Appendix

A Organization of the Appendix

The Appendix is organized as follows.

- Section A provides the organization of the appendix.
- Section B provides the proof of Theorem 1.
- Section C includes notations and proofs for the discussion in section 3.3.
- Section D includes the proofs and derivations for section 4.
- Section E presents additional related works.
- Section F shows additional experimental details and results, including basic information on each dataset and the computing infrastructure.

B Proof of Theorem 1

In this section, we provide the proof of Theorem 1. To simplify our discussion, we focus on the unconstrained best response, i.e. the case in which $\mathsf{F} = \mathsf{A}$. The proofs for the other two types of best response ($\mathsf{F} = \mathsf{M}$, $\mathsf{F} = \mathsf{I}$) follow the same arguments except that the inverse of $(S^{-1})_{\mathsf{I}}$ does not equal to S, but equals to $((S^{-1})_{\mathsf{I}})^{-1}$.

We first prove two lemmas that allow us to reformulate the best response as an optimization problem. The first states that the decision subject's goal is to maximize their utility, but they are unwilling to pay a cost greater than 2:

Lemma 1 (Decision Subject's Best-Response Function)**.** *Given a classifier $h : \mathcal{X} \to \{-1, +1\}$, a cost function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and a set of realizable feature vectors $\mathcal{X}^\dagger \subseteq \mathcal{X}$, the* best response *of a decision subject with features $x \in \mathcal{X}^\dagger$ is the solution to the following optimization program:*

$$\max_{x' \in \mathcal{X}^\dagger} \quad U(x, x') \quad \text{s.t.} \quad c(x, x') \leq 2$$

Proof. Since the classifier in our game outputs a binary decision (-1 or $+1$), decision subjects only have an incentive to change their features from x to x' when $c(x, x') \leq 2$. To see this, notice that an decision subject originally classified as -1 receives a default utility of $U(x, x) = f(x) - 0 = -1$ by presenting her original features x. Since costs are always non-negative, she can only hope to increase her utility by flipping the classifier's decision. If she changes her features to some x' such that $f(x') = +1$, then the new utility will be given by

$$U(x, x') = f(x') - c(x, x') = 1 - c(x, x')$$

Hence the decision subject will only change her features if $1 - c(x, x') \geq f(x) = -1$, or $c(x, x') \leq 2$. □

The next lemma turns the above maximization program into a minimization program, in which the decision subject seeks the minimum-cost change in x that crosses the decision boundary. If the cost exceeds 2, which is the maximum possible gain from adaptation, they would rather not modify any features.

Lemma 2. *Let $x^\star$ be an optimal solution to the following optimization problem:*

$$\begin{aligned} x^\star = \operatorname*{arg\,min}_{x' \in \mathcal{X}^*_{\mathsf{A}}(x)} & \quad c(x, x') \\ \text{s.t.} & \quad \operatorname{sign}(w^\intercal x') = 1 \end{aligned}$$

If no solution is returned, we say an $x^\star$ such that $c(x, x^\star) = \infty$ is returned. Define $\Delta(x)$ as follows:

$$\Delta(x) = \begin{cases} x^\star, & \text{if } c(x, x^\star) \leq 2 \\ x, & \text{otherwise} \end{cases}$$

Then $\Delta(x)$ *is an optimal solution to the optimization problem in Lemma 1.*

Proof. Recall that the utility function of the decision subject is $U(x, x') = f(x') - c(x, x')$, and that, by Lemma 1, they will only modify their features if the utility increases, i.e. if they achieve $f(x') = +1$ and while incurring cost $c(x, x') \leq 2$.

Consider two cases for $x' \neq x$:

1. When $c(x, x') > 2$, there are no feasible points for the optimization problem of Lemma 1.
2. When $c(x, x') \leq 2$, we only need to consider those feature vectors x' that satisfy $f(x') = 1$, because if $f(x') = -1$, the decision subject with features x would prefer not to change anything. Since maximizing $U(x, x') = f(x') - c(x, x')$ is equivalent to minimizing $c(x, x')$ if $f(x') = 1$, we conclude that when $c(x, x') \leq 2$, the optimum of the program of Lemma 1 is the same as the optimum of the program in Lemma 2.

□

Lemma 2 enables us to re-formulate the objective function as follows. Recall that $c(x, x') = \sqrt{(x_{\mathsf{A}} - {x_{\mathsf{A}}}')^\intercal S^{-1} (x_{\mathsf{A}} - {x_{\mathsf{A}}}')}$ where S^{-1} is symmetric positive definite. Thus S^{-1} has the following diagonalized form, in which Q is an orthogonal matrix and Λ^{-1} is a diagonal matrix:

$$S^{-1} = Q^\intercal \Lambda^{-1} Q = (\Lambda^{-\frac{1}{2}} Q)^\intercal (\Lambda^{-\frac{1}{2}} Q)$$

With this, we can re-write the cost function as

$$\begin{aligned} c(x, x') &= \sqrt{(x_{\mathsf{A}} - {x_{\mathsf{A}}}')^\intercal S^{-1} (x_{\mathsf{A}} - {x_{\mathsf{A}}}')} \\ &= \sqrt{(x_{\mathsf{A}} - {x_{\mathsf{A}}}')^\intercal (\Lambda^{-\frac{1}{2}} Q)^\intercal (\Lambda^{-\frac{1}{2}} Q)(x_{\mathsf{A}} - {x_{\mathsf{A}}}')} \\ &= \sqrt{(\Lambda^{-\frac{1}{2}} Q(x_{\mathsf{A}} - {x_{\mathsf{A}}}'))^\intercal (\Lambda^{-\frac{1}{2}} Q(x_{\mathsf{A}} - {x_{\mathsf{A}}}'))} \\ &= \|\Lambda^{-\frac{1}{2}} Q(x_{\mathsf{A}} - {x_{\mathsf{A}}}')\|_2 \end{aligned}$$

Meanwhile, the constraint in Lemma 2 can be written

$$\begin{aligned} \text{sign}(w \cdot x') &= \text{sign}(w_{\mathsf{A}} \cdot {x_{\mathsf{A}}}' + w_{\mathsf{IM}} \cdot x_{\mathsf{IM}}) \\ &= \text{sign}(w_{\mathsf{A}} \cdot {x_{\mathsf{A}}}' - (-w_{\mathsf{IM}} \cdot x_{\mathsf{IM}})) = 1 \end{aligned}$$

Hence the optimization problem can be reformulated as

$$\min_{{x_{\mathsf{A}}}' \in \mathcal{X}_A^*} \; \|(\Lambda^{-\frac{1}{2}} Q(x_{\mathsf{A}} - {x_{\mathsf{A}}}'))\|_2 \tag{7}$$

$$\text{s.t.} \;\; \text{sign}(w_{\mathsf{A}} \cdot {x_{\mathsf{A}}}' - (-w_{\mathsf{IM}} \cdot x_{\mathsf{IM}})) = 1 \tag{8}$$

The above optimization problem can be further simplified by getting rid of the sign$(\cdot)$:

Lemma 3. *If* ${x_{\mathsf{A}}}^{\mp}$ *is an optimal solution to Eq. (7) under constraint Eq. (8), then it must satisfy* $w_{\mathsf{A}} \cdot {x_{\mathsf{A}}}^{\mp} - (-w_{\mathsf{IM}} \cdot x_{\mathsf{IM}}) = 0$.

