The Role of Randomness and Noise in Strategic Classification

Mark Braverman
Department of Computer Science, Princeton University, USA
mbraverm@cs.princeton.edu

Sumegha Garg
Department of Computer Science, Princeton University, USA
sumeghag@cs.princeton.edu

Abstract

We investigate the problem of designing optimal classifiers in the "strategic classification" setting, where the classification is part of a game in which players can modify their features to attain a favorable classification outcome (while incurring some cost). Previously, the problem has been considered from a learning-theoretic perspective and from the algorithmic fairness perspective.

Our main contributions include

- Showing that if the objective is to maximize the efficiency of the classification process (defined as the accuracy of the outcome minus the sunk cost of the qualified players manipulating their features to gain a better outcome), then using randomized classifiers (that is, ones where the probability of a given feature vector to be accepted by the classifier is strictly between 0 and 1) is necessary.
- Showing that in many natural cases, the imposed optimal solution (in terms of efficiency) has the structure where players never change their feature vectors (and the randomized classifier is structured in a way, such that the gain in the probability of being classified as a '1' does not justify the expense of changing one's features).
- Observing that the randomized classification is not a *stable* best-response from the classifier's viewpoint, and that the classifier doesn't benefit from randomized classifiers without creating instability in the system.
- Showing that in some cases, a *noisier signal* leads to better equilibria outcomes — improving both accuracy and fairness when more than one subpopulation with different feature adjustment costs are involved. This is particularly interesting from a policy perspective, since it is hard to force institutions to stick to a particular randomized classification strategy (especially in a context of a market with multiple classifiers), but it is possible to alter the information environment to make the feature signals inherently noisier.

2012 ACM Subject Classification Theory of computation → Algorithmic game theory and mechanism design; Theory of computation → Machine learning theory

Keywords and phrases Strategic classification, noisy features, randomized classification, fairness

Digital Object Identifier 10.4230/LIPIcs.FORC.2020.9

Funding *Mark Braverman*: Research supported in part by the NSF Alan T. Waterman Award, Grant No. 1933331, a Packard Fellowship in Science and Engineering, and the Simons Collaboration on Algorithms and Geometry. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the National Science Foundation.

1 Introduction

Machine learning algorithms are increasingly being used to make decisions about the individuals in various areas such as university admissions, employment, health, etc. As the individuals gain information about the algorithms being used, they have an incentive to

1st Symposium on Foundations of Responsible Computing (FORC 2020).
Editor: Aaron Roth; Article No. 9; pp. 9:1–9:20
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

adapt their data so as to be classified desirably. For example, if a student is aware that a university heavily weighs SAT score in their admission process, she will be motivated to achieve a higher SAT score either through extensive test preparation or multiple tries. Such efforts by the students might not change their probability of being successful at the university, but are enough to fool the admissions' process. Therefore, under such "strategic manipulation" of one's data, the predictive power of the decisions are bound to decrease. One way to prevent such manipulation is by keeping the classification algorithms a secret, but this is not a practical solution to the problem, as some information is bound to leak over time and the transparency of these algorithms is a growing social concern. Thus, this motivates the study of algorithms that are optimal under "strategic manipulation". The problem of gaming in the context of classification algorithms is a well known problem and is increasingly gaining researchers' attention, for example, [8, 1, 9, 16, 4].

[2] and [8] modeled strategic classification as a Stackelberg competition– the algorithm (Jury) goes first and publishes the classifier, and then the individuals get to transform their data, after knowing the classifier, incurring certain costs to manipulate. The individuals would manipulate their features as long as the cost to manipulate is less than the advantage gained in getting the desirable classification. We assume that such manipulations don't change the actual qualifications of an individual. A natural question is: what classifier achieves optimal classification accuracy under the Stackelberg competition? These papers considered the task of strategic classification when the published classifier is deterministic. We study the role of randomness (and addition of noise to the features) in strategic classification and define the Stackelberg equilibrium for probabilistic classifiers, that assigns a real number in $[0, 1]$, to each individual and a classification outcome o, representing the probability of being classified as o.

As higher SAT scores are preferred by a university, the students would put an effort in increasing their SAT score, thereby, forcing the university to raise the score bar to optimize its accuracy (under the Stackelberg equilibrium). Due to this increased bar of acceptance, even the students who were above the true cutoff would have to put an extra effort to achieve a SAT score above this raised bar. And this effort is entirely the result of gaming in the classification system. We define the *cost of strategy* for a published classifier to be the total extra effort, it induced, amongst the qualified individuals of the population. Then, we define the *efficiency* of a published classifier to be its classification accuracy minus the cost of strategy under the Stackelberg equilibrium. A natural question here is: what classifier achieves the optimal efficiency? The efficiency of a published classifier represents the total impact of the classifier on all the agents in the Stackelberg equilibrium.

In normal classification problems it is never a good idea to use randomness, since one should always adhere to the best/utility maximizing action based on the prediction. Just as in games, randomness may lead to better solution in strategic classification, the paper aims to start understanding tradeoffs between efficiency losses due to randomness and efficiency gains through better equilibria induced by the randomized classifier.

Gaming in classification adds to the plethora of fairness concerns associated with classification algorithms, when the costs of manipulation are different across subpopulations. For example, a high weightage of SAT scores (for university admissions) favors the subgroups of the society that have the resources to enroll in test preparation or attempt the test multiple times. Further, varying costs across the subpopulations can lead to varied efforts put by identically qualified individuals, belonging to different subpopulations, to achieve the same outcome. [16] and [9] study the disparate effects of strategic classification on subpopulations (we will discuss these papers more in the related work section). [9] observes that a single classifier might have different classification errors on subpopulations due to the varying cost of manipulations. We also study the effect of strategic manipulation on the classification errors across subpopulations and how randomized classifiers or noisy features may reduce the disparate effects.

Strategic classification is a well known problem and there has been research in many other aspects of strategic classification, for example, learning the optimal classifier efficiently when the samples might also be strategic [8, 4], mechanism design under strategic manipulation [3, 5, 12], and studying the manipulation costs that actually change the inherent qualifications [14, 15]. The focus of this paper is theoretically demonstrating the role of randomness and noise in the strategic setting.

1.1 Our contributions

Above, we talked about how strategic manipulation can deteriorate the classification accuracy and lead to unfair classification. We investigate the different scenarios of the classification task that help in regaining the lost accuracy and fairness guarantees. Our entire work is based on *one-dimensional feature space.*

1.1.1 Randomized classifiers

Firstly, we formulate the strategic classification task, when the published classifier is randomized. Instead of publishing a single binary classifier (for 2 classification outcomes, 0 and 1), the Jury publishes a distribution of classifiers and promises to pick the final classifier from that distribution. Another interpretation is that the Jury assigns a value in $[0, 1]$ to each feature value, which represents the probability of an individual with this feature being classified as 1. The individuals manipulate their features, after knowing the set of classifiers but not the final classifier, incurring certain costs according to the *cost function.*

Not surprisingly, we show through examples that a probabilistic classifier can achieve strictly higher expected accuracy and efficiency than any binary classifier under strategic setting. Note that, without any strategic manipulation, a randomized classifier has no advantage over deterministic classifiers in terms of classification accuracy. The intuition is as follows: using randomness, the Jury can discourage the individuals from manipulating their features by making the advantage gained by any such a manipulation small enough.

For *simple* cost functions, we then characterize the randomized classifier that achieves optimal efficiency. We prove that such a classifier sets the probabilities (of being classified as 1) such that none of the individuals have an incentive to manipulate their feature. Given two features x and x' in the feature space, let $c(x, x')$ denote the cost of manipulating one's feature from x to x'. Informally, we say a cost function c is *simple* when all the costs are non-negative, the cost to manipulate to a "less" qualified feature is 0, and the costs are sub-additive, that is, manipulating your feature x directly to x'' is at least easier than first manipulating it to x' and then to x''. The characterization theorem, stated informally, is as follows:

▶ **Theorem 1** (Informal statement of Theorem 3)**.** *For simple cost functions, the most efficient randomized classifier is such that the best response of all the individuals is to reveal their true features.*

This characterization, in addition to being mathematically clean, allows us to infer the following: let A and B be two subpopulations (identical in terms of qualifications) such that the costs to manipulation are *higher* for individuals in A than in B, then the optimal efficiency obtained for the subpopulation A is greater than that in B.

