Classification with Few Tests through Self-Selection

Hanrui Zhang,[1] **Yu Cheng,** [2] **Vincent Conitzer** [1]
[1] Duke University
[2] University of Illinois at Chicago
hrzhang@cs.duke.edu, yucheng2@uic.edu, conitzer@cs.duke.edu

Abstract

We study test-based binary classification, where a principal either accepts or rejects agents based on the outcomes they get in a set of tests. The principal commits to a policy, which consists of all sets of outcomes that lead to acceptance. Each agent is modeled by a distribution over the space of possible outcomes. When an agent takes a test, he pays a cost and receives an independent sample from his distribution as the outcome. Agents can always choose between taking another test and stopping. They maximize their expected utility, which is the value of acceptance if the principal's policy accepts the set of outcomes they have and 0 otherwise, minus the total cost of tests taken.

We focus on the case where agents can be either "good" or "bad" (corresponding to their distribution over test outcomes), and the principal's goal is to accept good agents and reject bad ones. We show, roughly speaking, that as long as the good and bad agents have different distributions (which can be arbitrarily close to each other), the principal can always achieve perfect accuracy, meaning good agents are accepted with probability 1, and bad ones are rejected with probability 1. Moreover, there is a policy achieving perfect accuracy under which the maximum number of tests any agent needs to take is constant — in sharp contrast to the case where the principal directly observes samples from agents' distributions. The key technique is to choose the policy so that agents self-select into taking tests.

1 Introduction

We often classify based on the outcomes of *tests*. In a narrow sense, tests can take the form of exams, with numerical scores as *outcomes*. For example, a course often has one or more midterm exams and one final exam, and the intructor uses the outcomes of these exams to decide the final grades of (i.e., to *classify*) students. More generally, a test can be any activity that takes a certain amount of effort and produces a verifiable outcome. Examples include job interviews, research paper submissions, etc. These outcomes, presumably correlated with the true skills of the test takers (henceforth the *agents*), are then used by a *principal* to classify them — the collective feedback from different interviewers determines whether the interviewee gets the job, and the list of papers one has published strongly affects one's future opportunities as a researcher.

An agent's performance on tests is inevitably random — on any given day, a capable student may not perform well due to being tired or sick, due to bad luck in which questions were selected, or for reasons that we cannot identify. For this reason, a principal generally is willing to take into consideration multiple test outcomes when making decisions. It then matters how these tests are offered to agents. Oversimplifying, there are two ways of offering tests: *mandatory* tests and *optional* tests. With mandatory tests, the principal decides which tests each agent should take and/or how many times they should take them, as well as which (combinations of) outcomes an agent needs to have in order to be classified into a certain category. A straightforward example of mandatory tests is students taking exams in school, where typically all students are required to take all exams in a course, whose outcomes together determine the final grade of the student. On the other hand, with optional tests, the principal decides the latter (i.e., which outcomes suffice for classification into a certain category) but not the former (i.e., which and/or how many tests each agent should take). One example is (an oversimplified version of) the academic job market, where agents' publication records determine whether they are invited for an onsite interview, but agents can decide how often to put in the effort to prepare a new paper for submission to a conference or journal (i.e., to "take" and optional "test"). At first glance, it may appear that mandatory tests allow the principal tighter control over the classification process, and therefore would benefit the principal more than optional tests. As a consequence, the principal should enforce mandatory tests whenever possible (or economically feasible). However, the above intuitive reasoning does not appear to be fully backed by evidence from reality: optional tests continue to be implemented in high-stakes classification tasks such as US college admissions.[1] This raises the following question:

Are there any advantages of optional tests for classification over mandatory ones?

On top of that, in many other scenarios mandatory tests are simply unrealistic, and the principal has to rely on op-

[1] Applicants may take SAT and/or ACT tests, among others, as many times as they want.

tional tests for decision-making — for example, it is impossible for an academic hiring committee to *require* job applicants to submit their work to certain conferences in a prescribed way, e.g., one paper to NeurIPS'20 and one paper to ICML'21. In such cases, the principal would still like the classification process to be as accurate and efficient as possible. This leads us to the following question:

How can one design a classification process with optional tests in the most accurate and/or efficient way?

Our results. We give somewhat surprising answers to the above two questions in the case of binary classification of binary agents (elaborated below): we characterize the optimal design of a classification process with optional tests, and based on this show that classification with optional tests can be arbitrarily more efficient than optimal classification with mandatory tests for the same task.

To be more specific, we consider a setting where a principal either accepts or rejects agents based on the set of test outcomes that they get. Before tests are taken, the principal commits to a policy, which consists of all sets of outcomes that lead to acceptance. Each agent is modeled by a distribution over the space of possible outcomes, corresponding to how the agent tends to perform in a test. When an agent takes a test, he pays a cost and receives an independent sample from his distribution as the outcome. Agents can always choose between taking another test, and stopping. They maximize their expected utility, which is the value of acceptance if the principal's policy accepts the set of outcomes they have and 0 otherwise, minus the total cost of tests taken.

