aware evaluation mechanisms. *International Joint Conferences on Artificial Intelligence Organization*, 2020.

[22] Nika Haghtalab, Chara Podimata, and Kunhe Yang. Calibrated stackelberg games: Learning optimal commitments against calibrated agents. *Advances in Neural Information Processing Systems*, 36:61645–61677, 2023.

[23] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, 2016.

[24] Keegan Harris, Hoda Heidari, and Steven Z Wu. Stateful strategic regression. In *Advances in Neural Information Processing Systems*, 2021.

[25] Keegan Harris, Valerie Chen, Joon Kim, Ameet Talwalkar, Hoda Heidari, and Steven Z Wu. Bayesian persuasion for algorithmic recourse. *Advances in Neural Information Processing Systems*, 35:11131–11144, 2022.

[26] Zhiyu He, Saverio Bolognani, Florian Dörfler, and Michael Muehlebach. Decision-dependent stochastic optimization: The role of distribution dynamics. *arXiv preprint arXiv:2503.07324*, 2025.

[27] Guy Horowitz and Nir Rosenfeld. Causal strategic classification: A tale of two shifts. In *International Conference on Machine Learning*, pages 13233–13253. PMLR, 2023.

[28] 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.

[29] Brian Jacob and Steven D Levitt. Catching cheating teachers: The results of an unusual experiment in implementing theory, 2003.

[30] K-12 Dive Staff. Test-optional admissions policies reshape college preparation industry, 2023. URL https://www.k12dive.com/news/test-optional-college-prep-industry-shift/637252/. Accessed: 2025-07-30.

[31] Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)*, 2020.

[32] Carly Lapidus. The cost of test-optional policies: where do students shift their focus?, 2023. URL https://www.inklingsnews.com/b/2025/05/09/scores-in-policies-out-schools-reconsider-stances-on-standardized-testing/#:~:text=The%20trend%20of%20moving%20away,optional%20policies.%20This. Accessed: 2025-07-30.

[33] 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.

[34] Sagi Levanon and Nir Rosenfeld. Generalized strategic classification and the case of aligned incentives. In *International Conference on Machine Learning*, pages 12593–12618. PMLR, 2022.

[35] Making Caring Common Project. Turning the tide: Inspiring concern for others and the common good through college admissions. Technical report, Harvard Graduate School of Education, 2016. URL https://mcc.gse.harvard.edu/reports/turning-the-tide-college-admissions. Accessed: 2025-08-14.

[36] Garth P McCormick. Computability of global solutions to factorable nonconvex programs: Part i—convex underestimating problems. *Mathematical programming*, 10(1):147–175, 1976.

[37] John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2020.

[38] 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.

[39] 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.

[40] Haifeng Qiu, Veerapandiyan Veerasamy, Chao Ning, Qirun Sun, and Hoay Beng Gooi. Two-stage robust optimization for assessment of pv hosting capacity based on decision-dependent uncertainty. *Journal of Modern Power Systems and Clean Energy*, 12(6):2091–2096, 2024.

[41] 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.

[42] Aras Selvi, Dick den Hertog, and Wolfram Wiesemann. A reformulation-linearization technique for optimization over simplices. *Mathematical Programming*, 197(1):427–447, 2023.

[43] Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.

[44] Han Shao, Avrim Blum, and Omar Montasser. Strategic classification under unknown personalized manipulation. *Advances in Neural Information Processing Systems*, 36:26452–26484, 2023.

[45] Brandon Smart and Gustavo Carneiro. Bootstrapping the relationship between images and their clean and noisy labels. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5344–5354, 2023.

[46] 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.

[47] Laura Spitalniak. Wealthier students, those at private schools list more extracurriculars on college applications, April 2023. URL https://www.highereddive.com/news/application-advantage-extracurriculars-wealthy-white-asianstudents/648117/. Accessed: 2025-08-13.

[48] Ravi Sundaram, Anil Vullikanti, Haifeng Xu, and Fan Yao. Pac-learning for strategic classification. *Journal of Machine Learning Research*, 24(192):1–38, 2023.

[49] Wei Tang, Chien-Ju Ho, and Yang Liu. Linear models are robust optimal under strategic behavior. In *International Conference on Artificial Intelligence and Statistics*, pages 2584–2592. PMLR, 2021.

[50] Mohit Tawarmalani and Nikolaos V Sahinidis. A polyhedral branch-and-cut approach to global optimization. *Mathematical programming*, 103(2):225–249, 2005.

[51] The New York Times. 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.

[52] Benyamin Trachtenberg and Nir Rosenfeld. Strategic classification with non-linear classifiers. *arXiv preprint arXiv:2505.23443*, 2025.

[53] Qiaochu Wang, Yan Huang, Stefanus Jasin, and Param Vir Singh. Algorithmic transparency with strategic users. *Management Science*, 69(4):2297–2317, 2023. doi: 10.1287/mnsc.2022.4475.

[54] Alia Wong. The dark side of the college essay, 2021. URL https://www.theatlantic.com/education/archive/2021/09/college-essay-application-help/620155/. Accessed: 2025-07-30.

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

[56] 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*, 2022.

[57] Yikai Zhang, Songzhu Zheng, Pengxiang Wu, Mayank Goswami, and Chao Chen. Learning with feature-dependent label noise: A progressive approach. *arXiv preprint arXiv:2103.07756*, 2021.

[58] Yunfan Zhang, Feng Liu, Yifan Su, Yue Chen, Zhaojian Wang, and João PS Catalão. Two-stage robust optimization under decision dependent uncertainty. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1295–1306, 2022.

[59] Jianzhe Zhen, Danique de Moor, and Dick den Hertog. An extension of the reformulation-linearization technique to nonlinear optimization. *Available at Optimization Online*, 2021.

A Motivating Examples

A.1 Motivating Example: Strategic Cost Dependence in a Two-Stage Setting

A motivating example arises in school admissions. Empirical evidence suggests that shifting to test-optional policies alters the market for preparatory services—such as tutoring for standardized tests— while increasing emphasis on other components like extracurricular activities, which in turn affects the costs of modifying each application component for students [32, 47]. In this context, the institution (e.g., a university) can observe the current costs students incur to meet that year's admission criteria; this aligns with our assumption that the costs are known to the firm at the time of decision-making. The school also understands that its decisions during the admission season this year (i.e., the weight placed on different application components) will shape the strategic cost landscape for applicants in the following year. Although the exact nature of next year's costs is uncertain, the school can incorporate this decision-dependent uncertainty into its current admission policies to better anticipate long-term effects.

This awareness is not merely theoretical. Elite institutions have explicitly acknowledged that their admissions criteria impose significant long-term costs and influence strategic behaviors. For instance, it has been noted that students often arrive academically exhausted, attributing it to the widespread belief that excessive AP coursework is necessary for admission [15]. These reflections, echoed in the Harvard-based "Turning the Tide" report [35], illustrate institutional awareness that current admissions signals can fuel competitive overextension and inequitable cost burdens on future applicants. This institutional perspective is reinforced by empirical evidence: Grau and et al. [19] finds that admissions' policies directly influence high school students' academic effort, with schools emphasizing particular criteria—such as GPA or extracurricular achievement—inducing students to reallocate effort and preparation to align with those priorities.

A.2 Motivating Example: Decision-Dependent Uncertainty Set Modeling

To illustrate how decision-dependent uncertainty sets can be represented as a polyhedral set of the form (6),

$$\Omega(\beta) = \{\omega \in \mathbb{R}^d_+ : \mathbf{F}(\beta)\omega \leq \mathbf{h} + \mathbf{G}\beta\}.$$

Consider the college admissions example provided before: shifting emphasis from standardized tests to GPA or extracurriculars can alter the strategic landscape. For example, de-emphasizing SAT scores (e.g., through test-optional policies) may decrease demand for test prep (lowering SAT-related costs), while increasing demand for tutoring in coursework or essay coaching [30, 32], thus raising those prices. The constraint $\mathbf{F}(\beta)\omega \leq \mathbf{h}+\mathbf{G}\beta$ captures this shift, where $\mathbf{h}$ may represent a tutorial class's average price, $\mathbf{G}\beta$ reflects cost adjustments driven by the relative importance of different admissions criteria, and $\mathbf{F}(\beta)$ governs the structural relationships among cost components, altering how increases in one type of preparation (e.g., GPA-related tutoring) constrain or interact with others (e.g., extracurricular coaching).

