Sequential Strategic Screening

Lee Cohen [1] **Saeed Sharifi-Malvajerdi** [1] **Kevin Stangl** [1] **Ali Vakilian** [1] **Juba Ziani** [2]

Abstract

We initiate the study of strategic behavior in screening processes with *multiple* classifiers. We focus on two contrasting settings: a "conjunctive" setting in which an individual must satisfy all classifiers simultaneously, and a sequential setting in which an individual to succeed must satisfy classifiers one at a time. In other words, we introduce the combination of *strategic classification* with screening processes. We show that sequential screening pipelines exhibit new and surprising behavior where individuals can exploit the sequential ordering of the tests to "zig-zag" between classifiers without having to simultaneously satisfy all of them. We demonstrate an individual can obtain a positive outcome using a limited manipulation budget even when far from the intersection of the positive regions of every classifier. Finally, we consider a learner whose goal is to design a sequential screening process that is robust to such manipulations, and provide a construction for the learner that optimizes a natural objective.

1. Introduction

Screening processes (Arunachaleswaran et al., 2022; Blum et al., 2022; Cohen et al., 2020) involve evaluating and selecting individuals for a specific, pre-defined purpose, such as a job, educational program, or loan application. These screening processes are generally designed to identify which individuals are qualified for a position or opportunity, often using multiple sequential classifiers or tests. For example, many hiring processes involve multiple rounds of interviews; university admissions can involve a combination of standardized tests, essays, or interviews. They have substantial practical benefits, in that they can allow a complex decision to be broken into a sequence of smaller and cheaper steps; this allows, for example, to split a decision across multiple independent interviewers, or across smaller and easier-to-measure criteria and requirements.

Many of the decisions made by such screening processes are high stakes. For example, university admissions can affect an individual's prospects for their entire life. Loan decisions can have a long-term (sometimes even inter-generational) effect on a family's wealth or socio-economic status. When these decisions are high stakes, i.e. when obtaining a positive outcome is valuable or potentially life-changing or obtaining a negative outcome can be harmful, individuals may want to manipulate their features to trick the classifier into assigning them a positive outcome.

In machine learning, this idea is known as strategic classification, and was notably introduced and studied by (Brückner & Scheffer, 2011; Hardt et al., 2016). The current work aims to incorporate strategic classification within screening processes, taking a departure from the classical point of view in the strategic classification literature that focuses on a single classifier (see related work section).

The key novel idea of our model of *strategic screening processes (or pipelines)*, compared to the strategic classification literature, comes from the fact that i) an individual has to pass and manipulate her way through *several* classifiers, and ii) that we consider *sequential* screening pipelines.

In a sequential screening pipeline, once an individual (also called *Agent*) has passed a test or stage of this pipeline, she can "forget" about the said stage; whether or not she passes the next stage depends *only on her performance in that stage*. For example, a job candidate that has passed the initial human resources interview may not need to worry about convincing that interviewer, and can instead expand her effort solely into preparing for the first technical round of interviews. Alternatively, imagine a student 'cramming' for a sequence of final exams, where one has a finite capacity to study that is used up over a week of tests. One wants to achieve a minimum score on each test, with a minimum of effort, by studying in between each test.

Our goal in this work is to examine how considering a pipeline comprised of a sequence of classifiers affects and modifies the way a strategic agent manipulates her features to obtain a positive classification outcome, and how a learner

[*]Equal contribution [1]Toyota Technological Institute at Chicago, Chicago, IL, USA [2]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA. Correspondence to: Kevin Stangl <kevin@ttic.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

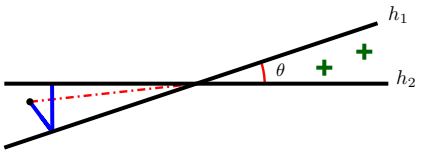

Figure 1. Suppose the agent is the disqualified (i.e., placed in the negative region of the conjunctions of h_1, h_2) point. A trivial manipulation strategy is to use the shortest *direct* path to the positive region, which is the dashed red path. However, the agent may also first manipulate slightly to pass h_1, then manipulate minimally again to pass h_2, as depicted with the blue solid path. This is what we call a zig-zag strategy.

(which we primarily call the *Firm*) should take this strategic behavior into account to design screening pipelines that are robust to such manipulation. In our model, 1) the firm deploys a sequential pipeline of classifiers, 2) the agent is given full knowledge of the pipeline and computes their optimal manipulation strategy, then 3) the agent goes through the screening pipeline and implements said optimal manipulation strategy in order to pass the tests sequentially, one at a time.

We make a distinction between the following two cases: 1) the firm deploys its classifiers sequentially which we refer to as a *sequential screening process*; 2) the firm deploys a single classifier whose positive classification region is the intersection of the positive regions of the classifiers that form the pipeline which we sometimes refer to as *simultaneous (or conjunctive) testing*—this single classifier is basically the *conjunction* or intersection of classifiers from the pipeline. The former corresponds to a natural screening process that is often used in practice and for which we give our main results, while the latter is primarily considered as a benchmark for our results for the sequential case.

Our Contributions. We show a perhaps surprising result: an agent can exploit the sequential nature of the screening process and move through the whole pipeline even when she started far from the intersection of the positive classification regions of all classifiers. In other words, the sequentiality of screening processes can *improve* an agent's ability to manipulate her way through multiple classifiers compared to the simultaneous screening. We name the resulting set of strategies for such an agent in the sequential case *"zig-zag" strategies*. In other words, whenever the agent does not manipulate straight to a point that is classified as positive by the conjunction of all classifiers, we call it a zig-zag strategy. An example of such a strategy that zig-zags between two classifiers is provided in Figure 1.

In Figure 1, since there is a small angle θ between the two tests, an agent at the bottom of the figure can zag right and then left as shown by the blue lines. In this case, the agent is classified as positive in every single step, and by making θ arbitrarily small, will have arbitrarily lower total cost (e.g., the cumulative ℓ_2 distance) compared to going directly to the intersection point of the classifiers. We provide concrete classifiers and an initial feature vector for such a case in Example 3.2.

In fact, in Section 3.2 we show that for a given point, as θ goes to zero, the ratio between the total cost of the zig-zag strategy and the cost of going directly to the intersection can become arbitrarily large. As we assume that conjunction of the classifiers captures the objective of the firm, using a pipeline can allow more disqualified people to get a positive outcome by manipulating their features. We show this in Figure 2: This figure shows the region of the agents space that can successfully manipulate to pass two linear tests in the two-dimensional setting, given a budget τ for manipulation. As shown by the figure, individuals in the green region of Figure 2.c can pass the tests in the sequential setting but would not be able to do so if they had to pass the tests simultaneously.

We further show how the optimal zig-zag strategy of an agent can be obtained computationally efficiently via a simple convex optimization framework in Section 3.3 and provide a closed-form characterization of this strategy in the special case of 2-dimensional features and a pipeline of exactly two classifiers in Section 3.4.

In Section 3.5 we consider a "monotonicity" condition under which, agents prefer to use the simple strategy which passes all classifiers simultaneously in a single move and does not zig-zag between classifiers.

Finally, in Section 4.1, we exhibit a defense strategy that maximizes true positives subject to not allowing any false positives. Interestingly, we show that under this strategy, deploying classifiers sequentially allows for a higher utility for the firm than using a conjunction of classifiers.

Related Work. Our work inscribes itself at the intersection of two recent lines of work. The first one studies how strategic behavior affects decision-making algorithms (e.g. regression or classification algorithms), and how to design decision rules that take into account or dis-incentivize strategic behavior. This line of work is extensive and comprised of the works of (Brückner & Scheffer, 2011; Hardt et al., 2016; Kleinberg & Raghavan, 2020; Braverman & Garg, 2020; Miller et al., 2020; Liu et al., 2020; Jagadeesan et al., 2021; Haghtalab et al., 2020; Meir et al., 2010; 2011; 2012; Dekel et al., 2010; Chen et al., 2018; Cummings et al., 2015; Khajehnejad et al., 2019; Ustun et al., 2019; Chen et al., 2020b; Björkegren et al., 2020; Dee et al., 2019; Perote & Perote-Pena, 2004; Ahmadi et al., 2021; Tang et al., 2021; Hu et al., 2019; Milli et al., 2019; Perdomo et al., 2020; Ghalme et al., 2021; Braverman & Garg, 2020; Ahmadi

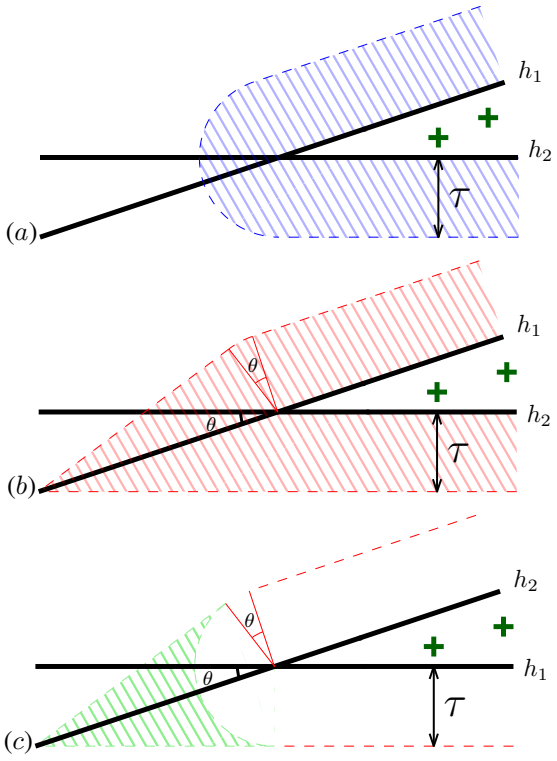

Figure 2. Each agent has a manipulation budget of τ and the cost function is ℓ_2 distance. Then, (a) shows the region of agents who afford to manipulate their feature vectors to pass both tests simultaneously, (b) shows the region of agents who afford to manipulate their feature vectors to pass the tests sequentially (i.e, first h_1, then h_2), and (c) shows the difference in these two regions.

et al., 2022; Bechavod et al., 2021; 2022; Shavit et al., 2020; Dong et al., 2018; Chen et al., 2020a; Harris et al., 2021).