We focus on the case where agents can be either "good" or "bad" (corresponding to two different distributions over test outcomes), and the principal's goal is to accept good agents and reject bad ones. We first characterize the optimal strategy of an agent in response to the principal's policy. Fixing the principal's policy, each agent faces a Markov Decision Process (MDP), where the state is the set of test outcomes that he has collected. In general, at any state of the MDP, the agent can always choose between taking another test and stopping, and the optimal strategy could be any function mapping each combination of outcomes to one of the two actions. Our first key observation is that without loss of generality, the agent's optimal strategy is either to keep taking tests until acceptance, or to leave immediately without taking any test. This is because intuitively, after taking some tests, the agent must have received some outcomes, which makes his situation at least as good as when he started in terms of the expected number of *future* tests he needs to take in order to be accepted; the cost of the tests already taken is sunk. So, if an agent ever chooses to start taking tests, he must be willing to keep taking tests until acceptance, since the cost of past tests should not affect his decision. This is making two assumptions: (1) the agent can choose not to submit some of the test outcomes, and (2) the agent already knows his own type (good or bad) at the beginning, and hence is not learning about himself from the test outcomes.

With agents' optimal strategy characterized, we consider the principal's problem, i.e., the design of her classification policy. We first study the case where the principal controls the cost of a test, by, for example, charging a registration fee. We show that in this case, as long as the good and bad agents have different distributions (which can be arbitrarily close to each other), the principal can always achieve perfect accuracy, meaning good agents are accepted with probability 1, and bad ones are rejected with probability 1. The key technique is to choose the policy so that agents self-select into (not) taking tests. Moreover, among perfectly accurate policies, we characterize the one with stochastically dominant efficiency in terms of the number of tests a good agent needs to take in order to be accepted. We show that quite surprisingly, under this policy, no agent ever has to take more than 2 tests. One may contrast this with the mandatory tests case, where the principal directly observes as many samples as she wants from agents' distributions — there, in order to classify correctly with probability $2/3$, the number of tests required can be arbitrarily large, as the distance between the good and bad distributions diminishes.

We then proceed to the case where the cost per test is fixed externally. With discrete outcomes, we show that perfect accuracy in general is no longer possible. We then consider the special case with continuous outcome distributions, or equivalently, where outcomes are associated with rich noise that is effectively continuous. We show that in a continuous world, with different good and bad distributions, perfect accuracy is again always possible. Moreover, we construct a perfectly accurate policy under which the maximum number of tests a good agent needs to take is $\lfloor 1/c \rfloor + 1$ where c is the fixed cost per test, and show this is essentially best possible for perfectly accurate policies. We also provide evidence that the above bound cannot be significantly improved even if we consider the expected number of tests.

Related work. Our results are along the line of work on strategic machine learning (Dalvi et al. 2004; Perote and Perote-Pena 2004; Dekel, Fischer, and Procaccia 2010; Brückner, Kanzow, and Scheffer 2012; Meir, Procaccia, and Rosenschein 2012; Cai, Daskalakis, and Papadimitriou 2015; Hardt et al. 2016; Roughgarden and Schrijvers 2017; Chen et al. 2018; Dong et al. 2018; Feng, Parkes, and Xu 2019; Chen, Liu, and Podimata 2020; Freeman et al. 2020; Krishnaswamy et al. 2021; Zhang, Cheng, and Conitzer 2021; Zhang and Conitzer 2021), as well as causality interpretations thereof (Bechavod et al. 2020; Perdomo et al. 2020; Shavit, Edelman, and Axelrod 2020). Most research on strategic machine learning, including ours, aims to tackle the potential (mis)alignment of interests between the learner (i.e., the principal) and entities about which information is being learned (i.e., agents). One key difference between our results and existing research is that previous work along this line typically focuses on *preventing* strategic manipulation of the classification process, while we *exploit* agents' incentives to make classification more accurate and efficient. Exceptions are the recent results by Kleinberg and Raghavan (2019) and Haghtalab et al. (2020), who study the improve-

ment in agents' true features (e.g., skills) that is encouraged by the classifier deployed by the principal. In contrast to their work, we consider a model where agents' features (i.e., the distributions associated with them) are fixed throughout the classification process.

Closely related to our results is the series of work by Zhang, Cheng, and Conitzer (2019b,a). There, too, agents observe samples from distributions associated with them, which they can then strategically transform and subsequently submit to the principal for classification. Our results differ from these results in that we consider a novel model where the number of samples generated is endogenous, depending on utility-maximizing agents' private information, whereas they assume the number of samples is exogenous (fixed externally).

There is a rich literature in economics on screening with tests, and the effect of self-selection therein (Mirrlees 1976; Spence 1978; Guasch and Weiss 1981; Nalebuff and Scharfstein 1987; Loh 1994). Most of those results consider one-time tests with clearly defined outcomes (e.g., pass or fail), and the principal (e.g., a hiring firm) often cares about maximizing revenue or social welfare, rather than achieving high accuracy or efficiency. In contrast, we consider repeated tests with an arbitrary outcome space, and the primary goal is to achieve high accuracy and efficiency.

A conceptually related topic is that of efficient statistics, where some basic and commonly studied problems are distinguishing, learning (Chan et al. 2014), and testing (Diakonikolas, Kane, and Nikishkin 2015; Valiant and Valiant 2017) distributions. Our results can be viewed as efficiently distinguishing distributions in the presence of strategic behavior.