B Related Work

Machine learning (ML) and AI increasingly influence decisions affecting strategic agents—individuals who adapt their observable features to secure favorable outcomes. This "strategic behavior" has sparked growing research interest. Early work assumes full information for both classifiers and agents, exploring settings where agents manipulate inputs without changing true qualifications (e.g., [23, 34, 38, 48]) or where they choose between manipulation and genuine improvement (e.g., [3, 4, 21, 24, 27, 53]).

More recent studies examine incomplete information. Some propose randomized classifiers to obscure decision boundaries and reduce gaming [6, 17, 48], with [17] extending results beyond linear models. Others analyze partial agent knowledge without randomization. For example, [10, 22, 25] explore strategic communication and calibrated forecasts, while [5, 18] study opacity and fairness implications of asymmetric information.

A separate line of work handles uncertainty in agent responses via online or sequential learning. These include learning under unknown manipulations [33, 44], unknown costs or preferences [1, 11], or distributional shifts [9, 26]. Other models address robustness to uncertain manipulation costs using worst-case optimization or distributionally robust methods [41, 43, 49].

Finally, studies show how decision rules shape manipulation incentives: lax oversight or platform design choices can reduce manipulation costs and increase gaming [2, 29]. However, most strategic ML models still assume fixed manipulation costs, overlooking how decisions themselves can alter them—highlighting the need to model decision-dependent manipulation costs. Unlike prior work, our approach explicitly accounts for this dynamic. We address the complexity of uncertainty in strategic behavior's costs, recognizing that it depends on the classifier's decisions.

C Approximations and Reformulation Details for the Two-Stage Problem

Throughout this section, we provide details on the approximation and the reformulation steps, along with intuitive interpretation and support for them in the context of strategic classification.

C.1 Approximation of the Dual Norm $\|\cdot\|_{*,\Sigma}$: Justification and Proof

Concretely, consider the empirical risk (5) appearing in the firm's optimization problem:

$$R_\Sigma(\beta) := \frac{1}{N}\sum_{i=1}^{N} \max\left\{0, 1 - y_i\left(\beta^\top x_i + u_*\|\beta\|_{*,\Sigma}\right)\right\}.$$

This risk contains terms of the form $\|\beta\|_{*,\Sigma}$, due to which the cost matrix $\Sigma(\omega)$ introduces an input-dependent geometry affecting the feasible region of the dual problem. Reformulating this norm constitutes a crucial first step toward achieving a linear second-stage problem in which uncertainty appears solely in the feasible region rather than in the objective. As discussed earlier, this transformation is essential for enabling the use of fixed extreme points of the dual second-stage problem, thereby facilitating bilevel decomposition and tractable optimization. Specifically, we use Lemma 1 to upper bound the dual-norm term.

Lemma 1. *There exists a constant $M > 0$ such that, $\|\beta\|_{*,\Sigma(\omega)} \le M\|\Sigma(\omega)^{-\frac{1}{2}}\|_\infty\|\beta\|_\infty$.*

Proof. From the *equivalence of norms theorem*, in finite-dimensional spaces, we know that for any two norms $\|\cdot\|_a$ and $\|\cdot\|_b$ on $\mathbb{R}^n$, there exist constants $m, M > 0$ such that:

$$m\|\beta\|_a \le \|\beta\|_b \le M\|\beta\|_a, \quad \forall \beta \in \mathbb{R}^n.$$

Noting that $\|\beta\|_{*,\Sigma(\omega)} = \|\Sigma(\omega)^{-\frac{1}{2}}\beta\|_*$ by definition, and invoking the equivalence of norms theorem for the ℓ_∞-norm, we have

$$\|\beta\|_{*,\Sigma(\omega)} \le M\|\Sigma(\omega)^{-\frac{1}{2}}\beta\|_\infty .$$

Additionally, recall that the induced matrix ∞-norm (the *maximum absolute row sum norm*) of a matrix $A \in \mathbb{R}^{d\times d}$ is defined as

$$\|A\|_\infty = \max_{1\le i\le d}\sum_{j=1}^{d}|a_{ij}|.$$

Using this definition, we observe that for the identity matrix $I_{d\times d}$,

$$\|\Sigma(\omega)^{-\frac{1}{2}}\beta\|_\infty \le \|\Sigma(\omega)^{-\frac{1}{2}}I_{d\times d}\beta\|_\infty .$$

Indeed, by expanding the norm directly,

$$\|\Sigma(\omega)^{-\frac{1}{2}}\beta\|_\infty = \max_i\left|\sum_{j=1}^{d}\Sigma(\omega)_{ij}^{-\frac{1}{2}}\,\beta_j\right|,$$

whereas

$$\|\Sigma(\omega)^{-\frac{1}{2}}I_{d\times d}\beta\|_\infty = \max_i\sum_{j=1}^{d}\left|\Sigma(\omega)_{ij}^{-\frac{1}{2}}\,\beta_j\right|.$$

Moreover, since the ℓ_∞-norm atisfies the *submultiplicativity* property, we have

$$\begin{aligned}\|\Sigma(\omega)^{-\frac{1}{2}}I_{d\times d}\beta\|_\infty &\le \|\Sigma(\omega)^{-\frac{1}{2}}\|_\infty\cdot\|I_{d\times d}\beta\|_\infty\\ &= \|\Sigma(\omega)^{-\frac{1}{2}}\|_\infty\cdot\|\beta\|_\infty.\end{aligned}$$

Applying this leads to the stated bound on the matrix-induced norm. □

Example. As a concrete example, consider the case where the primal norm adopted in the cost function is the commonly considered ℓ_2-norm. Then, the dual norm is also an ℓ_2-norm, and the following equivalence holds: $\|z\|_2 \le \sqrt{d}\|z\|_\infty$ for $z \in \mathbb{R}^d$. Thus, we derive the upper bound:

$$\|\beta\|_{*,\Sigma(\omega)} \le \sqrt{d}\ \|\Sigma(\omega)^{-\frac{1}{2}}\|_\infty\ \|\beta\|_\infty.$$

The upper bound in Lemma 1 provides a linearizable surrogate for the dual norm, which we will use to reformulate the recourse problem with a fixed dual feasible set suitable for decomposition. While this bounding argument is strictly necessary only for the second stage (since the cost matrix $\Sigma(\omega)$ depends on the unknown realization ω), we will also apply the same approximation to the first-stage term. Doing so ensures a consistent treatment of both stages and simplifies the subsequent linearization, making the overall problem more amenable to bilevel decomposition.

Intuitive justification for approximating the dual norm. From a technical perspective, using the proposed approximation of the dual-norm makes the uncertainty set of the second stage decision-independent, simplifying the analysis. We also provide some intuitive support for this approximation in the context of strategic classification.

As shown in Lemma 2.3 in Rosenfeld and Rosenfeld [41], for a cost function of the form $c(x, \hat{x}) = \phi(\|\hat{x} - x\|_{\Sigma(\omega^k)})$, the quantity $\|\beta'\|_{*,\Sigma(\omega^k)}$ represents the maximum possible score change resulting from an agent's strategic response to a fixed classifier β'. By approximating this term via the inequality $\|\beta'\|_{*,\Sigma(\omega)} \leq M\|\Sigma(\omega)^{-\frac{1}{2}}\|_\infty\|\beta'\|_\infty$, we obtain a conservative upper bound on the cost-aware strategic hinge loss. For negatively labeled (unqualified) agents, $y_i = -1$, the loss is upper bounded as: $\ell_{\Sigma(\omega),\text{s-hinge}}(\beta'^\top x_i, y_i) \leq \max\{0, 1 + (\beta'^\top x_i + u_* M\|\Sigma(\omega)^{-\frac{1}{2}}\|_\infty\|\beta'\|_\infty)\}$. In contrast, for positively labeled (qualified) agents, $y_i = 1$, the same approximation gives a lower bound on the hinge loss: $\ell_{\Sigma(\omega),\text{s-hinge}}(\beta'^\top x_i, y_i) \geq \max\{0, 1 - (\beta'^\top x_i + u_* M\|\Sigma(\omega)^{-\frac{1}{2}}\|_\infty\|\beta'\|_\infty)\}$.