1.1.2 Obstacles to using a randomized classifier

Till now, we have argued the benefits of using a probabilistic classifier. However, the degree to which it is possible to use or commit to a randomized strategy varies depending on the setting. There are two main drivers impeding the implementation of the most efficient Stacklberg equilibrium. Firstly, in many real-life classification settings, it might be unacceptable to use a probabilistic classifier, for example, due to legal restrictions (applicants with identical features must obtain identical outcomes). Secondly, for the more complicated scenario with multiple classifiers (such as college admissions), the effect of each Jury on the overall market is small, hence, diminishing the incentive to stick to a randomized strategy 'for the benefit of the market as a whole'. Informally, the best response of a single Jury, when the other classifiers commit to using a randomized classifier, is not a randomized classifier. And even if we got the Juries to commit to randomization, the final probabilities of classification depends on the number of classifiers (k) and hence, the implementation of the most efficient randomized classifier needs coordination between the multiple classifiers. Analyzing the equilibria for multiple classifiers is beyond the scope of this paper but we illustrate the instability of randomized classifier as follows. We show that unless Jury is able to commit to the published randomized classifier, such a classifier is not a stable solution to strategic classification. As mentioned above, randomization helps because of the following observation: if the difference between the probabilities, of being classified as 1 at *adjacent* features is small, the individuals have no incentive to manipulate their features. But, once the Jury knows that no one changed their feature, her best response, then, is to use the classifier that achieves best accuracy given the *true* features.

Formally, we show (Theorem 5) that for any published randomized classifier that achieves strictly higher accuracy compared to any deterministic classifier under Stackelberg equilibrium, Jury has an opportunity to improve its utility and get strictly better accuracy using a classifier different from the published.

The shortcomings of a randomized classifier can be redeemed by addition of noise to the features.

1.1.3 Addition of noise to the features

This brings us to our second scenario that uses noisy features for classification. Every individual has an associated private signal that identifies their qualification. The Jury sees a feature that is a noisy representation of this private signal. The individuals, after incurring certain cost, can effectively manipulate their private signal such that the features are a noisy representation of this updated private signal. Again, the assumption is that such a manipulation didn't change the true qualifications of an individual. We show, through an example of a cost function and a noise distribution, that in the strategic setting, using a deterministic classifier, the Jury achieves better accuracy when the features are noisy than any deterministic classifier in the noiseless case, that is, when Jury gets to see the private signal. This is counter-intuitive at first glance because under no strategic manipulation, noise can only decrease Jury's accuracy.

We also show examples where noisy features can help in achieving fairer outcomes across subpopulations. Let A and B be two subpopulations *identical* in qualifications but having different (but not extremely different) costs of manipulation (and $|A| \leq |B|$; A is a minority). We show, through an example, that no matter whether the minority has higher or lower costs of manipulation than the majority, it is at a disadvantage when Jury publishes a single deterministic classifier to optimize its overall accuracy (noiseless strategic setting). Here, by disadvantage, we mean that the minority has lower classification accuracy than the majority. Next, we show that the addition of appropriate noise to the private signals, in the same example, can ensure that Jury's best response classifier is fair across subpopulations. This is not that surprising as making the features completely noisy also lead to same outcomes for the subpopulations. However, such an addition of noise can also sometimes increase Jury's overall accuracy (improving both accuracy and fairness). We consider the case where the Jury would publish a single classifier for both the subpopulations (for e.g., either because A is a protected group and the Jury is not allowed to discriminate based on the subgroup membership or because the Jury has not yet identified these subpopulations and the differences in their cost functions). Informally, our results, can be stated as follows:

▶ **Theorem 2** (Informal statement of Theorems 6,7,8)**.** *Let A and B be two subpopulations that are* identical *in qualifications. Let $c_A \neq c_B$ be the cost functions for subpopulations A and B respectively. In Case 1, Jury gets to see the private signals and publishes a single deterministic classifier that achieves optimal overall accuracy (sum over the two subpopulations) under the Stackelberg equilibrium (for the cost functions c_A and c_B). In Case 2, the features are noisy representations of the private signal; Jury publishes a single deterministic classifier that achieves optimal overall accuracy under the Stackelberg equilibrium (knowing that the features are noisy). There exists an instantiation of the "identical qualifications" such that*

1. *If $|A| < |B|$, that is, A is a minority, for a wide set of costs functions c_A, c_B, A is always at a disadvantage when in Case 1.*
2. *There exists a setting of the "noise" (η) for each of the above cost functions, such that, Jury's best response in Case 2, is always fair, that is, achieves equal classification accuracy on the subpopulations.*
3. *There exists cost functions c_A, c_B from this wide set of cost functions, and corresponding noise η, such that Jury's accuracy in Case 2 is strictly better than in Case 1.*

This result has potentially interesting policy implications, since it is easier, both practically and legally, to commit to using noisier signals (for example by restricting the types of information available to the Jury) than to commit to disregarding pertinent information ex-post (as in randomized classification). Therefore, future mechanism design efforts involving strategic classification should carefully consider the mechanisms of information disclosure to the Jury.

1.2 Related Work

[8, 2] initiated the study of strategic classification through the lens of Stackelberg competition. [9, 16, 10] study the effects of strategic classification on different subpopulations and how it can exacerbate the social inequity in the world. [9] also made the observation that a single classifier would have varying classification accuracies across subpopulations with different costs of manipulation. [16] defined a concept called "social burden" of a classifier to be the sum of the minimum effort any qualified individual has to put in to be classified as 1. Thus, the subpopulations with higher costs of manipulation would have worse social burden and might be at a disadvantage. In such situations, intuitively, one would think that subsidizing the costs for the disadvantaged population might help. [9] showed that cost subsidy for disadvantaged individuals can sometimes lead to worse outcomes for the disadvantaged group.

In the present paper, we observe that the addition of noise, counter-intuitively, can help Jury's accuracy as well as serve the fairness concerns. There are many examples in game theory where loss of information helps an individual in strategic setting, for example, [6]. [11, 10] also studies the role of hiding information to serve fairness. [7] has a brief discussion at the end of the paper on making manipulated data more informative through addition of noise to the features (this was put online a couple of months after the first version of our paper was made online).

Another work related to Theorem 3 of the present paper is [13], which studies the scope of truthful mechanisms when the agents incur certain costs for misreporting their true type. In particular, the paper gives conditions, on the misreporting costs, that allow the revelation principle to hold, that is, any mechanism can be implemented by a truthful mechanism, where all the agents reveal their true types. The main difference between [13] and our paper is that the former allows the use of monetary transfers to the agents to develop truthful mechanisms and such transfers don't impact the objective value of the mechanism.

1.3 Organization

We formalize the model used for strategic classification in Section 2. In Section 3, we show how randomness helps in achieving better accuracy and efficiency. We also characterize the classifiers that achieve optimal efficiency for *simple* cost functions. In Section 4, we investigate the stability of randomized classifiers. In Section 5, we investigate the role of noisy features in strategic classification.

2 Preliminaries

In this paper, we concern ourselves with classification based on a one-dimensional feature space $\mathcal{X}$. In many of the examples, our feature space $\mathcal{X} \subseteq \mathbb{R}$ is discrete, hence, we use sum ($\sum$) in many of the definitions, but, these definitions are well-defined when $\mathcal{X}$ is taken to be continuous (for e.g., $\mathbb{R}$) by replacing sum ($\sum$) with integrals ($\int$) and probability distributions with probability density functions. We use the notation $\mathcal{N}(z, \sigma)$ to denote the gaussian distribution with mean z and standard deviation σ. We say a function $f : \mathcal{X} \to \{0, 1\}$ is a threshold function (classifier) with threshold τ if

$$f(x) = \begin{cases} 1 & \text{if } x \geq \tau \\ 0 & \text{otherwise} \end{cases}$$

We also use $1_{x \geq \tau}$ to denote a threshold function (classifier) with threshold τ. Sometimes, we will use $1_{x > \tau}$ that classifies x as 1 if and only if $x > \tau$.

2.1 The Model

Let $\mathcal{X}$ be the set of features. Let $\pi : \mathcal{X} \to [0, 1]$ be the probability distribution over the feature set realized by the individuals. Let $h : \mathcal{X} \to [0, 1]$ be the true probability of an individual being qualified (1) given the feature. We also refer to it as the true qualification function. Let $c(x, x')$ be the cost incurred by an individual to manipulate their feature from x to x' (We also use words, change and move, to refer to this manipulation). The classification is modeled as a sequential game where a Jury publishes a classifier (possibly probabilistic) $f : \mathcal{X} \to [0, 1]$ and contestants (individuals) can change their features (after seeing f) as long as they are ready to incur the cost of change. The previous papers in the area considered the task of strategic classification when the published classifier is deterministic binary classifier. Here, we formalize the Stackelberg prediction game for probabilistic classifiers.