Proof. We prove by contradiction. Let $x_{\mathsf{A}}^{\mp}$ is an optimal solution to Eq. (7) and suppose towards contraction that $w_{\mathsf{A}} x_{\mathsf{A}}^{\mp} > -w_{\mathsf{IM}} \cdot x_{\mathsf{IM}}$. Since the original feature vector x was classified as -1, we have

$$w_{\mathsf{A}} \cdot x_{\mathsf{A}}^{\mp} > -w_{\mathsf{IM}} \cdot x_{\mathsf{IM}}, \quad w_{\mathsf{A}} \cdot x_{\mathsf{A}} < -w_{\mathsf{IM}} \cdot x_{\mathsf{IM}}$$

By the continuity properties of linear vector space, there exists $\mu \in (0, 1)$ such that:

$$w_{\mathsf{A}} \left(\mu \cdot {x_{\mathsf{A}}}^{\mp} + (1 - \mu) x_{\mathsf{A}} \right) = -w_{\mathsf{IM}} \cdot x_{\mathsf{IM}}$$

Let ${x_{\mathsf{A}}}'' = \mu \cdot {x_{\mathsf{A}}}^{\mp} + (1 - \mu) x_{\mathsf{A}}$. Then $\mathrm{sign}(w_{\mathsf{A}} {x_{\mathsf{A}}}'' - (-w_{\mathsf{IM}} \cdot x_{\mathsf{IM}})) = 1$, i.e., x''_{A} also satisfies the constraint. Since ${x_{\mathsf{A}}}^{\mp}$ is an optimum of Eq. (7), we have

$$\|\Sigma^{-\frac{1}{2}} Q({x_{\mathsf{A}}}^{\mp} - x_{\mathsf{A}})\| \leq \|\Sigma^{-\frac{1}{2}} Q({x_{\mathsf{A}}}'' - x_{\mathsf{A}})\|$$

However, we also have:

$$\begin{aligned}
\|\Sigma^{-\frac{1}{2}} Q({x_{\mathsf{A}}}'' - x_{\mathsf{A}})\| &= \|\Sigma^{-\frac{1}{2}} Q(\mu \cdot {x_{\mathsf{A}}}^{\mp} + (1 - \mu) x_{\mathsf{A}} - x_{\mathsf{A}})\| \\
&= \|\Sigma^{-\frac{1}{2}} Q(\mu \cdot ({x_{\mathsf{A}}}^{\mp} - x_{\mathsf{A}}))\| \\
&= \mu \|\Sigma^{-\frac{1}{2}} Q({x_{\mathsf{A}}}^{\mp} - x_{\mathsf{A}})\| \\
&< \|\Sigma^{-\frac{1}{2}} Q({x_{\mathsf{A}}}^{\mp} - x_{\mathsf{A}})\|
\end{aligned}$$

contradicting our assumption that $x^{\mp}_{\mathsf{A}}$ is optimal. Therefore $x^{\mp}_{\mathsf{A}}$ must satisfy $w_{\mathsf{A}} x^{\mp}_{\mathsf{A}} = -w_{\mathsf{IM}} \cdot x_{\mathsf{IM}}$. ☐

As a result of Lemma 3, we can replace the constraint in Eq. (7) with its corresponding equality constraint without changing the optimal solution.[2] The decision subject's best-response program from Lemma 1 is therefore equivalent to

$$\min_{{x_{\mathsf{A}}}' \in \mathcal{X}^*_A} \|(\Lambda^{-\frac{1}{2}} Q(x_{\mathsf{A}} - {x_{\mathsf{A}}}'))\|_2 \tag{9}$$

$$\text{s.t.} \quad w_{\mathsf{A}} \cdot {x_{\mathsf{A}}}' - (-w_{\mathsf{IM}} \cdot x_{\mathsf{IM}}) = 0 \tag{10}$$

The following lemma gives us a closed-form solution for the above optimization problem:

Lemma 4. *The optimal solution to the optimization problem defined in Eq.* (9) *and Eq.* (10)

has the following closed form:

$${x_{\mathsf{A}}}^{\mp} = x_{\mathsf{A}} - \frac{w^{\mathsf{T}} x}{{w_{\mathsf{A}}}^{\mathsf{T}} S w_{\mathsf{A}}} S w_{\mathsf{A}}.$$

Proof. Notice that the above program has the form

$$\begin{aligned}
\min_{{x_{\mathsf{A}}}' \in {x_{\mathsf{A}}}^*} \quad & \|A {x_{\mathsf{A}}}' - b\|_2 \\
\text{s.t.} \quad & C {x_{\mathsf{A}}}' = d
\end{aligned}$$

where $A = \Lambda^{-\frac{1}{2}} Q$, $b = \Lambda^{-\frac{1}{2}} Q x_{\mathsf{A}}$, $C = {w_{\mathsf{A}}}^{\mathsf{T}}$, and $d = -{w_{\mathsf{IM}}}^{\mathsf{T}} x_{\mathsf{IM}}$. Note the following useful equalities:

$$\begin{aligned}
A^{\mathsf{T}} A &= (\Lambda^{-\frac{1}{2}} Q)^{\mathsf{T}} \Lambda^{-\frac{1}{2}} Q = S^{-1} \\
(A^{\mathsf{T}} A)^{-1} &= S \\
A^{\mathsf{T}} b &= (\Lambda^{-\frac{1}{2}} Q)^{\mathsf{T}} \Lambda^{-\frac{1}{2}} Q x_{\mathsf{A}} = S^{-1} x_{\mathsf{A}}
\end{aligned}$$

---

[2] A similar argument was made by Haghtalab et al. (2020) but here we provide a proof for a more general case, where the objective function is to minimize a weighted norm instead of simply $\|x_{\mathsf{A}} - {x_{\mathsf{A}}}'\|_2$.

The above is a norm minimization problem with equality constraints, whose optimum ${x_{\mathsf{A}}}^{\mp}$ has the following closed form Boyd & Vandenberghe (2004):

$$\begin{aligned}
{x_{\mathsf{A}}}^{\mp} &= (A^{\mathsf{T}}A)^{-1}\left(A^{\mathsf{T}}b - C^{\mathsf{T}}(C(A^{\mathsf{T}}A)^{-1}C^{\mathsf{T}})^{-1}(C(A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}b - d)\right)\\
&= S\left(S^{-1}x_{\mathsf{A}} - w_{\mathsf{A}}({w_{\mathsf{A}}}^{\mathsf{T}}Sw_{\mathsf{A}})^{-1}({w_{\mathsf{A}}}^{\mathsf{T}}S(S^{-1}x_{\mathsf{A}}) - (-{w_{\mathsf{IM}}}^{\mathsf{T}}x_{\mathsf{IM}}))\right)\\
&= x_{\mathsf{A}} - S\left(w_{\mathsf{A}}({w_{\mathsf{A}}}^{\mathsf{T}}Sw_{\mathsf{A}})^{-1}({w_{\mathsf{A}}}^{\mathsf{T}}x_{\mathsf{A}} + {w_{\mathsf{IM}}}^{\mathsf{T}}x_{\mathsf{IM}})\right)\\
&= x_{\mathsf{A}} - \frac{w^{\mathsf{T}}x}{{w_{\mathsf{A}}}^{\mathsf{T}}Sw_{\mathsf{A}}}Sw_{\mathsf{A}}
\end{aligned}$$

□

We can now compute the cost incurred by an individual with features x who plays their best response $x^{\mp}$:

$$\begin{aligned}
c(x, x\mp) &= \sqrt{(x_{\mathsf{A}} - {x_{\mathsf{A}}}^{\mp})^{\mathsf{T}}S^{-1}(x_{\mathsf{A}} - {x_{\mathsf{A}}}^{\mp})}\\
&= \sqrt{\left(\frac{w^{\mathsf{T}}x}{{w_{\mathsf{A}}}^{\mathsf{T}}Sw_{\mathsf{A}}}Sw_{\mathsf{A}}\right)^{\mathsf{T}}S^{-1}\left(\frac{w^{\mathsf{T}}x}{{w_{\mathsf{A}}}^{\mathsf{T}}Sw_{\mathsf{A}}}Sw_{\mathsf{A}}\right)}\\
&= \frac{|w^{\mathsf{T}}x|}{\sqrt{{w_{\mathsf{A}}}^{\mathsf{T}}Sw_{\mathsf{A}}}}
\end{aligned}$$

Hence an decision subject who was classified as -1 with feature vector x has the unconstrained best response

$$\Delta(x) = \begin{cases} x, & \text{if } \frac{|w^{\mathsf{T}}x|}{\sqrt{{w_{\mathsf{A}}}^{\mathsf{T}}Sw_{\mathsf{A}}}} \geq 2\\ \left[x_{\mathsf{A}} - \frac{w^{\mathsf{T}}x}{{w_{\mathsf{A}}}^{\mathsf{T}}Sw_{\mathsf{A}}}Sw_{\mathsf{A}} \mid x_{\mathsf{IM}}\right], & \text{otherwise}\end{cases}$$

which completes the proof of Theorem 1.

