Maximizing Welfare with Incentive-Aware Evaluation Mechanisms

Nika Haghtalab
Cornell University

Nicole Immorlica
Microsoft Research

Brendan Lucier
Microsoft Research

Jack Wang
Cornell University

Abstract

Motivated by applications such as college admission and insurance rate determination, we study a classification problem where the inputs are controlled by strategic individuals who can modify their features at a cost. A learner can only partially observe the features, and aims to classify individuals with respect to a quality score. The goal is to design a classification mechanism that maximizes the overall quality score in the population, taking any strategic updating into account.

When scores are linear and mechanisms can assign their own scores to agents, we show that the optimal classifier is an appropriate projection of the quality score. For the more restrictive task of binary classification via linear thresholds, we construct a (1/4)-approximation to the optimal classifier when the underlying feature distribution is sufficiently smooth and admits an oracle for finding dense regions. We extend our results to settings where the prior distribution is unknown and must be learned from samples.

1 Introduction

Machine learning classifiers are increasingly used to identify qualified individuals in areas such as education, hiring, and public health. This has inspired a line of work aimed at improving the performance and interpretability of classifiers for *identifying qualification and excellence* within a society given access to limited visible attributes of individuals. As these classifiers become widely deployed at a societal level, they can take on the additional role of *defining excellence and qualification*. That is, classifiers encourage people who seek to be "identified as qualified" to acquire attributes that are "deemed to be qualified" by the classifier. For example, a college admission policy that heavily relies on SAT scores will naturally encourage students to increase their SAT scores, which they might do by getting a better understanding of the core material, taking SAT prep lessons, cheating, etc. Progress in machine learning has not fully leveraged classifiers' role in incentivizing people to change their feature attributes, and at times has even considered it an inconvenience to the designer who must now take steps to ensure that their classifier cannot be "gamed" [18, 5, 10, 6, 3]. One of the motivations for such *strategy-proof classification* is Goodhart's law, which states "when a measure becomes a target, it ceases to be a good measure." Taking Goodhart's law to heart, one might view an individual's attributes to be immutable, and any strategic response to a classifier (changes in one's attributes) only serves to mask an agent's true qualification and thereby degrade the usefulness of the classifier.

What this narrative does not address is that in many applications of classification, one's qualifications can truly be improved by changing one's attributes. For example, students who improve their understanding of core material truly become more qualified for college education. These changes have the potential to raise the overall level of qualification and excellence in a society and should be encouraged by a social planner. In this work, we focus on this powerful and under-utilized role of classification in machine learning and ask how to

> *design a classifier on visible features that incentivizes individuals to improve a desired quality.*

For instance, in college admissions, the planner might wish to maximize the "quality" of candidates. Quality is a function of many features, such as persistence, creativity, GPA, past achievements, only a few of which may be directly observable by the planner. Nevertheless the planner designs an admission test on visible features to identify qualified individuals. To pass the test, ideally candidates improve their features and truly increase their quality as a result. In another motivating example, consider a designer who wants to increase average driver safety, which can depend on many features detailing every aspect of a driver's habits. The designer may only have access to

a set of visible features such as age, driver training, or even driving statistics like acceleration/deceleration speed (as recorded by cars' GPS devices). Then a scoring rule (that can affect a driver's insurance premium) attempts to estimate and abstract a notion of "safety" from these features. Drivers naturally adjust their behavior to maximize their score. In both cases, the mechanism does not just pick out high quality candidates or safe drivers in the underlying population, but it actually causes their distributions to change.

These observations motivate the following general problem. Agents are described by feature vectors in a high-dimensional feature space, and can change their innate features at a cost. There is a true function mapping feature vectors to a true score (binary or real-valued). The planner observes a low-dimensional projection of agents' features and chooses a scoring function from a fixed hypothesis class which maps this low-dimensional space to an observed score (binary or real-valued). Agents get value from having a high observed score (e.g., getting admitted to university or having a low car insurance premium), whereas the planner wishes to maximize the average true score of the population.

Our results depend on the form of the true function and the hypothesis class. As an example, we show in Appendix E that if the true function is binary, the hypothesis class is all binary functions, and the planner can observe the full feature vector of candidates, then the optimal hypothesis is the true function. In Section 3, we show a related characterization for when the true quality function is linear and the hypothesis class is all linear functions, in which case the optimal hypothesis is a projection of the true quality function on the visible subspace. Our most interesting results arise from the case when the true function is linear, the hypothesis class is all linear threshold functions. In this case, a simple projection does not work: we need to consider the distribution of agent (projected on the visible feature space) when choosing the hypothesis. In Section 4, we provide polynomial time approximation algorithms for finding the optimal linear threshold. In Section 5, we also provide sample complexity guarantees for learning the optimal hypothesis from samples only.

Our work builds upon the strategic machine learning literature introduce by Hardt et al. [10]. As in our work, agents are represented by feature vectors which can be manipulated at a cost. Hardt et al. [10] design optimal learning algorithms in the presence of these costly strategic manipulations. Hu et al.[14] and Milli et al.[19] extend Hardt et al.[10] by assuming different groups of agents have different costs of manipulation and study the disparate impact on their outcomes. Dong et al. [7] consider a setting in which the learner does not know the distribution of agents features or costs but must learn them through revealed preference. Importantly, in all these works, the manipulations do not change the underlying features of the agent and hence purely disadvantage the learning algorithm. Kleinberg and Raghavan [16] introduce a different model in which manipulations do change the underlying features. Some changes are advantageous, and the designer chooses a rule that incentivizes these while discouraging disadvantageous changes. Their main result is that simple linear mechanisms suffice for a single known agent (i.e., know initial feature vector). In contrast, we study a population of agents with a known distribution of feature vectors and optimize over the class of linear, or linear threshold, mechanisms. Alon et al. [1] also study extensions of Kleinberg and Raghavan [16] to multiple agents. In that work, agents differ in how costly it is for them to manipulate their features but they all have the same starting feature representation, but in our work, agents differ in their starting features while facing the same manipulation cost.

As noted by Kleinberg and Raghavan [16], their model is closely related to the field of contract design in economics. The canonical principal-agent model (see, e.g., [9, 21]) involves a single principle and a single agent. There is a single-dimensional output, say the crop yield of a farm, and the principal wishes to incentivize the agent to expend costly effort to maximize output. However, the principle can only observe output, and the mapping of effort to output is noisy. Under a variety of conditions, the optimal contract in such settings pays the agent an affine function of output [4, 8, 12], although the optimal contract in general settings can be quite complex [17]. Our model differs from this canonical literature in that both effort and output are multi-dimensional. In this regard, the work of Holstrom and Milgrom [13] is closest to ours. They also study a setting with a multi-dimensional feature space in which the principal observes only a low-dimensional representation. Important differences include that Holstrom and Milgrom [13] only study one type of agent whereas we allow agents to have different starting feature vectors, and they assume transferrable utility whereas in our setting payments are implicit and do not reduce the utility of

the principal.

2 Preliminaries

The Model. We denote the true features of an individual by $\mathbf{x} \in \mathbb{R}^n$, where the feature space $\mathbb{R}^n$ encodes all relevant features of a candidate, such as an individual's health history, biometrics, vaccination record, exercise and dietary habits, etc. We denote by $f : \mathbb{R}^n \to [0,1]$ the mapping from one's true features to their true quality. For example, a real-valued $f(\mathbf{x}) \in [0,1]$ can express the overall quality of candidates and a binary-valued $f(\mathbf{x})$ can determine whether $\mathbf{x}$ is meets a certain qualification level.

These true features of an individual may not be visible to the designer. Instead there is an $n \times n$ projection matrix P of rank k, i.e., $P^2 = P$, such that for any $\mathbf{x}$, $P\mathbf{x}$ represents the visible representation of the individual, such as vaccination record and health history, but not exercise and dietary habits. We define by $\mathrm{Img}(P) = \{\mathbf{z} \in \mathbb{R}^n \mid \mathbf{z} = P\mathbf{z}\}$ the set of all feature representations that are visible to the designer. We denote by $g : \mathbb{R}^n \to \mathbb{R}$ a mechanism whose outcome for any individual $\mathbf{x}$ depends only on $P\mathbf{x}$, i.e., the visible features of $\mathbf{x}$.[1] For example, $g(\mathbf{x}) \in \mathbb{R}$ can express the payoff an individual receives from the mechanism, $g(\mathbf{x}) \in [0,1]$ can express the probability that $\mathbf{x}$ is accepted by a randomized mechanism g, or a binary-valued $g(\mathbf{x}) \in \{0,1\}$ can express whether $\mathbf{x}$ is accepted or rejected deterministically.