Given f, we define the best response of a contestant with feature x[1], as follows

$$\Delta_f(x) = \text{argmax}_{y \in (\{x\} \cup \{x' | (f(x') - f(x)) > c(x,x')\})}(f(y)) \tag{1}$$

We will denote it by Δ when f is clear from the context. $\Delta(x)$ might not be well defined if there are multiple values of y that attains the maximum. In those cases, $\Delta(x)$ is chosen to be the smallest y amongst them. In words, you jump to another feature only if the cost of jumping is less than the advantage in being classified as 1.

We define the Jury's utility for publishing f ($U(f)$) as the classification accuracy with respect to $h(x)$. Thus, Jury's utility for publishing f is

$$\begin{aligned} U(f) &= \sum_{x \in \mathcal{X}} \pi(x)[f(\Delta(x)) \cdot h(x) + (1 - f(\Delta(x)) \cdot (1 - h(x))] \\ &= \sum_{x \in \mathcal{X}} \pi(x)[f(\Delta(x)) \cdot (2h(x) - 1) + 1 - h(x)] \end{aligned}$$

We define $C(f) = \sum_{x \in \mathcal{X}} \pi(x)[h(x) \cdot c(x, \Delta_f(x))]$ to be the cost of strategy for a published classifier f.

We define the efficiency of the classifier f ($E(f)$)[2] as follows:

$$\begin{aligned} E(f) &= U(f) - C(f) \\ &= \sum_{x \in \mathcal{X}} \pi(x)[f(\Delta(x)) \cdot h(x) + (1 - f(\Delta(x)) \cdot (1 - h(x))] - \sum_{x \in \mathcal{X}} \pi(x)[h(x) \cdot c(x, \Delta(x))] \\ &= \sum_{x \in \mathcal{X}} \pi(x)[f(\Delta(x)) \cdot h(x) + (1 - f(\Delta(x)) \cdot (1 - h(x)) - h(x) \cdot c(x, \Delta(x))] \end{aligned}$$

The focus of this paper is to demonstrate what role randomness and noise can play in strategic classification and not to give algorithms for learning the optimal or most efficient strategic classifier. We can present the ideas even by making the following assumptions on the cost function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$:

1. $c(x, x') \geq 0, \ \forall x, x' \in \mathcal{X}$.
2. $c(x', x) = 0, \ \forall x, x' \mid h(x') \geq h(x)$, that is, jumping to a lesser qualified feature is free.
3. $c(x, x'') \leq c(x, x') + c(x', x''), \ \forall x, x', x'' \in \mathcal{X}$, that is, the costs are sub-additive.
4. $c(x, x') \leq c(x, x''), \ \forall x, x', x'' \mid h(x'') \geq h(x')$, that is, jumping to a lesser qualified feature is easier.
5. $c(x', x'') \leq c(x, x''), \ \forall x, x', x'' \mid h(x') \geq h(x)$, that is, jumping from a lesser qualified feature is harder.

The last two points are implied by the first three, we wrote them as separate points for completeness. We call the cost function *simple* if it satisfies all the above assumptions.

By the virtue of the definition of simple cost functions, without loss of generality, we assume that h is monotonically increasing with the feature x, that is, $\forall x, x' \in \mathcal{X}, \ \ x' \geq x \implies h(x') \geq h(x)$.

Next, we mention a special kind of cost function that satisfies the assumptions: $c(x, x') = \max(a(x') - a(x), 0)$ where the function $a : \mathcal{X} \to \mathbb{R}$ is monotonically increasing in x, that is, $x' \geq x \implies a(x') \geq a(x)$.

[1] Such a best response model has been studied in the literature, for example, [17].

[2] We defined efficiency as $U(f) - C(f)$ for the simplicity of the presentation. Defining efficiency as $U(f) - \beta \cdot C(f)$ (for some $\beta > 0$) doesn't effect the theorems except for Theorem 3, which is no longer true for $\beta < 1$.

Given a cost function c, let

$$\mathrm{Lip}_1(c) = \{f \mid f : \mathcal{X} \to [0,1], f(x') - f(x) \leq c(x,x')\ \forall x, x' \in \mathcal{X}\}$$

Given the cost function c, we say f satisfies the Lipschitz constraint if $f \in \mathrm{Lip}_1(c)$. Note that any classifier $f \in \mathrm{Lip}_1(c)$ is monotonically increasing with x, that is, $x' \geq x \implies f(x') \geq f(x)$. This is because $\forall x' \geq x, f(x) - f(x') \leq c(x', x) = 0$. And $\forall x \in \mathcal{X}, \Delta_f(x) = x$, that is, no one changes their feature if f is the published classifier.

In Section 5, we generalize this model to the setting where the features are a noisy representation of an individual's private signal. An individual can make efforts to change their private signal but can't control the noise. The Jury only see the features and classifies an individual based on that. In Section 5, the fairness notion, we will concern ourselves with, is the classification accuracy of the published classifier across subpopulations.

3 Committed Randomness Helps both Utility and Efficiency

In this section, we compare the optimal utility and efficiency achieved by a deterministic binary classifier to a probabilistic classifier. Consider the following two scenarios:

Scenario 1: The Jury commits to using a binary classifier $f : \mathcal{X} \to \{0,1\}$. The best response function $\Delta_f : \mathcal{X} \to \mathcal{X}$, Jury's utility from publishing f $(U(f))$ and efficiency of the classifier f $(E(f))$ are defined as in Section 2.

Scenario 2: The Jury publishes a probabilistic classifier $f : \mathcal{X} \to [0,1]$ and commits to it. The best response function $\Delta_f : \mathcal{X} \to \mathcal{X}$, Jury's utility from publishing f $(U(f))$ and efficiency of the classifier f $(E(f))$ are as defined in Section 2. Note that this is equivalent to when Jury publishes a list of deterministic classifiers and chooses a classifier uniformly at random from them. Contestants update their feature without knowing which classifier gets picked up at the end.

The following example illustrates how randomization helps in getting strictly better utility and efficiency:

Let $\mathcal{X} = \{1,2\}$ and each feature contains half of the population. Let

$$h(x) = \begin{cases} 1 & \text{if } x = 2 \\ 0 & \text{otherwise} \end{cases}$$

Let the cost of changing the feature from 1 to 2 be 0.5. The the randomized classifier f defined as follows:

$$f(x) = \begin{cases} 1 & \text{if } x = 2 \\ 0.5 & \text{if } x = 1 \end{cases}$$

achieves an accuracy of 0.75. The contestants at $x = 2$ are happy as they are already being classified as 1 with probability 1. For the contestants at $x = 1$, $f(2) - f(1) = 0.5 = c(1,2)$ and hence, they don't have an incentive to manipulate their feature. As all the contestant retain their true features, the efficiency of f is also equal to 0.75. As the feature space is bounded, there are only three options for a deterministic classifier: keep the threshold at 1 and classify everyone as 1; keep the threshold at 2 and you end up classifying everyone as 1, as the contestants at 1 change their feature to 2; classify everyone as 0. All these classifiers have 0.5 accuracy and at most 0.5 efficiency.

In the mathematical example given above, the randomized classifier was set up such that none of the contestants had any incentive to change their feature. In the next subsection, we show that the most efficient classifier always looks like “this" for “simple" cost functions.

That is, if the cost function c satisfies the assumptions made in Section 2, then for every true qualification function h, there exists a function $f_h \in \mathrm{Lip}_1(c)$ that achieves the optimal efficiency.

3.1 Most Efficient Classifier for Simple Cost Functions

Recall, $E(f) = \sum_{x\in\mathcal{X}} \pi(x)[f(\Delta(x)) \cdot h(x) + (1 - f(\Delta(x)) \cdot (1 - h(x)) - h(x) \cdot c(x, \Delta(x))]$. Let $E^* = \max_{f:\mathcal{X}\to[0,1]} \sum_{x\in\mathcal{X}} \pi(x)[f(\Delta(x)) \cdot h(x) + (1 - f(\Delta(x)) \cdot (1 - h(x)) - h(x) \cdot c(x, \Delta(x))]$.