2 Preliminaries

In this section, we formally define the problem of test-based classification.

Agents, tests, and outcomes. Each agent is modeled by a distribution D over a space O of possible test outcomes (e.g., integers between 0 and 100, corresponding to numerical scores). Agents can choose to take as many tests as they want. When an agent with distribution D takes a test, he receives an outcome drawn from D independently of past test outcomes, and can then choose to continue taking tests, or to stop. The outcomes of all the tests taken (which form a multiset whose elements are from the outcome space O) will then be used by the principal for classification.

The principal and the policy. The principal, before agents decide whether or not to take tests, announces a policy $\mathcal{P}$ for classification. The policy in general is a collection of multisets, each of which consists of certain test outcomes from the outcome space O. For an agent with outcomes S, the policy $\mathcal{P}$ accepts the agent iff there is a multiset $T \in \mathcal{P}$ such that $T \subseteq S$ (i.e., the multiplicity of any element in T is no larger than that of the same element in S). In other words, the policy $\mathcal{P}$ provides a collection of options to agents, each of which is a multiset T of outcomes. An agent is accepted iff his multiset of test outcomes contains any of these options as a subset. A simple and natural example is when O is the set of integers between 0 and 100, and $\mathcal{P}$ contains a number of singleton multisets, each of which is an integer between 60 and 100, i.e.,

$$\mathcal{P} = \{\{i\} \mid 60 \leq i \leq 100\}.$$

This corresponds to the case where agents can repeatedly take exams, and are accepted (i.e., pass) iff they ever get a score of at least 60. Another example would be

$$\mathcal{P} = \{\{i\} \mid 60 \leq i \leq 100\} \cup \{\{i, j\} \mid 50 \leq i, j < 60\},$$

which is the same policy as before, except it now also suffices to score at least 50 *twice*.

How rational agents act in response to a policy. Fixing a policy $\mathcal{P}$, each agent faces an MDP, where the goal is to maximize his expected utility. Below we describe this MDP. Without loss of generality, being accepted gives agents value 1, and each test has a cost of $0 \leq c \leq 1$ (otherwise agents would never want to take any test). The states of the MDP, denoted $\mathcal{S}$, are all multisets over the outcome space O, corresponding to the set of outcomes the agent has collected so far. Initially, the state of the agent is the empty set $\emptyset$. At any state $S \in \mathcal{S}$, the agent can choose between two actions, taking another test (T) or leaving (L). If the agent chooses T, he pays cost c (i.e., receives reward $-c$), and transitions to a new state $S \cup \{o\}$ (note that this is a union of two multisets, where the multiplicity of any element in the union is the sum of those of the same element in the two operands), where $o \sim D$ is a random outcome drawn from D. If the agent chooses L, he receives reward 1 if his current multiset of outcomes S is accepted by the policy $\mathcal{P}$, and 0 otherwise; in either case, the MDP terminates immediately. Throughout the paper, we assume agents are perfectly rational and always play the utility-maximizing action. For simplicity, we assume agents always break ties in favor of leaving, i.e., when the two actions result in equal expected utility, they always play L. Our results still hold (with minor modifications) even if agents break ties adversarially.

The principal's goals. We consider two goals of the principal, accuracy and efficiency. We focus on the case where agents are either good (with distribution G) or bad (with distribution B), and the principal aims to accept as many good agents as possible, and reject as many bad ones as possible. The specific definition of accuracy is immaterial — as we will show, the principal can always achieve perfect accuracy (i.e., good agents are always accepted, and bad ones always rejected) as long as G and B are not identical. Given perfect accuracy, the principal may further hope to implement the classification in an efficient way, where agents take as few tests as possible. We consider two types of efficiency measures, the expected number of tests and the worst-case number of tests. The goal is to design perfectly accurate policies which (approximately) minimize either/both of these two measures (though in Section 4.3 we do consider how to minimize expected cost in our model under the constraint of perfect accuracy).

Control over the costs. In some scenarios, the cost of a test is controlled by the principal (e.g., when the dominant part of the cost is a registration fee set by the principal), while in others it is fixed externally (e.g., when the dominant part of the cost is time invested in traveling to the test site). We consider both cases in this paper. The flexible-cost case allows the principal refined control of the classification procedure, which, as we will show, implies more efficient policies in general.

3 Agents' Optimal Strategy: Self-Selection

We first characterize agents' best response to a policy, which effectively makes their decision space binary, and greatly simplifies the principal's problem.

Lemma 1. *Fixing a policy $\mathcal{P}$ and a cost per test c, the optimal expected reward of any agent is achieved by one of the following two strategies:*

- *Take no test (i.e., play L immediately) and leave with reward 0.*
- *Keep taking tests (i.e., playing T) until the set of outcomes collected is accepted by $\mathcal{P}$, and then play L.*

Moreover, the optimal strategy is unique iff the above two strategies result in strictly different expected rewards.