C Proofs of Propositions in Section 3.3

Notation. We make use of the following additional notation:

- $v^{(i)}$ denotes the i-th element of a vector v
- For any $\mathsf{F} \in \{\mathsf{A}, \mathsf{I}, \mathsf{M}\}$, $\Delta^{\mathsf{F}} \in \mathbb{R}^{d_{\mathsf{F}}}$ denotes the vector containing only features of type F within the best response $\Delta(x)$.
- $\mathbf{0}$ denotes the vector whose elements are all 0
- $A \succ B$ indicates that matrix $A - B$ is positive definite
- e_i denotes the vector containing 1 in its i-th component and 0 elsewhere

C.1 Proof of Proposition 1

Proof. Let $w_{\mathsf{M}}^{(m)} \neq 0$, and consider an decision subject with original features x who was classified as -1. By Theorem 1, the actionable sub-vector of x's unconstrained best response is

$$\Delta^{\mathsf{A}}(x) = \frac{w^{\mathsf{T}}x}{{w_{\mathsf{A}}}^{\mathsf{T}}Sw_{\mathsf{A}}}S \cdot w_{\mathsf{A}} = \frac{w^{\mathsf{T}}x}{{w_{\mathsf{A}}}^{\mathsf{T}}Sw_{\mathsf{A}}}\begin{bmatrix} S_{\mathsf{I}} & 0\\ 0 & S_{\mathsf{M}}\end{bmatrix}\begin{bmatrix} w_{\mathsf{I}}\\ w_{\mathsf{M}}\end{bmatrix} = \frac{w^{\mathsf{T}}x}{{w_{\mathsf{A}}}^{\mathsf{T}}Sw_{\mathsf{A}}}\begin{bmatrix} S_{\mathsf{I}} \cdot w_{\mathsf{I}}\\ S_{\mathsf{M}} \cdot w_{\mathsf{M}}\end{bmatrix}$$

And in particular,

$$\Delta^{\mathsf{M}}(x) = \frac{w^{\mathsf{T}}x}{{w_{\mathsf{A}}}^{\mathsf{T}}Sw_{\mathsf{A}}}S_{\mathsf{M}} \cdot w_{\mathsf{M}}$$

Since x was initially classified as -1, we have $w^{\mathsf{T}}x < 0$, which means $\frac{w^{\mathsf{T}}x}{w_{\mathsf{A}}Sw_{\mathsf{A}}} \neq 0$. For convenience, let $c = \frac{w^{\mathsf{T}}x}{w_{\mathsf{A}}Sw_{\mathsf{A}}}$. We have

$$\Delta^{\mathsf{M}}(x) - x_{\mathsf{M}} = cS_{\mathsf{M}}w_{\mathsf{M}} - x_{\mathsf{M}} = S_{\mathsf{M}}(cw_{\mathsf{M}} - {S_{\mathsf{M}}}^{-1}x_{\mathsf{M}})$$

Now examine the following:

$$\begin{aligned}(cw_{\mathsf{M}} - {S_{\mathsf{M}}}^{-1}x_{\mathsf{M}})^{(m)} &= cw_{\mathsf{M}}^{(m)} - (S_{\mathsf{M}}^{-1}x_{\mathsf{M}})^{(m)} \\ &= cw_{\mathsf{M}}^{(m)} - \sum_{i=1}^{d_{\mathsf{M}}}(S_{\mathsf{M}}^{-1})^{(im)}{x_{\mathsf{M}}}^{(m)}\end{aligned}$$

Recall that $cw_{\mathsf{M}}^{(m)} \neq 0$. Hence if $\sum_{i=1}^{d_{\mathsf{M}}}(S_{\mathsf{M}}^{-1})^{(im)} = 0$, or if

$$x_{\mathsf{M}}^{(m)} \neq \frac{cw_{\mathsf{M}}^{(m)}}{\sum_{i=1}^{d_{\mathsf{M}}}(S_{\mathsf{M}}^{-1})^{(im)}},$$

then $(cw_{\mathsf{M}} - {S_{\mathsf{M}}}^{-1}x_{\mathsf{M}})^{(m)} \neq 0$, and therefore $cw_{\mathsf{M}} - S_{\mathsf{M}}^{-1}x_{\mathsf{M}} \neq \mathbf{0}$. Since S_{M} is positive definite, it has full rank, which implies

$$\Delta^{\mathsf{M}}(x) - x_{\mathsf{M}} = S_{\mathsf{M}}(cw_{\mathsf{M}} - S_{\mathsf{M}}^{-1}x_{\mathsf{M}}) \neq \mathbf{0}$$

as required. With this, we have shown that when there exists a manipulated feature $x^{(m)}$ whose corresponding coefficient ${w_{\mathsf{A}}}^{(m)} \neq 0$, the classifier is vulnerable to changes in the manipulated features by the vast majority of decision subjects. ☐

C.2 Proof of Proposition 2

Proof. Consider a decision subject with features x such that $h(x) = -1$. Suppose x can flip this classification result by performing the improving best response $\Delta_{\mathsf{I}}(x)$, which implies that the cost of that action is no greater than 2 for this decision subject. We therefore have:

$$2 \geq c(x, \Delta_{\mathsf{I}}(x)) = \frac{|w^{\mathsf{T}}x|}{\sqrt{{w_{\mathsf{I}}}^{\mathsf{T}}S_{\mathsf{I}}w_{\mathsf{I}}}} > \frac{|w^{\mathsf{T}}x|}{\sqrt{{w_{\mathsf{I}}}^{\mathsf{T}}S_{\mathsf{I}}w_{\mathsf{I}} + {w_{\mathsf{M}}}^{\mathsf{T}}S_{\mathsf{M}}w_{\mathsf{M}}}} = \frac{|w^{\mathsf{T}}x|}{\sqrt{{w_{\mathsf{A}}}^{\mathsf{T}}Sw_{\mathsf{A}}}} = c(x, \Delta(x))$$

where the strict inequality is due to the fact that $S_{\mathsf{M}} \succ 0$ and $w_{\mathsf{M}} \neq \mathbf{0}$. As we have shown that $c(x, \Delta(x)) < 2$, we conclude whenever an decision subject can successfully flip her decision by the improving best response, she can also achieve it by performing the unconstrained best response.

On the other hand, consider the case when the unconstrained best response of a decision subject with features x^* has cost exactly 2:

$$2 = c(x^*, \Delta(x^*)) = \frac{|w^{\mathsf{T}}x^*|}{\sqrt{{w_{\mathsf{A}}}^{\mathsf{T}}Sw_{\mathsf{A}}}} = \frac{|w^{\mathsf{T}}x^*|}{\sqrt{{w_{\mathsf{I}}}^{\mathsf{T}}S_{\mathsf{I}}w_{\mathsf{I}} + {w_{\mathsf{M}}}^{\mathsf{T}}S_{\mathsf{M}}w_{\mathsf{M}}}} < \frac{|w^{\mathsf{T}}x^*|}{\sqrt{{w_{\mathsf{I}}}^{\mathsf{T}}S_{\mathsf{I}}w_{\mathsf{I}}}} = c(x^*, \Delta_{\mathsf{I}}(x^*))$$

where the strict inequality is due to the fact that $S_{\mathsf{M}} \succ 0$ and $w_{\mathsf{M}} \neq \mathbf{0}$. As we have shown that $c(x^*, \Delta_{\mathsf{I}}(x^*)) > 2$, we conclude that while the unconstrained best response is viable for this decision subject, the improving best response is not. ☐

C.3 Proof of Proposition 4

Proof. Let the cost covariance matrices for groups Φ and Ψ be

$$S_{\Psi}^{-1} = \begin{bmatrix} S_{\mathsf{I}}^{-1} & 0 \\ 0 & S_{\mathsf{M},\Phi}^{-1} \end{bmatrix}, \qquad S_{\Phi}^{-1} = \begin{bmatrix} S_{\mathsf{I}}^{-1} & 0 \\ 0 & S_{\mathsf{M},\Psi}^{-1} \end{bmatrix}$$

Here, we see that both groups have the same cost of changing improvable features, as represented in the cost submatrix ${S_{\mathsf{I}}}^{-1}$. However, the cost of manipulation for group Φ is higher than that of group Ψ, namely $S_{\mathsf{M},\Phi}^{-1} \succ S_{\mathsf{M},\Psi}^{-1}$.