This asymmetry has important implications: the approximation leads to a more conservative (i.e., robust) response to the manipulative behavior of unqualified agents, who pose a risk to classifier accuracy by altering features to obtain a false positive. At the same time, it is less conservative toward strategic behavior by already-qualified agents, whose feature manipulation only reinforces the correct decision. Therefore, this approximation supports designing classifiers that prioritize robustness against harmful manipulation, which aligns with real-world settings where accepting unqualified agents carries higher cost than misclassifying qualified ones.

C.2 Linear reformulation of the stage problems

Using the proposed approximation of the dual norm from Lemma 1, we can approximate the firm's optimization (7) as:

$$\begin{aligned}
&\min_{\beta\in\mathcal{B}} \big[\frac{1}{N}\sum_{i=1}^{N} \max\{0, 1 - y_i(\beta^\top x_i + u_* M\|\Sigma_0^{-\frac{1}{2}}\|_\infty\|\beta\|_\infty)\}\big] \\
&+ \max_{\omega\in\Omega(\beta)} \min_{\beta'\in\mathcal{B}'(\beta,\omega)} \\
&\big[\frac{1}{N}\sum_{i=1}^{N} \max\{0, 1 - y_i\big(\beta'^\top x_i + u_* M\|\Sigma(\omega)^{-\frac{1}{2}}\|_\infty\|\beta'\|_\infty\big)\}\big] \,, \qquad (8)
\end{aligned}$$

where, M is an appropriately selected constant depending on the p-norm adopted in the cost function. We next linearize each of the first and second stage objective functions.

There are now two further issues that prevent the applicability of the Bender C&CG algorithm: (1) the infinity norms and the max operator make the objective function nonlinear, and (2) the objective function of the second stage is decision-dependent (note the appearance of the matrix $\Sigma(\omega)$); for the applicability of the Bender C&CG algorithm, only the constraint set of the second stage (and not its objective function) can be decision-dependent. We address these issues in these subsections.

C.2.1 Linearizing and removing decision-dependence in the second-stage objective function

We begin with the second-stage objective function. First, the max operator in the objective function is handled via the introduction of slack variables $s_{2,i} \in \mathbb{R}_+$, subject to the constraints

$$s_{2,i} \geq 0, \text{and } s_{2,i} \geq 1 - y_i(\beta'^\top x_i + u_* M\|\Sigma(\omega)^{-\frac{1}{2}}\|_\infty\|\beta'\|_\infty).$$

Let $s_2 = (s_{2,1}, \ldots, s_{2,N}) \in \mathbb{R}_+^N$ denote the vector of slack variables at the second-stage.[1] Next, we linearize the ∞-norm terms, by introducing slack variables $t_2^q, t_2^\omega \in \mathbb{R}_+$, and impose the following constraints: $-t_2^q \leq \beta'_j \leq t_2^q,\ \forall j \in [1:d]$ and $-t_2^\omega \leq \sum_{r=1}^d [\Sigma(\omega)^{-\frac{1}{2}}]_{jr} \leq t_2^\omega,\ \forall j \in [1:d]$. This linearization of the infinity norms introduces a bilinear product on the constraints on the slack variables $s_{2,i}$; specifically,

$$s_{2,i} \geq 1 - y_i(\beta'^\top x_i + u_* M t_2^\omega \cdot t_2^q).$$

To preserve linearity, we adopt the McCormick envelope [36], introducing an auxiliary variable $z \in \mathbb{R}_+$ to represent the bilinear term $z = t_2^q \cdot t_2^\omega$, along with the corresponding envelope constraints. This relaxation is tight in LPs and widely used in nonlinear programming [42, 50, 59]. McCormick envelope requires bounds on t_2^q and t_2^ω. Intuitively, $\|\Sigma(\omega)^{-\frac{1}{2}}\|_\infty$ reflects the most cost-efficient direction of manipulation, while $\|\beta'\|_\infty$ measures classifier sensitivity; thus, their product captures the worst-case strategic impact on the firm loss. These bounds

[1] We note that despite the reformulation adding N (number of sample) constraints, this will not raise computational issues as the algorithm will solve this problem in minibatches.

can be derived from domain knowledge. For example, one may assume $\|\Sigma(\omega)^{-\frac{1}{2}}\|_\infty \leq \frac{1}{\underline{c}}$, where $\underline{c}$ is the minimal feature manipulation cost, and $\|\beta'\|_\infty \leq \bar{\beta}'$, where $\bar{\beta}'$ bounds the model coefficients. We denote these bounds by $t^\omega_{2,\max}$ and $t^q_{2,\max}$, respectively. Finally, we decompose the second-stage classifier's weight vector as $\beta' = \beta'^+ - \beta'^-$, where $\beta'^+, \beta'^- \in \mathbb{R}^d_+$ denote the positive and negative components of β, respectively; note that by adopting this choice, each of β'^+, β'^- is itself positive.

Putting these together, the linearized second-stage ("recourse") problem with second-stage feasible set $\mathcal{B}'(\beta, \omega)$ will be:

$$\min_{\beta'^+, \beta'^-, s_2, t^q_2, t^\omega_2, z} \frac{1}{N} \sum_{i=1}^N s_{2,i} \tag{9}$$

subject to:

$$s_{2,i} \geq 1 - y_i \left((\beta'^+ - \beta'^-)^\top x_i + u_* M z \right), i = 1, \ldots, N, \tag{(9).a}$$

$$-t^q_2 \leq \beta'^+_j - \beta'^-_j \leq t^q_2,\ j = 1, \ldots, d, \tag{(9).b}$$

$$-t^\omega_2 \leq \sum_{r=1}^d [\Sigma(\omega)^{-\frac{1}{2}}]_{jr} \leq t^\omega_2,\ j = 1, \ldots, d, \tag{(9).c}$$

$$\mathbf{B}_2(\beta'^+ - \beta'^-) \geq \mathbf{d} - \mathbf{B}_1(\beta^+ - \beta^-) - \mathbf{E}\omega, \tag{(9).d}$$

$$t^q_2 \leq t^q_{2,\max}, \quad t^\omega_2 \leq t^\omega_{2,\max}, \tag{(9).e}$$

$$z \leq t^\omega_{2,\max} t^q_2, \quad z \leq t^q_{2,\max} t^\omega_2, \tag{(9).f}$$

$$z \geq t^q_{2,\max} t^\omega_2 + t^\omega_{2,\max} t^q_2 - t^\omega_{2,\max} t^q_{2,\max}, \tag{(9).g}$$

$$\beta'^+_j, \beta'^-_j, s_{2,i}, t^q_2, t^\omega_2, z \geq 0, i = 1, \ldots, N, j = 1, \ldots, d. \tag{(9).h}$$

C.2.2 Linearizing the first-stage objective functions.

While not strictly necessary for the application of the Bender C&CG algorithm, we also linearize the first-stage objective function of (8) for consistency and in order to later simplify the algorithmic solution. This will involve linearizing the max operator with slack variables $s_1 = (s_{1,1}, \ldots, s_{1,N}) \in \mathbb{R}^N_+$, linearizing the infinity norm on β with slack variable $t_1 \in \mathbb{R}_+$, and representing the classifier's weight vector as $\beta = \beta^+ - \beta^-$, where $\beta^+, \beta^- \in \mathbb{R}^d_+$ denote the positive and negative components.