The proof of Lemma 1, as well as all other proofs, is deferred to the appendix. Again, the intuition is that if an agent ever wants to start taking tests, then after taking some tests, he will be in at least as favorable a position as at the beginning in terms of tests passed, and it was worth it to start then, so it must certainly be worth it to continue now (the cost of previous tests is sunk, and therefore irrelevant). One important implication is that, depending on the policy, the cost per test, and the agent's distribution, each agent either does not attempt to get accepted at all, or keeps trying and eventually gets accepted with probability 1. This indicates that, when provided the right incentives, self-selecting agents may perform the classification for the principal in a perfectly accurate way.

More specifically, for any policy $\mathcal{P}$ and distribution D over the outcome space O, let $T(\mathcal{P}, D)$ denote the (random) number of tests an agent with distribution D needs to take in order to be accepted by $\mathcal{P}$, i.e.,

$$T(\mathcal{P}, D) = \min\{t \mid \mathcal{P} \text{ accepts } \{o_1, \ldots, o_t\}\},$$

where $\{o_t\}_{t\geq 1}$ are iid draws from D. We have the following claim.

Lemma 2. *Fix a policy $\mathcal{P}$ and a cost per test c. An agent with distribution D will always keep taking tests until acceptance if*

$$c \cdot \mathbb{E}_{\{o_t\}\sim D^{\mathbb{Z}_+}}[T(\mathcal{P}, D)] < 1,$$

and leave immediately otherwise.

In the rest of the paper, we will heavily exploit Lemma 2.

4 The Flexible-Cost Case

We begin our investigation with the case where the cost of a test is set by the principal, which turns out to be simpler. For simplicity, we assume the outcome space $O = [k] = \{1, \ldots, k\}$ for some integer $k > 0$. For any distribution D over O (which can be either G or B), for any $S \subseteq O$, let $D(S) = \Pr_{o\sim D}[o \in S]$. As a shorthand, for any $o \in O$, let $D(o) = D(\{o\})$. All the results in this section can be easily generalized to arbitrary outcome spaces.[2]

4.1 Memoryless Policies Suffice for Accurate Classification

We first consider the possibility of accurate classification. In particular, for reasons that will be clear momentarily, we are interested in policy-cost pairs that achieve perfect accuracy, as defined below.

Definition 1 (Perfect Accuracy)**.** A policy-cost pair $(\mathcal{P}, c)$ is *perfectly accurate* for a good distribution G and a bad distribution B if the optimal strategies for good agents and bad agents respectively are to keep taking tests until acceptance and to leave immediately.

As a corollary of Lemma 2, a pair $(\mathcal{P}, c)$ is perfectly accurate iff

$$c \cdot \mathbb{E}[T(\mathcal{P}, G)] < 1 \leq c \cdot \mathbb{E}[T(\mathcal{P}, B)].$$

When the cost per test is controlled by the principal, we are further interested in policies that are perfectly implementable.

Definition 2 (Perfect Implementability)**.** A policy $\mathcal{P}$ is *perfectly implementable* for a good distribution G and a bad distribution B if there exists a cost per test c, such that the policy-cost pair $(\mathcal{P}, c)$ achieves perfect accuracy.

Given Lemma 2, we immediately have the following necessary and sufficient condition for perfect implementability.

Lemma 3. *Fix a good distribution G and a bad distribution B. A policy $\mathcal{P}$ is perfectly implementable iff*

$$\mathbb{E}[T(\mathcal{P}, G)] < \mathbb{E}[T(\mathcal{P}, B)].$$

Based on the above characterization, we show that as long as the good agents' distribution G is different from the bad agents' distribution B, there always exists a perfectly implementable policy which consists of only singleton sets of outcomes. In other words, the policy is memoryless, in that a new test outcome either immediately makes the agent accepted, or will be entirely ignored.

Theorem 1. *For any good distribution G and bad distribution B over the outcome space $O = [k]$ where $G \neq B$, there exists a set of outcomes $P \subseteq O$ such that the policy*

$$\mathcal{P} = \{\{o\} \mid o \in P\}$$

is perfectly implementable.

[2] For example, when O is an infinite, possibly continuous space (e.g., $O = \mathbb{R}$), one can discretize O into a finite number (which may depend on the desired precision) of regions such that the good and bad distributions after discretization are arbitrarily close to the respective original distributions.

We remark that such memoryless policies are widely deployed in practice, where the most common form is to set a threshold and accept an agent iff the highest score he ever gets passes that threshold. However, as we will show later, this is not the most efficient form of perfectly accurate policies.

4.2 Policy with Stochastically Dominant Efficiency

We now proceed to efficient policies. Below we characterize the perfectly implementable policy $\mathcal{P}$ with stochastically dominant efficiency for any good distribution G and bad distribution B. The number of tests required for a good agent to be accepted under this policy, $T(\mathcal{P}, G)$, stochastically dominates the same number, $T(\mathcal{P}', G)$, of any other perfectly implementable policy $\mathcal{P}'$. Furthermore, under this policy, any good agent is guaranteed to be accepted after taking at most 2 tests. As a result, this policy achieves the optimal expected number of tests, the optimal worst-case number of tests (which is 2), and optimality with respect to almost any reasonable measure of efficiency.