Let $\mathrm{cost}(\mathbf{x}, \mathbf{x}')$ represent the cost that an individual incurs when changing their features from $\mathbf{x}$ to $\mathbf{x}'$. We consider $\mathrm{cost}(\mathbf{x}, \mathbf{x}') = c\|\mathbf{x} - \mathbf{x}'\|_2$ for some $c > 0$. Given mechanism g, the total payoff $\mathbf{x}$ receives from changing its feature representation to $\mathbf{x}'$ is $U_g(\mathbf{x}, \mathbf{x}') = g(\mathbf{x}') - \mathrm{cost}(\mathbf{x}, \mathbf{x}')$. Let $\delta_g : \mathbb{R}^n \to \mathbb{R}^n$ denote the best response of an individual to g, i.e.,

$$\delta_g(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{x}'} U_g(\mathbf{x}, \mathbf{x}').$$

We consider a distribution $\mathcal{D}$ over feature vector in $\mathbb{R}^n$, representing individuals. Our goal is to design a mechanism g such that, when individuals from distribution $\mathcal{D}$ best respond to it, it yields the highest quality individuals on average. That is to find a mechanism $g \in \mathcal{G}$ that maximizes

$$\mathrm{Val}(g) = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\big[f(\delta_g(\mathbf{x}))\big]. \tag{1}$$

We often consider the gain in the quality of individuals compared to the average quality before deploying any mechanism, i.e., the baseline $\mathbb{E}[f(\mathbf{x})]$, defined by

$$\mathrm{Gain}(g) = \mathrm{Val}(g) - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[f(\mathbf{x})]. \tag{2}$$

Note that $g_{\mathrm{opt}} \in \mathcal{G}$ can be equivalently defined as the mechanism with highest $\mathrm{Gain}(g)$.

For example, f can indicate the overall health of an individual and $\mathcal{G}$ the set of governmental policies on how to set insurance premiums based on the observable features of individuals, e.g., setting lower premiums for those who received preventative care in the previous year. Then, g_{opt} corresponds to the policy that leads to the healthiest society on average. In Sections 3 and 4, we work with different design choices for f and $\mathcal{G}$, and show how to find (approximately) optimal mechanisms.

Types of Mechanisms. More specifically, in this work we consider a known true quality function f that is a *linear function* $\mathbf{w}_f \cdot \mathbf{x} - b_f$ for some vector $\mathbf{w}_f \in \mathbb{R}^n$ and $b_f \in \mathbb{R}$. Without loss of generality we assume that the domain and cost multiplier c are scaled so that $\mathbf{w}_f$ is a unit vector.

In Section 3, we consider the class of *linear mechanisms* $\mathcal{G}_{\mathrm{lin}}$, i.e., any $g \in \mathcal{G}_{\mathrm{lin}}$ is represented by $g(\mathbf{x}) = \mathbf{w}_g \cdot \mathbf{x} - b_g$, for a scalar $b_g \in \mathbb{R}$ and vector $\mathbf{w}_g \in \mathrm{Img}(P)$ of length $\|\mathbf{w}_g\|_2 \le R$ for some $R \in \mathbb{R}^+$. In Section 4, we consider the class of *linear threshold (halfspace) mechanisms* $\mathcal{G}_{0\text{-}1}$, where a $g \in \mathcal{G}_{0\text{-}1}$ is represented by $g(\mathbf{x}) = \mathrm{sign}(\mathbf{w}_g \cdot \mathbf{x} - b_g)$ for some unit vector $\mathbf{w}_g \in \mathrm{Img}(P)$ and scalar $b_g \in \mathbb{R}$.

Other Notation. $\|\cdot\|$ refers to the L_2 norm of a vector unless otherwise stated. For $\ell \ge 0$, the *ℓ-margin density* of a halfspace $\mathrm{sign}(\mathbf{w} \cdot \mathbf{x} - b)$ is

$$\mathrm{Den}^{\ell}_{\mathcal{D}}(\mathbf{w}, b) = \Pr_{\mathbf{x}\sim\mathcal{D}}\Big[\mathbf{w} \cdot \mathbf{x} - b \in [-\ell, 0]\Big].$$

This is the total density $\mathcal{D}$ assigns to points that are at distance at most ℓ from being included in the positive side of the halfspace. The *soft ℓ-margin density* of this halfspace is defined as

$$\text{S-Den}^{\ell}_{\mathcal{D}}(\mathbf{w}, b) = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\Big[(b - \mathbf{w}\cdot\mathbf{x})\mathbf{1}\big(\mathbf{w}\cdot\mathbf{x} - b \in [-\ell, 0]\big)\Big].$$

When ℓ and $\mathcal{D}$ are clear from the context, we suppress them in the above notation.

Throughout this paper, we assume that individuals have features that are within a ball of radius r of the origin, i.e., $\mathcal{D}$ is only supported on $\mathcal{X} = \{\mathbf{x} \mid \|\mathbf{x}\| \le r\}$. In Section 4.2, we work with distribution $\mathcal{D}$ that is additionally *σ-smooth*. $\mathcal{D}$ is σ-smoothed distribution if there is a corresponding distribution $\mathcal{P}$ over $\mathcal{X}$ such

[1] Equivalently, $g(\mathbf{x}) = g^{||}(P\mathbf{x})$ for an unrestricted function $g^{||} : \mathbb{R}^n \to R$.

that to sample $\mathbf{x}' \sim \mathcal{D}$ one first samples $\mathbf{x} \sim \mathcal{P}$ and then $\mathbf{x}' = \mathbf{x} + N(0, \sigma^2 I)$. Smoothing is a common assumption in theory of computer science where $N(0, \sigma^2 I)$ models uncertainties in real-life measurements. To ensure that the noise in measurements is small compared to radius of the domain r, we assume that $\sigma \in O(r)$.

3 Linear Mechanisms

In this section, we show how the optimal linear mechanism in $\mathcal{G}_{\text{lin}}$ is characterized by $g_{\text{opt}}(\mathbf{x}) = \mathbf{w}_g \cdot \mathbf{x}$ for $\mathbf{w}_g$ that is (the largest vector) in the direction of $P\mathbf{w}_f$. This leads to an algorithm with $O(n)$ runtime for finding the optimal mechanism. At a high level, this result shows that when the true quality function f is linear, the optimal linear mechanism is in the direction of the closest vector to $\mathbf{w}_f$ in the visible feature space. Indeed, this characterization extends to any true quality function that is a monotonic transformation of a linear function.

Theorem 3.1 (Linear Mechanisms)**.** *Let* $f(\mathbf{x}) = h(\mathbf{w}_f \cdot \mathbf{x} - b_f)$ *for some monotonic function* $h : \mathbb{R} \to \mathbb{R}$. *Let* $g(\mathbf{x}) = \frac{(P\mathbf{w}_f)R}{\|P\mathbf{w}_f\|_2} \cdot \mathbf{x}$. *Then,*

$$\text{Gain}(g) = \max_{g \in \mathcal{G}_{lin}} \text{Gain}(g).$$

The proof of Theorem 3.1 follows from basic linear algebra and is included in Appendix B for completeness. But let us provide an intuitive explanation of this proof. In response to mechanism $g(\mathbf{x}) = \mathbf{w}_g \cdot \mathbf{x} - b_g$, any movement of $\mathbf{x}$ that is orthogonal to $\mathbf{w}_g$ leaves the outcome of the mechanism unchanged but incurs a movement cost to the individual. Therefore, the best-response of any individual to g is to move only in the direction of $\mathbf{w}_g$. Each unit of movement in the direction of $\mathbf{w}_g$ translates to $\mathbf{w}_g \cdot \mathbf{w}_f$ units of movement in the direction of $\mathbf{w}_f$ and increases the mechanism designer's utility by $\mathbf{w}_g \cdot \mathbf{w}_f$. So, the mechanism's payoff is maximized by the longest vector in the direction of $\mathbf{w}_f$ in the visible feature space. [2]

It is interesting to note that designing an optimal *linear* mechanism (Theorem 3.1) does not use any information about the distribution of instances in $\mathcal{D}$, rather directly projects $\mathbf{w}_f$ on the subspace of visible features. We see in Section 4 that this is not the case for *linear threshold* mechanisms, where the distribution of feature attributes plays a central role in choosing the optimal mechanism.

[2] For general norms defined by a convex symmetric set $\mathcal{K}$, $\mathbf{w}_g \in \mathcal{K}$ that maximizes $\mathbf{w}_g \cdot \mathbf{w}_f$ corresponds to the dual of $\mathbf{w}_f$ with respect to $\|\cdot\|_{\mathcal{K}}$

4 Linear Threshold Mechanisms

In this section, we consider the class of linear threshold mechanisms $\mathcal{G}_{0\text{-}1}$ and explore the computational aspects of finding $g_{\text{opt}} \in \mathcal{G}_{0\text{-}1}$. Note that any $g(\mathbf{x}) = \text{sign}(\mathbf{w}_g \cdot \mathbf{x} - b_g)$ corresponds to the transformation of a linear mechanism from $\mathcal{G}_{\text{lin}}$ that only rewards those whose quality passes a threshold. As in the case of linear mechanisms, individuals only move in the direction of $\mathbf{w}_g$ and every unit of their movement improves the payoff of the mechanism by $\mathbf{w}_g \cdot \mathbf{w}_f$. However, an individual's payoff from the mechanism improves only if its movement pushed the feature representation from the 0-side of a linear threshold to +1-side. Therefore only those individuals whose feature representations are close to the decision boundary of g but on the 0-side will be incentivized to improve their feature representation.