The second line of work is separate and aims to understand how decisions compose and affect each other in decision-making and screening pipelines (Cohen et al., 2020; Bower et al., 2017; Blum et al., 2022; Arunachaleswaran et al., 2022; Dwork et al., 2020; Dwork & Ilvento, 2018). These works study settings in which *multiple* decisions are made about an individual or an applicant. (Harris et al., 2021) has a similar motivation to ours in studying how multiple rounds of interaction change strategic dynamics, however, the linearity of their model allows them to treat time-steps independently while our agents can benefit from using information on the subsequent steps of the pipeline.

However, and to the best of our knowledge, there is little work bringing these two fields together and studying strategic behavior in the context of decision *pipelines* comprised of *multiple* classifiers. This is where the contribution of the current work lies.

Interestingly, there are interesting connections between our model with classical work in learning intersections of half-spaces (Klivans & Servedio, 2004; Klivans & Sherstov, 2009). In our model, we think of the half-spaces as *known* in advance, so our model differs in that agents do not need to *learn* half-spaces. However, future work could instead consider a learner who must learn the intersection of half-spaces while simultaneously considering the effect of strategic behavior, a complex learning problem. Further, there is a subtle distinction that agents in our work that agents may modify their features to pass half-spaces sequentially, but without needing to be in the intersection of all half-spaces; the crux of our contribution is in fact to show that sequentiality often leads to very different agent behavior than modifying features to reach the intersection of the classifiers' positive region.

The sequentiality of our framework is related to the line of work on convex body chasing (Sellke, 2020; Friedman & Linial, 1993; Bubeck et al., 2019; Argue et al., 2021; Guan et al., 2022; Bansa et al., 2018; Bubeck et al., 2020), but once again, a distinction between our paper and this line of work is that agents know all classifiers in advance and does not need to plan for an adversary.

Finally, perhaps closest to our work is the line of work on Online Convex Optimization (OCO) with switching costs and known loss functions. These works also assume that (1) the (single) agent observes the loss function before picking a point at each round or even observes the next (fixed size) loss functions sequence, and (2) the cost functions are dependent on the previous point x_t, (e.g., ℓ_2 distance between the current and the previous point). However, our work differs in some of the specific assumptions we make (for example, an agent cannot choose their initial features, while one can choose the starting point in Online Convex Optimization with switching costs and known loss functions (Shi et al., 2020; Li et al., 2021; Cesa-Bianchi et al., 2013)), but more importantly, our main focus is different: beyond characterizing the optimal strategy for a strategic agent, we are interested in i) understanding how sequentiality affects and potentially increases agents' ability to strategize and ii) developing screening pipelines that are robust to strategic behavior.

2. Our Model

Formally, individuals (or agents) are represented by a set of features $x \in \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{R}^d$, for $d \geq 1$. The firm has a fixed sequence of binary tests or classifiers $h_1, h_2, \ldots, h_k : \mathcal{X} \to \{0, 1\}$ that are deployed to select qualified individuals while screening out unqualified individuals. Here, an outcome of 1 (positive) corresponds to an acceptance, and an outcome of 0 (negative) corresponds to a rejection. Once a person is rejected by a test they leave the pipeline.

In the whole paper, we assume that the classifiers are linear and defined by half-spaces; i.e. $h_i(x) = 1 \iff w_i^\top x \geq b_i$ for some vector $w_i \in \mathbb{R}^d$ and real threshold $b_i \in \mathbb{R}$. Equivalently, we often write $h_i(x) = \mathbb{1}\left[w_i^\top x \geq b_i\right]$.[1]

In this work we assume that the true qualifications of individuals are determined by the conjunction of the classifiers adopted by the firm in the pipeline, i.e. an agent x is qualified if and only if $h_i(x) = 1$ for all i. In other words, the firm has designed a pipeline that makes no error in predicting individuals' qualifications *absent strategic behavior*.

However, in the presence of strategic behavior, individuals try to manipulate their feature vectors to become positively classified by the classifiers simply because they receive a positive utility from a positive outcome. Similar to prior works, throughout this work, we assume a "white box" model meaning agents know the parameters for each classifier. More precisely, the firm commits to using a sequential screening process consisting of classifiers $h_1, h_2 \dots h_k$, and each agent knows the parameters of each hypothesis, the order of the tests, her own feature value x, and the cost to manipulate to any other point in the input space.

An agent's cost function is modeled by a function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ that takes two points $x, \hat{x}$ and outputs the cost of moving from x to $\hat{x}$. One can think of x as the initial feature vector of an agent and $\hat{x}$ as the manipulated features. In the sequential setting that we consider, we take the cost of manipulation to be the cumulative cost across every single manipulation. In particular, for a manipulation path $x^{(0)} \to x^{(1)} \to x^{(2)} \to \dots \to x^{(k)}$ taken by an agent whose true feature values are $x^{(0)}$, the cost of manipulation is given by $\sum_{i=1}^{k} c(x^{(i-1)}, x^{(i)})$. We assume such manipulations do not change nor improve one's true qualifications[2] and we discuss how the firm mitigates this effect of manipulation.

In turn, the firm's goal is to have an accurate screening process whose predictions are as robust to and unaffected by such strategic: the firm modifies its classifiers $h_1, \cdots, h_k$ to $\tilde{h}_1, \cdots, \tilde{h}_k$ so that the output of $\tilde{h}_1, \cdots, \tilde{h}_k$ on manipulated agents' features can identify the qualified agents optimally with respect to a given "accuracy measure"; we will consider two such measures in Section 4.

[1] While more general classes of classifiers could be considered, linear classifiers are a natural starting point to study strategic classification. This linearity assumption arises in previous work, e.g. (Kleinberg & Raghavan, 2020; Tang et al., 2021; Ahmadi et al., 2022) to only name a few.

[2] E.g., in a loan application, such manipulations could be opening a new credit card account: doing so may temporarily increase an agent's credit score, but does not change anything about an agent's intrinsic financial responsibility and ability to repay the loan.

2.1. Agent's Manipulation

We proceed by formally defining the minimal cost of manipulation, which is the minimal cost an agent has to invest to pass all classifiers, and the best response of an agent for both sequential and simultaneous testing.

Definition 2.1 (Manipulation Cost: Sequential)**.** Given a sequence of classifiers $h_1, \dots, h_k$, a global cost function c, and an agent $x^{(0)} \in \mathcal{X}$, the manipulation cost of an agent in the sequential setting is defined as the minimum cost incurred by her to pass all the classifiers sequentially, i.e.,

$$c^*_{seq}\left(x^{(0)}, \{h_1, \dots, h_k\}\right) = \min_{x^{(1)}, \dots, x^{(k)} \in \mathcal{X}} \sum_{i=0}^{k-1} c(x^{(i)}, x^{(i+1)}) \quad \text{s.t.} \quad h_i(x^{(i)}) = 1 \ \ \forall i \in [k].$$

The *best response* of $x^{(0)}$ to the sequential testing $h_1, \dots, h_k$ is the path $x^{(1)}, \dots, x^{(k)}$ that minimizes the objective.

Definition 2.2 (Manipulation Cost: Conjunction or Simultaneous)**.** Given a set of classifiers $\{h_1, \dots, h_k\}$, a global cost function c, and an agent x, the manipulation cost of an agent in the conjunction setting is defined as the minimum cost incurred by her to pass all the classifiers at the same time, i.e.,

$$c^*_{conj}\left(x, \{h_1, \dots, h_k\}\right) = \min_{z \in \mathcal{X}} \ c(x, z) \quad \text{s.t.} \quad h_i(z) = 1 \ \forall i \in [k].$$

The *best response* of x to the conjunction of $h_1 \dots, h_k$ is the z that minimizes the objective.

3. Best Response of Agents in a Screening Process with Oblivious Defender

In this section, we study the manipulation strategy of an agent. In particular, we present algorithms to compute optimal manipulation strategies efficiently. For brevity, some of the proofs are relegated to the appendix. We make the following assumption on the cost function in most of the section, unless explicitly noted otherwise:

Assumption 3.1. The cost of moving from x to $\hat{x}$ is given by $c(x, \hat{x}) = \|\hat{x} - x\|_2$, where $\|.\|_2$ denotes the standard Euclidean norm.

3.1. Optimal Strategies in the Conjunction Case

As a warm-up to our zig-zag strategy in Section 3.3, we first consider the optimal strategy for our benchmark, which is the case of the simultaneous conjunction of k classifiers. In the case where agents are supposed to pass a collection

of linear classifiers simultaneously, the best response of an agent $x \in \mathbb{R}^d$ is given by solving the following optimization problem

$$\begin{aligned} \min_{z} \; & c(x, z) \\ \text{s.t.} \; & w_i^\top z \geq b_i \; \forall i \in [k]. \end{aligned} \tag{1}$$

which is a convex program as long as c is convex in z.