Theorem 2. *Let $P \subseteq O = [k]$ be a set of outcomes such that*

$$P \in \operatorname{argmax}_{S \subseteq O : G(S) > B(S)} G(S).$$

The policy

$$\mathcal{P} = \{\{o\} \mid o \in P\} \cup \{\{o_1, o_2\} \mid o_1, o_2 \in O\}$$

is perfectly implementable, and stochastically dominates any other perfectly implementable policy $\mathcal{P}'$, in the sense that for any $t \in \mathbb{Z}_+$,

$$\Pr[T(\mathcal{P}, G) \le t] \ge \Pr[T(\mathcal{P}', G) \le t].$$

The policy $\mathcal{P}$ constructed in Theorem 2 accepts any set of outcomes which either contains some outcome in $P \subseteq O$, or has cardinality at least 2. In other words, $\mathcal{P}$ accepts an agent if the first outcome he receives is in P, or he ever takes 2 tests. One may contrast Theorem 2 with the setting where the principal, rather than the agent himself, chooses the number of tests each agent needs to take. Suppose, rather than deploying a policy and letting agents themselves choose whether or not to take tests, we directly observe iid samples from an unknown distribution D, which can be either G or B — this corresponds to the case where we simply ask each agent to take as many tests as we want. There, how many samples one needs to observe in order to tell with confidence whether D is G or B depends on the total variation distance between G and B, defined below.

Definition 3 (Total Variation Distance)**.** The total variation distance $d_{\mathrm{TV}}(D_1, D_2)$ between two distributions D_1 and D_2 over O is defined as

$$d_{\mathrm{TV}}(D_1, D_2) = \sup_{S \subseteq O} (D_1(S) - D_2(S)).$$

Observe that $G \neq B$ iff $d_{\mathrm{TV}}(G, B) > 0$. It is folklore that in order to identify D with probability at least $2/3$, one needs $\Omega(d_{\mathrm{TV}}(G, B)^{-2})$ iid samples from D. Moreover, it is easy to see that as long as the supports of G and B overlap, one can never be completely sure with any finite number of samples. Theorem 2, on the other hand, essentially says that whenever $d_{\mathrm{TV}}(G, B) > 0$, the principal never needs to observe more than 2 samples in order to distinguish G and B, and good agents never need to take more than 2 tests. In other words, by incentivizing self-selection, the principal is able to reduce the number of tests required dramatically, from $\Omega(d_{\mathrm{TV}}(G, B)^{-2})$ to 2, and at the same time improve the accuracy to 1. Perhaps even more surprisingly, this is done by giving agents more freedom to choose the number of tests they take. Of course, this is feasible only because agents themselves know their distribution at the outset; if nobody knows the distribution, $\Omega(d_{\mathrm{TV}}(G, B)^{-2})$ tests would still be required. This partially explains the practical success of classification with optional tests: they can be arbitrarily more efficient than mandatory tests enforced by the principal, especially when good and bad agents' distributions are closer to each other and therefore are harder to distinguish.

4.3 Cost Efficiency of Policies

While the policy in Theorem 2 is efficient in terms of the *number of tests*, it could impose a total expected *cost* on good agents that is quite close to the benefit of being accepted. Depending on the circumstances, cost efficiency may be considered more important, and indeed, often the classification procedure can be implemented in much less costly ways, intuitively for the following reasons. First, when G and B are hard to distinguish, it is natural that good agents need to spend significant effort in order to distinguish themselves from bad ones. But, in many real-world scenarios, (most) good agents are considerably different from (most) bad ones. In such cases, good agents pay much less cost, since the principal only needs to make bad agents marginally unwilling to take tests. Second, there can be a tradeoff between efficiency (i.e., the number of tests taken) and cost. If the principal is willing to make good agents take more than 2 tests, then she can design a more selective policy (i.e., making it hard to pass) that creates a sharper separation between good and bad agents, and set a lower cost per test to achieve perfect accuracy. Below we formalize this intuition, and characterize the optimal cost efficiency possible, subject to perfect accuracy, for memoryless policies.

Theorem 3. *Fix any good distribution G and bad distribution B over the outcome space $O = [k]$ where $G \neq B$. There exists a memoryless policy*

$$\mathcal{P} = \{\{o\} \mid o \in P\}$$

for some $P \subseteq O$, and a cost per test c, such that $(\mathcal{P}, c)$ is perfectly accurate, and the expected total cost paid by good agents is

$$c \cdot \mathbb{E}[T(\mathcal{P}, G)] = \min_{o \in O} B(o)/G(o).$$

Moreover, no policy-cost pair $(\mathcal{P}', c')$ satisfies (1) $\mathcal{P}'$ is memoryless, and (2) the expected total cost paid by good agents is

$$c' \cdot \mathbb{E}[T(\mathcal{P}', G)] < \min_{o \in O} B(o)/G(o).$$

The above theorem says that the optimal cost efficiency achievable by memoryless policies is determined by the

minimum ratio between B and G over the test outcome space. We also remark that cost efficiency directly implies robustness against bad agents who value acceptance more than good agents.[3] Fixing any good distribution G and bad distribution B, when good agents have value 1 and bad agents have value $v \geq 1$ for acceptance, there exists a perfectly implementable policy iff the optimal cost efficiency achievable when all agents have value 1 is better than v^{-1}, i.e., there exists a policy-cost pair $(\mathcal{P}, c)$ such that

$$c \cdot \mathbb{E}[T(\mathcal{P}, G)] < v^{-1} \quad \text{and} \quad c \cdot \mathbb{E}[T(\mathcal{P}, B)] \geq 1.$$

In fact, given such a cost efficient pair $(\mathcal{P}, c)$, $(\mathcal{P}, v \cdot c)$ is a perfectly accurate pair when bad agents have value $v \geq 1$ for acceptance.