The resulting linearized first-stage problem with first-stage feasible set $\mathcal{B}$ is:

$$\min_{\beta^+, \beta^-, s_1, t_1} \frac{1}{N} \sum_{i=1}^N s_{1,i} \tag{11}$$

subject to:

$$s_{1,i} \geq 1 - y_i \left((\beta^+ - \beta^-)^\top x_i + u_* M \|\Sigma_0^{-\frac{1}{2}}\|_\infty t_1 \right), \quad i = 1, \ldots, N, \tag{(11).a}$$

$$-t_1 \leq \beta^+_j - \beta^-_j \leq t_1,\ j = 1, \ldots, d, \tag{(11).b}$$

$$\mathbf{A}(\beta^+ - \beta^-) \geq \mathbf{b}, \tag{(11).c}$$

$$\beta^+_j, \beta^-_j, t_1, s_{1,i} \geq 0,\ \ i = 1, \ldots, N,\ \ j = 1, \ldots, d. \tag{(11).d}$$

C.3 Linear tri-level reformulation of the firm's optimization problem

We now present the complete linear reformulation of our two-stage robust optimization problem under decision-dependent uncertainty. The resulting tri-level linear optimization problem is:

$$\min_{v_1 \in \mathcal{S}_{t1}} \quad \frac{1}{N} \sum_{i=1}^N s_{1,i} + \max_{\omega \in \Omega(\beta)} \min_{v_2 \in \mathcal{S}_{t2}(\omega)} \frac{1}{N} \sum_{i=1}^N s_{2,i}. \tag{13}$$

Here, the first-stage and second-stage decision vectors are:

$$v_1 := (\beta^+, \beta^-, s_1, t_1), \quad v_2 := (\beta'^+, \beta'^-, s_2, t^\omega_2, t^q_2, z),$$

the feasible set for the first-stage problem is:

$$\mathcal{S}_{t1} = \left\{ \beta^+, \beta^-, s_1, t_1 \middle| \text{constraints ((11).a) – ((11).d)} \right\},$$

and, the second-stage feasible set for any given ω is:

$$\mathcal{S}_{t2}(\omega) = \big\{\beta'^{+}, \beta'^{-}, s_2, z, t_2^q, t_2^\omega \mid \text{constraints ((9).a)–((9).h)}\big\}.$$

To solve this reformulated problem, we can now adopt the *Benders Column-and-constraint Generation (C&CG) algorithm* of Zeng and Wang [55]. The application of this algorithm is subject to mild assumptions on (13), which we also assume: (i) For any $\beta \in \mathcal{B}$, $\Omega(\beta) \neq \emptyset$; (ii) $\Omega(\beta)$ is bounded, i.e., for any $\beta \in \mathcal{B}$, $\omega_j < \infty \; \forall j$; (iii) The program $\min\{\frac{1}{N}\sum_{i=1}^{N} s_{1,i} + \frac{1}{N}\sum_{i=1}^{N} s_{2,i} : v_1 \in \mathcal{S}_{t1}, \omega \in \Omega(\beta), v_2 \in \mathcal{S}_{t2}(\omega)\}$ has a finite optimal value. Assumption (i) is required to maintain a valid two-stage framework. For example, if for a first-stage firm's decision β^k the uncertainty set turns out to be empty ($\Omega(\beta^k) = \emptyset$), then the second-stage cost matrix $\Sigma(\omega)$ is not defined, and the second-stage problem becomes undefined. Consequently, the two-stage problem is trivially unbounded. Assumption (ii) ensures boundedness of the random factor ω. This limits the influence of the first-stage decision on the manipulation cost in the second stage. In other words, it is not possible to drive ω to infinity, which implies, by the definition of $\Sigma(\omega)$ and our choice of $g(\omega)$, that the second-stage cost cannot be zero, and manipulation cannot be free. Finally, assumption (iii) helps detect infeasibility of (13) through its associated relaxation. By the definition of most two-stage optimization problems, a first-stage decision β^k is feasible if the second-stage problem is feasible for all $\omega \in \Omega(\beta^k)$, and infeasible otherwise. Thus, the two-stage problem is infeasible if no feasible first-stage decision exists. Hence, (13) is infeasible if its relaxation is infeasible.

D Master Problem Reformulation Derivations

According to Zeng and Wang [55], a single-level reformulation of the linear tri-model in (13) is required to define the master problem. In this appendix, we will step by step derive this master problem following Zeng and Wang [55] approach and reformulations. Specifically, this is achieved by first dualizing the linearized second-stage ("resource") problem in (9) (Appendix D.1) to formulate the bilevel reformulation, leveraging the extreme points of the fixed feasible region. As noted by [55], the resulting bilevel reformulation has a lower level complex disjoint bilinear program, which can be solved by enumerating on the extreme points and rays of the dual feasible region. This results in a linear bilevel reformulation (Appendix D.2). Moreover, this linear bilevel reformulation has lower-level linear programs for each extreme point and extreme ray (of the dual second-stage ("recourse") feasible set), whose KKT condition-based sets are enumerated to formulate the single-level reformulation (Appendix D.3). By considering a subset of these extreme points and rays, we achieve a relaxation of the large-scale single-level reformulation and a lower bound on its optimal value, that is, the "Master problem" (Appendix D.4).

D.1 Dualizing the second-stage ("recourse") problem

As mentioned before, we start by dualizing the second-stage ("recourse") problem to have the bilevel reformulation. Let $X \in \mathbb{R}^{N\times d}$, where each row is $x_i^\top$, for $i = 1, 2, \cdots, N$, and $Y \in \mathbb{Z}^N$. Given $s_1, s_2 \in \mathbb{R}^N$. The row sum of $\Sigma(\omega)^{-\frac{1}{2}}$ in constraint ((9).c), can be written as

$$\sum_{j=1}^{d} Q(g(\omega))_{ij}^{-\frac{1}{2}} \cdot \Sigma_{0,ij}^{-\frac{1}{2}}.$$

This is a sum of linear functions in ω. For example, if both Σ_0, and $Q(g(\cdot))$ are diagonal matrices, the row sum simplifies to $g(\omega_i)^{-\frac{1}{2}} \cdot \sigma_{\Sigma_0,i}^{-\frac{1}{2}}$. Under our assumption that $g(\omega)^{-\frac{1}{2}}$ is linear in ω, the row sum is therefore linear in ω as well. More generally, this row sum can be expressed compactly in vector form as

$$a_g \Sigma_0^{-\frac{1}{2}} \omega,$$

where a_g encodes the coefficients induced by $g(\omega)^{-\frac{1}{2}}$. Below, we rewrite the second-stage problem in vector form, annotating each constraint with its corresponding dual variable.

$$\min \frac{\overrightarrow{1}}{N} s_2 \tag{14}$$

Subject to

$$
\begin{array}{lrll}
s_2 & +(Y \odot X)(\beta^{'+} - \beta^{'-}) + (u_* MY)z & \geq \overrightarrow{1}. & (\pi_0 \in \mathbb{R}^N) \\
& -(\beta^{'+} - \beta^{'-}) + \overrightarrow{1} t_2^q & \geq \overrightarrow{0}. & (\pi_1 \in \mathbb{R}^d) \\
& (\beta^{'+} - \beta^{'-}) + \overrightarrow{1} t_2^q & \geq \overrightarrow{0}. & (\pi_2 \in \mathbb{R}^d) \\
& \overrightarrow{1} t_2^\omega & \geq -a_g \Sigma_0^{-1/2} \omega. & (\pi_3 \in \mathbb{R}^d) \\
& \overrightarrow{1} t_2^\omega & \geq a_g \Sigma_0^{-1/2} \omega. & (\pi_4 \in \mathbb{R}^d) \\
& \mathbf{B}_2(\beta^{'+} - \beta^{'-}) & \geq \mathbf{d} - \mathbf{B}_1(\beta^+ - \beta^-) - \mathbf{E}\omega. & (\pi_5 \in \mathbb{R}^n) \\
& - t_2^q & \geq -t_{2,max}^q. & (\pi_6 \in \mathbb{R}) \\
& - t_2^\omega & \geq -t_{2,max}^\omega. & (\pi_7 \in \mathbb{R}) \\
& - z + t_{2,max}^\omega t_2^q. & \geq 0. & (\pi_8 \in \mathbb{R}) \\
& - z + t_{2,max}^q t_2^\omega & \geq 0. & (\pi_9 \in \mathbb{R}) \\
& z - t_{2,max}^q t_2^\omega - t_{2,max}^\omega t_2^q & \geq -t_{2,max}^\omega t_{2,max}^q. & (\pi_{10} \in \mathbb{R}) \\
& s_2, \beta^{'+}, \beta^{'-}, z, t_2^\omega, t_2^q & \geq 0. &
\end{array}
$$

Here, n denotes the number of constraints that characterize the feasible adjustment space of the classifier in the second stage, as outlined in Section 3.