Lemma 4.1. *For* $g(\mathbf{x}) = \text{sign}(\mathbf{w}_g \cdot \mathbf{x} - b_g) \in \mathcal{G}_{0\text{-}1}$,

$$\delta_g(\mathbf{x}) = \mathbf{x} - \mathbf{1}\left(\mathbf{w}_g \cdot \mathbf{x} - b_g \in \left[-\frac{1}{c}, 0\right]\right)(\mathbf{w}_g \cdot \mathbf{x} - b_g)\mathbf{w}_g.$$

The fact that only individuals within a small margin of the decision boundary of g are incentivized to improve their features leads to very different dynamics and challenges compared to the linear mechanism case. For example, $\mathbf{w}_g \propto P\mathbf{w}_f$ no longer leads to a good mechanism as it may only incentivize a small fraction of individuals to improve their features, as seen in the following example.

Example 4.2. *As in Figure 1, consider* $\mathbf{w}_f = \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right)$ *and* P *that projects a vector on its second and third coordinates. Consider a distribution* $\mathcal{D}$ *that consists of* $\mathbf{x} = (0, 0, x_3)$ *for* $x_3 \sim Unif[-r, r]$. *Any* g *only incentivizes individuals who are at distance at most* $\ell = 1/c$ *from the the decision boundary, highlighted by the shaded regions. Mechanism* $g(\mathbf{x}) = \text{sign}(x_2 - \ell)$ *incentivizes everyone to move* ℓ *unit in the direction of* x_2 *and leads to total utility of* $\mathbb{E}[f(\delta_g(\mathbf{x}))] = \ell/\sqrt{3}$. *But,* $g'(\mathbf{x}) = \text{sign}(\mathbf{w}'_g \cdot \mathbf{x} - b'_g)$ *for unit vector* $\mathbf{w}'_g \propto P\mathbf{w}_f$ *only incentivizes* $\sqrt{2}\ell/2r$ *fraction of the individuals, on average each moves only* $\ell/2 \cdot \sqrt{2/3}$ *units in the direction of* $\mathbf{w}_f$. *Therefore,* $\text{Gain}(g') \leq \frac{\ell^2}{2r\sqrt{3}} \ll \text{Gain}(g)$ *when unit cost* $c \gg \frac{1}{r}$.

Using characterization of an individual's best-response to g from Lemma 4.1, for any true quality function $f(\mathbf{x}) = \mathbf{w}_f \cdot \mathbf{x} - b_f$ and $g(\mathbf{x}) = \text{sign}(\mathbf{w}_g \cdot \mathbf{x} -$

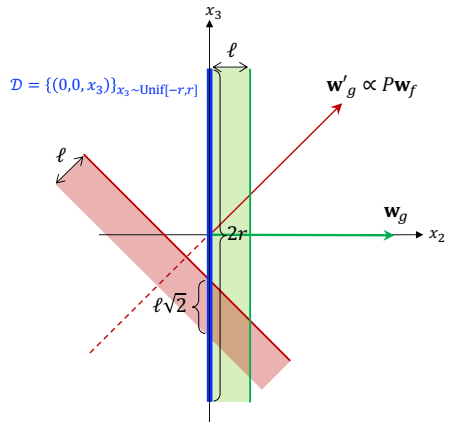

Figure 1: Mechanism $g'(\mathbf{x}) = \text{sign}(\mathbf{w}'_g \cdot \mathbf{x} - b_g)$ for any $\mathbf{w}'_g \propto P\mathbf{w}_f$ is far from optimal.

b_g), we have

$$\text{Gain}(g) = (\mathbf{w}_g \cdot \mathbf{w}_f) \cdot \text{S-Den}^{1/c}(\mathbf{w}_g, b_g). \quad (3)$$

Then to find $g_{\text{opt}} \in \mathcal{G}_{\text{0-1}}$, we need to simultaneously optimize the total density of instances that fall within the margin of the classifier, their average distance to the decision boundary, and the correlation between $\mathbf{w}_g$ and $\mathbf{w}_f$. One of the challenges involved in this optimization task is that the problem of finding a linear threshold with dense margin is NP-Hard [**?**]. In this section, we will show that under multiple settings a mechanism with near optimal gain can be computed in polynomial time. In Section 4.1, we show that one can achieve a $1/(4rc)$-approximation to the optimal mechanism. This is a good approximation guarantee when the unit cost c is not too large relative to the radius of the domain. However, in most cases the cost to change ones' feature is much larger than a constant fraction of the radius. In Section 4.2, we look for a constant approximation for any value of the cost c. We revisit the challenges to optimizing the non-convex and non-smooth objective of Equation 3 and ask whether having an oracle for optimizing the margin density of a linear threshold can help us in finding a near optimal mechanism in $\mathcal{G}_{\text{0-1}}$. We show that indeed such an oracle gives a 1/4-approximation whenever distribution $\mathcal{D}$ is smooth.

4.1 1/(4rc)-Approximation

In this section, we give a simple multiplicative approximation algorithm that works well when the cost unit c is not too large compared to the radius of the domain r. At a high level, when the cost is small more individuals choose to improve their features. This leads to large margin density for even seemingly naïve choices of the mechanism. We show that a simple algorithm that commits to using mechanisms with $\mathbf{w}_g \propto P\mathbf{w}_f$ and only optimizes over the bias term b_g returns a reasonably good solution.

Theorem 4.3 (1/(4rc)-approximation)**.** *Consider the polynomial time algorithm that returns the best g from $\mathcal{G} = \{\text{sign}(\frac{P\mathbf{w}_f}{\|P\mathbf{w}_f\|} \cdot \mathbf{x} - b_g) | b_g = i/2c, \forall i \in [\lceil 2rc \rceil + 1]\}$. Then, $\text{Gain}(g) \geq \frac{1}{4rc}\text{Gain}(g_{\text{opt}})$.*

Proof Sketch. Let $\mathcal{G}' = \{\text{sign}(\frac{P\mathbf{w}_f}{\|P\mathbf{w}_f\|} \cdot \mathbf{x} - b_g) | b_g = i/c, \forall i \in [\lceil rc \rceil]\}$. Note that the $1/c$-left margin of these halfspaces cover the support of $\mathcal{D}$. So, there is $g' \in \mathcal{G}'$ such that $\text{Den}^{1/c}(\mathbf{w}_{g'}, b_{g'}) \geq 1/cr$. Consider instances in the $1/c$-left margin of g. If more than half of them are at distance more than $1/2c$, then the total $\text{S-Den}^{1/c}(\mathbf{w}_{g'}, b_{g'}) \geq 1/(4rc^2)$. Otherwise, more than half of these instances are at distance more than $1/2c$ from the decision boundary of $g''(\mathbf{x}) = \text{sign}(\mathbf{w}_{g'}, b_{g'} + 1/2c)$ and as a result $\text{S-Den}^{1/c}(\mathbf{w}_{g'}, b_{g'} + 1/2c) \geq 1/(4rc^2)$. Therefore, $\max\{\text{Gain}(g'), \text{Gain}(g'')\} \geq (\mathbf{w}_f \cdot \frac{P\mathbf{w}_f}{\|P\mathbf{w}_f\|})/(4rc^2)$. On the other hand, $\text{Gain}(g_{\text{opt}}) \leq (\mathbf{w}_f \cdot \mathbf{w}_{g_{\text{opt}}})/c$. Since $\mathbf{w}_f \cdot \mathbf{w}_g$ is maximized by the unit vector $\mathbf{w}_g = \frac{P\mathbf{w}_f}{\|P\mathbf{w}_f\|}$, this algorithm returns a $1/(4rc)$ approximation to the gain of g_{opt}. □

4.2 Oracle-based 1/4-Approximation

One of the challenges in computing a near optimal mechanism is finding a halfspace whose margin captures a large fraction of instances, i.e., has large margin density. Many variants of this problem have been studied in the past and are known to be hard. For example, the densest subspace, the densest halfspace, densest cube, and the densest ball are all known to be hard to approximate [2, 11, 15]. Yet, finding dense regions is a routine unsupervised learning task for which there are existing optimization tools that are known to perform well in practice. In this section, we assume that we have access to such an optimization tool, which we call *a density optimization oracle.*

Definition 4.1 (Density Optimization Oracle)**.** *Oracle $\mathcal{O}$ takes any distribution $\mathcal{D}$, margin ℓ, a set $\mathcal{K} \subseteq \mathbb{R}^n$, takes $O(1)$ time and returns*

$$\mathcal{O}(\mathcal{D}, \ell, \mathcal{K}) \in \underset{\substack{\mathbf{w} \in \mathcal{K}, \|\mathbf{w}\|=1 \\ b \in \mathbb{R}}}{\text{argmax}} \text{Den}^{\ell}_{\mathcal{D}}(\mathbf{w}, b). \quad (4)$$

Another challenge in computing a near optimal mechanism is that $\text{Gain}(\cdot)$ is a non-smooth function. As a result, there are distributions for which small changes to $(\mathbf{w}, b)$, e.g., to improve $\mathbf{w} \cdot \mathbf{w}_f$, could result in a completely different gain. However, one of the properties of real-life distributions is that there is some amount of noise in the data, e.g., because measurements are not perfect. This is modeled by *smoothed distributions*, as described in Section 2. Smooth distributions provide an implicit regularization that smooths the expected loss function over the space of all solutions.