5 The Fixed-Cost Case

Now we proceed to the more challenging setting where the cost per test c is fixed externally. We show that in such cases, perfect accuracy in general requires stronger conditions on the good and bad distributions. However, as we argue below, these conditions are still rather reasonable for practical purposes.

5.1 Accurate Classification Requires Continuous Information

When the cost per test is set by the principal, Theorem 1 states that perfect accuracy can be achieved by some policy-cost pair as long as the good and bad distributions are different. However, this is not true when the cost $0 < c < 1$ is fixed, as illustrated in the following example.

Example 1. Suppose the cost per test is fixed at $c = 0.9$. The outcome space $O = \{1, 2\}$, the good distribution G assigns probability $G(1) = G(2) = 0.5$, and the bad distribution B assigns $B(1) = 0$ and $B(2) = 1$. Suppose there is a policy $\mathcal{P}$ such that $(\mathcal{P}, c)$ is perfectly accurate. Then, in order for good agents to take tests, by Lemma 2,

$$\mathbb{E}[T(\mathcal{P}, G)] < 10/9,$$

and since $T(\mathcal{P}, G)$ is distributed over $\mathbb{Z}_+$, elementary calculation gives

$$\Pr[T(\mathcal{P}, G) = 1] > 8/9.$$

As a result, it must be the case that $\{1\} \in \mathcal{P}$ and $\{2\} \in \mathcal{P}$ simultaneously. However, this implies

$$\Pr[T(\mathcal{P}, B) = 1] = 1 \Longrightarrow \mathbb{E}[T(\mathcal{P}, B)] < 10/9.$$

So bad agents will also take tests and get accepted under $(\mathcal{P}, c)$, a contradiction. In other words, no policy $\mathcal{P}$ exists such that $(\mathcal{P}, c)$ is perfectly accurate.

The above example shows that perfect accuracy cannot be achieved with an infeasibly high cost per test, even if the outcome space is extremely simple (i.e., binary) and the good and bad distributions are clearly different. Nevertheless, the impossibility of perfect accuracy comes almost solely from the discreteness in the outcomes — intuitively, accepting only one of the two outcomes does not provide enough motivation for good agents to take tests, while accepting both provides too much motivation, so that every agent wants to take tests regardless of his distribution.

Real-world tests, however, are often intrinsically (approximately) continuous. In a narrow sense, test outcomes, in the form of numerical scores, usually range from 0 to 100, where presumably an agent can get any integer score in between with positive probability. As argued above, in a broader sense, a test could be any activity which takes a certain effort and produces a verifiable outcome. Besides numerical test scores, such an outcome could take the form of a course project, a research paper, or an oral presentation. These outcomes are essentially continuous, in the sense that, for example, no two oral presentations are exactly the same, even if they are given by the same presenter using the same slides. Even for relatively discrete outcome spaces, an outcome is often accompanied by arbitrarily rich noise, which makes outcomes effectively continuous.[4] For example, in a simplistic model, a paper submitted to a conference can be either accepted or rejected, so one could argue the outcome of such a submission is binary. However, it is extremely unlikely that two different papers (as PDF files) share the same hash value, which can effectively be viewed as continuous noise that we can add to the outcome, thereby making the outcome space continuous. This (hash value) part of the enriched outcome may not be correlated with the type of the agent, but that will not matter for our purposes.

Based on the above observations, in the rest of this section, we assume the outcome space O, as well as the good distribution G and the bad distribution B, is continuous. This could model continuity in the outcome distribution itself, or noise, or the two aspects in combination. More specifically, without loss of generality, we assume $O = [0, k]$ for some positive integer $k \in \mathbb{Z}_+$, and the good distribution G (resp. the bad distribution B) is constant when restricted to the interval $[i-1, i]$ for any $i \in [k] = \{1, \ldots, k\}$. We call such distributions piecewise constant.[5] One way to interpret this is that there are k possible outcomes. A good (resp. bad) agent receives the i-th outcome with probability $G([i-1, i])$ (resp. $B([i-1, i])$). Moreover, there is continuous noise x independent of the outcome and the agent type, uniformly distributed over $[0, 1]$, so the final combination of the outcome and the noise, $i - x$, has distribution G (resp. B). As a shorthand, for any $i \in [k]$, let $G(i) = G([i-1, i])$, and $B(i) = B([i-1, i])$. While for ease of presentation we focus on this specific model, in fact, our results apply to general distributions satisfying certain continuity conditions.[6]

[3] The case where good agents value acceptance more is no harder than the case where all agents have the same value for acceptance.