▶ **Theorem 3.** *For every monotone true qualification function $h : \mathcal{X} \to [0, 1]$, probability distribution $\pi : \mathcal{X} \to [0, 1]$ over the features, simple cost function c, there exists $g \in Lip_1(c)$ such that $E(g) = E^*$.*

Proof. Let f be an efficiency maximizing classifier. We argue that $g : \mathcal{X} \to [0, 1]$ defined as

$$g(x) = \max_y\{f(y) - c(x, y)\}$$

is in $\mathrm{Lip}_1(c)$ and satisfies $E(g) \geq E(f)$. Let $\delta_f(x) = \mathrm{argmax}_y\{f(y) - c(x, y)\}$. When f is clear from the context, we will drop the subscript on δ. Using definition of δ, $g(x) \in [0, 1]$ as $\forall x, y \in \mathcal{X},\ f(y)-c(x, y) \leq f(y) \leq 1$ $(c(x, y) \geq 0)$ and $\max_y\{f(y)-c(x, y\} \geq f(x)-c(x, x) \geq 0$. For all $x, x' \in \mathcal{X}$,

$$\begin{aligned} g(x') - g(x) &= f(\delta(x')) - c(x', \delta(x')) - f(\delta(x)) + c(x, \delta(x)) \\ &= f(\delta(x')) - c(x, \delta(x')) - f(\delta(x)) + c(x, \delta(x)) + (c(x, \delta(x')) - c(x', \delta(x'))) \\ &\leq c(x, \delta(x')) - c(x', \delta(x')) \leq c(x, x') \qquad \text{(sub-additivity)} \end{aligned}$$

The first inequality follows the definition of δ, that is, $\forall y \in \mathcal{X}, f(\delta(x)) - c(x, \delta(x)) \geq f(y) - c(x, y)$. Therefore, $f(\delta(x')) - c(x, \delta(x')) - f(\delta(x)) + c(x, \delta(x)) \leq 0$. The second inequality follows from the fact that the cost function c is simple and satisfies the sub-additivity condition. This proves that $g \in \mathrm{Lip}_1(c)$. This implies, as observed previously, $\forall x \in \mathcal{X}, \Delta_g(x) = x$. Next, we show that $E(g) \geq E(f)$ and hence $E(g) = E^*$. Efficiency of the classifier g is

$$\begin{aligned} E(g) &= \sum_{x\in\mathcal{X}} \pi(x)[g(\Delta_g(x)) \cdot h(x) + (1 - g(\Delta_g(x)) \cdot (1 - h(x)) - h(x) \cdot c(x, \Delta_g(x))] \\ &= \sum_{x\in\mathcal{X}} \pi(x)[2 \cdot g(x) \cdot h(x) - g(x) - h(x) + 1] \end{aligned}$$

Efficiency of the classifier f is

$$\begin{aligned} E(f) &= \sum_{x\in\mathcal{X}} \pi(x)[f(\Delta_f(x)) \cdot h(x) + (1 - f(\Delta_f(x)) \cdot (1 - h(x)) - h(x) \cdot c(x, \Delta_f(x))] \\ &= \sum_{x\in\mathcal{X}} \pi(x)[2f(\Delta(x)) \cdot h(x) - f(\Delta(x)) - h(x) + 1 - h(x) \cdot c(x, \Delta(x))] \end{aligned}$$

$E(g) - E(f) = \sum_{x\in\mathcal{X}} \pi(x)[(g(x) - f(\Delta(x))) \cdot (2h(x) - 1) + h(x) \cdot c(x, \Delta(x))]$

▷ Claim 4. $\forall x,\ \ [(g(x) - f(\Delta(x))) \cdot (2h(x) - 1) + h(x) \cdot c(x, \Delta(x))] \geq 0$.

Please refer to Appendix A for the proof of the claim. It's straightforward to see that $E(g) - E(f) \geq 0$ using the above claim. Therefore, we showed a classifier $g \in \mathrm{Lip}_1(c)$ such that $E(g) = E^*$. ◀

In words, *when we are concerned with the efficiency of the published classifier, the optimal is achieved by a probabilistic classifier that has zero cost of strategy and gives individuals no incentive to change their feature.*

4 Are Randomized Classifiers in Equilibrium from Jury's Perspective?

As discussed in the Section 1, there are many obstacles to implementing a randomized classifier in the strategic setting. In this section, we illustrate the instability caused by the use of randomized classifiers (which becomes increasingly important while considering multiple classifiers). In Section 3, we saw that a randomized classifier can achieve better accuracy and efficiency than any binary classifier. While maximizing efficiency, we further showed that the optimally efficient classifier is such that every contestant reveals their true feature. Once the Jury knows the contestants' true features, she can be greedy and classify the individuals using a threshold function with $\tau = \min\{x \mid h(x) \geq \frac{1}{2}\}$ as the threshold to achieve the best accuracy. Therefore, unless the Jury commits to using randomness, she has an incentive of not sticking to the promised randomized classifier. The question is: what's the best accuracy/efficiency achieved by a classifier that is in equilibrium even from Jury's perspective? We formalize this equilibrium concept as follows (the true qualification function h and the cost function c are fixed):

1. Jury publishes a randomized classifier $f : \mathcal{X} \to [0,1]$.
2. Contestants, knowing f, changes their feature from x to $\Delta_f(x)$.
3. f is in equilibrium from Jury's perspective if given that the contestants changed their features according to the best response function Δ_f, f achieves the best classification accuracy, that is, for all classifiers $g \in \mathcal{X} \to [0,1]$,

$$\sum_{x \in \mathcal{X}} \pi(x)[f(\Delta_f(x)) \cdot h(x) + (1 - f(\Delta_f(x)) \cdot (1 - h(x))] \qquad (2)$$
$$- \sum_{x \in \mathcal{X}} \pi(x)[g(\Delta_f(x)) \cdot h(x) + (1 - g(\Delta_f(x)) \cdot (1 - h(x))] \geq 0$$

Using next theorem, we show that for any randomized classifier that is in equilibrium from Jury's perspective, there exists a binary classifier that achieves at least the same accuracy.

▶ **Theorem 5.** *Given a monotone true qualification function h, probability distribution π over the features, and a simple cost function c, let $f^* : \mathcal{X} \to \{0,1\}$ be the classifier that optimizes Jury's utility over the deterministic classifiers under Stackelberg equilibrium. Let $f : \mathcal{X} \to [0,1]$ be a randomized classifier such that $U(f) > U(f^*)$, then f is not in an equilibrium from Jury's perspective (the notion defined above).*

Please refer to Appendix B for the proof.

Disclaimer: f' as defined above might also not be in equilibrium from Jury's perspective. The above theorem illustrates the following point: *Jury doesn't benefit from randomized classifiers without creating instability in the system.*

Can we somehow exploit this power of randomness while overcoming the obstacles to randomized classification? The answer is yes – make the features noisy.

5 Noisy Features Give the System Free Randomness

We formalize the setting with noisy features as follows: every individual has a private signal $y \in \mathcal{X}$. The true qualification function $h : \mathcal{X} \to [0,1]$ depends on y, that is, $h(y)$ is the probability of an individual being qualified (1) given that its private signal is y. Given a private signal y, a feature is drawn randomly from the distribution $p_y : \mathcal{X} \to [0,1]$, that is, $p_y(x)$ is the probability that an individual's feature is x when their private signal is y. If $\mathcal{X} = \mathbb{R}$, the right intuition for p_y is it being $\mathcal{N}(y, \sigma)$ where $\mathcal{N}(y, \sigma)$ is the gaussian distribution with mean y and standard deviation σ. Let $\pi : \mathcal{X} \to [0,1]$ be the probability distribution over the private signals y realized by the individuals.

Let $c(y, y')$ be the cost incurred by the contestant to change their private signal from y to y'. The contestants can put effort to change their private signals but the feature would still be drawn randomly using the updated private signal.

The classification is again modeled as a sequential game where a Jury publishes a deterministic classifier $f : \mathcal{X} \to \{0,1\}$. We restricts ourselves to deterministic classifiers due to the observations made in Section 4. Contestants change their private signals as long as they are ready to incur the cost of change. Given a private signal y, let $q_f(y)$ denote the probability of a contestant, with private signal y, being classified as 1 when f is the classifier. Therefore, $q_f(y) = \sum_{x \in \mathcal{X}} p_y(x) \cdot f(x)$.

Given f, the best response of a contestant with private signal y is given as,

$$\Delta_f(y) = \text{argmax}_{z \in \{y\} \cup \{y' | q_f(y') - q_f(y) > c(y,y')\}} (q_f(z)) \tag{3}$$

We will denote it by Δ when f is clear from the context. $\Delta(y)$ might not be well defined if there are multiple values of z that attains the maximum. In those cases, $\Delta(y)$ is chosen to be the smallest z amongst them. In words, you jump to another private signal only if the cost of jumping is less than the advantage in being classified as 1. Even though f is deterministic, due to noisy features, the effective classifier given the private signal y (q_f) is probabilistic. Therefore, we will see below that the noise allows us similar advantages as that of a probabilistic classifier.

The accuracy and efficiency of the classifier f are defined as follows:

$$U(f) = \sum_{y \in \mathcal{X}} \pi(y)[q_f(\Delta(y)) \cdot h(y) + (1 - q_f(\Delta(y)) \cdot (1 - h(y))]$$

$$E(f) = \sum_{y \in \mathcal{X}} \pi(y)[q_f(\Delta(y)) \cdot h(y) + (1 - q_f(\Delta(y)) \cdot (1 - h(y))] - \sum_{y \in \mathcal{X}} \pi(y)[h(y) \cdot c(y, \Delta(y))]$$

We assume that h is monotonically increasing with y and the cost function c is simple. Next, we will demonstrate how noisy features can lead fairer outcomes and even increase Jury's accuracy.