The following is the **LP** dual of the linearized second-stage ("recourse") problem:

$$
\begin{aligned}
\max \mathbf{1}^\top \pi_0 + a_g(\Sigma_0^{-\frac{1}{2}}\omega)^\top \pi_3 - a_g(\Sigma_0^{-\frac{1}{2}}\omega)^\top \pi_4 + (d - \mathbf{B}_1(\beta^+ - \beta^-))^\top \pi_5 \\
- (\mathbf{E}\omega)^\top \pi_5 - t_{2,max}^q \, \pi_6 - t_{2,max}^\omega \, \pi_7 - t_{2,max}^\omega t_{2,max}^q \, \pi_{10}
\end{aligned} \tag{16}
$$

Subject to

$$
\begin{aligned}
& \pi_0 \leq \frac{\overrightarrow{1}}{N}. && \text{((16).a)} \\
& (Y \odot X)^\top \pi_0 + \mathbf{B}_2^\top \pi_5 - \pi_1 + \pi_2 \leq 0. && \text{((16).b)} \\
& -(Y \odot X)^\top \pi_0 - \mathbf{B}_2^\top \pi_5 + \pi_1 - \pi_2 \leq 0. && \text{((16).c)} \\
& u_* M(Y)^\top \pi_0 - \pi_8 - \pi_9 + \pi_{10} \leq 0. && \text{((16).d)} \\
& \overrightarrow{1} \pi_1 + \overrightarrow{1} \pi_2 - \pi_6 + t_{2,max}^\omega \pi_8 - t_{2,max}^\omega \pi_{10} \leq 0. && \text{((16).e)} \\
& \overrightarrow{1} \pi_3 + \overrightarrow{1} \pi_4 - \pi_7 + t_{2,max}^q \pi_9 - t_{2,max}^q \pi_{10} \leq 0. && \text{((16).f)} \\
& \pi_0, \pi_2, \pi_1, \pi_3, \pi_4, \pi_5, \pi_6, \pi_7, \pi_8, \pi_9, \pi_{10} \geq 0.
\end{aligned}
$$

Let π denote the set of all dual variables, with each π_j for $j = 0, .., 10$ is in the proper dimension. Note that the dual feasible region is independent of both the uncertainty in ω and the first stage decision $(\beta^+ - \beta^-)$. Let the feasible set of the dual-recourse problem be denoted by

$$
\Pi = \Big\{ \pi \geq 0, \ \textbf{Equations } \text{((16).a)-((16).f)} \Big\}.
$$

D.2 Bilevel Reformulation

Given the dual second-stage problem from the previous section, we now proceed by writing the bilevel reformulation of the linear tri-level problem (8).

$$
\begin{aligned}
\min_{\mathcal{S}_{t1}} \frac{1}{N} \sum_{i=1}^{N} s_{1,i} + \max \Big\{ & \mathbf{1}^\top \pi_0 + a_g(\Sigma_0^{-\frac{1}{2}}\omega)^\top \pi_3 - a_g(\Sigma_0^{-\frac{1}{2}}\omega)^\top \pi_4 + (d - \mathbf{B}_1(\beta^+ - \beta^-))^\top \pi_5 \\
& -(\mathbf{E}\omega)^\top \pi_5 - t_{2,max}^q \, \pi_6 - t_{2,max}^\omega \, \pi_7 - t_{2,max}^\omega t_{2,max}^q \, \pi_{10} : \omega \in \Omega(\beta), \pi \in \Pi \Big\}.
\end{aligned} \tag{18}
$$

Note, as discussed before, this bilevel reformulation has a lower-level complex disjoint bilinear program. Zeng and Wang [55] reformulate this as a linear bilevel reformulation by enumerating on the extreme points and rays of Π. Let $\mathcal{P}_\Pi, \mathcal{R}_\Pi$ be the set of extreme points and extreme rays of Π, respectively, with $K_p = |\mathcal{P}_\Pi|$

and $K_r = |\mathcal{R}_\Pi|$. By enumerating, we can further get a simpler but large-scale linear bilevel reformulation as follows:

$$\min \frac{1}{N}\sum_{i=1}^{N} s_{1,i} + \eta \tag{19}$$

Subject to

$$v_1 \in \mathcal{S}_{t1}, \tag{(19).a}$$

$$\Big\{\eta \geq \mathbf{1}^\top \pi_0 + (d - \mathbf{B}_1(\beta^+ - \beta^-))^\top \pi_5 - t^q_{2,max}\,\pi_6 - t^\omega_{2,max}\,\pi_7 - t^\omega_{2,max} t^q_{2,max}\,\pi_{10} \tag{(19).b}$$

$$+ \max_{\omega\in\Omega(\beta)} \{a_g(\Sigma_0^{-\frac{1}{2}}\omega)^\top \pi_3 + a_g(-\Sigma_0^{-\frac{1}{2}}\omega)^\top \pi_4 + (-\mathbf{E}\omega)^\top \pi_5\} : \omega \in \Omega(\beta), \pi \in \Pi\Big\} \forall \pi \in \mathcal{P}_\Pi,$$

$$\Big\{\mathbf{1}^\top \gamma_0 + (d - \mathbf{B}_1(\beta^+ - \beta^-))^\top \gamma_5 - t^q_{2,max}\,\gamma_6 - t^{\tilde{\mathbf{v}}}_{2,max}\,\gamma_7 - t^{\tilde{\mathbf{v}}}_{2,max} t^q_{2,max}\,\gamma_{10} \tag{(19).c}$$

$$+ \max_{\tilde{\mathbf{v}}\in\Omega(\beta)} \{a_g(\Sigma_0^{-\frac{1}{2}}\tilde{\mathbf{v}})^\top \gamma_3 + a_g(-\Sigma_0^{-\frac{1}{2}}\tilde{\mathbf{v}})^\top \gamma_4 + (-\mathbf{E}\tilde{\mathbf{v}})^\top \gamma_5\} \leq 0\Big\} \forall \gamma \in \mathcal{R}_\Pi.$$

Note that the variable $\tilde{\mathbf{v}}$ is an alias of ω. Given the lower-level LPs appears in ((19).b) and ((19).c), let $\mathbf{LP}(\beta, U) : \max\{a_g(\Sigma_0^{-\frac{1}{2}}\omega)^\top U_3 + a_g(-\Sigma_0^{-\frac{1}{2}}\omega)^\top U_4 + (-\mathbf{E}\omega)^\top U_5 : \omega \in \Omega(\beta)\}$. Using the KKT conditions let $\mathcal{O}\Omega(\beta, \pi^k)$ denotes the optimal solution set of $\mathbf{LP}(\beta, \pi^k)$. Then,

$$\mathcal{O}\Omega(\beta, \pi^k) = \begin{Bmatrix} \mathbf{F}(\beta)\,\omega^k \leq \mathbf{h} + \mathbf{G}(\beta^+ - \beta^-), \\ \mathbf{F}(\beta)^\top \lambda^k \geq +a_g\Sigma_0^{-\frac{1}{2}}\pi_3^k - a_g\Sigma_0^{-\frac{1}{2}}\pi_4^k - \mathbf{E}\pi_5^k \\ \lambda^k \odot \Big(\mathbf{h} + \mathbf{G}(\beta^+ - \beta^-) - \mathbf{F}(\beta)\,\omega^k\Big) = 0, \\ \omega^k \odot \Big(\mathbf{F}(\beta)^\top \lambda^k - a_g\Sigma_0^{-\frac{1}{2}}\pi_3^k + a_g\Sigma_0^{-\frac{1}{2}}\pi_4^k + \mathbf{E}^\top \pi_5^k\Big) = 0 \\ \omega^k,\ \lambda^k \geq 0 \end{Bmatrix}, \tag{21}$$

where λ^k represents the dual variable of constraints $\Omega(\beta)$. Similarly, we can define $\mathcal{OV}(\beta, \gamma^l)$ to be the set of optimal solution for $\mathbf{LP}(\beta, \gamma^l)$ using KKT conditions.