In the special case in which $d = 2$ and $k = 2$, i.e. when feature vectors are two-dimensional and an agent must be positively classified by the conjunction of two linear classifiers $h_1(x) = \mathbb{1}(w_1^\top x \geq b_1)$ and $h_2(x) = \mathbb{1}(w_2^\top x \geq b_2)$, we provide a closed form characterization of an agent's strategy.

We assume that the two classifiers are *not* parallel to each other because if $w_2 = kw_1$ for some $k \in \mathbb{R}$, then one can show that either the acceptance regions of h_1 and h_2 do not overlap, or the optimal strategy of an agent is simply the orthogonal projection onto the intersection of the acceptance regions of h_1 and h_2.

We further assume, without loss of generality, that $b_1 = b_2 = 0$ because if either b_1 or b_2 is nonzero, one can use the change of variables $x' \triangleq x + s$ to write the classifiers as $h_1(x') = \mathbb{1}(w_1^\top x' \geq 0)$ and $h_2(x') = \mathbb{1}(w_2^\top x' \geq 0)$. Here s is the solution to $\{w_1^\top s = -b_1, w_2^\top s = -b_2\}$.

For any $w \in \mathbb{R}^2$ with $\|w\|_2 = 1$, let $P_w(x)$ and $d_w(x)$ be the orthogonal projection of x onto the region $\{y \in \mathbb{R}^2 : w^\top y \geq 0\}$, and its orthogonal distance to the same region, respectively. We have

$$P_w(x) \triangleq \begin{cases} x & \text{if } w^\top x \geq 0 \\ x - (w^\top x)w & \text{if } w^\top x < 0 \end{cases},$$

$$d_w(x) \triangleq \begin{cases} 0 & \text{if } w^\top x \geq 0 \\ |w^\top x| & \text{if } w^\top x < 0 \end{cases}.$$

Given this setup, the best response characterization of an agent x can be given as follows. If $h_1(x) = h_2(x) = 1$ then $z = x$. Otherwise, the best response is either the orthogonal projection onto the acceptance region of h_1 or h_2, or moving directly to the intersection of the classifiers ($\vec{0}$):

1. If $h_1(P_{w_2}(x)) = 1$, then $z = P_{w_2}(x)$ and the cost of manipulation is $c^*_{conj}\left(x^{(0)}, \{h_1, h_2\}\right) = d_{w_2}(x)$.
2. If $h_2(P_{w_1}(x)) = 1$, then $z = P_{w_1}(x)$ and the cost of manipulation is $c^*_{conj}\left(x^{(0)}, \{h_1, h_2\}\right) = d_{w_1}(x)$.
3. if $h_1(P_{w_2}(x)) = h_2(P_{w_1}(x)) = 0$ then $z = \vec{0}$ and the cost of manipulation is $c^*_{conj}\left(x^{(0)}, \{h_1, h_2\}\right) = \|x\|_2$.

Given a budget τ, agents who can manipulate with a cost of at most τ to pass the two tests simultaneously, i.e. $\{x^{(0)} : c^*_{conj}\left(x^{(0)}, \{h_1, h_2\}\right) \leq \tau\}$ is highlighted in Figure 2.a.

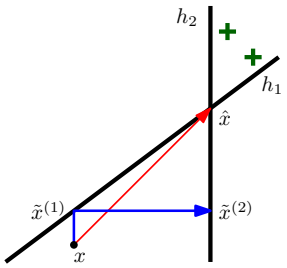

Figure 3. An example for a zig-zag strategy being better for an agent that starts at x in the sequential case than moving in a single step. Here, an agent would prefer to first manipulate to $\tilde{x}^{(1)}$ then to $\tilde{x}^{(2)}$ (the blue arrows) instead of straightforwardly moving from x to $\hat{x}$ as would be optimal in the conjunction case (the red arrow).

3.2. A Zig-Zag Manipulation on Sequential Classification Pipelines

Here, we make the observation that the sequential nature of the problem can change how an agent will modify her features in order to pass a collection of classifiers, compared to the case when said classifiers are deployed simultaneously. We illustrate this potentially counter-intuitive observation via the following simple example:

Example 3.2. Consider a two-dimensional setting. Suppose an agent going up for classification has an initial feature vector $x = (0, 0)$. Suppose the cost an agent faces to change her features from x to a new vector $\hat{x}$ is given by $\|\hat{x} - x\|_2$. Further, imagine an agent must pass two classifiers: $h_1(x) = \mathbb{1}\{4x_2 - 3x_1 \geq 1\}$, and $h_2(x) = \mathbb{1}\{x_1 \geq 1\}$, where x_i is the $i-$th component of x.

It is not hard to see, by triangle inequality, that if an agent is facing a conjunction of h_1 and h_2, an agent's cost is minimized when $\hat{x} = (1, 1)$ (this is in fact the intersection of the decision boundaries of h_1 and h_2), in which case the cost incurred by an agent is $\sqrt{1+1} = \sqrt{2}$ (see the red manipulation in Figure 3).

However, if the classifiers are offered sequentially, i.e. h_1 then h_2, consider the following feature manipulation: first, the agent sets $\tilde{x}^{(1)} = (0, 1/4)$, in which case she passes h_1 and incurs a cost of $1/4$. Then, the agent sets $\tilde{x}^{(2)} = (1, 1/4)$; the cost to go from $\tilde{x}^{(1)}$ to $\tilde{x}^{(2)}$ is $\|_2(1, 1/4) - (0, 1/4)\| = 1$ (see the blue manipulation in Figure 3). In turn, the total cost of this manipulation to pass (i.e., get a positive classification on) both classifiers is at most $1 + 1/4 = 5/4$, and is always better than the $\sqrt{2}$ cost for the conjunction of classifiers! □

Intuitively, here, the main idea is that in the "conjunction of classifiers" case, an agent must manipulate her features a single time in a way that satisfies all classifiers at once. However, when facing a sequence of classifiers $h_1, \ldots, h_k$, once an agent has passed classifier h_{i-1} for any given i,

it can "forget" classifier h_{i-1} and manipulate its features to pass h_i while *not being required to pass* h_{i-1} anymore. In turn, the potential manipulations for an agent in the sequential case are less constrained than in the conjunction of classifiers case. This result is formalized below:

Claim 3.3. Let $h_1, \ldots, h_k$ be a sequence of k linear classifiers. For any agent with initial feature vector $x \in \mathbb{R}^d$ ($d \geq 1$), $c^*_{conj}(x, \{h_1, \ldots, h_k\}) \geq c^*_{seq}(x, \{h_1, \ldots, h_k\})$.

Intuitively, the above claim follows from the observation that any best response solution to the conjunction case in particular still passes all classifiers and has the same cost in the sequential case.

However, there can be a significant gap between how much budget an agent needs to spend in the conjunctive versus in the sequential case to successfully pass all classifiers (for illustration, see Figure 2). In fact, we show below that the multiplicative gap between the conjunctive and sequential manipulation cost can be unbounded, even in the two-dimensional setting:

Lemma 3.4. *Consider $d = 2$. For any constant $M > 0$, there exists two linear classifiers h_1 and h_2 and an initial feature vector $x^{(0)}$ such that $\frac{c^*_{conj}(x^{(0)}, \{h_1, h_2\})}{c^*_{seq}(x^{(0)}, \{h_1, h_2\})} \geq M$.*

Proof. Pick $x^{(0)} = (0, 0)$. Let $\gamma > 0$ be a real number. Consider $h_1(x) = \mathbb{1}\left\{\frac{x_1}{\gamma} + x_2 \geq 1\right\}$ and $h_2(x) = \mathbb{1}\left\{\frac{x_1}{\gamma} - x_2 \geq 1\right\}$. Let $\hat{x}$ be the agent's features after manipulation. To obtain a positive classification outcome, the agent requires both $\hat{x}_1 \geq \gamma(1 - \hat{x}_2)$ and $\hat{x}_1 \geq \gamma(1 + \hat{x}_2)$. Since one of $1 - \hat{x}_2$ or $1 + \hat{x}_2$ has to be at least 1, this implies $\hat{x}_1 \geq \gamma$. In turn, $c(x, \{h_1, h_2\}) = \|\hat{x}\| \geq \gamma$.

However, in the sequential case, a manipulation that passes h_1 is to set $x^{(1)} = (0, 1)$. Then a manipulation that passes h_2, starting from $x^{(1)}$, is to set $x^{(2)} = (0, -1)$. The total cost is $\|(0, 1) - (0, 0)\| + \|(0, -1) - (0, 1)\| = 1 + 2 = 3$. In particular, $\frac{c^*_{conj}(x, \{h_1, \ldots, h_k\})}{c^*_{seq}(x, \{h_1, \ldots, h_k\})} \geq \gamma/3$. The result is obtained by setting $\gamma = 3M$. □

3.3. An Algorithmic Characterization of an agent's Optimal Strategy in the Sequential Case

In this section, we show that in the sequential setting, an agent can compute her optimal sequences of manipulations efficiently. Consider any initial feature vector $x^{(0)} \in \mathbb{R}^d$ for an agent. Further, suppose an agent must pass k linear classifiers $h_1, \ldots, h_k$. For $i \in [k]$, we write once again $h_i(x) = \mathbb{1}[w_i^\top x \geq b_i]$ the i-th classifier that an agent must get a positive classification on. Here and for this subsection only, we relax our assumption on the cost function to be more general, and not limited to ℓ_2 costs:

Assumption 3.5. The cost $c(x, \hat{x})$ of moving from feature vector x to feature vector $\hat{x}$ is convex in $(x, \hat{x})$.