Using these two assumptions, i.e., access to a density optimization oracle and smoothness of the distribution, we show that there is a polynomial time algorithm that achieves a $1/4$ approximation to the gain of g_{opt}. At a high level, smoothness of $\mathcal{D}$ allows us to limit our search to those $\mathbf{w}_g \in \mathbb{R}^n$ for which $\mathbf{w}_f \cdot \mathbf{w}_g = v$ for a small set of discrete values v. For each v, we use the density oracle to search over all $\mathbf{w}_g$ such that $\mathbf{w}_f \cdot \mathbf{w}_g = v$ and return a candidate with high margin density. Using a technique similar to the proof sketch of Theorem 4.3, we show how a mechanism with high margin density will lead us to (a potentially different) mechanism with high soft-margin density and as a result a near optimal gain. This approach is illustrated in more detail in Algorithm 1.

Theorem 4.4 (1/4-approximation)**.** *Consider a distribution over $\mathcal{X}$ and the corresponding σ-smoothed distribution $\mathcal{D}$ for some $\sigma \in O(r)$. Then, for any small enough ϵ, Algorithm 1 runs in time $\text{poly}(d, 1/\epsilon)$, makes $O(\frac{1}{\epsilon^4} + \frac{r^2}{\epsilon^2\sigma^2})$ oracle calls and returns $g \in \mathcal{G}_{0\text{-}1}$, such that*

$$\text{Gain}(g) \geq \frac{1}{4}\text{Gain}(g_{\text{opt}}) - \epsilon.$$

In the remainder of this section, we give a proof sketch of Theorem 4.4. We defer the detailed proof of this theorem to Appendix C. At a high level, the proof of Theorem 4.4 relies on three technical pieces.

Proof Sketch of Theorem 4.4. Recall from Equation 3, $\text{Gain}(g_{\text{opt}})$ is the product of two values $(\mathbf{w}_{g_{\text{opt}}} \cdot \mathbf{w}_f)$ and $\text{S-Den}(\mathbf{w}_{g_{\text{opt}}}, b_{g_{\text{opt}}})$. The first lemma shows that we can do a grid search over the value of $(\mathbf{w}_{g_{\text{opt}}} \cdot \mathbf{w}_f)$. In other words, there is a predefined grid on the values of $\mathbf{w} \cdot \mathbf{w}_f$ for which there is a $\mathbf{w}$ for which $\mathbf{w} \cdot \mathbf{w}_f \approx \mathbf{w}_{g_{\text{opt}}} \cdot \mathbf{w}_f$. This is demonstrated in Figure 4.2.

This lemma is formally described below and its proof is deferred to Appendix C.2.

Algorithm 1 $(1/4 - \epsilon)$ Approximation for $\mathcal{G}_{0\text{-}1}$

Input: σ-smoothed distribution $\mathcal{D}$ with radius r before perturbation,
Vector $\mathbf{w}_f$,
Projection matrix P,
Desired accuracy ϵ.
Output: a linear threshold mechanism g
Let $\epsilon' = \min\{\epsilon^4, \epsilon^2\sigma^4/r^4\}$.
Let $\mathcal{C} = \emptyset$.
for $\eta = 0, \epsilon'\|P\mathbf{w}_f\|, 2\epsilon'\|P\mathbf{w}_f\|, \ldots, \|P\mathbf{w}_f\|$ **do**
Let $\mathcal{K}_\eta = \text{Img}(P) \cap \{\mathbf{w} \mid \mathbf{w} \cdot \mathbf{w}_f = \eta\}$.
Use the density optimization oracle to compute

$$(\mathbf{w}^\eta, b^\eta) \leftarrow \mathcal{O}(\mathcal{D}, \frac{1}{c}, \mathcal{K}_\eta)$$

Let $\mathcal{C} \leftarrow \mathcal{C} \cup \{(\mathbf{w}^\eta, b^\eta), (\mathbf{w}^\eta, b^\eta + \frac{1}{2c})\}$
end for
return $g(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x} - b)$, where

$$(\mathbf{w}, b) \leftarrow \operatorname*{argmax}_{(\mathbf{w},b)\in\mathcal{C}} (\mathbf{w} \cdot \mathbf{w}_f)\text{S-Den}_{\mathcal{D}}^{1/c}(\mathbf{w}, b).$$

Lemma 4.5 (Discretization)**.** *For any two unit vectors $\mathbf{w}_1, \mathbf{w}_2 \in \text{Img}(P)$, and any ϵ, such that $\mathbf{w}_1 \cdot \mathbf{w}_2 \leq 1 - 2\epsilon$, there is a unit vector $\mathbf{w} \in \text{Img}(P)$, such that $\mathbf{w} \cdot \mathbf{w}_2 = \mathbf{w} \cdot \mathbf{w}_1 + \epsilon$ and $\mathbf{w} \cdot \mathbf{w}_1 \geq 1 - \epsilon$.*

The second technical lemma shows that approximating $\mathbf{w}_g$ by a close vector, $\mathbf{w}$, and b_g by a close scalar, b, only has a small effect on its soft margin density. That is, the soft margin density is *Lipschitz*. For this claim to hold, it is essential for the distribution to be smooth. The key property of a σ-smoothed distribution $\mathcal{D}$—corresponding to the original distribution $\mathcal{P}$—is that $\text{Den}_{\mathcal{D}}^{\ell}(\mathbf{w}_g, b_g)$ includes instances $\mathbf{x} \sim \mathcal{P}$ that are not in the margin of $(\mathbf{w}_g, b_g)$. As shown in Figure 4.2, these instances also contribute to the soft margin density of any other $\mathbf{w}$ and b, as long as the distance of $\mathbf{x}$ to halfplanes $\mathbf{w} \cdot \mathbf{x} = b$ and $\mathbf{w}_g \cdot \mathbf{x} = b_g$ are comparable. So, it is sufficient to show that the distance of any $\mathbf{x}$ to two halfplanes with a small angle is approximately the same. Here, angle between two unit vectors is defined as $\theta(\mathbf{w}, \mathbf{w}') = \arccos(\mathbf{w} \cdot \mathbf{w}')$. Lemma 4.5 leverages this fact to prove that soft margin density is Lipschitz smooth. The proof of this lemma is deferred to Appendix C.3.

Lemma 4.6 (Soft Margin Lipschitzness)**.** *For any*

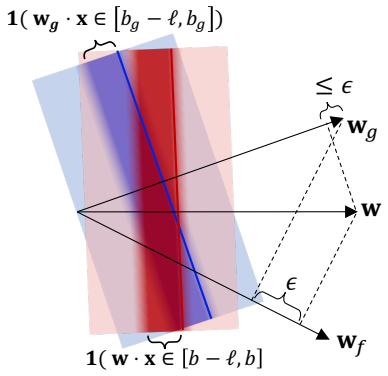

Figure 2: The $\mathbf{w}_g \cdot \mathbf{w}_f$ can be increased by any small amount ϵ to match a specific value, when $\mathbf{w}_f$ is shifted by a $\leq \epsilon$ angle to vector $\mathbf{w}$. When a distribution is smooth, the margin density of any two mechanisms $(\mathbf{w}, b)$ and $(\mathbf{w}_g, b_g)$ are close when $\theta(\mathbf{w}, \mathbf{w}_g)$ and $|b - b_g|$ are small. The lightly shaded area shows the region that instances outside of the margin contribute to (soft-)margin density.