We are now equipped to compare the costs for the two decision subjects:

$$\begin{aligned} c(x_{\phi}, \Delta(x_{\phi})) &= \frac{|w^{\mathsf{T}} x_{\phi}|}{\sqrt{{w_{\mathsf{A}}}^{\mathsf{T}} S_{\Phi} w_{\mathsf{A}}}} = \frac{|w^{\mathsf{T}} x|}{\sqrt{{w_{\mathsf{I}}}^{\mathsf{T}} S_{\mathsf{I}} w_{\mathsf{I}} + {w_{\mathsf{M}}}^{\mathsf{T}} \cdot S_{\mathsf{M},\Phi} \cdot w_{\mathsf{M}}}} \\ c(x_{\psi}, \Delta(x_{\psi})) &= \frac{|w^{\mathsf{T}} x_{\psi}|}{\sqrt{{w_{\mathsf{A}}}^{\mathsf{T}} S_{\Psi} w_{\mathsf{A}}}} = \frac{|w^{\mathsf{T}} x|}{\sqrt{{w_{\mathsf{I}}}^{\mathsf{T}} S_{\mathsf{I}} w_{\mathsf{I}} + {w_{\mathsf{M}}}^{\mathsf{T}} \cdot S_{\mathsf{M},\Psi} \cdot w_{\mathsf{M}}}} \end{aligned}$$

Since $S_{\mathsf{M},\Phi}^{-1} \succ S_{\mathsf{M},\Psi}^{-1}$, we have $S_{\mathsf{M},\Phi} \prec S_{\mathsf{M},\Psi}$. And since $w_{\mathsf{M}} \neq \mathbf{0}$, this implies $0 < {w_{\mathsf{M}}}^{\mathsf{T}} S_{\mathsf{M},\Phi} w_{\mathsf{M}} < {w_{\mathsf{M}}}^{\mathsf{T}} \cdot S_{\mathsf{M},\Psi} \cdot w_{\mathsf{M}}$. As a result, $c(x_{\phi}, \Delta(x_{\phi})) > c(x_{\psi}, \Delta(x_{\psi}))$ as required. ☐

C.4 Additional Analysis

Proposition 7 (Correlations between Features May Reduce Cost). *For any cost matrix S^{-1} and any nontrivial classifier h, there exist indices $k, \ell \in [d_{\mathsf{A}}]$ and $\tau \in \mathbb{R}$ such that every feature vector x has lower best-response cost under the cost matrix $\tilde{S}^{-1}$ given by*

$$\tilde{S}_{ij}^{-1} = \tilde{S}_{ji}^{-1} = \begin{cases} S_{ij}^{-1} + \tau, & \text{if } i = k, j = \ell \\ S_{ij}^{-1}, & \text{otherwise} \end{cases}$$

than under S^{-1}; that is, $c_{\tilde{S}^{-1}}(x, \Delta(x)) < c_{S^{-1}}(x, \Delta(x))$ for all x.

Proof. Consider any cost matrix $S^{-1} \in \mathbb{R}^{d_{\mathsf{A}} \times d_{\mathsf{A}}}$ and any nontrivial classifier h (i.e. h does not assign every x the same prediction). Since S^{-1} is positive definite, so is its inverse S, and all of their diagonal entries are positive. And since h is nontrivial, it must contain a nonzero coefficient $w_i \neq 0$. Additionally, let w_j be any other coefficient.

Let $\tilde{S}^{-1} = S^{-1} + \tau(e_i e_j^{\mathsf{T}} + e_j e_i^{\mathsf{T}})$ for some constant $\tau \in \mathbb{R}$ to be set later. We claim that there exists τ such that the best-response adaptation always costs less under $\tilde{S}^{-1}$ than S^{-1}. To do so, we compute the inverse of $\tilde{S}^{-1}$ and invoke the closed-form cost expression given by Theorem 1.

To begin computing the inverse, note that by the Sherman-Morrison-Woodbury formula Golub & Van Loan (2013),

$$\begin{aligned}
\tilde{S} = \left(\tilde{S}^{-1}\right)^{-1} &= S - \tau S \begin{bmatrix} e_i & e_j \end{bmatrix} \left(I + \tau \begin{bmatrix} e_j^\top \\ e_i^\top \end{bmatrix} S \begin{bmatrix} e_i & e_j \end{bmatrix} \right)^{-1} \begin{bmatrix} e_j^\top \\ e_i^\top \end{bmatrix} S && (11) \\
&= S - \tau S \begin{bmatrix} e_i & e_j \end{bmatrix} \left(I + \tau \begin{bmatrix} S_{ij} & S_{jj} \\ S_{ii} & S_{ij} \end{bmatrix} \right)^{-1} \begin{bmatrix} e_j^\top \\ e_i^\top \end{bmatrix} S && (12) \\
&= S - \tau S \begin{bmatrix} e_i & e_j \end{bmatrix} \left[\tau \left(\frac{1}{\tau} I + \begin{bmatrix} S_{ij} & S_{jj} \\ S_{ii} & S_{ij} \end{bmatrix} \right) \right]^{-1} \begin{bmatrix} e_j^\top \\ e_i^\top \end{bmatrix} S && (13) \\
&= S - \tau S \begin{bmatrix} e_i & e_j \end{bmatrix} \tau^{-1} \begin{bmatrix} \frac{1}{\tau} + S_{ij} & S_{jj} \\ S_{ii} & \frac{1}{\tau} + S_{ij} \end{bmatrix}^{-1} \begin{bmatrix} e_j^\top \\ e_i^\top \end{bmatrix} S && (14) \\
&= S - S \begin{bmatrix} e_i & e_j \end{bmatrix} \underbrace{\begin{bmatrix} \frac{1}{\tau} + S_{ij} & S_{jj} \\ S_{ii} & \frac{1}{\tau} + S_{ij} \end{bmatrix}}_{T}^{-1} \begin{bmatrix} e_j^\top \\ e_i^\top \end{bmatrix} S && (15)
\end{aligned}$$

Clearly, we can ensure that T is invertible by setting τ so that $\det(T) \neq 0$. But as the following lemmas show, we can actually say much more: $\det(T)$ can be made either positive or negative, and moreover, both can be accomplished with a choice of $\tau > 0$ or $\tau < 0$. This flexibility in choosing τ will become crucial later.

First, we need the following useful fact about positive definite matrices:

Lemma 5 (Off-diagonal entries of a positive definite matrix)**.** *If $A \in \mathbb{R}^{n \times n}$ is symmetric positive definite, then for all $i, j \in [n]$, $\sqrt{A_{ii} A_{jj}} > |A_{ij}|$.*

Proof. By positive definiteness, we have, for any nonzero $\alpha, \beta \in \mathbb{R}$,

$$(\alpha e_i + \beta e_j)^\top A (\alpha e_i + \beta e_j) = \alpha^2 A_{ii} + \beta^2 A_{jj} + 2\alpha\beta A_{ij} > 0$$

For a choice of $\alpha = -A_{ij}$ and $\beta = A_{ii}$, we have

$$A_{ij}^2 A_{ii} + A_{ii}^2 A_{jj} - 2A_{ij}^2 A_{ii} = A_{ii}(A_{ii} A_{jj} - A_{ij}^2) > 0$$

Since $A_{ii} > 0$, we must have $A_{ii} A_{jj} - A_{ij}^2 > 0$, from which the claim follows. ☐

Now we can characterize the possible settings of τ and $\det(T)$:

Lemma 6 (Possible settings of τ)**.** *There exist $\tau_{\max}, \tau_{\min} > 0$ such that the following hold:*

1. $\det(T) > 0$ *for any $\tau \in \mathbb{R}$ such that $\tau_{\max} \geq |\tau| > 0$.*
2. $\det(T) < 0$ *for any $\tau \in \mathbb{R}$ such that $\tau_{\min} \leq |\tau|$.*