This is a relatively straightforward and mild assumption; absent convexity, computing the best feature modifications for even a single step can be a computationally intractable problem. The assumption covers but is not limited to a large class of cost functions of the form $c(x, \hat{x}) = \|\hat{x} - x\|$, for *any* norm $\|.\|$. It can also encode cost functions where different features or directions have different costs of manipulation; an example is $c(x, \hat{x}) = (\hat{x} - x)^\top A (\hat{x} - x)$ where A is a positive definite matrix, as used in (Shavit et al., 2020; Bechavod et al., 2022).

In this case, an agent's goal, starting from her initial feature vector $x^{(0)}$, is to find a sequence of feature modifications $x^{(1)}$ to $x^{(k)}$ such that: 1) for all $i \in [k]$, $h_i(x^{(i)}) = 1$. I.e., x^i passes the i-th classifier; and 2) the total cost $\sum_{i=1}^k c(x^{(i-1)}, x^{(i)})$ of going from $x^{(0)} \to x^{(1)} \to x^{(2)} \to \ldots \to x^{(k)}$ is minimized. This can be written as the following optimization problem:

$$\begin{aligned} \min_{x^{(1)}, \ldots, x^{(k)}} \quad & \sum_{i=1}^{k} c(x^{(i-1)}, x^{(i)}) \\ \text{s.t.} \quad & w_i^\top x^{(i)} \geq b_i \; \forall i \in [k]. \end{aligned} \tag{2}$$

Claim 3.6. Program (2) is convex in $(x^{(1)}, \ldots, x^{(k)})$.

In turn, we can solve the problem faced by an agent's computationally efficiently, through standard convex optimization techniques.

3.4. A Closed-Form Characterization in the 2-Classifier, 2-Dimensional Case

We now provide closed-form characterization of an agent's best response in the sequential case, under the two-dimensional two-classifier ($d = k = 2$) setting that we considered in Section 3.1. Here, we take the cost function to be the standard Euclidean norm, i.e. $c(x, \hat{x}) = \|\hat{x} - x\|_2$, as per Assumption 3.1.

Theorem 3.7. *Consider two linear classifiers $h_1(x) = \mathbb{1}(w_1^\top x \geq 0)$ and $h_2(x) = \mathbb{1}(w_2^\top x \geq 0)$ where $\|w_i\|_2 = 1$ for $i \in \{1, 2\}$ and an agent $x^{(0)} \in \mathbb{R}^2$ such that $h_1(x^{(0)}) = 0$ and $h_2(P_{w_1}(x^{(0)})) = 0$. Let $0 < \theta < \pi$ be the angle between (the positive region of) the two linear classifiers; i.e. θ is the solution to $\cos\theta = -w_1^\top w_2$. Then:*

1. *If $|\tan\theta| > \|P_{w_1}(x^{(0)})\|_2 / d_{w_1}(x^{(0)})$, then the best response for an agent is to pick $x^{(2)} = x^{(1)} = \vec{0}$. In this case, the cost of manipulation is $c^*_{seq}(x^{(0)}, \{h_1, h_2\}) = \|x^{(0)}\|_2$.*
2. *If $|\tan\theta| \leq \|P_{w_1}(x^{(0)})\|_2 / d_{w_1}(x^{(0)})$, then the best*

response is given by

$$x^{(1)} = \left(1 - \frac{d_{w_1}(x^{(0)})}{\|P_{w_1}(x^{(0)})\|_2} |\tan\theta|\right) P_{w_1}(x^{(0)})$$

and $x^{(2)} = P_{w_2}(x^{(1)})$*, and the cost of manipulation is given by*

$$c^*_{seq}\left(x^{(0)}, \{h_1, h_2\}\right) = d_{w_1}(x^{(0)})|\cos\theta| + \|P_{w_1}(x^{(0)})\|_2 \sin\theta.$$

The proof of this theorem is provided in the Appendix. First, note that once the first feature modification has happened and an agent has passed classifier h_1 and is at $x^{(1)}$, the theorem states that an agent picks $x^{(2)}$ to simply be the orthogonal projection onto the positive region of h_2. This is because the cost for going from $x^{(1)}$ to $x^{(2)}$ is simply the l_2 distance between them, in which case picking $x^{(2)}$ to be the orthogonal projection of $x^{(1)}$ on h_2 minimizes that distance. The main contribution and challenge of Theorem 3.7 are therefore to understand how to set $x^{(1)}$ and what is the minimum amount of effort that an agent expands to do so.

Now let's examine different cases in Theorem 3.7. Note that we assumed $h_1(x^{(0)}) = 0$ and $h_2(P_{w_1}(x^{(0)})) = 0$, i.e. that an agent is not in the positive region for the first test and $P_{w_1}(x^{(0)})$ is not in the positive region for the second test, because otherwise, the solution is trivial. In fact, if $h_1(x^{(0)}) = 1$, then the solution is simply staying at $x^{(0)}$ for the first test and then projecting orthogonally onto the positive region of h_2 to pass the second test:

$$x^{(1)} = x^{(0)},\ x^{(2)} = P_{w_2}(x^{(1)})$$
$$c^*_{seq}\left(x^{(0)}, \{h_1, h_2\}\right) = d_{w_2}(x^{(0)})$$

This corresponds to region R_1 of agents in Figure 4. If $h_1(x^{(0)}) = 0$, but $h_2(P_{w_1}(x^{(0)})) = 1$, then the best response solution is simply the orthogonal projection onto the positive region of h_1:

$$x^{(2)} = x^{(1)} = P_{w_1}(x^{(0)})$$
$$c^*_{seq}\left(x^{(0)}, \{h_1, h_2\}\right) = d_{w_1}(x^{(0)})$$

This corresponds to region R_4 of agents in Figure 4. Additionally, the first case in the closed-form solutions in Theorem 3.7 corresponds to the region of the space where agents prefer to travel directly to the intersection of the two classifiers than deploying a zig-zag strategy: this corresponds to region R_3 in Figure 4. The second case corresponds to the region where agents do find that a zig-zag strategy is less costly and gives the algebraic characterization of the optimal zig-zag strategy. This region for an agent is denoted by R_2 in Figure 4. Also, as shown by Figure 4.b, the zig-zag strategy of agents in R_2 has the following geometric characterization: pick $x^{(1)}$ on h_1 such that the line passing through $x^{(0)}$ and $x^{(1)}$ has angle θ with the line perpendicular to h_1.

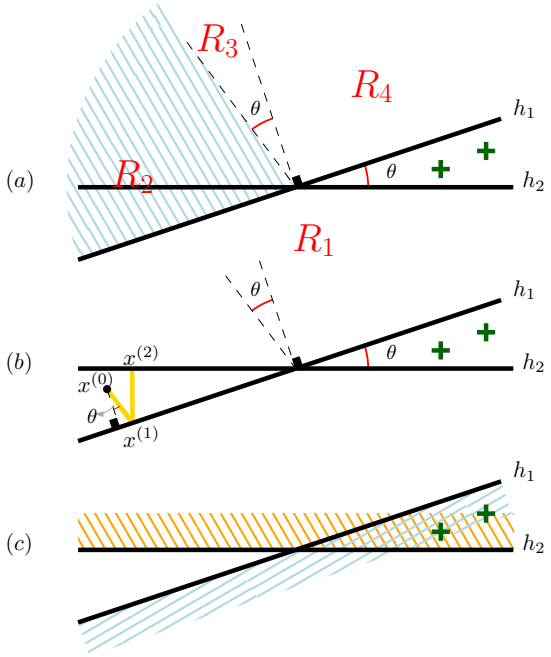

Figure 4. (a) Different cases for how agents best respond: agents in R_1 stay at their location to pass the first test and project onto h_2 to pass the second. Agents in R_2 deploy a zig-zag strategy. Agents in R_3 move to the intersection of h_1 and h_2. Agents in R_4 project onto h_1. (b) Geometric characterization of the zig-zag strategy: the line passing through $x^{(0)}$ and $x^{(1)}$ has angle θ with the line perpendicular to h_1. (c) This figure highlights the positive regions of h_1, h_2, and their intersection.

Given a budget τ, agents who can manipulate with a cost of at most τ to pass the two tests in the sequential setting, i.e. $\{x^{(0)} : c^*_{seq}\left(x^{(0)}, \{h_1, h_2\}\right) \leq \tau\}$ is highlighted in Figure 2.b.

We conclude this section by showing that if $\theta \geq \pi/2$, then agents incur the same cost in the sequential setting as they would under the conjunction setting. In other words, agents can deploy the strategy that they would use if they had to pass the two tests simultaneously. The proof of this theorem is provided in the Appendix.

Theorem 3.8. *If $\pi/2 \leq \theta < \pi$, then for every agent $x^{(0)}$ there exists optimal strategies $x^{(1)}$ and $x^{(2)}$ s.t. $x^{(1)} = x^{(2)}$, i.e., $c^*_{seq}\left(x^{(0)}, \{h_1, h_2\}\right) = c^*_{conj}\left(x^{(0)}, \{h_1, h_2\}\right)$.*

3.5. Monotonicity

We now consider a monotonicity property that excludes the possibility of a zig-zag strategy arising. A similar property is noted in (Milli et al., 2019).

Definition 3.9 (Feature Monotone Classifiers)**.** Classifier $h_i : \mathbb{R}^d \to \{0, 1\}$ is *monotone* if for every individual x that is classified as positive by h_i, any feature-wise increase in the features of x results in a positive classification by h_i. Formally,

$$\forall x \in \mathbb{R}^d : h_i(x) = 1 \Rightarrow h_i(x + \alpha) = 1 \quad \forall \alpha \in (\mathbb{R}_{\geq 0})^d.$$

Note that this monotonicity property may not hold in some classification problems. For example, when applying for a mortgage for $\$100,000$, presumably monotonically increasing income means one is more credit-worthy. However, if an individual reports a $\$3$ million a year income for a loan of $\$100,000$, such a large income could instead indicate fraudulent income reporting or remarkably poor financial planning since presumably such a high net worth individual should not need such a small loan.