5.1 Noisy Features achieve Fairer Equilibriums

Consider two subpopulations A and B. For simplicity, these subpopulations are a partition of the individuals in the universe. Let s_A denote the probability an individual from the universe is in subpopulation A. Similarly, s_B ($s_A = 1 - s_B$). Let $h_A : A \to [0,1]$ be the true qualification function for the subpopulation A. Similarly, h_B. Let $c_A : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be the cost function for the subpopulation A, that is, $c_A(y, y')$ is the cost of changing the private signal from y to y' for an individual in A. Similarly, c_B is defined. Let $\pi_A : A \to [0,1]$ and π_B be the probability distribution over the private signals realized by the subpopulations A and B respectively.

Given a published deterministic classifier $f : \mathcal{X} \to \{0,1\}$, the best response of the contestant in subpopulation A with private signal y ($\Delta_f^A(y)$) is defined using c_A as the cost function. Similarly, for subpopulation B, let $\Delta_f^B(y)$ denote the best response of the contestant in subpopulation B with private signal y and when the published classifier is f. We use $U_A(f)$ and $U_B(f)$ to denote the accuracy of the classifier f on the respective subpopulations.

We consider the setting where $h_A = h_B = h$ and $\pi_A = \pi_B = \Pi$, but the cost functions c_A and c_B are different. In this section, we use the symbol Π to denote the probability distribution over the private signals to avoid confusion with the Archimedes' constant π.

In our first example, we show that even though the subpopulations are identical with respect to their qualifications, different costs can lead to unfair classification when classification is based on private signals. Through our second example, we show that the use of noisy features, for strategic classification, can lead to increase in the overall accuracy of classification as well as give fair classification. We evaluate the fairness of a classifier f quantitively using the difference between the accuracies, that is, $|U_A(f) - U_B(f)|$.

Let's start with the example. $\mathcal{X} = \mathbb{R}$. Let the true qualification function for both the subpopulations be as follows: $h(y) = \begin{cases} 1 & \text{if } y > d \\ \frac{y}{2d} + \frac{1}{2} & \text{if } y \in [-d, d] \\ 0 & \text{if } y < -d \end{cases}$, where d is a fixed large enough positive real number. Let the probability density function on the private signals realized by the subpopulations be as follows: $\Pi(y) = \frac{e^{-\frac{y^2}{2t^2}}}{\sqrt{2\pi}t}$, that is, the gaussian distribution with mean 0 and standard deviation t. Again, t is fixed positive real number. We assume $d >> t$.

Let σ_A and σ_B be positive real numbers. The cost function for a subpopulation $S \in \{A, B\}$ is defined as follows (with $(y' - y)^+ = \max\{y' - y, 0\}$):

$$c_S(y, y') = \frac{(y' - y)^+}{\sqrt{2\pi}\sigma_S} \tag{4}$$

We start with the setting where the features are the private signals and not a noisy representation of them.

Remark: If the Jury is allowed to publish different classifiers for the two subpopulations, then she can achieve "the best possible accuracy" on both the subpopulations. It's easy to see that the classifier $f_S : \mathcal{X} \to \{0, 1\}$, defined as follows, achieves as much accuracy as a classifier under no strategic manipulation of the features can achieve on the subpopulation $S \in \{A, B\}$: $f_S(y) = \begin{cases} 1 & \text{if } y \geq \sqrt{2\pi}\sigma_S \\ 0 & \text{otherwise} \end{cases}$.

All the contestants in a subpopulation S, with $0 < y < \sqrt{2\pi}\sigma_S$ report their private signals to be $\sqrt{2\pi}\sigma_S$ as cost of this change is < 1 whereas the advantage gained in the probability of being classified as 1 is 1. For all the contestants with private signal $y \leq 0$, the cost of change is too high (≥ 1) and thus, they report their true private signals. Therefore, the classifier f_S ends up classifying everyone with private signal $y > 0$ as 1 which is the accuracy maximizing classification under the "no strategic manipulation" setting.

How strategic classification leads to unfairness: When $\sigma_A \neq \sigma_B$, the optimal classifiers for the subpopulations A and B are different and hence, when we choose a single classifier for both the subpopulations, we are bound to loose on the accuracy of at least one of the subpopulations. Through an example (Theorem 6), we suggest that: *while maximizing the overall accuracy over the universe, the minority group might be at a disadvantage irrespective of whether their costs to change the private signals are higher or lower than the majority subpopulation.* Without loss of generality, we assume that A is the minority subpopulation, that is, $s_A \leq s_B$. In many real life scenarios, the Jury would publish a single classifier for both the subpopulations either because A is a protected group and the Jury is not allowed to discriminate based on the subgroup membership or because the Jury has not yet identified these subpopulations and the differences in their cost functions.

▶ **Theorem 6.** *Let A and B be two subpopulations such that the true qualification functions, h_A, h_B, the probability density functions, π_A, π_B and the cost functions c_A, c_B are as instantiated above.*

Assuming $|\sigma_A - \sigma_B| \leq \frac{t}{\sqrt{2\pi}}$, let f^* be the deterministic classifier that maximizes Jury's utility ($U(f)$), if $s_A < s_B$ and $\sigma_A \neq \sigma_B$ (the cost functions are different), then $U_A(f^*) < U_B(f^*)$, that is, the minority is at a disadvantage, even though their qualifications were identical ($h_A = h_B$, $\pi_A = \pi_B$).

Please refer to Appendix C for the proof.

Next we show that, when the features are appropriately noisy, the optimal classifier from Jury's perspective is fair to the subpopulations. The intuition is as follows: if the noise is large enough such that none of contestants in either of the subpopulations want to manipulate their private signals, then the cost differences become irrelevant and hence, the optimal classifier achieves equal accuracy on both the subpopulations. You would think that this addition of noise would compromise Jury's utility. Subsequently, we show that adding noise might also improve the overall accuracy of the Jury's optimal classifier, therefore, addition of noise can make everyone happier. The latter is a continuation to the results at the start of Section 5 about the usefulness of noise to the Jury under strategic classification.

Noisy features lead to fairer outcomes: Now, we analyze the setting with noisy features and prove the following theorem. The true qualification function h, cost functions (c_A and c_B) and the probability density function Π are as defined for the first example. Let $\sigma = \max\{\sigma_A, \sigma_B\}$. Given a private signal y, the features x are distributed according to the gaussian with mean y and standard deviation σ. The probability density function for the feature x given the private signal y is $p_y(x) = \frac{e^{-\frac{(x-y)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}$.

▶ **Theorem 7.** *Let A and B be two subpopulations such that the true qualification functions, h_A, h_B, the probability density functions, π_A, π_B and the cost functions c_A, c_B are as instantiated above. When the features are drawn with a gaussian noise of mean 0 and standard deviation σ, such that, $\sigma \geq \sigma_A, \sigma_B$, if f^* is the deterministic classifier that maximizes Jury's utility ($U(f)$), then f^* is fair, that is, $U_A(f^*) = U_B(f^*)$.*

Please refer to Appendix D for the proof.

Theorem 7 would hold for when we are concerned with multiple subpopulations as long as $\sigma \geq \sigma_S$ for every relevant subpopulation S. In words, using noisy features we *can* ensure that the best response of a Jury, maximizing her own utility, is fair to all the subpopulations that are identical in terms of qualifications but different in terms of the costs to manipulate the private signals, as long as the costs of manipulation for a subpopulation are not too small.

Noisy features can also improve Jury's utility: Next, we show that further in some cases, *the addition of noise to the features is not only beneficial for ensuring fairness but might also achieve better overall accuracy under strategic classification compared to when a noiseless signal is used.*

Retaining the instantiations of h_A, h_B, π_A, π_B, c_A, c_B and σ as above, consider the following two scenarios: 1. Jury bases her classifier on the private signal y. 2. The features are drawn with a gaussian noise of mean 0 and standard deviation σ and Jury bases her classifier on the features (x).

Let f_0^* and f_σ^* be the optimal classifiers under strategic classification in the two scenarios respectively. Let $U(f_0^*)$ be the overall classification accuracy (Jury's utility) under Scenario 1 and $U(f_\sigma^*)$ be the overall classification accuracy (Jury's utility) under Scenario 2. We assume that the subpopulations are equally populated, that is, $s_A = s_B$ for simplicity of calculations in the next theorem.