Proof. To prove the first claim, note that having

$$\det(T) = \left(\frac{1}{\tau} + S_{ij}\right)^2 - S_{ii} S_{jj} > 0$$

is equivalent to

$$\left|\frac{1}{\tau} + S_{ij}\right| > \sqrt{S_{ii} S_{jj}}$$

It suffices to choose τ such that

$$\left|\frac{1}{\tau}\right| - |S_{ij}| > \sqrt{S_{ii}S_{jj}}$$
$$\frac{1}{|\tau|} > \sqrt{S_{ii}S_{jj}} + |S_{ij}|$$

So any τ such that $0 < |\tau| < \left(\sqrt{S_{ii}S_{jj}} + |S_{ij}|\right)^{-1}$ results in $\det(T) > 0$. Analogously, for the second claim, a sufficient condition for $\det(T) < 0$ is that

$$\frac{1}{|\tau|} < \sqrt{S_{ii}S_{jj}} - |S_{ij}|$$

By Lemma 5, the right-hand side is positive. Hence it suffices to pick any τ such that

$$|\tau| > \left(\sqrt{S_{ii}S_{jj}} - |S_{ij}|\right)^{-1}.$$

□

With this lemma in place, we can describe the difference between the inverses of S^{-1} and $\tilde{S}^{-1}$. Denote this matrix by $E = S - \tilde{S}$. We show the following:

Lemma 7 (Difference between inverse cost matrices)**.** *The k, ℓ-th entry of E has the following form:*

$$E_{k\ell} = \frac{1}{\det(T)}\left(E'_{k\ell} + \frac{1}{\tau}E''_{k\ell}\right)$$

where $E'_{k\ell}$ and $E''_{k\ell}$ do not depend on τ.

Proof. Assume that τ has been chosen so that $\det(T) \neq 0$, as Lemma 6 showed to be possible. We then have

$$T^{-1} = \frac{1}{\det(T)}\begin{bmatrix} \frac{1}{\tau} + S_{ij} & -S_{jj} \\ -S_{ii} & \frac{1}{\tau} + S_{ij} \end{bmatrix}$$

Thus continuing from equation 15, we have

$$\tilde{S} = S - \frac{1}{\det(T)} S \underbrace{\begin{bmatrix} e_i & e_j \end{bmatrix} \begin{bmatrix} \frac{1}{\tau} + S_{ij} & -S_{jj} \\ -S_{ii} & \frac{1}{\tau} + S_{ij} \end{bmatrix} \begin{bmatrix} e_j^\intercal \\ e_i^\intercal \end{bmatrix}}_{V} S$$

It can be verified that V is a $d_{\mathsf{A}} \times d_{\mathsf{A}}$ matrix whose only nonzero entries are

$$V_{ii} = -S_{jj}, \qquad V_{jj} = -S_{ii}, \qquad V_{ij} = V_{ji} = \frac{1}{\tau} + S_{ij}$$

Next we evaluate the $d_{\mathsf{A}} \times d_{\mathsf{A}}$ matrix SVS. For any $k, \ell \in [d_{\mathsf{A}}]$, we have

$$\begin{aligned}
(SVS)_{k\ell} &= \sum_{i'=1}^{d_{\mathsf{A}}} \sum_{j'=1}^{d_{\mathsf{A}}} S_{ki'}V_{i'j'}S_{j'\ell} \\
&= S_{ki}V_{ii}S_{i\ell} + S_{ki}V_{ij}S_{j\ell} + S_{kj}V_{ji}S_{i\ell} + S_{kj}V_{jj}S_{j\ell} && (V \text{ has four nonzero entries}) \\
&= V_{ii}S_{ki}S_{i\ell} + V_{jj}S_{kj}S_{j\ell} + V_{ij}(S_{ki}S_{j\ell} + S_{kj}S_{i\ell}) && (V_{ij} = V_{ji}) \\
&= -S_{jj}S_{ki}S_{i\ell} - S_{ii}S_{kj}S_{j\ell} + \left(\frac{1}{\tau} + S_{ij}\right)(S_{ki}S_{j\ell} + S_{kj}S_{i\ell}) \\
&= \underbrace{-S_{jj}S_{ki}S_{i\ell} - S_{ii}S_{kj}S_{j\ell} + S_{ij}(S_{ki}S_{j\ell} + S_{kj}S_{i\ell})}_{E'_{k\ell}} + \frac{1}{\tau}\underbrace{(S_{ki}S_{j\ell} + S_{kj}S_{i\ell})}_{E''_{k\ell}}
\end{aligned}$$

which proves the claim. □

We now compute the marginal best-response cost incurred due to the difference between the inverse cost matrices, $E = S - \tilde{S}$. We have

$$
\begin{aligned}
{w_{\mathsf{A}}}^{\mathsf{T}} E w_{\mathsf{A}} &= \sum_{k=1}^{d_{\mathsf{A}}} \sum_{\ell=1}^{d_{\mathsf{A}}} w_k w_\ell E_{k\ell} \\
&= \frac{1}{\det(T)} \sum_{k=1}^{d_{\mathsf{A}}} \sum_{\ell=1}^{d_{\mathsf{A}}} w_k w_\ell \left(E'_{k\ell} + \frac{1}{\tau} E''_{k\ell} \right) && \text{(by Lemma 7)} \\
&= \frac{1}{\det(T)} \left[\underbrace{\sum_{k=1}^{d_{\mathsf{A}}} \sum_{\ell=1}^{d_{\mathsf{A}}} w_k w_\ell E'_{k\ell}}_{E'} + \frac{1}{\tau} \underbrace{\sum_{k=1}^{d_{\mathsf{A}}} \sum_{\ell=1}^{d_{\mathsf{A}}} w_k w_\ell E''_{k\ell}}_{E''} \right]
\end{aligned}
$$

By Lemma 6, there exists $\tau \neq 0$ such that

$$
\operatorname{sign}(\det(T)) = -\operatorname{sign}(E') \quad \text{and} \quad \operatorname{sign}(\tau) = -\operatorname{sign}(\det(T)) \cdot \operatorname{sign}(E'')
$$

Such a choice of τ results in ${w_{\mathsf{A}}}^{\mathsf{T}} E w_{\mathsf{A}} < 0$. Finally by Theorem 1, we have for all x that

$$
c_{\tilde{S}^{-1}}(x, \Delta_{\tilde{S}^{-1}}(x)) = \frac{|w^{\mathsf{T}} x|}{\sqrt{{w_{\mathsf{A}}}^{\mathsf{T}} \tilde{S} w_{\mathsf{A}}}} = \frac{|w^{\mathsf{T}} x|}{\sqrt{{w_{\mathsf{A}}}^{\mathsf{T}} S w_{\mathsf{A}} - {w_{\mathsf{A}}}^{\mathsf{T}} E w_{\mathsf{A}}}} < \frac{|w^{\mathsf{T}} x|}{\sqrt{{w_{\mathsf{A}}}^{\mathsf{T}} S w_{\mathsf{A}}}} = c_{S^{-1}}(x, \Delta_{S^{-1}}(x))
$$

which completes the proof. $\square$

D Proofs and Derivations in Section 4

D.1 Proof of Proposition 5

Proof. We want to show that the standard strategic risk conditioned on an unchanged true label is upper-bounded by the first term in our model designer's objective, $R_{\mathsf{M}}(h)$:

$$
\mathop{\mathbb{E}}_{x \sim \mathcal{D}} [\mathbb{1}[h(x_*) \neq y] \mid \Delta(y) = y] \leq \mathop{\mathbb{E}}_{x \sim \mathcal{D}} [\mathbb{1}(h(x_*^{\mathsf{M}}) \neq y)]
$$

We assume that the manipulating best response is more likely to result in a positive prediction than the unconstrained best response, given that the true labels do not change:

$$
\mathop{\mathbb{E}}_{x \sim \mathcal{D}} [\mathbb{1}[h(x_*) \neq y] \mid \Delta(y) = y] \leq \mathop{\mathbb{E}}_{\mathcal{D}} [\mathbb{1}[h(x_*^{\mathsf{M}}) \neq y] \mid \Delta_{\mathsf{M}}(y) = y] \tag{16}
$$