distribution over $\mathcal{X}$ and its corresponding σ-smooth distribution $\mathcal{D}$, for $\epsilon \leq \frac{1}{3R^2}$, any $\nu < 1$, for $R = 2r + \sigma\sqrt{2\ln(2/\nu)}$, and any $\ell \leq O(R)$, if $\theta(\mathbf{w}_1, \mathbf{w}_2) \leq \epsilon\sigma^2$ and $|b_1 - b_2| \leq \epsilon\sigma$, we have

$$\left|\text{S-Den}^{\ell}_{\mathcal{D}}(\mathbf{w}_1, b_1) - \text{S-Den}^{\ell}_{\mathcal{D}}(\mathbf{w}_2, b_2)\right| \leq O(\nu + \epsilon R^2).$$

Lemmas 4.5 and 4.6 show that the gain of any mechanism is approximated by the gain mechanisms $g(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x} - b)$ for which $\mathbf{w} \cdot \mathbf{w}_f$ falls on a grid of small size. The final step to bring these results together is to show that we can find a mechanism (between those on the grid) that has a large soft margin density. To do this, we show that the mechanism (with potentially small changes to it) that has the highest margin density has at least 1/4 of the soft density of the optimal mechanism. The proof technique of this lemma is similar to that of Theorem 4.3, which involves analyzing whether more than half of instances within an ℓ-margin of a mechanism are at distance at most $\ell/2$ of the margin, in which case soft margin density of the mechanism is at least 1/4 of its density. Otherwise, shifting the bias of the mechanism by $\ell/2$ results in a mechanism whose soft margin is a least 1/4 of the original margin density.

Lemma 4.7 (Margin Density Approximates Soft–Margin Density)**.** *For any distribution $\mathcal{D}$ over $\mathbb{R}^n$, a class of unit vectors $\mathcal{K} \subseteq \mathbb{R}^n$, margin $\ell > 0$, and a set of values N, let $(\mathbf{w}_\eta, b_\eta) = \mathcal{O}(\mathcal{D}, \ell, \mathcal{K} \cap \{\mathbf{w} \cdot \mathbf{w}_f = \eta\})$ for $\eta \in N$, and let $(\hat{\mathbf{w}}, \hat{b})$ be the mechanism with maximum $\eta\text{Den}^{\ell}_{\mathcal{D}}(\mathbf{w}_\eta, b_\eta)$ among these options. Then,*

$$\max\left\{\text{Gain}(\hat{\mathbf{w}}, \hat{b}), \text{Gain}(\hat{\mathbf{w}}, \hat{b} + \frac{\ell}{2})\right\} \geq \frac{1}{4} \max_{\substack{\mathbf{w} \in \mathcal{K} \\ \mathbf{w} \cdot \mathbf{w}_f \in N \\ b \in \mathbb{R}}} \text{Gain}(\mathbf{w}, b).$$

Putting these lemmas together, one can show that Algorithm 1, finds a $\frac{1}{4}$ approximation. □

5 Learning Optimal Mechanisms From Samples

Up to this point, we have assumed that distribution $\mathcal{D}$ is known to the mechanism designer and all computations, such as measuring margin density and soft margin density, can be directly performed on $\mathcal{D}$. However, in most applications the underlying distribution over feature attributes is unknown or the mechanism only has access to a small set of historic data. In this section, we show that all computations that are performed in this paper can be done instead on a small set of samples that are drawn i.i.d. from distribution $\mathcal{D}$. Note that (since no mechanism is yet deployed at this point) these samples represent the initial non-strategic feature attributes.

We first note that the characterization of the optimal *linear* mechanism (Theorem 3.1) is independent of distribution $\mathcal{D}$, that is, even with no samples from $\mathcal{D}$ we can still design the optimal linear mechanism. On the other hand, the optimal *linear threshold* mechanism (Theorem 4.4) heavily depends on the distribution of instances, e.g., through the margin and soft-margin density of $\mathcal{D}$. To address sample complexity of learning a linear threshold mechanism, we use the concept of *pseudo-dimension*. This is the analog of VC-dimension for real-valued functions.

Definition 5.1 (Pseudo-dimension)**.** *Consider a set of real-valued functions $\mathcal{F}$ on the instance space $\mathcal{X}$. A set of instances $x^{(1)}, \ldots, x^{(n)} \in \mathcal{X}$ is* shattered *by $\mathcal{F}$, if there are values $v^{(1)}, \ldots, v^{(n)}$ such for any subset $T \subseteq [n]$ there is a function $f_T \in \mathcal{F}$ for which $f(x^{(i)}) \geq v^{(i)}$ if and only if $i \in T$. The size of the largest set that is shattered by $\mathcal{F}$ is the* pseudo-dimension *of $\mathcal{F}$.*

It is well-known that the pseudo-dimension of a function class closely (via upper and lower bounds) controls the number of samples that are needed for learning a good function. This is characterized by the uniform-convergence property as shown below.

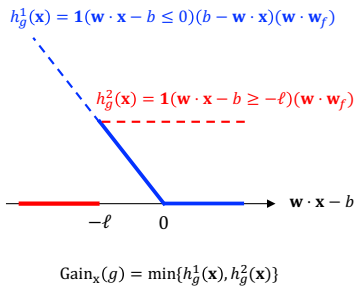

Figure 3: $\text{Gain}_{\mathbf{x}}(g)$ is a minimum of $h_g^1(\mathbf{x})$ and $h_g^2(\mathbf{x})$.

Theorem 5.1 (Uniform Convergence [20]). *Consider the class of functions $\mathcal{F}$ such that $f : \mathcal{X} \to [0, H]$ with pseudo-dimension d. For any distribution $\mathcal{D}$ and a set S of $m = \epsilon^{-2}H^2(d + \ln(1/\epsilon))$ i.i.d. randomly drawn samples from $\mathcal{D}$, with probability $1 - \delta$, for all $f \in \mathcal{F}$*

$$\left| \frac{1}{m} \sum_{x \in S} f(x) - \mathop{\mathbb{E}}_{x \sim \mathcal{D}}[f(x)] \right| \le \epsilon.$$

As is commonly used in machine learning, uniform convergence implies that choosing any mechanism based on its performance on the sample set leads to a mechanism whose expected performance is within 2ϵ of the optimal mechanism for the underlying distribution.

Lemma 5.2 (Pseudo-dimension). *For each $g = sign(\mathbf{w}_g \cdot \mathbf{x} - b_g)$ and $\mathbf{x}$ let $\text{Gain}_{\mathbf{x}}(g)$ be the gain of g just on instance $\mathbf{x}$ and let $\text{Den}_{\mathbf{x}}^{\ell}(g) = \mathbf{1}(\mathbf{w}_g \cdot \mathbf{x} - b_g \in [-\ell, 0])$. The class of real-valued functions $\{\text{Gain} \circ g\}_{g \in \mathcal{G}_{0\text{-}1}}$ has a pseudo-dimension that is at most $O(\text{rank}(P))$ Moreover, the class of functions $\{\mathbf{x} \to \text{Den}_{\mathbf{x}}^{\ell}(g)\}_{g \in \mathcal{G}_{0\text{-}1}}$ has a pseudo-dimension (equivalently VC dimension) $O(\text{rank}(P))$.*

Proof. For $g \in \mathcal{G}_{0\text{-}1}$, note that

$$\text{Gain}_{\mathbf{x}}(g) = \mathbf{1}\left(\mathbf{w}_g \cdot \mathbf{x} - b_g \in \left[-\frac{1}{c}, 0\right]\right)(b_g - \mathbf{w}_g \cdot \mathbf{x})(\mathbf{w}_f \cdot \mathbf{w}_g).$$

Note that we can write $\text{Gain}_{\mathbf{x}}(g) = \min\{h_1(\mathbf{x}), h_2(\mathbf{x})\}$ for $h_g^1(\mathbf{x}) := \mathbf{1}(\mathbf{w}\cdot\mathbf{x} - b \le 0)(b - \mathbf{w}\cdot\mathbf{x})(\mathbf{w}\cdot\mathbf{w}_f)$ and $h_g^2(\mathbf{x}) := \mathbf{1}(\mathbf{w}\cdot\mathbf{x} - b \ge -\ell)(\mathbf{w}\cdot\mathbf{w}_f)$. As show in Figure 3, h_g^1 and h_g^2 are both monotone functions of $\mathbf{w}\cdot\mathbf{x} - b$.

Pollard [20] shows that the set of all linear functions in a rank k subspace has pseudo-dimension k. Pollard [20] also shows that the set of monotone transformations of these functions has pseudo-dimension $O(k)$. Therefore, the set of functions $\{h_g^1(\mathbf{x})\}_{g \in \mathcal{G}_{0\text{-}1}}$ and $\{h_g^2(\mathbf{x})\}_{g \in \mathcal{G}_{0\text{-}1}}$, each have pseudo-dimension $\text{Rank}(P)$. It is well-known that the set of all minimums of two functions from two classes each with pseudo-dimension d has a pseudo-dimension of $O(d)$. Therefore, $\{\mathbf{x} \to \text{Gain}_{\mathbf{x}}(g)\}_{g \in \mathcal{G}_{0\text{-}1}}$ has pseudo-dimension of at most $O(\text{rank}(P))$. A similar construction shows that $\{\mathbf{x} \to \text{Den}_{\mathbf{x}}^{\ell}(g)\}_{g \in \mathcal{G}_{0\text{-}1}}$ has VC dimension of at most $O(\text{rank}(P))$. □

We give a bound on the number of samples sufficient for finding a near optimal mechanism in $\mathcal{G}_{0\text{-}1}$.