▶ **Theorem 8.** *There exists qualification functions, h_A, h_B, the probability density functions over the private signals, π_A, π_B, the cost functions c_A, c_B and $\sigma > 0$ such that, $U(f_\sigma^*) > U(f_0^*)$, that is, the Jury gets better classification accuracy when the features are drawn with a gaussian noise of mean 0 and standard deviation σ. Here, the subpopulations have identical qualifications ($h_A = h_B$, $\pi_A = \pi_B$) but different cost functions.*

Please refer to Appendix E for the proof. This theorem corroborates the idea that not only the subpopulations, but even the Jury might prefer noisy features. In the above example, for simplicity, we assumed $s_A = s_B$. Therefore, the optimal classifier was fair even in the noiseless setting. But a slight tweak in s_A so that $s_A < s_B$ wouldn't change Jury's utility, in Scenario 1, by much and thus, would give an example where the noiseless setting has both unfairness and lesser overall classification accuracy.

In this paper, we study the interaction of noise with strategic classification through some simple examples, and leave the task of generalizing these results for future research.

6 Discussion

The problem of classification (and the strategic classification problem it entails) is of tremendous importance both practically (affecting pretty much every industry) and theoretically (with implications ranging from algorithms to policy and law). Therefore, clarifying the role randomness plays in this specific family of games is an important goal. Just as in games, randomness may lead to better solution in strategic classification. Moreover, in many important settings (such as college admissions in some jurisdictions), the classifier is required to be deterministic by law — which is not a handicap for algorithmic classification, but is a handicap for strategic one. In addition, we proved that, in many natural cases, any randomized classifier (based on one-dimension) that achieves strictly better accuracy than the optimal deterministic one is not stable from the classifier's standpoint, thus illustrating the difficulty of implementing a randomized classifier in a more complicated scenario with multiple classifiers (such as college admissions). This motivates the use of noisy features as a commitment device, which can improve both accuracy and fairness, and is also practically possible (for example by restricting the types of information available to the classifier).

References

1. Emrah Akyol, Cedric Langbort, and Tamer Basar. Price of transparency in strategic machine learning. *arXiv preprint arXiv:1610.08210*, 2016.
2. Michael Brückner and Tobias Scheffer. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 547–555. ACM, 2011.
3. Yiling Chen, Chara Podimata, Ariel D Procaccia, and Nisarg Shah. Strategyproof linear regression in high dimensions. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 9–26. ACM, 2018.
4. Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70. ACM, 2018.
5. Kfir Eliaz and Ran Spiegler. The model selection curse. *American Economic Review: Insights.*
6. Richard Engelbrecht-Wiggans. On the value of private information in an auction: ignorance may be bliss. *BEBR faculty working paper; no. 1242*, 1986.

7 Alex Frankel and Navin Kartik. Improving information from manipulable data. *arXiv preprint arXiv:1908.10330*, 2019.

8 Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122. ACM, 2016.

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

10 Nicole Immorlica, Katrina Ligett, and Juba Ziani. Access to population-level signaling as a source of inequality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 249–258. ACM, 2019.

11 Sampath Kannan, Aaron Roth, and Juba Ziani. Downstream effects of affirmative action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 240–248. ACM, 2019.

12 Andrew Kephart and Vincent Conitzer. Complexity of mechanism design with signaling costs. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 357–365, 2015.

13 Andrew Kephart and Vincent Conitzer. The revelation principle for mechanism design with reporting costs. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 85–102, 2016.

14 Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 825–844. ACM, 2019.

15 John Miller, Smitha Milli, and Moritz Hardt. Strategic adaptation to classifiers: A causal perspective. *arXiv preprint arXiv:1910.10362*, 2019.

16 Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 230–239. ACM, 2019.

17 Christopher A Wilkens, Ruggiero Cavallo, Rad Niazadeh, and Samuel Taggart. Mechanism design for value maximizers. *arXiv preprint arXiv:1607.04362*, 2016.

A Proof of Claim 4

Recalling, $g(x) = f(\delta(x)) - c(x, \delta(x))$. Using definition of δ, we know that

$$g(x) = f(\delta(x)) - c(x, \delta(x)) \geq f(\Delta(x)) - c(x, \Delta(x)) \tag{5}$$

And, using definition of Δ, we can show that

$$f(\Delta(x)) \geq g(x) \tag{6}$$

This is because, either $f(\delta(x)) - c(x, \delta(x)) = f(x)$ and as $f(\Delta(x)) \geq f(x)$, we get the inequality. Or, $f(\delta(x)) - c(x, \delta(x)) > f(x)$, which implies that x has an incentive to change its feature to $\delta(x)$. Therefore, by the definition of Δ, $f(\Delta(x)) \geq f(\delta(x)) \geq f(\delta(x)) - c(x, \delta(x))$. The expression in the claim can be rewritten as

$$\begin{aligned}&(g(x) - f(\Delta(x))) \cdot (2h(x) - 1) + h(x) \cdot c(x, \Delta(x)) \\ &\qquad = (g(x) - f(\Delta(x))) \cdot (h(x) - 1) + h(x) \cdot (g(x) - f(\Delta(x)) + c(x, \Delta(x)))\end{aligned}$$

As $g(x) - f(\Delta(x)) \leq 0$ from Equation 6 and $g(x) - f(\Delta(x)) + c(x, \Delta(x)) \geq 0$ from Equation 5, the inequality follows from the fact that $0 \leq h(x) \leq 1$. This proves the claim.

B Proof of Theorem 5

Equation 2 implies that for all classifiers $g \in \mathcal{X} \to [0,1]$,

$$\sum_{x\in\mathcal{X}} \pi(x)[(f(\Delta_f(x)) - g(\Delta_f(x))) \cdot (2h(x)-1)] \geq 0$$
$$\implies \sum_{y\in\mathcal{X}} (f(y)-g(y)) \cdot \sum_{x:\Delta_f(x)=y} \pi(x)(2h(x)-1) \geq 0$$

Therefore, if f is in equilibrium from the Jury's perspective, for all $y \in \mathcal{X}$ such that $f(y) \in (0,1)$, $\sum_{x:\Delta_f(x)=y} \pi(x)(2h(x)-1) = 0$ otherwise Jury can choose $g(y) = 1$ (or 0) depending on whether $\sum_{x:\Delta_f(x)=y} \pi(x)(2h(x)-1) > 0$ (or < 0) to increase her accuracy. Therefore, accuracy of the classifier f is given by

$$\begin{aligned} U(f) &= \sum_{x\in\mathcal{X}} \pi(x)[f(\Delta_f(x)) \cdot (2h(x)-1) + (1-h(x))] \\ &= \sum_{y\in\mathcal{X}} f(y) \cdot \sum_{x:\Delta_f(x)=y} \pi(x)(2h(x)-1) + \sum_{x\in\mathcal{X}} \pi(x)(1-h(x)) \\ &= \sum_{y:f(y)=1} \sum_{x:\Delta_f(x)=y} \pi(x)(2h(x)-1) + \sum_{x\in\mathcal{X}} \pi(x)(1-h(x)) \end{aligned}$$

Consider a binary classifier $f' : \mathcal{X} \to \{0,1\}$ defined as follows: $f(x) \in [0,1) \implies f'(x) = 0$ and $f(x) = 1 \implies f'(x) = 1$. We can show that $U(f') \geq U(f)$. The contestants who change their features when f' is the published classifier is a subset of $\{x \in \mathcal{X} \mid f(\Delta_f(x)) \in (0,1]\}$ and as $\sum_{x:f(\Delta_f(x))\in(0,1)} \pi(x)(2h(x)-1) = 0$, the accuracy of f' can only increase. This is because: $\forall x \in \mathcal{X}$ if $f(\Delta_f(x)) = 0$, then $f'(\Delta_{f'}(x)) = 0$ as otherwise if x changed its feature under f', it had an incentive to change under f too.

If $x' > x$, $f(\Delta_f(x')), f(\Delta_f(x)) \in (0,1)$ and x changes its feature under f', then x' has the incentive to change too as $c(x',x) = 0$, and hence, the subset of $\{x \in \mathcal{X} \mid f(\Delta_f(x)) \in (0,1)\}$ that change their features under f' can only do a positive addition to the utility (h is monotonically increasing with x and $\sum_{x:f(\Delta_f(x))\in(0,1)} \pi(x)(2h(x)-1) = 0$). And, the contestants (x) who changed their features under f such that $f(\Delta_f(x)) = 1$ would also change their features under f' such that $f'(\Delta_{f'}(x)) = 1$ (as $f'(x) \leq f(x)$) and are already included in the calculation of $U(f)$.