[4] This has also been observed, e.g., in (Zhang, Cheng, and Conitzer 2019b).

[5] While this appears to be a more restrictive definition than the common notion of piecewise constant distributions, observe that without loss of generality, one can always scale the pieces and the distributions simultaneously, such that the pieces are of the same length.

[6] For example, it is known that all Lebesgue measurable functions (including all continuous ones) are approximated by step functions (i.e., piecewise constant ones) up to any precision. This

5.2 Accurate Classification with Continuous Outcomes

Under the continuity assumption, we now show that perfect accuracy is possible with fixed cost per test, whenever the good and bad distributions, G and B, are not identical. Moreover, as in the variable cost case, perfect accuracy again can be achieved using a memoryless policy.

Theorem 4. *When the outcome space $O = [0, k]$, for any cost per test $0 < c < 1$, and good and bad distributions G and B (where $G \neq B$) that are constant on $[i-1, i]$ for any $i \in [k]$, there exists a policy $\mathcal{P}$ such that $(\mathcal{P}, c)$ is perfectly accurate for G and B. Moreover, $\mathcal{P}$ consists of only singleton sets of outcomes.*

5.3 Nearly Optimal Policies

As illustrated by Theorem 2, memoryless policies do not generally achieve optimal efficiency when the cost per test is set by the principal. The same intuition applies to the fixed cost case as well. Below, we construct a policy for any piecewise constant and distinct good and bad distributions which requires at most $\lfloor 1/c \rfloor + 1$ tests, where c is the cost per test. We then show that the policy we construct has (1) optimal worst case efficiency, and (2) approximately optimal expected efficiency when the good and bad distributions are not trivially different.

Theorem 5. *When the outcome space is $O = [0, k]$, for any cost per test $0 < c < 1$, and good and bad distributions G and B (where $G \neq B$) that are constant on $[i-1, i]$ for any $i \in [k]$, there exists a policy $\mathcal{P}$ such that $(\mathcal{P}, c)$ is perfectly accurate for G and B. Moreover,*

$$\Pr[T(\mathcal{P}, G) \leq \lfloor 1/c \rfloor + 1] = 1.$$

Unlike Theorem 2, the above policy-cost pair is not guaranteed to dominate all other perfectly accurate pairs. However, it is in fact optimal in terms of the maximum number of tests a good agent may have to take before getting accepted, as long as the good and bad distributions share the same support.

Proposition 1. *For any piecewise constant good and bad distributions G and B where G and B share the same support, if a policy-cost pair $(\mathcal{P}, c)$ is perfectly accurate, then*

$$\Pr[T(\mathcal{P}, G) < \lfloor 1/c \rfloor + 1] < 1.$$

In other words, the maximum number of tests a good agent may have to take is at least $\lfloor 1/c \rfloor + 1$.

Proposition 1 states that the policy constructed in Theorem 5 is in fact optimal in terms of the maximum number of tests any good agent may have to take. However, it is unclear whether one can do significantly better[7] in terms of the expected number of tests, especially when the good and bad distributions are sufficiently different. We do show that, even when good and bad distributions are far apart (i.e., when $d_{\mathrm{TV}}(G, B) = \Omega(1)$), there still exist good and bad distributions G and B such that the expected number of tests required is at least $1/2c$. In other words, the policy constructed in Theorem 5 is also asymptotically optimal (in a worst-case sense) in terms of the expected number of tests.

Proposition 2. *There exist piecewise constant good and bad distributions G and B where $d_{\mathrm{TV}}(G, B) \geq 0.1$, such that if a policy-cost pair $(\mathcal{P}, c)$ is perfectly accurate, then*

$$\mathbb{E}[T(\mathcal{P}, G)] \geq \frac{1}{2c}.$$

Again, one can contrast Theorem 5 with the case where the principal directly determines how many tests an agent takes. In that case, as discussed above, the number of tests required to correctly identify an agent's distribution D is $\Omega(d_{\mathrm{TV}}(G, B)^{-2})$, whereas Theorem 5 requires good agents to take at most about $1/c$ tests and guarantees perfect accuracy, regardless of how close G and B are to each other. In other words, unless G and B are far away and a small error probability is acceptable, allowing agents to choose between taking tests and leaving immediately is far more efficient than enforcing a certain number of tests.

Finally, we remark that with a fixed cost per test, there is no tradeoff between efficiency (i.e., the number of tests taken by good agents) and cost efficiency (i.e., the expected total cost paid by good agents) — they are always proportional to each other.

6 Conclusion and Future Research

In this paper, we characterize the accuracy and efficiency of classification with optional tests. Our results partially explain the practical success of optional tests, and provide a principled way of designing accurate and efficient classification processes. In particular, we show how much better one can do with self-selection than in comparable settings without self-selection that were studied recently, even when we augment those models with self-selection in the simplest possible way. Our results also easily generalize to some richer settings. For example, if taking the test might make one better at the test next time (due to practice), this retains the key property that once an agent starts taking tests, that agent will continue until the agent succeeds. For future directions, one could relax some of the assumptions to obtain more robustness in the design of classification processes. For example, test outcomes might be strategically transformed (as studied in (Zhang, Cheng, and Conitzer 2019a)), the cost per test might be unknown, and agents might not be completely sure about their own distributions before taking tests.

gives a way of generalizing our results to all Lebesgue measurable density functions.