Theorem 5.3 (Sample Complexity). *For any small enough ϵ and δ, there is*

$$m \in O\left(\frac{1}{c^2\epsilon^2}\left(\text{rank}(P) + \ln(1/\delta)\right)\right)$$

such that for $S \sim \mathcal{D}^m$ i.i.d. samples with probability $1 - \delta$, $\hat{g} \in \mathcal{G}_{0\text{-}1}$ that optimizes the empirical gain on S has $\text{Gain}(\hat{g}) \ge \text{Gain}(g_{\text{opt}}) - \epsilon$. Furthermore, with probability $1 - \delta$ when Algorithm 1 is run on S, the outcome $\hat{g}$ has $\text{Gain}(\hat{g}) \ge \frac{1}{4}\text{Gain}(g_{\text{opt}}) - \epsilon$.

The proof of this theorem is included in Appendix D and follows directly from Lemma 5.2, Theorem 5.1 and the fact that Algorithm 1 only uses $\mathcal{D}$ to compute the margin density of the gain of a classifier.

6 Discussion

Up to now, we have only considered the setting of Equation 1. There are, however, other reasonable objectives that one may want to consider. In some cases, our goal is to design a classification function g that indicates whether an individual would be accepted or rejected by the function. In such cases, we may be interested in designing function g that, when people best-respond to it, yield the highest quality accepted individuals on average. That is, for a set of boolean valued functions $\mathcal{G}$, $\max_{g \in \mathcal{G}} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}\left[f(\delta_g(\mathbf{x})) \mid g(\mathbf{x}) = 1\right]$.

In our work, we consider the L_2 cost function (i.e., $\text{cost}(\mathbf{x}, \mathbf{x}') = c\|\mathbf{x} - \mathbf{x}'\|_2$). This specification models scenarios in which an individual can improve multiple features most efficiently by combining effort rather than working on each feature individually (L_1 norm). For example, simultaneously improving writing, vocabulary, and critical analysis of a student more by for example reading novels, is more effective than spending effort to improve vocabulary by, say, memorizing vocab lists, and then improving the other attributes. It would be interesting to see if similar results hold

with different specifications of cost, including non-metric costs (e.g., ones with learning curves whereby the first 10% improvement is cheaper than the next 10%), and different costs for different types of individuals.

Finally, throughout this paper, we have assumed knowledge of the true mapping of features to qualities (i.e., f). In many settings, one might not know this mapping, or even the full set of features. Instead, the designer only observes the quality of individuals after they respond to their incentives (i.e., $(f(\delta_g(\mathbf{x})))$), and the projection of their new feature set (i.e., $P\delta_g(\mathbf{x})$). It would be very interesting to consider how one can design effective classifiers in this setting.

References

[1] Tal Alon, Magdalen Dobson, Ariel D Procaccia, Inbal Talgam-Cohen, and Jamie Tucker-Foltz. Multiagent evaluation mechanisms. 2019.

[2] Shai Ben-David, Nadav Eiron, and Hans Ulrich Simon. The computational complexity of densest region detection. *Journal of Computer and System Sciences*, 64(1):22–47, 2002.

[3] Yang Cai, Constantinos Daskalakis, and Christos Papadimitriou. Optimum statistical estimation with strategic data sources. In *Proceedings of the 28th Conference on Computational Learning Theory (COLT)*, pages 280–296, 2015.

[4] Gabriel Carroll. Robustness and linear contracts. *American Economic Review*, 105(2):536–563, 2015.

[5] Yiling Chen, Chara Podimata, Ariel D Procaccia, and Nisarg Shah. Strategyproof linear regression in high dimensions. In *Proceedings of the 19th ACM Conference on Economics and Computation (EC)*, pages 9–26. ACM, 2018.

[6] Ofer Dekel, Felix Fischer, and Ariel D. Procaccia. Incentive compatible regression learning. *Journal of Computer and System Sciences*, 76(8):759–777, 2010.

[7] Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 19th ACM Conference on Economics and Computation (EC)*, pages 55–70, 2018.

[8] Paul Dütting, Tim Roughgarden, and Inbal Talgam-Cohen. Simple versus optimal contracts. In *Proceedings of the 20th ACM Conference on Economics and Computation (EC)*, pages 369–387, 2019.

[9] Sanford J Grossman and Oliver D Hart. An analysis of the principal-agent problem. *Econometrica*, 51(1):7–45, 1983.

[10] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 7th ACM Conference on Innovations in Theoretical Computer Science Conference (ITCS)*, pages 111–122, 2016.

[11] Moritz Hardt and Ankur Moitra. Algorithms and hardness for robust subspace recovery. In *Conference on Computational Learning Theory (COLT)*, pages 354–375, 2013.

[12] Bengt Holmstrom and Paul Milgrom. Aggregation and linearity in the provision of intertemporal incentives. *Econometrica*, 55(2):303–328, 1987.

[13] Bengt Holmstrom and Paul Milgrom. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization*, 7:24–52, 1991.

[14] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the 2nd ACM Conference on Fairness, Accountability, and Transparency*, pages 259–268, 2019.

[15] David S. Johnson and Franco P Preparata. *Theoretical Computer Science*, 6(1):93–107, 1978.

[16] Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? In *Proceedings of the 20th ACM Conference on Economics and Computation (EC)*, pages 825–844, 2019.

[17] R Preston Mcafee and John McMillan. Bidding for contracts: A principal-agent analysis. *The RAND Journal of Economics*, 17(3):326–338, 1986.

[18] Reshef Meir, Ariel D Procaccia, and Jeffrey S Rosenschein. Algorithms for strategyproof classification. *Artificial Intelligence*, 186:123–156, 2012.

[19] Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the 2nd ACM Conference on Fairness, Accountability, and Transparency*, pages 230–239, 2019.

[20] D. Pollard. *Convergence of Stochastic Processes*. Springer Series in Statistics. 2011.

[21] Stephen A Ross. The economic theory of agency: The principal's problem. *American Economic Review*, 63(2):134–139, 1973.

A Useful Properties of Smoothed Distributions

Lemma A.1. *For any σ-smoothed distribution $\mathcal{D}$ and any unit vector $\mathbf{w}$ and any range $[a, b]$,*

$$\Pr_{\mathbf{x}\sim\mathcal{D}}[\mathbf{w}\cdot\mathbf{x}\in[a,b]] \leq \frac{|b-a|}{\sigma\sqrt{2\pi}}.$$

Proof. First note that for any Gaussian distribution with variance $\sigma^2 I$ over $\mathbf{x}$, the distribution of $\mathbf{w}\cdot\mathbf{x}$ is a 1-dimensional Gaussian with variance σ^2. Since any σ-smoothed distribution is a mixture of many Gaussians, distribution over $\mathbf{w}\cdot\mathbf{x}$ is also the mixture of many Gaussians. It is not hard to see that the density of any one dimensional Gaussian is maximized in range $[a, b]$ if its centered at $(a-b)/2$. Therefore,

$$\begin{aligned}
\Pr_{\mathbf{x}\sim\mathcal{D}}[\mathbf{w}\cdot\mathbf{x}\in[a,b]] &\leq \Pr_{x\sim N(0,\sigma^2)}\left[\frac{a-b}{2}\leq x\leq\frac{b-a}{2}\right] \\
&\leq \frac{2}{\sigma\sqrt{2\pi}}\int_0^{(b-a)/2}\exp\left(-\frac{z^2}{2\sigma^2}\right)\,dz \\
&\leq \frac{(b-a)}{\sigma\sqrt{2\pi}}.
\end{aligned}$$

□

Lemma A.2. *For any distribution over $\{\mathbf{x} \mid \|\mathbf{x}\| \leq r\}$ and its corresponding σ-smoothed distribution $\mathcal{D}$, and $H \geq 2r + \sigma\sqrt{2\ln(1/\epsilon)}$, we have*

$$\Pr_{\mathbf{x}\sim\mathcal{D}}\left[\|\mathbf{x}\|\geq H\right]\leq\epsilon.$$