D.3 Single Level Reformulation

The bilevel reformulation in the previous section (Appendix D.2) is also equivalent to the single-level optimization problem in this section. Using the sets $\mathcal{O}\Omega$ and $\mathcal{OV}$ defined in the previous section, we can write a single-level reformulation,

$$\min \frac{1}{N}\sum_{i=1}^{N} s_{1,i} + \eta \tag{22}$$

Subject to

$$v_1 \in \mathcal{S}_{t1},$$

$$\Big\{\eta \geq \mathbf{1}^\top \pi_0^k + (d - \mathbf{B}_1(\beta^+ - \beta^-))^\top \pi_5^k - t^q_{2,max}\,\pi_6^k - t^{\omega^k}_{2,max}\,\pi_7^k - t^{\omega^k}_{2,max} t^q_{2,max}\,\pi_{10}^k$$

$$+ a_g(\Sigma_0^{-\frac{1}{2}}\omega^k)^\top \pi_3^k + a_g(-\Sigma_0^{-\frac{1}{2}}\omega^k)^\top \pi_4^k - (\mathbf{E}\omega)^\top \pi_5^k\Big\},\ k = 1, \cdots, K_p,$$

$$(\omega^k, \lambda^k) \in \mathcal{O}\Omega(\beta, \pi^k),\ k = 1, \cdots, K_p$$

$$\Big\{\mathbf{1}^\top \gamma_0^l + (d - \mathbf{B}_1(\beta^+ - \beta^-))^\top \gamma_5^l - t^q_{2,max}\,\gamma_6^l - t^{\tilde{\mathbf{v}}^l}_{2,max}\,\gamma_7^l - t^{\tilde{\mathbf{v}}^l}_{2,max} t^q_{2,max}\,\gamma_{10}^l$$

$$+ a_g(\Sigma_0^{-\frac{1}{2}}\tilde{\mathbf{v}}^l)^\top \gamma_3^l - a_g(\Sigma_0^{-\frac{1}{2}}\tilde{\mathbf{v}}^l)^\top \gamma_4^l - (\mathbf{E}\tilde{\mathbf{v}}^{\mathbf{l}})^\top \gamma_5^l\} \leq 0\Big\},\ l = 1, \cdots, K_r$$

$$(\tilde{\mathbf{v}}^l, \xi^l) \in \mathcal{OV}(\beta, \gamma_1^l),\ l = 1, \cdots, K_r$$

D.4 The Master Problem

As mentioned before, the single-level reformulation includes the use of all extreme points and rays of the dual-recourse feasible region (Π). Hence, for a subset of these extreme points and the extreme rays sets, we have a relaxation of the single-level reformulation ((22)) which is the "Master problem". Let $\hat{\mathcal{P}}_\Pi \subseteq \mathcal{P}_\Pi$, and $\hat{\mathcal{R}}_\Pi \subseteq \mathcal{R}_\Pi$ be the subset of the extreme points and rays, receptively. We can lower bound the problems in the following **"Master problem"**:

$$\min \frac{1}{N}\sum_{i=1}^{N} s_{1,i} + \eta \tag{23}$$

Subject to

$$v_1 \in \mathcal{S}_{t1}, \tag{(23).a}$$

$$\begin{aligned}\big\{\eta \geq \mathbf{1}^\top \pi_0 + (d - \mathbf{B}_1(\beta^+ - \beta^-))^\top \pi_5 - t^q_{2,max}\,\pi_6 - t^\omega_{2,max}\,\pi_7 - t^\omega_{2,max} t^q_{2,max}\,\pi_{10} \\ + a_g(\Sigma_0^{-\frac{1}{2}}\omega^\pi)^\top \pi_3 + a_g(-\Sigma_0^{-\frac{1}{2}}\omega^\pi)^\top \pi_4 - (\mathbf{E}\omega^\pi)^\top \pi_5\big\} \quad \forall \pi \in \hat{\mathcal{P}}_\Pi,\end{aligned} \tag{(23).b}$$

$$(\omega^\pi, \lambda^\pi) \in \mathcal{O}\Omega(\beta, \pi^\pi),\ \forall \pi \in \hat{\mathcal{P}}_\Pi \tag{(23).c}$$

$$\begin{aligned}\big\{\mathbf{1}^\top \gamma_0 + (d - \mathbf{B}_1(\beta^+ - \beta^-))^\top \gamma_5 - t^q_{2,max}\,\gamma_6 - t^{\tilde{\mathbf{v}}}_{2,max}\,\gamma_7 - t^{\tilde{\mathbf{v}}}_{2,max} t^q_{2,max}\,\gamma_{10} \\ + a_g(\Sigma_0^{-\frac{1}{2}}\tilde{\mathbf{v}}^\gamma)^\top \gamma_3 - a_g(\Sigma_0^{-\frac{1}{2}}\tilde{\mathbf{v}}^\gamma)^\top \gamma_4 - (\mathbf{E}\tilde{\mathbf{v}}^\gamma)^\top \gamma_5\big\} \leq 0 \quad \forall \gamma \in \hat{\mathcal{R}}_\Pi.\end{aligned} \tag{(23).d}$$

$$(\tilde{\mathbf{v}}^\gamma, \xi^\gamma) \in \mathcal{OV}(\beta, \gamma_1^\gamma),\ \forall \gamma \in \hat{\mathcal{R}}_\Pi \tag{(23).e}$$

E Formulate the Subproblems of C&CG

Subproblem 1 (SP1) The first subproblem (SP1) is formulated to check the feasibility of the current first-stage β^*, which by definition is feasible if the recourse problem is feasible to all scenarios in the uncertainty set $\Omega(\beta^*)$.

$$\begin{aligned}\eta_f(\beta^*) = \max_{\omega \in \Omega(\beta^*)} \min\ & \mathbf{1}^\top \tilde{\mathbf{S}}_0 + \mathbf{1}^\top \tilde{\mathbf{S}}_1 + \mathbf{1}^\top \tilde{\mathbf{S}}_2 + \mathbf{1}^\top \tilde{\mathbf{S}}_3 + \mathbf{1}^\top \tilde{\mathbf{S}}_4 + \mathbf{1}^\top \tilde{\mathbf{S}}_5 + \tilde{\mathbf{S}}_6 + \tilde{\mathbf{S}}_7 \\ & + \tilde{\mathbf{S}}_8 + \tilde{\mathbf{S}}_9 + \tilde{\mathbf{S}}_{10}\end{aligned} \tag{25}$$

S.t.