In fact, in case $k = 2$, the angle θ measures the "alignment" between the classifiers. In the above example, the classifiers may not be aligned. Increases in income are desirable to show financial responsibility; yet, beyond a certain point (for example, when the income becomes much larger than the desired loan), income may become an indicator of poor financial planning or fraudulent transactions. In some hiring settings, having sufficient qualifications is desirable; yet, over-qualification can often be grounds for rejection of a job application.

Theorem 3.10. *Let $h_1, \ldots, h_k$ be a sequence of monotone classifiers, and let the initial feature vector $x^{(0)}$ be such that $h_i(x^{(0)}) = 0$ for every $i \in [k]$. Assume the cost function can be written as $c(x, \hat{x}) = \|\hat{x} - x\|$ for some norm $\|.\|$. Then, we have that*

$$c^*_{seq}\left(x^{(0)}, \{h_1, \ldots, h_k\}\right) = c^*_{conj}\left(x^{(0)}, \{h_1, \ldots, h_k\}\right).$$

Theorem 3.10 in particular implies that under our monotonicity assumption and for a large class of reasonable cost functions, an agent has no incentive to zig-zag in the sequential case and in fact can simply follow the same strategy as in the simultaneous or conjunctive case. This insight immediately extends even when $x^{(0)}$ is positively classified by some but not all of the h_i's as any best response is guaranteed to increase the feature values and thus will maintain the positive classification results of these classifiers.

3.6. Myopic or Greedy Strategy

A natural question that reader might have is how the cost of the zig-zag strategy compares to the cost of a greedy strategy that simply manipulates to the nearest passing point of the current test. One advantage of a greedy strategy is that an agent only needs to know what the next classifier they face is, rather than the entire screening pipeline in advance.

Given that the agent has full information about the pipeline, the zig-zag manipulation is by definition the optimal strategy and the greedy strategy can be sub-optimal. In the two-classifier two-dimensional case that we consider in our paper, our theorem states that the zig-zag manipulation is the unique optimal manipulation and that this manipulation is different from the greedy manipulation (see Figure 4(b)). In fact, for $k = 2$, the additive gap between the cost of the zig-zag strategy and the greedy strategy can be shown to be $(1 - \cos(\theta)) \cdot r$ where θ is the angle between the two classifiers and r is the distance of the agent from the first classifier.

One can also show that the gap is unbounded when k grows large: previous work (Friedman & Linial, 1993) shows an unbounded gap between the movement cost of being greedy and directly going to the closest point at the intersection of the half-spaces. Because the optimal zig-zag strategy cannot do worse than directly reaching this closest point, the gap between zig-zag and greedy is also unbounded.

4. Manipulation Resistant Defenses

Up to this point in the paper, we have focused mainly on the existence and feasibility of a zig-zag manipulation strategy from the perspective of an agent. We now shift gears and discuss the firm's decision space. We are interested in understanding how the firm can modify its classifiers to maintain a high level of accuracy (if possible), despite the strategic manipulations of an agent. To this end, we assume there is a joint distribution of features and labels $\mathcal{D}$ over $\mathcal{X} \times \{0, 1]\}$. Interestingly, previous works (Brückner & Scheffer, 2011; Hardt et al., 2016) show hardness results for finding optimal strategic classifiers, where the objective is finding a single classifier h that attains the strategic maximum accuracy.

Now, we can introduce the defender's game for a typical strategic classification problem.

$$\begin{aligned} \min_{h \in \mathcal{H}} \quad & P_{(x,y)\sim\mathcal{D}}[h(z^*(x)) \neq y] \\ \text{s.t.} \quad & z^*(x) = \arg\max_z \; h(z) - c(x, z) \end{aligned} \tag{3}$$

In our paper, h is actually given by the sequential composition of classifiers in the screening process and $c(x, z)$ is the sum of manipulation costs per stage. The objective function in this optimization problem is a direct generalization of 0-1 loss for normal learning problems, only complicated by the strategic behavior of an agent.

As Brückner & Scheffer (2011) observe, this is a bi-level optimization problem and is NP-hard (Jeroslow, 1985) to compute, even when constraints and objectives are linear. Interestingly, Hardt et al. (2016) also show a hardness of approximation result for general metrics. Because of these past hardness results, we instead focus on a more tractable defense objective.

4.1. Conservative Defense

Here, we consider a different objective motivated by the hiring process in firms, in which avoiding false positives and not hiring unqualified candidates can be seen as arguably more important than avoiding false negatives and not missing out on good candidates. This objective, described below, has been previously studied in the context of strategic classification, in particular in (Ahmadi et al., 2022).

Definition 4.1 (No False Positive Objective)**.** Given the manipulation budget τ and the initial linear classifiers $h_1, \cdots, h_k$, the goal of the firm is to design a modified set of linear classifiers $\tilde{h}_1, \cdots, \tilde{h}_k$ that maximize the true positive rate of the pipeline on manipulated feature vectors subject to no false positives. Recall that the ground truth is determined by the conjunction of $h_1, \cdots, h_k$ on unmanipulated feature vectors of agents.

Without loss of generality, we assume the pipeline is non-trivial: the intersection of acceptance regions of $h_1, \cdots, h_k$ is non-empty.

We prove that, under standard assumptions on linear classifiers of the firm, a defense strategy that "shifts" all classifiers by the manipulation budget, is the optimal strategy for the firm in both pipeline and conjunction settings. We formally define the defense strategy as follows:

Definition 4.2 (Conservative Strategy)**.** Given the manipulation budget τ, the firm conservatively assumes that each agent has a manipulation budget of τ per test. For each test $h_i(x) = \mathbb{1}[w_i^\top x \geq b_i]$, the firm replaces it by a "τ-shifted" linear separator $\tilde{h}_i(x) = \mathbb{1}[w_i^\top x \geq b_i + \tau]$). In this section, without loss of generality, we assume that all w_i's have ℓ_2-norm equal to one.

Our statement holds when the linear classifiers satisfy the following "general position" type condition.

Definition 4.3. We say a collection of linear classifiers $\mathcal{H} = \{h_1(x) = \mathbb{1}[w_1^\top x \geq b_1], \cdots, h_k(x) = \mathbb{1}[w_k^\top x \geq b_k]\}$ with $w_1, \cdots, w_k \in \mathbb{R}^d$ are in "general position" if for any $i \in [k]$, the intersection of $\{x | w_i^\top x = b_i\}$ and $\{x | \bigwedge_{j \in [k], j \neq i} h_j(x) = 1\}$ lies in a $(d-1)$-dimensional subspace but in no $(d-2)$-dimensional subspace. In $\mathbb{R}^2$, this condition is equivalent to the standard general position assumption (i.e., no three lines meet at the same point). Moreover, this condition implies that no test in $\mathcal{H}$ is "redundant", i.e., for every $i \in [k]$, the positive region of $\mathcal{H}$ (i.e., $\bigwedge_{h \in \mathcal{H}} \{x | h(x) = 1\}$) is a proper subset of the positive region of $\mathcal{H} \setminus h_i$. See Figure 5 for an example in $\mathbb{R}^2$.

Now, we are ready to state the main result of this section.

Theorem 4.4. *Consider a set of linear classifiers $\mathcal{H} = \{h_1, \cdots, h_k\}$ that are in "general position" (as in Definition 4.3). Moreover, suppose that each agent has a manipulation budget of τ. Then, in both the conjunction and sequential settings, the conservative defense is a strategy that maximizes true positives subject to zero false positives.*

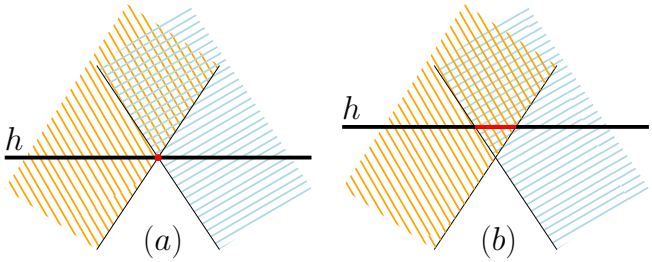

Figure 5. In (a), the intersection of h with the positive half plane of the other two classifiers that are in blue and gray shadows is a point which is of zero dimension. This case is not in the general position and h is a redundant classifier. However, in (b), the intersection of h with the described positive regions is a line segment, a one-dimensional object. Here, h is not redundant.

The proof is provided in Appendix D.1. Note that while the conservative defense strategy has the maximum possible true positive subject to zero false positive in both simultaneous and sequential settings, by Claim 3.3, the conservative defense achieves a higher true positive rate in the sequential setting compared to the simultaneous case. Informally, from the firm's point of view, *under manipulation, the sequential setting is a more efficient screening process*.

5. Discussion

We have initiated the study of *Strategic Screening*, combining screening problems with strategic classification. This is a natural and wide-spread problem both in automated and semi-automated decision making. We believe these examples and our convex program can aid in the design and monitoring of these screening processes. Substantial open questions remain regarding fairness implications (Appendix B) of the defender's solution and exactly how susceptible real world pipelines are to zig-zagging.