We therefore have:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{x \sim \mathcal{D}} [\mathbb{1}(h(x_*^{\mathsf{M}}) \neq y)] \\
&= \mathop{\mathbb{E}}_{x \sim \mathcal{D}} [\mathbb{1}(h(x_*^{\mathsf{M}}) \neq y) \mid \Delta_{\mathsf{M}}(y) \neq y] \cdot \Pr[\Delta_{\mathsf{M}}(y) \neq y] \\
&\qquad + \mathop{\mathbb{E}}_{x \sim \mathcal{D}} [\mathbb{1}(h(x_*^{\mathsf{M}}) \neq y) \mid \Delta_{\mathsf{M}}(y) = y] \cdot \Pr[\Delta_{\mathsf{M}}(y) = y] \\
&= \mathop{\mathbb{E}}_{x \sim \mathcal{D}} [\mathbb{1}(h(x_*^{\mathsf{M}}) \neq y) \mid \Delta_{\mathsf{M}}(y) = y] && (\Pr[\Delta_{\mathsf{M}}(y) = y] = 1) \\
&\geq \mathop{\mathbb{E}}_{x \sim \mathcal{D}} [\mathbb{1}(h(x_*) \neq y) \mid \Delta(y) = y] && \text{(by equation 16)}
\end{aligned}
$$

$\square$

D.2 Proof of Proposition 6

Proof. Let $\mathcal{D}^*$ be the distribution induced by deploying classifier h. By the covariate shift assumption, $\Pr_{\mathcal{D}^*}(Y = y|X = x) = \Pr_{\mathcal{D}}(Y = y|X = x)$. Therefore

$$\begin{aligned}
\Pr_{x\sim\mathcal{D}^*}[y(x) = +1] &= \mathop{\mathbb{E}}_{\mathcal{D}^*}[\mathbb{1}[y(x) = +1]] \\
&= \int \mathbb{1}[y(x) = +1] \Pr_{\mathcal{D}^*}(X = x)dx \\
&= \int \mathbb{1}[y(x) = +1] \frac{\Pr_{\mathcal{D}^*}(X = x)}{\Pr_{\mathcal{D}}(X = x)} \Pr_{\mathcal{D}}(X = x)dx \\
&= \int \mathbb{1}[y(x) = +1]\omega_h(x) \Pr_{\mathcal{D}}(X = x)dx \\
&= \mathop{\mathbb{E}}_{\mathcal{D}}[\omega_h(x)\mathbb{1}[y(x) = +1]]
\end{aligned}$$

This implies

$$\Pr_{x\sim\mathcal{D}^*}[y(x) = +1] \geq \Pr_{x\sim\mathcal{D}}[y(x) = +1] \iff \mathop{\mathbb{E}}_{\mathcal{D}}[(\omega_h(x) - 1)\mathbb{1}[y(x) = +1]] \geq 0 \tag{17}$$

By similar reasoning, we have

$$\Pr_{x\sim\mathcal{D}^*}[h(x) = +1] = \mathop{\mathbb{E}}_{\mathcal{D}^*}[\mathbb{1}[h(x) = +1]] = \mathop{\mathbb{E}}_{\mathcal{D}}[\omega_h(x)\mathbb{1}[h(x) = +1]]$$

which implies

$$\Pr_{x\sim\mathcal{D}^*}[h(x) = +1] \geq \Pr_{x\sim\mathcal{D}}[h(x) = +1] \iff \mathop{\mathbb{E}}_{\mathcal{D}}[(\omega_h(x) - 1)\mathbb{1}[h(x) = +1]] \geq 0 \tag{18}$$

It is easy to verify that $\mathbb{E}_{x\sim\mathcal{D}}[\omega_h(x)] = 1$, and this gives us

$$\mathop{\mathbb{E}}_{\mathcal{D}}[(\omega_h(x) - 1)\mathbb{1}[y(x) = +1]] = \mathrm{Cov}_{\mathcal{D}}(\omega_h(x), \mathbb{1}[y(x) = +1]) \tag{19}$$

$$\mathop{\mathbb{E}}_{\mathcal{D}}[(\omega_h(x) - 1)\mathbb{1}[h(x) = +1]] = \mathrm{Cov}_{\mathcal{D}}(\omega_h(x), \mathbb{1}[h(x) = +1]) \tag{20}$$

By equation 17, equation 18, and equation 19, the condition

$$\Pr_{x\sim\mathcal{D}^*}[h(x) = +1] \geq \Pr_{x\sim\mathcal{D}}[h(x) = +1] \iff \Pr_{x\sim\mathcal{D}^*}[y(x) = +1] \geq \Pr_{x\sim\mathcal{D}}[y(x) = +1]$$

is equivalent to the condition

$$\mathrm{Cov}_{\mathcal{D}}(\omega_h(x), \mathbb{1}[y(x) = +1]) \geq 0 \iff \mathrm{Cov}_{\mathcal{D}}(\omega_h(x), \mathbb{1}[h(x) = +1]) \geq 0$$

□

D.3 Derivations for the model designer's objective function

Now that we have obtained a closed-form expression for both the unconstrained and improving best response from the decision subjects, we can analyze the objective function for the model designer, and the model that would be deployed at equilibrium. Recall that the objective function for the model designer is

$$\min_{w\in\mathbb{R}^{d+1}} \quad \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\mathbb{1}(h(\Delta_{\mathsf{M}}(x)) \neq y)] + \lambda \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\mathbb{1}(h(\Delta_{\mathsf{I}}(x)) \neq +1)]$$

By Theorem 1, $h(\Delta_{\mathsf{M}}(x))$ has the closed form

$$\begin{aligned} h(\Delta_{\mathsf{M}}(x)) &= \begin{cases} +1 & \text{if } \; w \cdot x \geq -2\sqrt{{w_{\mathsf{M}}}^{\mathsf{T}} S_{\mathsf{M}} w_{\mathsf{M}}} \\ -1 & \text{otherwise} \end{cases} \\ &= 2 \cdot \mathbb{1}\left[w \cdot x \geq -2\sqrt{{w_{\mathsf{M}}}^{\mathsf{T}} S_{\mathsf{M}} w_{\mathsf{M}}} \right] - 1 \end{aligned}$$

and similarly,

$$h(\Delta_{\mathsf{I}}(x)) = 2 \cdot \mathbb{1}\left[w \cdot x \geq -2\sqrt{{w_{\mathsf{I}}}^{\mathsf{T}} S_{\mathsf{I}} w_{\mathsf{I}}} \right] - 1$$

The model designer's objective can then be re-written as follows:

$$\begin{aligned} &\mathbb{E}_{x \sim D} \left[\mathbb{1}[h(\Delta_{\mathsf{M}}(x)) \neq y] + \lambda \mathbb{1}[h(\Delta_{\mathsf{I}}(x)) \neq +1] \right] \\ =&\mathbb{E}_{x \sim \mathcal{D}} \left[1 - \frac{1}{2}(1 + h(\Delta_{\mathsf{M}}(x)) \cdot y) + \lambda(1 - \frac{1}{2}(1 + h(\Delta_{\mathsf{I}}(x)) \cdot 1)) \right] \\ =&\mathbb{E}_{x \sim \mathcal{D}} \left[\frac{1}{2}(1 + \lambda) - \frac{1}{2} h(\Delta_{\mathsf{M}}(x)) \cdot y - \frac{\lambda}{2} h(\Delta_{\mathsf{I}}(x)) \right] \end{aligned}$$

Removing the constants, the objective function becomes:

$$\begin{aligned} &\min_{w} \mathbb{E}_{x \sim \mathcal{D}} \left[\lambda - h(\Delta_{\mathsf{M}}(x)) \cdot y - \lambda h(\Delta_{\mathsf{I}}(x)) \right] \\ = &\min_{w} \mathbb{E}_{x \sim \mathcal{D}} \left[- \left(2 \cdot \mathbb{1}\left[w \cdot x \geq -2\sqrt{{w_{\mathsf{M}}}^{\mathsf{T}} S_{\mathsf{M}} w_{\mathsf{M}}} \right] - 1 \right) \cdot y(x) - 2\lambda \cdot \mathbb{1}\left[w \cdot x \geq -2\sqrt{{w_{\mathsf{I}}}^{\mathsf{T}} S_{I} w_{\mathsf{I}}} \right] \right] \end{aligned}$$