$$\begin{aligned}
&\tilde{\mathbf{S}}_0 + s_{2,i} + y_i\left((\beta'^+ - \beta'^-)^\top x_i + u_* M z\right) \geq 1, \quad i = 1, \ldots, N,\\
&\tilde{\mathbf{S}}_1 - (\beta'^+ - \beta'^-)_j \geq -t^q_2, \quad j = 1, \ldots, d,\\
&\tilde{\mathbf{S}}_2 + (\beta'^+ - \beta'^-)_j \geq -t^q_2, \quad j = 1, \ldots, d,\\
&\tilde{\mathbf{S}}_3 - \sum_j [\Sigma(\omega)^{-\frac{1}{2}}]_{ij} \geq -t^\omega_2, \quad j = 1, \ldots, d,\\
&\tilde{\mathbf{S}}_4 + \sum_j [\Sigma(\omega)^{-\frac{1}{2}}]_{ij} - t^\omega_2 \geq -t^\omega_2, \quad j = 1, \ldots, d,\\
&\tilde{\mathbf{S}}_5 + \mathbf{B}_2(\beta'^+ - \beta'^-) \geq \mathbf{d} - \mathbf{B}_1\beta^* - \mathbf{E}\omega,\\
&\tilde{\mathbf{S}}_6 - t^q_2 \geq -t^q_{2,max}\\
&\tilde{\mathbf{S}}_7 - t^\omega_2 \geq -t^\omega_{2,max}\\
&\tilde{\mathbf{S}}_8 - z \geq -t^\omega_{2,max} t^q_2\\
&\tilde{\mathbf{S}}_9 - z \geq -t^q_{2,max} t^\omega_2\\
&\tilde{\mathbf{S}}_{10} + z \geq t^q_{2,max} t^\omega_2 + t^\omega_{2,max} t^q_2 - t^\omega_{2,max} t^q_{2,max}\\
&\tilde{\mathbf{S}}_j, z, s_{2,i}, \beta'^+, \beta'^-, t^q_2, t^\omega_2 \geq 0, \text{ for } i = 1, 2, ..., N \text{ , for } j = 1, 2, .., 12.
\end{aligned}$$

Accordingly, the linearized tri-model in Equation (8) and all its equivalences are feasible for β^* if and only if **SP1** objective value is zero ($\eta_f(\beta^*) = 0$).

[Case A]: When $\eta_f(\boldsymbol{\beta}^*) = 0$, meaning that the second-stage ("recourse") problem is feasible for all $\omega \in \Omega(\boldsymbol{\beta}^*)$, We then solve the second subproblem (SP2) to assess the worst-case performance of β^* by identifying the worst-case realization ω_s^* and its corresponding recourse cost $\eta_s(\beta^*)$.

$$\text{(SP2)} \quad \eta_s(\beta^*) = \max_{\omega \in \Omega(\beta^*)} \min \left\{ \frac{1}{N} \sum_{i=1}^{N} s_{2,i} : \text{Equation ((9).a)-((9).h))} \right\} \tag{26}$$

(SP2) can be addressed by reformulating the minimization problem via its KKT conditions or equivalently through its dual problem. In both computational approaches, the optimal dual variables, denoted by π^*, correspond to an extreme point of Π. From the lower-level linear program in the bilevel reformulation, denoted as $\mathbf{LP}(\beta, U)$, it follows that

$$\begin{aligned} \eta_s(\beta^*) = \mathbf{1}^\top \pi_0^* + (d - \mathbf{B}_1(\beta^{+*} - \beta^{-*}))^\top \pi_5^* - t_{2,max}^q \, \pi_6^* - t_{2,max}^\omega \, \pi_7^* - t_{2,max}^\omega t_{2,max}^q \, \pi_{10}^* \\ + \text{LP}(\beta^*, \pi^*). \end{aligned} \tag{27}$$

[Case B]: Conversely, if $\eta_f(\boldsymbol{\beta}^*) > 0$, the optimal solution to (25), denoted by ω_f^*, renders the second-stage ("recourse") problem infeasible. In this situation, we solve the third subproblem **(SP3)**, which corresponds to the dual of the second-stage ("recourse") problem evaluated at ω_f^*.

$$\begin{aligned} \text{(SP3)} \quad \max \Big\{ \mathbf{1}^\top \pi_0 + (d - \mathbf{B}_1(\beta^{+*} - \beta^{-*}))^\top \pi_5 - t_{2,max}^q \, \pi_6 - t_{2,max}^\omega \, \pi_7 - t_{2,max}^\omega t_{2,max}^q \, \pi_{10} \\ + a_g(\Sigma_0^{-\frac{1}{2}} \omega_f^*)^\top \pi_3 + a_g(-\Sigma_0^{-\frac{1}{2}} \omega_f^*)^\top \pi_4 - (\mathbf{E}\omega_f^*)^\top \pi_5 : \pi \in \Pi \Big\} \end{aligned} \tag{28}$$

Note that **(SP3)** is unbounded with respect to (β^*, ω_f^*), its solution identifies an extreme ray of Π, denoted by γ^*, along which the objective value diverges to infinity. By convention, the corresponding worst-case recourse cost $\eta_s(\beta^*)$ is set to $+\infty$.

F Benders C&CG Algorithm

When the two-stage robust optimization (2-Stg RO) problem has a decision-independent uncertainty (DIU) set, classical Column-and-Constraint Generation (C&CG) iteratively solves a master problem by adding one recourse problem per identified worst-case scenario. As shown in [55], for 2-Stg RO with decision-dependent uncertainty (DDU) and a single-level reformulation (Appendix D.3), this strategy can be extended using a parametric framework.

The resulting **Benders C&CG algorithm** dynamically generates worst-case scenarios and the corresponding dual-based optimality or feasibility cuts, refining the master problem over iterations. Below are the algorithmic steps:

1. **Initialization**: Set lower bound LB $= -\infty$, upper bound UB $= +\infty$, iteration index $k = 1$, cut sets $\hat{\mathcal{P}}_\Pi, \hat{\mathcal{R}}_\Pi = \emptyset$, and choose a convergence tolerance $\epsilon > 0$.
2. **Master Problem (MP)**: Solve the master problem in (23) (Appendix D.4) to obtain candidate solution (v_1^k, η^k). Set LB $= \frac{1}{N} \sum_{i=1}^{N} s_{1,i}^k + \eta^k$.
3. **Subproblem 1 (SP1)**: For given β^k, solve subproblem (SP1) in (25) (Appendix E) to compute $\eta_f(\beta^k)$ and corresponding scenario ω_f^k.
4. **Cut Generation**:
   - **(Case A)**: If $\eta_f(\beta^k) = 0$, solve subproblem (SP2) in (26) to obtain $\eta_s(\beta^k)$, scenario ω_s^k, optimal dual solution π^k. Update $\hat{\mathcal{P}}_\Pi \leftarrow \hat{\mathcal{P}}_\Pi \cup \{\pi^k\}$. Add optimality cuts from ((23).b)–((23).c).
   - **(Case B)**: If $\eta_f(\beta^k) > 0$, solve subproblem (SP3) in (28) to obtain extreme ray γ^k, and set $\eta_s(\beta^k) = +\infty$. Update $\hat{\mathcal{R}}_\Pi \leftarrow \hat{\mathcal{R}}_\Pi \cup \{\gamma^k\}$. Add feasibility cuts from ((23).d)–((23).e).
5. **Upper Bound Update**: Set UB$^k = \frac{1}{N} \sum_{i=1}^{N} s_{1,i}^k + \eta_s(\beta^k)$, and update UB $= \min\{\text{UB}, \text{UB}^k\}$.
6. **Convergence Check**: If UB $-$ LB $\leq \epsilon$, terminate and return β^k as the optimal first-stage solution. Otherwise, increment $k \leftarrow k + 1$ and return to Step 2.

G Details of the Numerical Experiments

We evaluate our two-stage robust strategic classification framework with decision-dependent uncertain (DDU) manipulation costs, as described in Equation (13).

We generate a synthetic dataset with two-dimensional features $x \in \mathbb{R}^2$, sampled uniformly from $[-10, 10]$. Labels are assigned according to the linear boundary $x_1 + 5x_2 = 2$; specifically, $y = 1$ if $x_1 + 5x_2 > 2$, and $y = -1$ otherwise. To model noisy environments, labels are flipped according to *feature-dependent noise* [45, 57] with probability

$$\mathbb{P}_{\text{noise}}(x) = 0.5 \exp\left(-\left(\tfrac{x_1+5x_2}{5}\right)^2\right),$$

so that points closer to the boundary are more likely to be mislabeled. The dataset consists of 5000 points. To manage computational complexity, training uses mini-batch sampling with a batch size 100. All results are averaged over 25 independent training runs with different random datasets.

Agents receiving negative classification outcomes may strategically manipulate their features to achieve a positive outcome, gaining benefit $u = 1$ in both stages. The manipulation cost is defined via the ℓ_2-norm with a non-decreasing function $\phi(r) = 0.1r$.