Acknowledgements

We would like to thank Avrim Blum, Saba Ahmadi, and the reviewers for their helpful comments on earlier drafts of this paper. Kevin Stangl was supported in part by the National Science Foundation under grant CCF-2212968, by the Simons Foundation under the Simons Collaboration on the Theory of Algorithmic Fairness, and by the Defense Advanced Research Projects Agency under cooperative agreement HR00112020003. The views expressed in this work do not necessarily reflect the position or the policy of the Government and no official endorsement should be inferred. Approved for public release; distribution is unlimited.

References

Ahmadi, S., Beyhaghi, H., Blum, A., and Naggita, K. The strategic perceptron. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pp. 6–25, 2021.

Ahmadi, S., Beyhaghi, H., Blum, A., and Naggita, K. On classification of strategic agents who can both game and improve. In *Symposium on Foundations of Responsible Computing (FORC)*, volume 218, pp. 3:1–3:22, 2022.

Argue, C., Gupta, A., Tang, Z., and Guruganesh, G. Chasing convex bodies with linear competitive ratio. *Journal of the ACM (JACM)*, 68(5):1–10, 2021.

Arunachaleswaran, E. R., Kannan, S., Roth, A., and Ziani, J. Pipeline interventions. *Mathematics of Operations Research*, 2022.

Bansa, N., Böhm, M., Eliáš, M., Koumoutsos, G., and Umboh, S. W. Nested convex bodies are chaseable. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1253–1260. SIAM, 2018.

Bechavod, Y., Ligett, K., Wu, S., and Ziani, J. Gaming helps! learning from strategic interactions in natural dynamics. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1234–1242, 2021.

Bechavod, Y., Podimata, C., Wu, S., and Ziani, J. Information discrepancy in strategic learning. In *International Conference on Machine Learning (ICML)*, pp. 1691–1715, 2022.

Björkegren, D., Blumenstock, J. E., and Knight, S. Manipulation-proof machine learning. *arXiv preprint arXiv:2004.03865*, 2020.

Blum, A., Stangl, K., and Vakilian, A. Multi stage screening: Enforcing fairness and maximizing efficiency in a pre-existing pipeline. In *Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 1178–1193, 2022.

Bower, A., Kitchen, S. N., Niss, L., Strauss, M. J., Vargas, A., and Venkatasubramanian, S. Fair pipelines. *CoRR*, abs/1707.00391, 2017.

Braverman, M. and Garg, S. The role of randomness and noise in strategic classification. In *Foundations of Responsible Computing (FORC)*, volume 156 of *LIPIcs*, pp. 9:1–9:20, 2020.

Brückner, M. and Scheffer, T. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 547–555, 2011.

Bubeck, S., Lee, Y. T., Li, Y., and Sellke, M. Competitively chasing convex bodies. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pp. 861–868, 2019.

Bubeck, S., Klartag, B., Lee, Y. T., Li, Y., and Sellke, M. Chasing nested convex bodies nearly optimally. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1496–1508. SIAM, 2020.

Cesa-Bianchi, N., Dekel, O., and Shamir, O. Online learning with switching costs and other adaptive adversaries. *Advances in Neural Information Processing Systems*, 26, 2013.

Chen, Y., Podimata, C., Procaccia, A. D., and Shah, N. Strategyproof linear regression in high dimensions. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 9–26, 2018.

Chen, Y., Liu, Y., and Podimata, C. Learning strategy-aware linear classifiers. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:15265–15276, 2020a.

Chen, Y., Wang, J., and Liu, Y. Strategic recourse in linear classification. *arXiv preprint arXiv:2011.00355*, 2020b.

Cohen, L., Lipton, Z. C., and Mansour, Y. Efficient candidate screening under multiple tests and implications for fairness. In *1st Symposium on Foundations of Responsible Computing, FORC 2020, June 1-3, 2020*, 2020.

Cummings, R., Ioannidis, S., and Ligett, K. Truthful linear regression. In *Conference on Learning Theory*, pp. 448–483. PMLR, 2015.

Dee, T. S., Dobbie, W., Jacob, B. A., and Rockoff, J. The causes and consequences of test score manipulation: Evidence from the new york regents examinations. *American Economic Journal: Applied Economics*, 11(3):382–423, July 2019. doi: 10.1257/app.20170520.

Dekel, O., Fischer, F., and Procaccia, A. D. Incentive compatible regression learning. *Journal of Computer and System Sciences*, 76(8):759–777, 2010.

Dong, J., Roth, A., Schutzman, Z., Waggoner, B., and Wu, Z. S. Strategic classification from revealed preferences. In *Conference on Economics and Computation*, pp. 55–70, 2018.

Dwork, C. and Ilvento, C. Fairness under composition. In *10th Innovations in Theoretical Computer Science Conference (ITCS 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

Dwork, C., Ilvento, C., and Jagadeesan, M. Individual fairness in pipelines. In *1st Symposium on Foundations of Responsible Computing*, 2020.

Friedman, J. and Linial, N. On convex body chasing. *Discrete & Computational Geometry*, 9(3):293–321, 1993.

Ghalme, G., Nair, V., Eilat, I., Talgam-Cohen, I., and Rosenfeld, N. Strategic classification in the dark. In *International Conference on Machine Learning*, pp. 3672–3681. PMLR, 2021.

Guan, Y., Pan, L., Shishika, D., and Tsiotras, P. Chasing convex bodies generated by an adversary. *arXiv preprint arXiv:2209.13606*, 2022.

Haghtalab, N., Immorlica, N., Lucier, B., and Wang, J. Z. Maximizing welfare with incentive-aware evaluation mechanisms. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pp. 160–166, 2020.

Hardt, M., Megiddo, N., Papadimitriou, C., and Wootters, M. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pp. 111–122, 2016.

Harris, K., Heidari, H., and Wu, S. Z. Stateful strategic regression. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:28728–28741, 2021.

Hu, L., Immorlica, N., and Vaughan, J. W. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 259–268, 2019.

Jagadeesan, M., Mendler-Dünner, C., and Hardt, M. Alternative microfoundations for strategic classification. In *International Conference on Machine Learning*, pp. 4687–4697. PMLR, 2021.

Jeroslow, R. G. The polynomial hierarchy and a simple model for competitive analysis. *Math. Program.*, 32 (2):146–164, 1985. doi: 10.1007/BF01586088. URL https://doi.org/10.1007/BF01586088.

Khajehnejad, M., Tabibian, B., Schölkopf, B., Singla, A., and Gomez-Rodriguez, M. Optimal decision making under strategic behavior. *arXiv preprint arXiv:1905.09239*, 2019.

Kleinberg, J. and Raghavan, M. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)*, 8(4):1–23, 2020.

Klivans, A. R. and Servedio, R. A. Learning intersections of halfspaces with a margin. In *Learning Theory: 17th Annual Conference on Learning Theory, COLT 2004, Banff, Canada, July 1-4, 2004. Proceedings 17*, pp. 348–362. Springer, 2004.

Klivans, A. R. and Sherstov, A. A. Cryptographic hardness for learning intersections of halfspaces. *Journal of Computer and System Sciences*, 75(1):2–12, 2009.

Li, Y., Qu, G., and Li, N. Online optimization with predictions and switching costs: Fast algorithms and the fundamental limit. *IEEE Transactions on Automatic Control*, 2021.

Liu, L. T., Wilson, A., Haghtalab, N., Kalai, A. T., Borgs, C., and Chayes, J. The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 381–391, 2020.

Meir, R., Procaccia, A. D., and Rosenschein, J. S. On the limits of dictatorial classification. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, pp. 609–616, 2010.

Meir, R., Almagor, S., Michaely, A., and Rosenschein, J. S. Tight bounds for strategyproof classification. In *10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2011), Taipei, Taiwan, May 2-6, 2011, Volume 1-3*, pp. 319–326, 2011.

Meir, R., Procaccia, A. D., and Rosenschein, J. S. Algorithms for strategyproof classification. *Artificial Intelligence*, 186:123–156, 2012.

Miller, J., Milli, S., and Hardt, M. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*, pp. 6917–6926. PMLR, 2020.

Milli, S., Miller, J., Dragan, A. D., and Hardt, M. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 230–239, 2019.

Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. Performative prediction. In *International Conference on Machine Learning*, pp. 7599–7609. PMLR, 2020.

Perote, J. and Perote-Pena, J. Strategy-proof estimators for simple regression. *Mathematical Social Sciences*, 47(2): 153–176, 2004.

Sellke, M. Chasing convex bodies optimally. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1509–1518. SIAM, 2020.

Shavit, Y., Edelman, B., and Axelrod, B. Causal strategic linear regression. In *International Conference on Machine Learning (ICML)*, pp. 8676–8686, 2020.

Shi, G., Lin, Y., Chung, S.-J., Yue, Y., and Wierman, A. Online optimization with memory and competitive control. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020.

Tang, W., Ho, C.-J., and Liu, Y. Linear models are robust optimal under strategic behavior. In *International Conference on Artificial Intelligence and Statistics*, pp. 2584–2592. PMLR, 2021.

Ustun, B., Spangher, A., and Liu, Y. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 10–19, 2019.

A. Broader Impacts Analysis

The ICML 2023 Paper Guidelines at https://icml.cc/Conferences/2023/PaperGuidelines has a checklist of best paper practices intended to promote responsible machine learning research. Most of these factors are either not applicable to our work or we clearly satisfy these requirements.

This work is primarily theoretical. Throughout the paper, we state full assumptions of all theoretical results and provide complete proofs. The main questions we want to discuss in this appendix are possible negative social impacts and the limitations of our work.