Re-organizing the above equations, we can turn the model designer's *constrained* optimization problem in equation 6 into the following *unconstrained* problem:

$$\min_{w \in \mathbb{R}^d} \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[- \left(2 \cdot \mathbb{1}\left[w^{\mathsf{T}} x \geq -2\sqrt{\Omega_{\mathsf{M}}} \right] - 1 \right) \cdot y - 2\lambda \cdot \mathbb{1}\left[w^{\mathsf{T}} x \geq -2\sqrt{\Omega_{\mathsf{I}}} \right] \right] \tag{21}$$

The optimization problem in equation 21 is intractable since both the objective and the constraints are non-convex. To overcome this difficulty, we train our classifier by replacing the 0-1 loss function with a convex surrogate loss $\sigma(x) = \log\left(\frac{1}{1+e^{-x}}\right)$. This results in the following ERM problem:

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left[- \sigma\left(y_i \cdot (w^{\mathsf{T}} x_i + 2\sqrt{\Omega_{\mathsf{M}}}) \right) - \lambda \cdot \sigma(w^{\mathsf{T}} x_i + 2\sqrt{\Omega_{\mathsf{I}}}) \right] \tag{22}$$

Conditionally Actionable Features. In practice, individuals can often only change some features in either a positive or negative direction, but not both. However, modeling this restriction on the decision subject's side precludes a closed-form solution. Instead, we strongly disincentivize such moves in the model designer's objective function. The idea is that if the model designer is punished for encouraging an illegal action, the announced classifier will not incentivize such moves from decision subjects. The result is that decision subjects encounter an *implicit* direction constraint on the relevant variables. To that end, we construct a vector $\mathsf{dir} \in \{-1, 0, +1\}^d$ where dir_i represents the prohibited direction of change for the corresponding feature x_i; that is, $\mathsf{dir}_i = +1$ if x_i should not be allowed to increase, -1 if it should not decrease, and 0 if there are no direction constraints. We then append the following penalty term to the model designer's objective in Eq. (6):

$$-\eta \cdot \sum_{i=1}^{d} \max(\mathsf{dir}_i \cdot (\Delta(x) - x)_i, 0) \tag{23}$$

where $\eta > 0$ is a hyperparameter representing the weight given to this penalty term. Eq. (23) penalizes the weights of partially actionable features so that decision subjects would prefer to move towards a certain direction.

E Additional Related Work

Strategic Classification. There has been extensive research on strategic behavior in classification Hardt et al. (2016a); Cai et al. (2015); Chen et al. (2018); Dong et al. (2018); Dekel et al. (2010); Chen et al. (2020). Hardt et al. (2016a) was the first to formalize strategic behavior in classification based on a sequential two-player game (i.e. a Stackelberg game) between decision subjects and classifiers. Since then, other similar Stackelberg formulations have been studied Balcan et al. (2015). Dong et al. (2018) considers the setting in which decision subjects arrive in an online fashion and the learner lacks full knowledge of decision subjects' utility functions. More recently, Chen et al. (2020) proposes a learning algorithm with non-smooth utility and loss functions that adaptively partitions the learner's action space according to the decision subject's best responses.

Recourse. The concept of *recourse* in machine learning was first introduced in Ustun et al. (2019). There, an integer programming solution was developed to offer actionable recourse from a linear classifier. Our work builds on theirs by considering strategic actions from decision subjects, as well as by aiming to incentivize honest improvement. Venkatasubramanian & Alfano (2020) discusses a more adequate conceptualization and operationalization of recourse. Karimi et al. (2020a) provides a thorough survey of algorithmic recourse in terms of its definitions, formulations, solutions, and prospects. Inspired by the concept of recourse, Dean et al. (2020) develops a reachability problem to capture the ability of models to accommodate arbitrary changes in the interests of individuals in recommender systems. Bellamy et al. (2018) builds toolkits for actionable recourse analysis. Furthermore, Gupta et al. (2019) studies how to mitigate disparities in recourse across populations.

Causal Modeling of Features. A flurry of recent papers have demonstrated the importance of understanding causal factors for achieving fairness in machine learning Wang et al. (2019); Bhatt et al. (2020); Bechavod et al. (2020); Miller et al. (2020); Shavit et al. (2020). Miller et al. (2020) studies distinctions between gaming and improvement from a causal perspective. Shavit et al. (2020) provides efficient algorithms for simultaneously minimizing predictive risk and incentivizing decision subjects to improve their outcomes in a linear setting. In addition, Karimi et al. (2020b) develops methods for discovering recourse-achieving actions with high probability given limited causal knowledge. In contrast to these works, we explicitly separate improvable features from manipulated features when maximizing decision subjects' payoffs.

Incentive Design. Like our work, Kleinberg & Raghavan (2020) discusses how to incentivize decision subjects to improve a certain subset of features. Next, Haghtalab et al. (2020) shows that an appropriate projection is an optimal linear mechanism for strategic classification, as well as an *approximate* linear threshold mechanism. Our work complements theirs by providing appropriate linear classifiers that balance accuracy and improvement. Liu et al. (2020) considers the equilibria of a dynamic decision-making process in which individuals from different demographic groups invest rationally, and compares the impact of two interventions: decoupling the decision rule by group and subsidizing the cost of investment.

Algorithmic Fairness in Machine Learning. Our work contributes to the broad study of algorithmic fairness in machine learning. Most common notions of group fairness include disparate impact Feldman et al. (2015), demographic parity Agarwal et al. (2018), disparate mistreatment Zafar et al. (2019), equality of opportunity Hardt et al. (2016b) and calibration Chouldechova (2017). Among them, disparities in the recourse fraction can be viewed as equality of false positive rate (FPR) in the strategic classification setting. Disparities in costs and flipsets are also relevant to counterfactual fairness Kusner et al. (2017) and individual fairness Dwork et al. (2012). Similar to our work, von Kügelgen et al. (2020) also consider the intervention cost of recourse in flipping the prediction across subgroups, investigating the fairness of recourse from a causal perspective.

E.1 Agent's Best Response with Partially Actionable Features

Let feature i represents those features that should only be non-increasing, and feature j represents those features that should only be non-decreasing. Then the constraint can be represented as:

$$y_i \leq 0 \Leftrightarrow e_i^{\mathsf{T}} y \leq 0$$
$$y_j \geq 0 \Leftrightarrow e_j^{\mathsf{T}} y \geq 0$$

Assume that there are n_- features that can only be changed negatively, and there are n_+ features that can only be changed increasingly. We can further combine those new constraints into a matrix form like $Ey \leq 0$. The other constraint can be re-written as:

$$w^{\mathsf{T}} y - b' \geq 0 \Leftrightarrow -w^{\mathsf{T}} y \leq -b',$$

therefore the optimization problem can be rewritten as:

$$\begin{aligned} \min \quad & \frac{1}{2} y^{\mathsf{T}} Q y \\ s.t. \quad & \underbrace{\begin{bmatrix} E \\ -w^{\mathsf{T}} \end{bmatrix}}_{A} y \leq \underbrace{\begin{bmatrix} 0 \\ -b' \end{bmatrix}}_{b} \end{aligned}$$

where A is of the form:

$$A = \begin{bmatrix} I_{n_- \times n_-} & 0 & 0 \\ 0 & -I_{n_+ \times n_+} & 0 \\ \smile\smile & -w^{\mathsf{T}} & \smile\smile \end{bmatrix}_{(n_+ + n_- + 1) \times n}$$

F Additional Experimental Details and Results

In this section, we provide additional experimental results. In particular, we provide the full results with mean and standard deviation in Table 4.

F.1 Basic information of each dataset

Table 3: Basic information of each dataset.

Dataset	Size	Dimension	Prediction Task
`credit`	$20,000$	16	To predict if a person can repay their credit card loan.
`adult`	$48,842$	14	To predict whether income exceeds $50K/yr$ based on census data.
`german`	$1,000$	26	To predict whether a person is good or bad credit risk.
`spam`	4601	57	To predict if an email is a spam or not.