First-stage cost. The cost matrix is fixed and given by $\Sigma_0 = \text{diag}(2, 5)$, where manipulation of feature x_2 is 2.5 times more expensive than manipulating x_1. The corresponding cost function is:

$$c(x, \hat{x}) = \phi\big(\|\Sigma^{1/2}(\hat{x} - x)\|_2\big) = 0.1\|\Sigma_0^{1/2}(\hat{x} - x)\|_2.$$

Decision-dependent uncertainty. To construct the decision-dependent uncertainty set, we fix the interaction matrix $\mathbf{F}$ to be constant and independent of the first-stage decision β. As discussed in Section 2, this choice ensures that the first-stage decision β only influences the *bounds* of the uncertainty set via $\mathbf{G}\beta$, rather than altering the relative interaction between ω_1 and ω_2. The resulting decision-dependent uncertainty set is therefore $\Omega(\beta) = \{\omega \in \mathbb{R}^2_+ : \mathbf{F}\omega \leq \mathbf{h} + \mathbf{G}\beta\}$, with

$$\mathbf{F} = \begin{bmatrix} 1 & 2 \\ 3 & 1 \end{bmatrix}, \quad \mathbf{h} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} 5 & 2 \\ 1 & 5 \end{bmatrix}.$$

The diagonal entries of $\mathbf{G}$ are chosen larger than the off-diagonal entries, reflecting that β_i primarily governs the cost of manipulating its corresponding feature x_i, rather than x_j with $i \neq j$. Finally, we set $\mathbf{h} = (1, 1)^\top$ to capture average manipulation cost, which in the school admission analogy corresponds to the average expense of accessing cheating resources (e.g., buying leaked exams).

For instance, $\omega_1 + 2\omega_2 \leq 1 + 5\beta_1 + 2\beta_2$ and $3\omega_1 + \omega_2 \leq 1 + \beta_1 + 5\beta_2$, where the right-hand sides directly scale with β. Thus, higher weights on particular features expand the feasible set of manipulation costs, reflecting that strategic agents can more easily exploit heavily weighted features.

Second-stage cost. The second-stage classifier faces the cost $c(x, \hat{x}) = 0.1\|\Sigma(\omega)^{1/2}(\hat{x} - x)\|_2$, where $\Sigma(\omega) = \text{diag}(g(\omega_1), g(\omega_2)) \cdot \Sigma, \quad g(\omega) = \frac{1}{\omega^2}$. Smaller ω_i increase $g(\omega_i)$, raising the manipulation cost of feature i. Together with $\Omega(\beta)$, if $\beta_i > 0$ (feature valued positively by the classifier), the greater the β_i, the ω_i has a larger feasible upper bound, making it cheaper to manipulate when $\omega_i > 1$. Conversely, if $\beta_i < 0$, the greater the β_i in the negative, the ω_i tends to have a smaller upper bound, increasing the manipulation cost.

First-stage training procedures: (i) **Decision-dependent aware classifier (β^{DD}),** obtained using our two-stage robust optimization framework with DDU costs. (ii) **Decision-independent (unaware) classifier (β^{DI}),** trained using mini-batch gradient descent (batch size 100) to minimize the strategic hinge loss of Rosenfeld and Rosenfeld [41] under fixed cost Σ_0. Note that convexity is essential for the gradient descent method. Therefore, we include a regularization term, specifically $R_{\Sigma_0}(\beta) + \lambda_{reg} u_* \|\beta\|_{*,\Sigma_0}$, since the cost-aware strategic hinge loss $\ell_{\Sigma,\text{s-hinge}}$ in (4) is generally non-convex. By Proposition 4.3 of Rosenfeld and Rosenfeld [41], adding this regularizer and choosing $\lambda_{\text{reg}} \geq \mathbb{P}_{P_{XY}}(Y = 1)$, guarantees convexity.

Evaluation Protocol. We assess performance using the following metrics: (i) Average total 0–1 loss across both stages, (ii) Per-stage average 0–1 loss, (iii) Average qualified and unqualified (default-accepted), (iv) Total manipulations, and (v) Qualified and Unqualified manipulations accepted. Reported error bars correspond to the standard error of the mean, computed as the sample standard deviation across repeated runs divided by the square root of the number of repetitions.

For testing, we generate 25 instances per run, each consisting of 100 new test points sampled from the 5000-point dataset. In the second stage, the true ω^k is drawn from $\Omega(\beta^{\text{DD}})$ or $\Omega(\beta^{\text{DI}})$ depending on the model. The second-stage optimal classifiers β'^{DD} (aware) and β'^{DI} (unaware) are both obtained using mini-batch gradient

descent on the strategic hinge loss in (4) with extra regularizer for convexity as discussed before, with cost matrix $\Sigma(\omega^k)$. Each model is trained across 10 replicas and averaged for stability. Final performance of the second stage is averaged over 10 realizations of ω, with 25 test instances per realization.

In our experiment, we are not restricting the value of β and β', meaning we don't have any of $\mathbf{A}\beta \geq \mathbf{b}$, and $\mathbf{B_2}\beta' \geq \mathbf{d} - \mathbf{B_2}\beta - \mathbf{E}\omega$.

We implemented the Benders C&CG algorithm using Python and GurobiPy. Results show that the decision-dependent (DD) classifier alters manipulation costs in the second stage in a way that improves accuracy compared to its decision-independent (DI) counterpart.

As reported in Table 1, the second-stage average total 0–1 loss for the DD classifier is significantly lower than that of the DI baseline. This stems from the DD classifier's design, which explicitly considers how first-stage decisions affect later manipulation costs. Specifically, Concretely, the average second-stage 0–1 loss under DD is 5.22, compared to 43.92 for DI. Similarly, the average total manipulations are 1.42 under DD and 43.66 under DI. Among manipulated agents, the DD classifier accepts only 0.6 qualified and 0.82 unqualified individuals, versus 8.53 and 35.11 under DI.

This reduction occurs because the DD classifier anticipates the effect of its decisions, making manipulation costlier, thereby making manipulation more difficult. Figure 1 illustrates this: $\Omega(\beta^{\text{DI}})$ contains much larger values of ω than $\Omega(\beta^{\text{DD}})$, reflecting cheaper manipulation under DI. Moreover, Figure 2 shows that $\omega \in \Omega(\beta^{\text{DD}})$ is always strictly less than 1. Thus, it both limits the reduction in second-stage costs and ensures manipulation remains consistently expensive.

In contrast, Table 1 also reveals that the DD classifier performs worse in the first stage. Its average 0–1 loss is 25.36, compared to 22.80 for DI classifier. This trade-off arises because the DD classifier deliberately sacrifices first-stage accuracy to mitigate second-stage manipulation, whereas the DI classifier—being unaware—optimizes solely for immediate outcomes. For instance, under DD, the average total manipulations reach 25.24 (22.44 unqualified and 2.8 qualified agents), whereas DI yields 22.80 manipulations (18.92 unqualified and 3.80 qualified). This caused the dependency-aware model to consistently achieves lower total averages across all stages compared to the dependency-unaware model. Specifically, the overall average total 0–1 loss under dependency-aware is 30.58, substantially lower than 66.72 for the dependency-unaware model. Likewise, the overall average total manipulations are reduced to 26.66 under dependency-aware, compared to66.46 for the dependency-unaware. These results confirm that by accounting for decision-dependent manipulation costs, the dependency-aware model effectively reduces both classification loss and the extent of manipulations across stages.

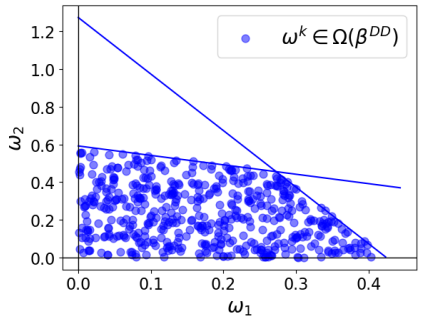

Figure 2: Decision-dependent set $\Omega(\beta^{DD})$