A.1. Social Impacts

Social impacts related to strategic screening likely exist in real systems, since it seems probable variants of the zig-zag strategy are in use by people in the wild. Our work makes the existence of such a strategy clear, and introduces some initial approaches to mitigate such a manipulation strategy.

This line of research could clarify how to design screening processes which are more resistant to strategic manipulation. This may perhaps help avoid some of the well-known fairness harms of strategic classification e.g. (Milli et al., 2019; Hu et al., 2019) due to disparate abilities to manipulate features across different groups. We believe such considerations to be a promising and important avenue for future research.

A.2. Limitations of Our Work

We discussed two key limitations briefly in the main body of the paper. One limitation comes in that since our paper is game theoretic in nature, we need to provide a model of the utility functions of the firm/agents and the manipulation cost. The current assumptions we make on these functions are consistent with other papers in the strategic classification literature, but may not fully reflect practical phenomena and considerations. An interesting direction on this front would be to study wider classes of model in future work, and to validate assumptions on costs and utilities using real data. Another limitation is the full information assumption that agents perfectly understand and can perfectly reason about the classifiers; it may be of interest to extend our result to models of incomplete understanding of or ability to reason about the firm's decision rule such as those of (Jagadeesan et al., 2021; Ghalme et al., 2021; Bechavod et al., 2022).

B. Fairness and Strategic Screening

Some of the works cited in the related work section consider fairness considerations in the space of strategic manipulation, stemming either from unequal abilities to manipulate (Milli et al., 2019; Hu et al., 2019) or unequal access to information about the classifiers (Bechavod et al., 2022) across different groups. We do not consider these connections in our work, but these considerations are of significant interest and a natural direction for further research, especially due to the importance of making fair decision in high-stake, life altering contexts. We finish with a few interesting examples for this.

Disparities might arise both in the conjunction and in the sequential setting, with or without defense. consider the classifiers presented in Example 3.2 and an instance in which candidates belong to two groups, G^1 and G^2 with initial feature vector distributed identically and characterized by different total manipulation budgets, $\sqrt{2} = \tau^2 > \tau^1 = 5/4$. The narrative of the fairness disparities in the conjunction case is a simple generalization of the single classifiers case (e.g., (Hardt et al., 2016))- If the distribution is such that a significant fraction of individuals (from both groups) starts at a feature vector that is classified by both classifiers as 0 and that requires $\sqrt{2}$ manipulation cost to reach their intersection— only the individuals form G_2 will be able to manipulate. For the sequential case, consider a distribution with a large enough fraction of individuals starting at $(0, 0)$. Example 3.2 demonstrates that only individuals from G_2 will have sufficient budget to manipulate (using the zig-zag strategy). If the firm applies the conservative defense, individuals from G_1 that should have been classified as positive might not have sufficient budget to manipulate their way to acceptance, which in turn implies higher false negative rates. This indicates, similarly to prior results in strategic classification (e.g., (Hu et al., 2019)), how the members of the advantaged group are more easily admitted or hired.

C. Proofs of Section 3

The following is a restatement of Claim 3.3.

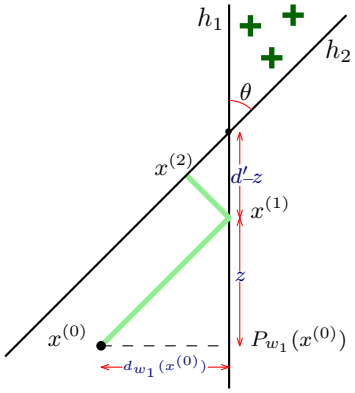

Figure 6. This figure shows how we reduced the optimization problem in Equation 4 to the one in Equation 5.

Claim C.1. Let $h_1, \ldots, h_k$ be a sequence of k linear classifiers. For any agent with initial feature vector $x \in \mathbb{R}^d$ ($d \geq 1$), $c^*_{conj}(x, \{h_1, \ldots, h_k\}) \geq c^*_{seq}(x, \{h_1, \ldots, h_k\})$.

Proof of Claim 3.3. Let c be the agent's cost function. Let $\hat{x}$ be a vector such that $h_i(\hat{x}) = 1$ for all $i \in [k]$, and such that $c(x, \hat{x}) \leq \tau$ where τ is the manipulation budget available to the agent. Since $\hat{x}$ satisfies $h_i(\hat{x}) = 1$ for all $i \in [k]$, the feature modification $x \to \hat{x}$ gives a positive classification outcome to the agent in the sequential case. Further, the cost of this manipulation is $c(x, \hat{x}) + 0 + \ldots + 0 = c(x, \hat{x})$. In turn, for any feasible one-shot manipulation that passes all classifiers in the conjunctive case, there exists a feasible sequential manipulation that passes all classifiers in the sequential case which could be of a lower cost; this concludes the proof. □

Theorem 3.7. *Consider two linear classifiers $h_1(x) = \mathbb{1}(w_1^\top x \geq 0)$ and $h_2(x) = \mathbb{1}(w_2^\top x \geq 0)$ where $\|w_i\|_2 = 1$ for $i \in \{1, 2\}$ and an agent $x^{(0)} \in \mathbb{R}^2$ such that $h_1(x^{(0)}) = 0$ and $h_2(P_{w_1}(x^{(0)})) = 0$. Let $0 < \theta < \pi$ be the angle between (the positive region of) the two linear classifiers; i.e. θ is the solution to $\cos\theta = -w_1^\top w_2$. Then:*

1. *If $|\tan\theta| > \|P_{w_1}(x^{(0)})\|_2 / d_{w_1}(x^{(0)})$, then the best response for an agent is to pick $x^{(2)} = x^{(1)} = \vec{0}$. In this case, the cost of manipulation is $c^*_{seq}\left(x^{(0)}, \{h_1, h_2\}\right) = \|x^{(0)}\|_2$.*

2. *If $|\tan\theta| \leq \|P_{w_1}(x^{(0)})\|_2 / d_{w_1}(x^{(0)})$, then the best response is given by*

$$x^{(1)} = \left(1 - \frac{d_{w_1}(x^{(0)})}{\|P_{w_1}(x^{(0)})\|_2} |\tan\theta|\right) P_{w_1}(x^{(0)})$$

and $x^{(2)} = P_{w_2}(x^{(1)})$, and the cost of manipulation is given by

$$\begin{aligned} &c^*_{seq}\left(x^{(0)}, \{h_1, h_2\}\right) \\ &= d_{w_1}(x^{(0)}) |\cos\theta| + \|P_{w_1}(x^{(0)})\|_2 \sin\theta. \end{aligned}$$

Proof of Theorem 3.7. Given classifiers h_1 and h_2, the best response of an agent $x^{(0)}$ is a solution to the following optimization problem, as noted in Section 3.3:

$$\begin{aligned} &c^*_{seq}\left(x^{(0)}, \{h_1, h_2\}\right) \\ &= \min_{x^{(1)}, x^{(2)}} \left\{ \|x^{(0)} - x^{(1)}\|_2 + \|x^{(1)} - x^{(2)}\|_2 : w_1^\top x^{(1)} \geq 0, w_2^\top x^{(2)} \geq 0 \right\} \end{aligned}$$

First, we remark that given any $x^{(1)}$, the optimal choice of $x^{(2)}$ is the orthogonal projection of $x^{(1)}$ on classifier f_2. Therefore, the best response can be written as:

$$c^*_{seq}\left(x^{(0)}, \{h_1, h_2\}\right) = \min_{x^{(1)} \in \mathbb{R}^2} \left\{ \|x^{(0)} - x^{(1)}\|_2 + d_{w_2}\left(x^{(1)}\right) : w_1^\top z \geq 0 \right\} \tag{4}$$

To simplify notations, we will denote $x \triangleq x^{(0)}$. Under the assumptions of the theorem (more specifically, $h_1(x) = 0$ and $h_2(P_{w_1}(x)) = 0$), Equation (4) can be rewritten as an optimization over a one-dimensional variable:

$$\min_{0 \le z \le d'_{w_1}(x)} \left\{ g(z) \triangleq \sqrt{d^2_{w_1}(x) + z^2} + (d'_{w_1}(x) - z) \sin\theta \right\} \tag{5}$$

where $d'_{w_1}(x) \triangleq \|P_{w_1}(x)\|_2$ – see Figure 6 for a graphical justification of this rewriting. Note that $g(z)$ achieves its minimum either at the boundaries or at the point where $g'(z) = 0$. Therefore, we have that the minimum is one of the following:

$$\begin{aligned}
&z = 0 \Longrightarrow g(z) = d_{w_1}(x) + d'_{w_1}(x) \sin\theta \\
&z = d'_{w_1}(x) \Longrightarrow g(z) = \sqrt{d^2_{w_1}(x) + d'^2_{w_1}(x)} = \|x\|_2 \\
&z = d_{w_1}(x) |\tan\theta| \Longrightarrow g(z) = d_{w_1}(x) \cos\theta + d'_{w_1}(x) \sin\theta \ (g'(z) = 0)
\end{aligned}$$