[7] It is certainly possible to do somewhat better. For example, when $G([0,1]) = B([1,2]) = 0.9$, $G([1,2]) = B([0,1]) = 0.1$, and the cost per test $c = 0.5$, $\mathcal{P} = \{\{o\} \mid o \in [0,1]\}$ accepts a good agent after $10/9 < \lfloor 1/c \rfloor + 1 = 3$ tests in expectation. $(\mathcal{P}, c)$ is perfectly accurate, because a bad agent will require 10 tests in expectation, so a bad agent will not attempt the test.

Acknowledgements

The authors are thankful for support from NSF under award IIS-1814056.

Ethics Statement

Our results help explain the practical success of classification with optional tests. By characterizing the optimal policy with optional tests, our results provide a principled way of designing classification criteria which are provably accurate and efficient. Such criteria could be applied in classification tasks with potential social impact, e.g., college admissions. Of course, our results could be used for classification with harmful objectives and exacerbate the damage.

References

Bechavod, Y.; Ligett, K.; Wu, Z. S.; and Ziani, J. 2020. Causal Feature Discovery through Strategic Modification. *arXiv preprint arXiv:2002.07024* .

Brückner, M.; Kanzow, C.; and Scheffer, T. 2012. Static prediction games for adversarial learning problems. *The Journal of Machine Learning Research* 13(1): 2617–2654.

Cai, Y.; Daskalakis, C.; and Papadimitriou, C. 2015. Optimum statistical estimation with strategic data sources. In *Conference on Learning Theory*, 280–296.

Chan, S.-O.; Diakonikolas, I.; Servedio, R. A.; and Sun, X. 2014. Efficient density estimation via piecewise polynomial approximation. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, 604–613.

Chen, Y.; Liu, Y.; and Podimata, C. 2020. Learning Strategy-Aware Linear Classifiers. *arXiv preprint arXiv:1911.04004* .

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

Dalvi, N.; Domingos, P.; Sanghai, S.; and Verma, D. 2004. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 99–108.

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

Diakonikolas, I.; Kane, D. M.; and Nikishkin, V. 2015. Testing identity of structured distributions. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, 1841–1854. Society for Industrial and Applied Mathematics.

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

Feng, Z.; Parkes, D. C.; and Xu, H. 2019. The intrinsic robustness of stochastic bandits to strategic manipulation. *arXiv preprint arXiv:1906.01528* .

Freeman, R.; Pennock, D. M.; Podimata, C.; and Vaughan, J. W. 2020. No-Regret and Incentive-Compatible Prediction with Expert Advice. *arXiv preprint arXiv:2002.08837* .

Guasch, J. L.; and Weiss, A. 1981. Self-selection in the labor market. *The American Economic Review* 71(3): 275–284.

Haghtalab, N.; Immorlica, N.; Lucier, B.; and Wang, J. 2020. Maximizing Welfare with Incentive-Aware Evaluation Mechanisms. In *29th International Joint Conference on Artificial Intelligence*.

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

Kleinberg, J.; and Raghavan, M. 2019. How Do Classifiers Induce Agents to Invest Effort Strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation*, 825–844.

Krishnaswamy, A.; Li, H.; Rein, D.; Zhang, H.; and Conitzer, V. 2021. Classification with Strategically Withheld Data. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Loh, E. S. 1994. Employment probation as a sorting mechanism. *ILR Review* 47(3): 471–486.

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

Mirrlees, J. A. 1976. Optimal tax theory: A synthesis .

Nalebuff, B.; and Scharfstein, D. 1987. Testing in models of asymmetric information. *The Review of Economic Studies* 54(2): 265–277.

Perdomo, J. C.; Zrnic, T.; Mendler-Dünner, C.; and Hardt, M. 2020. Performative prediction. *arXiv preprint arXiv:2002.06673* .

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

Roughgarden, T.; and Schrijvers, O. 2017. Online prediction with selfish experts. In *Advances in Neural Information Processing Systems*, 1300–1310.

Shavit, Y.; Edelman, B.; and Axelrod, B. 2020. Learning From Strategic Agents: Accuracy, Improvement, and Causality. *arXiv preprint arXiv:2002.10066* .

Spence, M. 1978. Job market signaling. In *Uncertainty in economics*, 281–306. Elsevier.

Valiant, G.; and Valiant, P. 2017. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing* 46(1): 429–455.

Zhang, H.; Cheng, Y.; and Conitzer, V. 2019a. Distinguishing Distributions When Samples Are Strategically Transformed. In *Advances in Neural Information Processing Systems*, 3187–3195.

Zhang, H.; Cheng, Y.; and Conitzer, V. 2019b. When samples are strategically selected. In *International Conference on Machine Learning*, 7345–7353.

Zhang, H.; Cheng, Y.; and Conitzer, V. 2021. Automated Mechanism Design for Classification with Partial Verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhang, H.; and Conitzer, V. 2021. Incentive-Aware PAC Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.