Proof. Consider the Gaussian distribution $N(\mathbf{c}, \sigma^2 I)$ for some $\|\mathbf{c}\| \leq r$. Then for any $\|\mathbf{x}\| \geq H$, it must be that the distance of $\mathbf{x}$ to $\mathbf{c}$ is at least $H - 2r$. Therefore,

$$\Pr_{\mathbf{x}\sim\mathcal{D}}[\|\mathbf{x}\|\geq H]\leq\Pr_{x\sim N(0,\sigma^2)}[x\geq H-2r]\leq\exp\left(-\frac{(H-2r)^2}{2\sigma^2}\right)\leq\epsilon.$$

□

Lemma A.3. *For any distribution over $\{\mathbf{x} \mid \|\mathbf{x}\| \leq r\}$ and let $p(\cdot)$ denote the density of the corresponding σ-smoothed distribution. Let $\mathbf{x}$ and $\mathbf{x}'$ be such that $\|\mathbf{x}-\mathbf{x}'\| \leq D$. Then,*

$$\frac{p(\mathbf{x})}{p(\mathbf{x}')}\leq\exp\left(\frac{D(2r+\|\mathbf{x}\|+\|\mathbf{x}'\|)}{2\sigma^2}\right).$$

Proof. Consider distribution $N(\mathbf{c}, \sigma^2 I)$ for some $\|\mathbf{c}\| \leq r$. Without loss of generality assume that $\|\mathbf{c}-\mathbf{x}\| \leq \|\mathbf{c}-\mathbf{x}'\|$, else $p(\mathbf{x})/p(\mathbf{x}') \leq 1$. We have

$$\begin{aligned}
\frac{p(\mathbf{x})}{p(\mathbf{x}')} &= \frac{\exp\left(\frac{-\|\mathbf{c}-\mathbf{x}\|^2}{2\sigma^2}\right)}{\exp\left(\frac{-\|\mathbf{c}-\mathbf{x}'\|^2}{2\sigma^2}\right)} \\
&= \exp\left(\frac{\|\mathbf{c}-\mathbf{x}'\|^2-\|\mathbf{c}-\mathbf{x}\|^2}{2\sigma^2}\right) \\
&= \exp\left(\frac{(\|\mathbf{c}-\mathbf{x}'\|-\|\mathbf{c}-\mathbf{x}\|)\,(\|\mathbf{c}-\mathbf{x}'\|+\|\mathbf{c}-\mathbf{x}\|)}{2\sigma^2}\right) \\
&\leq \exp\left(\frac{D(2r+\|\mathbf{x}\|+\|\mathbf{x}'\|)}{2\sigma^2}\right).
\end{aligned}$$

Since $\mathcal{D}$ is a σ-smoothed distribution, it is a mixture of many Gaussians with variance $\sigma^2 I$ centered within L_2 ball of radius r. Summing over the density for all these Gaussian proves the claim. □

B Proofs from Section 3

Lemma B.1. *For any linear* $g(\mathbf{x}) = \mathbf{w}_g \cdot \mathbf{x} - b_g$ *and cost function* $\text{cost}(\mathbf{x}, \mathbf{x}') = c\|\mathbf{x} - \mathbf{x}'\|_2$,

$$\delta_g(\mathbf{x}) = \mathbf{x} + \alpha \mathbf{w}_g,$$

where $\alpha = 0$ *if* $c > \|\mathbf{w}_g\|$ *and* $\alpha = \infty$ *if* $c \leq \|\mathbf{w}_g\|$.

Proof. Consider an image of a vector on $\mathbf{w}_g$ and its orthogonal space, note that any $\mathbf{x}'$ can be represented by $\mathbf{x}' - \mathbf{x} = \alpha \mathbf{w}_g + \mathbf{z}$, where $\mathbf{z} \cdot \mathbf{w}_g = 0$ and $\alpha \in \mathbb{R}$. Then, the payoff that a player with features $\mathbf{x}$ gets from playing $\mathbf{x}'$ is

$$\begin{aligned} U_g(\mathbf{x}, \mathbf{x}') &= \mathbf{w}_g \cdot \mathbf{x}' - b_g - c\|\mathbf{x} - \mathbf{x}'\| \\ &= \mathbf{w}_g \cdot (\alpha \mathbf{w}_g + \mathbf{z}) - b_g - c\|\alpha \mathbf{w}_g + \mathbf{z}\| \\ &= \alpha \|\mathbf{w}_g\|^2 - b_g - \alpha c \|\mathbf{w}_g\| - |\alpha| c \|\mathbf{z}\|, \end{aligned}$$

where the last transition is due to the fact that $\mathbf{z}$ is orthogonal to $\mathbf{w}_g$.

It is clear from the above equation that the player's utility is optimized only if $\mathbf{z} = \mathbf{0}$ and $\alpha \geq 0$, that is, $\delta_g(\mathbf{x}) = \mathbf{x} + \alpha \mathbf{w}_g$ for some $\alpha \geq 0$. Therefore,

$$\delta_g(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{x}'} \ \mathbf{w}_g \cdot \mathbf{x} - b_g - c\|\mathbf{x} - \mathbf{x}'\| = \mathbf{x} + \left(\operatorname*{argmax}_{\alpha \geq 0} \ \alpha \|\mathbf{w}_g\| \left(\|\mathbf{w}_g\| - c \right) \right) \mathbf{w}_g.$$

Note that $\alpha \|\mathbf{w}_g\| \left(\|\mathbf{w}_g\| - c \right)$ is maximizes at $\alpha = 0$ if $c > \|\mathbf{w}_g\|$ and is maximized at $\alpha = \infty$ if $c \leq \|\mathbf{w}_g\|$. □

B.1 Proof of Theorem 3.1

For ease of exposition we refer to parameters of g_{opt} by $\mathbf{w}^*$ and b^*. Let $\mathbf{w}_g = \frac{(P\mathbf{w}_f)R}{\|P\mathbf{w}_f\|_2}$ and $g(\mathbf{x}) = \mathbf{w}_g \cdot \mathbf{x}$. By the choice of $\mathbf{w}_g$ and definition of $\mathcal{G}$, we have $\|\mathbf{w}_g\| = R \geq \|\mathbf{w}^*\|$ and $\mathbf{w}_g \in \text{Img}(P)$.

By Lemma B.1 and the fact that $\|\mathbf{w}_g\| \geq \|\mathbf{w}^*\|$ there are $\alpha_g, \alpha_{g_{\text{opt}}} \in \{0, \infty\}$, such that $\alpha_g \geq \alpha_{g_{\text{opt}}}$, $\delta_{g_{\text{opt}}}(\mathbf{x}) = \mathbf{x} + \alpha_{g_{\text{opt}}} \mathbf{w}_g$, and $\delta_g(\mathbf{x}) = \mathbf{x} + \alpha_g \mathbf{w}_g$. Furthermore, we have

$$\mathbf{w}_f \cdot \mathbf{w}^* \leq \|\mathbf{w}_f\| \|\mathbf{w}^*\| \leq \|\mathbf{w}_f\| R = \mathbf{w}_f \cdot \frac{R(P\mathbf{w}_f)}{\|P\mathbf{w}_f\|} = \mathbf{w}_f \cdot \mathbf{w}_g,$$

where the first transition is by Cauchy-Schwarz, the second transition is by the definition of $\mathcal{G}$, and the last two transitions are by the definition of $\mathbf{w}_g$. Using the monotonicity of h and the fact that $\alpha_g \geq \alpha_{g_{\text{opt}}}$ and $\mathbf{w}_f \cdot \mathbf{w}_g \geq \mathbf{w}_f \cdot \mathbf{w}^*$, we have

$$\begin{aligned} \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}} \left[f(\delta_{g_{\text{opt}}}(\mathbf{x}) \right] &= \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}} \left[h\left(\mathbf{w}_f \cdot \mathbf{x} + \alpha_g \mathbf{w}_f \cdot \mathbf{w}^* - b_f \right) \right] \\ &\leq \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}} \left[h\left(\mathbf{w}_f \cdot \mathbf{x} + \alpha_{g_{\text{opt}}} \mathbf{w}_f \cdot \mathbf{w}_g - b_f \right) \right] \\ &= \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}} \left[f(\delta_g(\mathbf{x})) \right]. \end{aligned}$$

This proves the claim.

C Proofs from Section 4

C.1 Proof of Lemma 4.1

By definition, the closest $\mathbf{x}'$ to $\mathbf{x}$ such that $g(\mathbf{x}') = 1$ is the projection of $\mathbf{x}$ on g, defined by $\mathbf{x}' = \mathbf{x} - (\mathbf{w} \cdot \mathbf{x} - b)\mathbf{w}$. By definition

$$\text{cost}(\mathbf{x}, \mathbf{x}') = c\|\mathbf{x} - \mathbf{x}'\|_2 = c\left(\mathbf{w} \cdot \mathbf{x} - b \right).$$

The claim follows by noting that a player with features $\mathbf{x}$ would report $\delta_g(\mathbf{x}) = \mathbf{x}'$ only if $\text{cost}(\mathbf{x}, \mathbf{x}') \leq 1$, otherwise $\delta_g(\mathbf{x}) = \mathbf{x}$.