We can show that if $|\tan\theta| > d'_{w_1}(x)/d_{w_1}(x)$, then the minimizer $z^\star = d'_{w_1}(x)$, meaning $x^{(2)} = x^{(1)} = \vec{0}$, and that

$$c^*_{seq}(x, \{h_1, h_2\}) = \|x\|_2$$

and if $|\tan\theta| \le d'_{w_1}(x)/d_{w_1}(x)$, then the minimizer $z^\star = d_{w_1}(x)|\tan\theta|$ which implies

$$x^{(1)} = \left(1 - \frac{d_{w_1}(x^{(0)})}{\|P_{w_1}(x^{(0)})\|_2} |\tan\theta|\right) P_{w_1}(x^{(0)})$$

and $x^{(2)} = P_{w_2}(x^{(1)})$, and that

$$c^*_{seq}(x, \{h_1, h_2\}) = d_{w_1}(x)|\cos\theta| + d'_{w_1}(x)\sin\theta$$

Therefore, putting the two cases together,

$$c^*_{seq}(x, \{h_1, h_2\}) = \begin{cases} \|x\|_2 & \text{if } |\tan\theta| > d'_{w_1}(x)/d_{w_1}(x) \\ d_{w_1}(x)|\cos\theta| + d'_{w_1}(x)\sin\theta & \text{if } |\tan\theta| \le d'_{w_1}(x)/d_{w_1}(x) \end{cases}$$

□

Theorem 3.8. *If $\pi/2 \le \theta < \pi$, then for every agent $x^{(0)}$ there exists optimal strategies $x^{(1)}$ and $x^{(2)}$ s.t. $x^{(1)} = x^{(2)}$, i.e., $c^*_{seq}\left(x^{(0)}, \{h_1, h_2\}\right) = c^*_{conj}\left(x^{(0)}, \{h_1, h_2\}\right)$.*

Proof. Let $(x^{(1)}, x^{(2)} = P_{w_2}(x^{(1)}))$ be an optimal strategy of the agent in the sequential setting. Suppose $x^{(1)} \neq x^{(2)}$. We have that

$$\begin{aligned}
w_1^\top x^{(2)} &= w_1^\top \left(x^{(1)} - (w_2^\top x^{(1)}) w_2\right) \\
&= w_1^\top x^{(1)} - (w_2^\top x^{(1)})(w_1^\top w_2)
\end{aligned}$$

But note that $w_1^\top x^{(1)} \ge 0$ because $x^{(1)}$ passes the first classifier by definition, $w_2^\top x^{(1)} \le 0$ because $x^{(1)} \neq x^{(2)}$, and $w_1^\top w_2 \ge 0$ because $\pi/2 \le \theta < \pi$. Therefore, $w_1^\top x^{(2)} \ge 0$ which implies $h_1(x^{(2)}) = 1$. However, if $h_1(x^{(2)}) = 1$, then the following manipulation: $y^{(0)} = x^{(0)}$ and $y^{(1)} = y^{(2)} = x^{(2)}$ passes both tests and that its cost is: $\|x^{(2)} - x^{(0)}\|_2 \le \|x^{(2)} - x^{(1)}\|_2 + \|x^{(1)} - x^{(0)}\|_2$ by the triangle inequality. Given the optimality of $(x^{(1)}, x^{(2)})$, we conclude that $(y^{(1)}, y^{(2)})$ is another optimal strategy that the agent can deploy. □

Theorem 3.10. *Let $h_1, \ldots, h_k$ be a sequence of monotone classifiers, and let the initial feature vector $x^{(0)}$ be such that $h_i(x^{(0)}) = 0$ for every $i \in [k]$. Assume the cost function can be written as $c(x, \hat{x}) = \|\hat{x} - x\|$ for some norm $\|.\|$. Then, we have that*

$$c^*_{seq}\left(x^{(0)}, \{h_1, \ldots, h_k\}\right) = c^*_{conj}\left(x^{(0)}, \{h_1, \ldots, h_k\}\right).$$

Proof. Let $f_{1,\ldots,k} : \mathbb{R}^d \to \{0,1\}$ denote the function that returns the conjunction of all the classifiers, i.e., $f_{1,\ldots,k}(x) = h_1(x) \wedge \ldots \wedge h_k(x)$.

Let $z^*_{1,\ldots,k}(x^0)$ denote the point on $f_{1,\ldots,k}$ that minimizes the cost, i.e., $z^*_{1,\ldots,k}(x^0) = \text{argmin}_{x^{(1)}} \|x^{(0)} - x^{(1)}\|_p$. Note that by definition, points on $f_{1,\ldots,k}$ are classified as positive by all classifiers $h_1, \ldots, h_k$ (i.e., $z^*_{1,\ldots,k}(x^0)$ this is the best response for the conjunction case).

It follows from the triangle inequality that any $x^{(1)}$ such that $h_1(x^{(1)}) \wedge \ldots \wedge h_k(x^{(1)}) = 1$ has cost $c(x^{(0)}, x^{(1)}) \geq c(x^{(0)}, z^*_{1,\ldots,k}(x^0))$.

We proceed by induction on the number of classifiers. For the induction base, consider $k = 1$. Clearly, in this case moving to $z^*_{1,\ldots,k}(x)$ yields the best response.

For the induction step, assume that for every initial point x', and every $k-1$ monotone classifiers $h_2, \ldots, h_k$ it holds that

$$\|x' - z^*_{2,\ldots,k}(x')\|_p \leq \|x' - z_2\|_2 + \ldots + \|z_{k-1} - z_k\|_p.$$

for every $z_2, \ldots, z_k \in \mathbb{R}^d$ such that $h_i(z_i) = 1$.

Adding the additional classifier in the beginning, h_1 and considering the initial point, x. Assume by contradiction that there exists a path $x = z_0, z_1 \ldots, z_k$ such that $h_i(z_i) \geq 0$ for every $i \in [k]$ and that

$$\begin{aligned} c^*_{seq}(x, \{h_1, \ldots, h_k\}) &= \|x - z_1\|_p + \ldots + \|z_{k-1} - z_k\|_p \\ &< \|x - z^*_{1,\ldots,k}(x)\|_p. \end{aligned} \tag{6}$$

Since the path from z_1 to z_k is a best response for $h_2, \ldots, h_k$ when the initial feature vector z_1, by setting $x' = z_1$ we can apply the induction step we and replace this path by $x, z_1, z^*_{2,\ldots,k}(x')$ without increasing the sum of manipulations. If $f_{1,\ldots,k}(z^*_{2,\ldots,k}(z_1)) = 1$, we have that $\|x - z_1\|_p + \|z_1 - z^*_{2,\ldots,k}(z_1)\|_p \leq \|x - z^*_{1,\ldots,k}(x)\|_p$ due to the triangle inequality and the definition of $z^*_{1,\ldots,k}(x)$ and this is a contradiction to Eq. 6.

So assume $f_{1,\ldots,k}(z^*_{2,\ldots,k}(z_1)) = 0$. Since $h_i(z^*_{2,\ldots,k}(z_1)) = 1$ for every $i \geq 2$ by definition, we have that $h_1(z^*_{2,\ldots,k}) = 0$. As $h_1(z_1) = 1$, we can define $z' \in \mathbb{R}^d$ such that

$$z'[j] = \max\{z^*_{2,\ldots,k}(z_1)[j], z_1[j]\},$$

and from monotonicity it follows that $f_{2,\ldots,k}(z') = 1$.

Finally, we have that $\|x - z_1\|_p + \|z_1 - z'\|_p < \|x - z_1\|_p + \|z_1 - z^*_{2,\ldots,k}(z_1)\|_p$, which is a contradiction to the minimiality of $z^*_{2,\ldots,k}(z_1)$ and thus to the minimality of $z_2, \ldots, z_k$. □

D. Proofs of Section 4

D.1. Conservative Defense Proofs

Theorem 4.4. *Consider a set of linear classifiers $\mathcal{H} = \{h_1, \cdots, h_k\}$ that are in "general position" (as in Definition 4.3). Moreover, suppose that each agent has a manipulation budget of τ. Then, in both the conjunction and sequential settings, the conservative defense is a strategy that maximizes true positives subject to zero false positives.*

Proof of Theorem 4.4. First, we prove that conservative defense achieve zero false positive in both cases. To show this, by Claim 3.3, it suffices to show it for the sequential setting only. Consider an agent x who initially (i.e., before manipulation) is not in the positive region of conjunctions of $h_1, \cdots h_k$; i.e., $\Pi_{j\in[k]} h_j(x^{(0)}) = 0$. Hence, there exists a classifier h_i such that $w_i^\top x^{(0)} < b_i$. Now, let $x^{(i)} : x^{(0)} + \epsilon_i$ denote the (manipulated) location of x right before stage i. Since the total manipulation budget of x is τ, $w_i^\top x^{(i)} \leq w_i^\top x^{(0)} + w_i^\top \epsilon_i < b_i + \tau$ (the choice of ε_i that maximizes $w_i^\top \epsilon_i$ is $\epsilon_i = \tau w_i$, and $w_i^\top (\tau w_i) = \tau$ since $\|w_i\|_2 = 1$). Hence, $\tilde{h}(x^{(i)}) = 0$ and agent x cannot pass the modified pipeline $\tilde{h}_1, \cdots, \tilde{h}_k$.

Next, consider test i and let Δ^i denote the subspace of points (i.e., agents) in the intersection of $\{x | h_i(x) = 0\}$ and $\bigwedge_{j\in[k], j\neq i}\{x | h_j(x) = 1\}$. By the general position assumption, Δ^i is a $(d-1)$-dimensional subspace and is a subset of the $(d-1)$-dimensional hyperplane corresponding to $w_i^\top x = b_i$. Then, there exists only a unique linear separator which is at distance exactly τ from Δ^i (and is in the positive side of h_i); $\hat{h}_i(x) := \mathbb{1}[w_i^\top x \geq b_i + \tau]$. Given that any defense strategy

with zero false positive has to classify an agents in Δ^i as negative, it is straightforward to verify that any "feasible" modified linear separator h_i' (i.e., achieving zero false positive) results in true positive rate less than or equal to the one replaces h_i' with $\hat{h}_i$. □