C Proof of Theorem 6

Jury publishes a deterministic classifier and as there's no noise involved, without loss of generality, we can assume that f is a threshold classifier on the space $\mathcal{X}$ (as c_A and c_B are simple cost functions). This assumption is justified in Section 3. Given the classifier $f : \mathcal{X} \to \{0,1\}$ with threshold τ, the best response of a contestant in the subpopulation $S \in \{A,B\}$ is given as follows:

$$\Delta_f^S(y) = \begin{cases} y & \text{if } y \geq \tau \\ \tau & \text{if } \tau - \sqrt{2\pi}\sigma_S < y < \tau \\ y & \text{if } y \leq \tau - \sqrt{2\pi}\sigma_S \end{cases}$$

The accuracy of the classifier f for the subpopulation S is given as follows:

$$U_S(f) = \int_{-\infty}^{\infty} \Pi(y)[f(\Delta^S(y)) \cdot (2h(y)-1) + (1-h(y))]dy$$

Let $c = \int_{-\infty}^{\infty} \Pi(y)[(1-h(y))]dy$ which is independent of the subpopulation and the classifier. Therefore, $U_S(f) = \left(\int_{-\infty}^{\infty} \Pi(y)[f(\Delta^S(y)) \cdot (2h(y)-1)]dy\right) + c$.

For the convenience of calculations, we will replace $h(y)$ with the following function,

$$h'(y) = \frac{y}{2d} + \frac{1}{2}$$

As d is large and Π is a gaussian centered at 0, this change barely affects the utility values. To be precise, the difference in the utility calculations for any classifier f while using h' instead of h is bounded by

$$\begin{aligned}
\left|\int_{-\infty}^{\infty} \Pi(y)[f(\Delta^S(y)) \cdot 2(h(y)-h'(y))]dy\right| &\leq 2\int_{-\infty}^{\infty} \Pi(y)[f(\Delta^S(y))|h(y)-h'(y)|]dy \\
&\leq 2\int_{-\infty}^{\infty} \Pi(y) \cdot |h(y)-h'(y)|dy \\
&= 4\int_{d}^{\infty} \Pi(y) \cdot \left(\frac{y}{2d} - \frac{1}{2}\right) dy \\
&\leq 2\int_{d}^{\infty} \frac{e^{-\frac{y^2}{2t^2}}}{\sqrt{2\pi}t} \cdot \frac{y}{d}\, dy \quad = 2\frac{te^{-\frac{d^2}{2t^2}}}{\sqrt{2\pi}d}
\end{aligned}$$

As we take d ($d >> t$) to be large enough, we would be able to ignore this difference. From now onwards, we use h' as the "true qualification function".

Therefore, the accuracy of the classifier f over the subpopulation $S \in \{A, B\}$ can be approximated by

$$\begin{aligned}
U_S(f) &= \left(\int_{-\infty}^{\infty} \Pi(y)[f(\Delta^S(y)) \cdot (2h'(y)-1)]dy\right) + c = \left(\int_{-\infty}^{\infty} \Pi(y) \cdot f(\Delta^S(y)) \cdot \frac{y}{d}\, dy\right) + c \\
&= \left(\int_{\tau-\sqrt{2\pi}\sigma_S}^{\infty} \frac{e^{-\frac{y^2}{2t^2}}}{\sqrt{2\pi}t} \cdot \frac{y}{d}\, dy\right) + c = \frac{t}{\sqrt{2\pi}d} e^{-(\tau-\sqrt{2\pi}\sigma_S)^2/2t^2} + c
\end{aligned}$$

The second last equality follows from the definition of Δ_f^S and the fact that f classifies everyone, with the updated private signal greater than or equal to τ, as 1 and 0 otherwise.

The overall accuracy of the classifier f is given by

$$\begin{aligned}
U(f) &= s_A \cdot U_A(f) + s_B \cdot U_B(f) \\
&= s_A \cdot \frac{t}{\sqrt{2\pi}d} e^{-(\tau-\sqrt{2\pi}\sigma_A)^2/2t^2} + s_B \cdot \frac{t}{\sqrt{2\pi}d} e^{-(\tau-\sqrt{2\pi}\sigma_B)^2/2t^2} + c
\end{aligned} \tag{7}$$

It's clear from the expression that the accuracy for the subpopulation A is maximized at $\tau_A = \sqrt{2\pi}\sigma_A$ and that of B is maximized at $\tau_B = \sqrt{2\pi}\sigma_B$. Consider the case when $s_A < s_B$. As $\tau_A \neq \tau_B$, and $U_B(f)$ has a larger weight in the expression, intuitively, while optimizing the overall accuracy, τ would try to achieve better accuracy for the subpopulation B, irrespective of whether $\sigma_A > \sigma_B$ or $\sigma_A < \sigma_B$, leading to unfairness across the subpopulations (A being at a disadvantage).

It's complicated to calculate the optimal τ, below we give a proof of the fact that the optimal τ would be such that $U_A(f) < U_B(f)$. To find the optimal value of τ, we differentiate $U(f)$ with respect τ as follows:

$$\frac{dU(f)}{d\tau} = s_A \cdot \frac{dU_A(f)}{d\tau} + s_B \cdot \frac{dU_B(f)}{d\tau}$$
$$= -\frac{1}{\sqrt{2\pi}td}\left(s_A \cdot (\tau - \sqrt{2\pi}\sigma_A) \cdot e^{-(\tau-\sqrt{2\pi}\sigma_A)^2/2t^2} + s_B \cdot (\tau - \sqrt{2\pi}\sigma_B) \cdot e^{-(\tau-\sqrt{2\pi}\sigma_B)^2/2t^2}\right)$$

Therefore, $\frac{dU(f)}{d\tau} = 0$

$$\implies s_A \cdot (\tau - \sqrt{2\pi}\sigma_A) \cdot e^{-(\tau-\sqrt{2\pi}\sigma_A)^2/2t^2} + s_B \cdot (\tau - \sqrt{2\pi}\sigma_B) \cdot e^{-(\tau-\sqrt{2\pi}\sigma_B)^2/2t^2} = 0$$
$$\implies \left|\frac{(\tau - \sqrt{2\pi}\sigma_A) \cdot e^{-(\tau-\sqrt{2\pi}\sigma_A)^2/2t^2}}{(\tau - \sqrt{2\pi}\sigma_B) \cdot e^{-(\tau-\sqrt{2\pi}\sigma_B)^2/2t^2}}\right| > 1 \qquad (s_B > s_A)$$

As $ze^{-\frac{z^2}{2t^2}}$ is maximized at $z = t$, as long as $|\sigma_A - \sigma_B| \leq \frac{t}{\sqrt{2\pi}}$ (implying $|\tau - \sqrt{2\pi}\sigma_S| \leq t$ for $S \in \{A, B\}$), the overall accuracy is maximized at a threshold τ such that $|\tau - \sqrt{2\pi}\sigma_A| > |\tau - \sqrt{2\pi}\sigma_B|$ and hence, $U_A(f^*) < U_B(f^*)$, where f^* is the optimal classifier from Jury's perspective. The assumption, $|\sigma_A - \sigma_B| \leq \frac{t}{\sqrt{2\pi}}$, can be interpreted as the subpopulations being different but not extremely different, which is reasonable assumption in many real life scenarios.

D Proof of Theorem 7

Again, we will replace the function h with h' (as in proof of Theorem 6) while loosing an insignificant amount in all the calculations ($d >> t, \sigma$). Let $\Pi' : \mathcal{X} \to [0, 1]$ be the probability density function over the features realized by each of the subpopulations. Let $H(x)$ ($H : \mathcal{X} \to [0, 1]$) represent the probability of an individual being qualified (1) given that the Jury sees feature x. These functions are same for both the subpopulations. As the Jury only sees the feature and not the private signal, her accuracy is information-theoretically limited by these functions as we will describe below. Firstly, $\Pi' : \mathcal{X} \to [0, 1]$ is given as follows:

$$\Pi'(x) = \int_{-\infty}^{\infty} \Pi(y) \cdot p_y(x)dy = \int_{-\infty}^{\infty} \frac{e^{-\frac{y^2}{2t^2}}}{\sqrt{2\pi}t} \cdot \frac{e^{-\frac{(x-y)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}dy$$
$$= \int_{-\infty}^{\infty} \frac{e^{-\frac{x^2}{2(\sigma^2+t^2)}}}{\sqrt{2\pi}t} \cdot \frac{e^{-(y-\frac{xt^2}{\sigma^2+t^2})^2/(2\frac{\sigma^2t^2}{\sigma^2+t^2})}}{\sqrt{2\pi}\sigma}dy$$
$$= \frac{e^{-\frac{x^2}{2(\sigma^2+t^2)}}}{\sqrt{2\pi}t \cdot \sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-(y-\frac{xt^2}{\sigma^2+t^2})^2/(2\frac{\sigma^2t^2}{\sigma^2+t^2})}dy$$
$$= \frac{e^{-\frac{x^2}{2(\sigma^2+t^2)}}}{\sqrt{2\pi}t \cdot \sqrt{2\pi}\sigma}\sqrt{2\pi\frac{\sigma^2t^2}{\sigma^2+t^2}} = \frac{e^{-\frac{x^2}{2(\sigma^2+t^2)}}}{\sqrt{2\pi(\sigma^2+t^2)}}$$

Therefore, the probability density function over the features realized by the subpopulations, with $\mathcal{N}(0, \sigma)$ gaussian noise, is itself a gaussian with mean 0 and $\sqrt{(\sigma^2 + t^2)}$ standard deviation.