F.2 Specific Cost Matrix

We specify the cost matrix as follows:

$$S_{ij}^{-1} = \begin{cases} 1, & \text{if } i = j \text{ and } i \in \mathsf{I} \\ 0.2, & \text{if } i = j \text{ and } j \in \mathsf{M} \\ 1, & \text{if the cost of changing features } i \\ & \text{and } j \text{ are \textit{negatively} correlated} \\ -1, & \text{if the cost of changing features } i \\ & \text{and } j \text{ are \textit{positively} correlated} \\ 0, & \text{otherwise} \end{cases}$$

Table 4: Performance metrics for all methods over 4 real data sets with non-diagonal cost matrix. We report the mean and standard deviation for 5-fold cross validation. The constructive adaptation (CA) consistently achieves a high accuracy at deployment while providing the highest improvement rates across all four datasets.

		METHODS			
Dataset	**Metrics**	ST	DF	MP	CA
CREDIT	*test error*	29.52 ± 0.37	29.66 ± 0.40	29.65 ± 0.41	29.60 ± 0.44
	deploy error	31.25 ± 0.56	29.66 ± 0.40	29.41 ± 0.32	29.49 ± 0.38
	improvement rate	46.35 ± 3.81	44.71 ± 4.75	36.76 ± 0.53	48.27 ± 5.50
ADULT	*test error*	23.05 ± 0.47	33.55 ± 0.73	24.94 ± 0.52	27.22 ± 0.65
	deploy error	38.64 ± 4.46	33.55 ± 0.73	26.85 ± 0.59	29.34 ± 0.45
	improvement rate	30.92 ± 3.31	60.63 ± 29.40	36.70 ± 1.62	63.79 ± 7.80
GERMAN	*test error*	30.85 ± 0.82	36.10 ± 1.97	33.25 ± 1.44	34.70 ± 2.15
	deploy error	33.40 ± 1.78	36.10 ± 1.97	34.60 ± 1.94	34.25 ± 1.78
	improvement rate	41.20 ± 5.77	42.10 ± 9.07	33.50 ± 2.53	56.10 ± 6.40
SPAMBASE	*test error*	7.11 ± 0.52	10.18 ± 0.45	11.52 ± 0.12	14.37 ± 0.24
	deploy error	22.40 ± 3.14	10.18 ± 0.45	12.92 ± 0.58	14.70 ± 0.36
	improvement rate	40.04 ± 13.06	32.46 ± 14.63	26.42 ± 4.80	43.98 ± 6.18

We use the `credit` dataset as a demonstration of how we specify the non-diagonal element in the cost matrix. For two feature variables that have a positive correlation, e.g., *CheckingAccountBalance* and *SavingsAccountBalance*, we assign -1 to the corresponding elements in the cost matrix S. For two feature variables that have a negative correlation, e.g., *CheckingAccountBalance* and *MissedPayments*, we assign $+1$ to the corresponding elements in the cost matrix.

F.3 Computing Infrastructure

We conducted all experiments on a 3 GHz 6-Core Intel Core i5 CPU. All our methods have relatively modest computational cost and can be trained within a few minutes.

F.4 Results for non-diagonal cost matrix

In real life, the specification of the cost matrix might require examining the causal correlations among different features. We consider a non-diagonal cost matrix setup based on common knowledge and describe the rationale as below. For two feature variables that have a positive correlation, e.g., *CheckingAccountBalance* and *SavingsAccountBalance*, we assign -1 to the corresponding elements in the cost matrix. For two feature variables that have a negative correlation, e.g., *CheckingAccountBalance* and *MissedPayments*, we assign +1 to the corresponding elements in the cost matrix. We also note that the non-diagonal cost matrix must be invertible under our assumption on the cost of modifying features. We provide more detailed results for each dataset in Table 4, which shows the means and standard deviations of different metrics. Compared to the empirical results of using a diagonal matrix, we achieve similar results with respect to the three evaluation criteria across all four methods.

F.5 Additional Experimental Results for Non-Linear models

We also work with a three-layer neural network to validate the effectiveness of the oracle best response in Algorithm 1. We note that the LIME program needs to learn a local linear model for each instance, which is very time-consuming. Therefore, we downsample only 10% of data examples from the `credit` dataset. We follow the same setting as the linear classifier experiments. We compare our method with the static classifier in Table 5. We find out for this non-linear model setting, our approach has a higher improvement rate while preventing manipulations with the deploy error 27.72% vs. 35.64%.

Table 5: Performance metrics for non-linear models.

	Methods	
Metrics	ST	CA
test error	30.72%	30.01%
deploy error	35.64%	27.72%
improvement rate	0.99%	2.97%

F.6 Flipsets

We also construct flipsets for individuals in the `german` dataset using the closed-form solution Eq. (3) under our trained classifier. The individual characterized as a "bad consumer" (-1) is supposed to decrease their missed payments in order to flip their outcome of the classifier with respect to a non-diagonal cost matrix. In contrast, even though the individual improves their loan rate or liable individuals, the baseline classifier will still reject them. We also provide flipsets for conditionally actionable features on the `credit` dataset in Table 7. The individual will undesirably reduce their education level when the classifier is unaware of the partially actionable features. In contrast, the individual decreases their total overdue months instead when the direction penalty is imposed during training.

Table 6: Flipset for a person denied credit by ManipulatedProof on the `german` dataset. The red up arrows ↑ represent increasing the values of features, while the red down arrows ↓ represent decreasing.

Feature	Type	Original	LightTouch	ManipulatedProof
LoanRateAsPercentOfIncome	I	3	3	2 ↓
NumberOfOtherLoansAtBank	I	1	1	1
NumberOfLiableIndividuals	I	1	0 ↓	2 ↑
CheckingAccountBalance $\geq$ 0	I	0	0	0
CheckingAccountBalance $\geq$ 200	I	0	0	0
SavingsAccountBalance $\geq$ 100	I	0	0	0
SavingsAccountBalance $\geq$ 500	I	0	0	0
MissedPayments	I	1	0 ↓	1
NoCurrentLoan	I	0	0	0
CriticalAccountOrLoansElsewhere	I	0	0	0
OtherLoansAtBank	I	0	0	0
OtherLoansAtStore	I	0	0	0
HasCoapplicant	I	0	0	0
HasGuarantor	I	0	0	0
Unemployed	I	0	0	0
LoanDuration	M	48	47 ↓	47 ↓
PurposeOfLoan	M	0	0	0
LoanAmount	M	4308	4307 ↓	4307 ↓
HasTelephone	M	0	0	0
Gender	U	0	0	0
ForeignWorker	U	0	0	0
Single	U	0	0	0
Age	U	24	24	24
YearsAtCurrentHome	U	4	4	4
OwnsHouse	U	0	0	0
RentsHouse	U	1	1	1
YearsAtCurrentJob $\leq$ 1	U	1	1	1
YearsAtCurrentJob $\geq$ 4	U	0	0	0
JobClassIsSkilled	U	1	1	1
GoodConsumer	-	-1	$+1$ ↑	-1

Table 7: Flipset for an individual on Credit dataset with partially actionable features. The red up arrows ↑ represent any increasing values, while the red down arrows ↓ represent any decreasing values.

Feature	Type	dir	Original	$\eta = 0$	$\eta = 100$
EducationLevel	I	$+1$	3	2 ↓	3
TotalOverdueCounts	I	0	1	1	1
TotalMonthsOverdue	I	0	1	1	0 ↓
MaxBillAmountOverLast6Months	M	0	0	0	0
MaxPaymentAmountOverLast6Months	M	0	0	0	0
MonthsWithZeroBalanceOverLast6Months	M	0	0	0	0
MonthsWithLowSpendingOverLast6Months	M	0	6	5 ↓	6
MonthsWithHighSpendingOverLast6Months	M	0	0	0	0
MostRecentBillAmount	M	0	0	0	0
MostRecentPaymentAmount	M	0	0	0	0
Married	U	0	1	1	1
Single	U	0	0	0	0
$Age \leq 25$	U	0	0	0	0
$25 \leq Age \leq 40$	U	0	0	0	0
$40 \leq Age < 60$	U	0	0	0	0
$Age \geq 60$	U	0	1	1	1
HistoryOfOverduePayments	U	0	1	1	1
NoDefaultNextMonth	-	-	-1	$+1$ ↑	$+1$ ↑