The qualification function given the features, H, is given as follows:

$$H(x) = \frac{1}{\Pi'(x)} \int_{-\infty}^{\infty} \Pi(y) \cdot p_y(x) \cdot h(y)dy$$

We replace h with h', thus replacing H with H' as defined below:

$$
\begin{aligned}
H'(x) &= \frac{1}{\Pi'(x)}\int_{-\infty}^{\infty} \Pi(y)\cdot p_y(x)\cdot h'(y)dy = \frac{1}{\Pi'(x)}\int_{-\infty}^{\infty} \frac{e^{-\frac{y^2}{2t^2}}}{\sqrt{2\pi}t}\cdot\frac{e^{-\frac{(x-y)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}\cdot\left(\frac{y}{2d}+\frac{1}{2}\right)dy \\
&= \frac{1}{2}+\frac{1}{\Pi'(x)}\int_{-\infty}^{\infty}\frac{e^{-\frac{x^2}{2(\sigma^2+t^2)}}}{\sqrt{2\pi}t}\cdot\frac{e^{-(y-\frac{xt^2}{\sigma^2+t^2})^2/(2\frac{\sigma^2t^2}{\sigma^2+t^2})}}{\sqrt{2\pi}\sigma}\cdot\frac{y}{2d}\ dy \\
&= \frac{1}{2}+\frac{1}{2d\cdot\Pi'(x)}\frac{e^{-\frac{x^2}{2(\sigma^2+t^2)}}}{\sqrt{2\pi}t\cdot\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-(y-\frac{xt^2}{\sigma^2+t^2})^2/(2\frac{\sigma^2t^2}{\sigma^2+t^2})}\cdot y\ dy \\
&= \frac{1}{2}+\frac{1}{2d\cdot\Pi'(x)}\frac{e^{-\frac{x^2}{2(\sigma^2+t^2)}}}{\sqrt{2\pi}t\cdot\sqrt{2\pi}\sigma}\cdot\sqrt{2\pi\frac{\sigma^2t^2}{\sigma^2+t^2}}\cdot\frac{xt^2}{\sigma^2+t^2} \\
&= \frac{1}{2}+\frac{t^2}{\sigma^2+t^2}\frac{x}{2d}
\end{aligned}
$$

Therefore, when there's no strategic manipulation, Jury would classify any individual with feature $x>0$ as 1 and 0 otherwise. This is because, $H'(x)>\frac{1}{2}$ if and only if $x>0$ and the Jury would classify a feature as 1 if and only if, in expectation, the individuals with that feature are more likely to be qualified. This is true irrespective of whether an individual is from the subpopulation A or B because these subpopulations are identical in terms of qualifications, that is, $h_A=h_B=h$ and $\pi_A=\pi_B=\Pi$.

We show that for the cost functions defined above, if Jury publishes $f=1_{x>0}$, as the classifier, then none of the contestants in both the subpopulations A and B have an incentive to change their private signal (under $\mathcal{N}(0,\sigma)$ gaussian noise). Hence, the Jury gets the best possible accuracy from these features and the classification is fair. For a subpopulation $S\in\{A,B\}$, let $q_f^S(y)$ denote the probability of a contestant, with private signal y, being classified as 1 when f is the classifier. Therefore,

$$
q_f^S(y)=\int_{-\infty}^{\infty} f(x)\cdot p_y(x)dx=\int_{0}^{\infty}\frac{e^{-\frac{(x-y)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}dx
$$

For a subpopulation $S\in\{A,B\}$, let's calculate the advantage that a contestant, with private signal y, gets by changing its signal to y' ($y'>y$, otherwise $q_f^S(y')\le q_f^S(y)$):

$$
\begin{aligned}
q_f^S(y')-q_f^S(y) &= \int_{0}^{\infty}\frac{e^{-\frac{(x-y')^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}dx-\int_{0}^{\infty}\frac{e^{-\frac{(x-y)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}dx \quad =\int_{-y'}^{\infty}\frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}dx-\int_{-y}^{\infty}\frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}dx \\
&= \int_{-y'}^{-y}\frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}dx \quad \le\int_{-y'}^{-y}\frac{1}{\sqrt{2\pi}\sigma}dx \quad =\frac{y'-y}{\sqrt{2\pi}\sigma}
\end{aligned}
$$

As $\sigma=\max\{\sigma_A,\sigma_B\}$ and recalling the definitions of the cost functions c_A and c_B (Equation 4), we get that

$$
q_f^A(y')-q_f^A(y)\le c_A(y,y') \qquad \text{and} \qquad q_f^B(y')-q_f^B(y)\le c_B(y,y')
$$

Therefore, none of the contestants in any of the subpopulations have an incentive to change their private signals. The accuracy of the classifier f on the subpopulation A is given as

$$\begin{aligned}
U_A(f) &= \left(\int_{-\infty}^{\infty} \Pi(y)[q_f^A(\Delta_f^A(y)) \cdot (2h(y)-1)]dy\right) + c \\
&= \left(\int_{-\infty}^{\infty} \Pi(y) \int_0^{\infty} \frac{e^{-\frac{(x-y)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} dx \cdot (2h(y)-1)dy\right) + c \\
&= \left(\int_0^{\infty} \left(\int_{-\infty}^{\infty} \Pi(y) \frac{e^{-\frac{(x-y)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} \cdot (2h(y)-1)dy\right) dx\right) + c \\
&= \left(\int_0^{\infty} \Pi'(x) \cdot (2H(x)-1)dx\right) + c
\end{aligned}$$

Replacing H with H' without loosing much in the approximation, we get that

$$U_A(f) = \left(\int_0^{\infty} \frac{e^{-\frac{x^2}{2(\sigma^2+t^2)}}}{\sqrt{2\pi(\sigma^2+t^2)}} \cdot \frac{t^2}{\sigma^2+t^2} \frac{x}{d} dx\right) + c = \frac{t^2}{\sqrt{2\pi(\sigma^2+t^2)} \cdot d} + c$$

Similarly for $U_B(f)$ and hence, $U(f) = U_B(f) = U_A(f) = \frac{t^2}{\sqrt{2\pi(\sigma^2+t^2)} \cdot d} + c$.

E Proof of Theorem 8

We retain the instantiations of h_A, h_B, π_A, π_B, c_A, c_B and σ as above. As seen above, in Scenario 2, $1_{x>0}$ is the classifier that optimizes Jury's utility and hence, $U(f_\sigma^*) = \frac{t^2}{\sqrt{2\pi(\sigma^2+t^2)} \cdot d} + c$. Actually, it's approximately equal to this but the error is extremely small ($e^{-\Omega(d)}$, $d >> t, \sigma$).

In Scenario 1, the utility of any threshold classifier (f) with τ as the threshold is given by Equation 7 (without loss of generality, we can optimize over threshold classifiers). Therefore,

$$U(f) = s_A \cdot \frac{t}{\sqrt{2\pi}d} e^{-(\tau-\sqrt{2\pi}\sigma_A)^2/2t^2} + s_B \cdot \frac{t}{\sqrt{2\pi}d} e^{-(\tau-\sqrt{2\pi}\sigma_B)^2/2t^2} + c$$

When $s_A = s_B = \frac{1}{2}$ and we assume that $|\sigma_A - \sigma_B| \le \frac{t}{\sqrt{2\pi}}$, it's easy enough to see that the above expression is maximized at $\tau = \frac{\sqrt{2\pi}\sigma_A + \sqrt{2\pi}\sigma_B}{2}$. Therefore, the optimal classification accuracy in Scenario 1, is

$$U(f_0^*) = \frac{t}{\sqrt{2\pi}d} e^{-(\frac{\sqrt{2\pi}\sigma_A - \sqrt{2\pi}\sigma_B}{2})^2/2t^2} + c$$

For $\sigma_B = \sigma$, $\sigma_A = 0.1\sigma$, $t = 0.9\sqrt{2\pi}\sigma$, $U(f_\sigma^*) > U(f_0^*)$.