C.2 Proof of Lemma 4.5

The proof idea is quite simple: Only a small move in angle is required to achieve an ϵ change in the projection of on $\mathbf{w}_2$. Here, we prove this rigorously. Let $\mathbf{w} = \alpha\mathbf{w}_1 + \beta\mathbf{w}_2$ for α and β that we will described shortly. Note that $\text{Img}(P)$ includes any linear combination of $\mathbf{w}_1$ and $\mathbf{w}_2$. Therefore, $\mathbf{w} \in P$. Let $\omega = \mathbf{w}_1 \cdot \mathbf{w}_2$,

$$\alpha = \sqrt{\frac{1-(\omega+\epsilon)^2}{1-\omega^2}},$$

and

$$\beta = \sqrt{1-\alpha^2+\alpha^2\omega^2} - \alpha\omega.$$

By definition of α and β, $\|\mathbf{w}\| = \sqrt{\alpha^2+\beta^2+2\alpha\beta\omega} = 1$. That, is $\mathbf{w}$ is indeed a unit vector. Next, we show that $\mathbf{w}\cdot\mathbf{w}_2 = \mathbf{w}_1\cdot\mathbf{w}_2 + \epsilon$. We have

$$\mathbf{w}\cdot\mathbf{w}_2 = \alpha\omega+\beta = \sqrt{\frac{\omega^2\left(1-(\epsilon+\omega)^2\right)}{1-\omega^2} - \frac{1-(\epsilon+\omega)^2}{1-\omega^2} + 1} = \omega+\epsilon = \mathbf{w}_1\cdot\mathbf{w}_2+\epsilon.$$

It remains to show that $\mathbf{w}_1\cdot\mathbf{w} \geq 1-\epsilon$. By definition of $\mathbf{w}$, we have

$$\mathbf{w}\cdot\mathbf{w}_1 = \alpha+\beta\omega = \left(1-\omega^2\right)\sqrt{\frac{(\epsilon+\omega)^2-1}{\omega^2-1}} + \omega(\epsilon+\omega).$$

One can show that the right hand-side equation is monotonically decreasing in ω Since $\omega < 1-2\epsilon$, we have that

$$\begin{aligned}
\mathbf{w}\cdot\mathbf{w}_1 &\geq \left(1-(1-2\epsilon)^2\right)\sqrt{\frac{1-(1-\epsilon)^2}{1-(1-2\epsilon)^2}} + (1-2\epsilon)(1-\epsilon) \\
&= (1-\epsilon)\left(2\epsilon\left(\sqrt{\frac{2-\epsilon}{1-\epsilon}}-1\right)+1\right) \\
&\geq 1-\epsilon.
\end{aligned}$$

This completes the proof.

C.3 Proof of Lemma 4.6

Without loss of generality, let the angle between $\mathbf{w}_1$ and $\mathbf{w}_2$ to be exactly $\beta = \epsilon\sigma^2$. Also without loss of generality assume $b_1 > b_2$ and let $\mathbf{w}_1 = (1,0,\ldots,0)$ and $\mathbf{w}_2 = (\cos(\beta),\sin(\beta),0,\ldots,0)$. Consider the following rotation matrix

$$M = \begin{pmatrix} \cos(\beta) & \sin(\beta) & 0 & \cdots & 0 \\ -\sin(\beta) & \cos(\beta) & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

that rotates every vector by angle β on the axis of $\mathbf{w}_1, \mathbf{w}_2$. Note that $\mathbf{x} \to P\mathbf{x}$ is a bijection. Furthermore, for any $\mathbf{x}$, $\mathbf{w}_1\cdot\mathbf{x} = \mathbf{w}_2\cdot(M\mathbf{x})$ and the $\|\mathbf{x}\| = \|M\mathbf{x}\|$.

We use the mapping $\mathbf{x}\to P\mathbf{x}$ to upper bound the difference between the soft-margin density of $(\mathbf{w}_1, b_1)$ and $(\mathbf{w}_2, b_2)$. At a high level, we show that the total contribution of $\mathbf{x}$ to soft-margin density of $(\mathbf{w}_1, b_1)$ is close to the total contribution of $M\mathbf{x}$ to the soft-margin density of $(\mathbf{w}_2, b_2)$.

For this to work, we first show that almost all instances $\mathbf{x}$ in the ℓ-margin of $(\mathbf{w}_1, b_1)$ translate to instances $M\mathbf{x}$ in the ℓ-margin of $(\mathbf{w}_2, b_2)$, and vice versa. That is,

$$\Pr_{\mathbf{x}\sim\mathcal{D}}\left[\mathbf{w}_1\cdot\mathbf{x}-b_1 \in [-\ell,0] \text{ but } \mathbf{w}_2\cdot M\mathbf{x}-b_2 \notin [-\ell,0], \|\mathbf{x}\| \leq R\right] \leq \Pr_{\mathbf{x}\sim\mathcal{D}}\left[\mathbf{w}_1\cdot\mathbf{x}-b_1 \in [-\sigma\epsilon,0]\right]$$

$$\leq \frac{\epsilon}{\sqrt{2\pi}}, \tag{5}$$

where the first transition is due to that fact that $\mathbf{w}_1 \cdot \mathbf{x} = \mathbf{w}_2 \cdot M\mathbf{x}$, so only instances for which $\mathbf{w}_1 \cdot \mathbf{x} - b_1 \in [-\sigma\epsilon, 0]$ contribute to the above event. The last transition is by Lemma A.1. Similarly,

$$\Pr_{\mathbf{x}\sim\mathcal{D}}\Big[\mathbf{w}_1 \cdot \mathbf{x} - b_1 \notin [-\ell, 0] \text{ but } \mathbf{w}_2 \cdot M\mathbf{x} - b_2 \in [-\ell, 0], \|\mathbf{x}\| \leq R\Big] \leq \frac{\epsilon}{\sqrt{2\pi}}. \tag{6}$$

Next, we show that the contribution of $\mathbf{x}$ to the soft-margin of $(\mathbf{w}_1, b_1)$ is close to the contribution of $M\mathbf{x}$ to the soft-margin of $(\mathbf{w}_2, b_2)$.

$$|(b_1 - \mathbf{w}_1 \cdot \mathbf{x}) - (b_2 - \mathbf{w}_2 \cdot M\mathbf{x})| \leq |b_1 - b_2| \leq \sigma\epsilon. \tag{7}$$

Moreover, $\mathbf{x}$ is also close to $M\mathbf{x}$, since M rotates any vector by angle β only. That is, for all $\mathbf{x}$, such that $\|\mathbf{x}\| \leq R$, we have

$$\|\mathbf{x} - M\mathbf{x}\| \leq \|\mathbf{x}\|\|I - M\|_F \leq 2\sqrt{2}\sin(\beta/2)R \leq \sqrt{2}\beta R.$$

Now consider the density of distribution $\mathcal{D}$ indicated by $p(\cdot)$. By Lemma A.3 and the fact that $\|\mathbf{x}\| = \|M\mathbf{x}\|$, we have that for any $\|\mathbf{x}\| \leq R$,

$$\frac{p(\mathbf{x})}{p(M\mathbf{x})} \leq \exp\left(\frac{\sqrt{2}\beta R \cdot (6r + 2\sigma\sqrt{2\ln(2/\epsilon)})}{2\sigma^2}\right) \leq \exp\left(\frac{3\beta R^2}{\sigma^2}\right) < 1 + 6\beta\frac{R^2}{\sigma^2} = 1 + 6\epsilon R^2 \tag{8}$$

where the penultimate transition is by the fact that $\exp(x) \leq 1 + 2x$ for $x < 1$ and $\frac{3\beta R^2}{\sigma^2} \leq 1$ for $\epsilon \leq 1/3R^2$. Lastly, by Lemma A.2, all but ν fraction of the points in $\mathcal{D}$ are within distance R of the origin. Putting these together, for $\Omega = \{\mathbf{x} \mid \|\mathbf{x}\| \leq R\}$, we have

$$\begin{aligned}
&\left|\text{S-Den}_{\mathcal{D}}(\mathbf{w}_1, b_1) - \text{S-Den}_{\mathcal{D}}(\mathbf{w}_2, b_2)\right| \\
&= \left|\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\Big[(b_1 - \mathbf{w}_1 \cdot \mathbf{x})\mathbf{1}(\mathbf{w}_1 \cdot \mathbf{x} - b_1 \in [-\ell, 0])\Big] - \mathop{\mathbb{E}}_{M\mathbf{x}\sim\mathcal{D}}\Big[(b_2 - \mathbf{w}_2 \cdot M\mathbf{x})\mathbf{1}(\mathbf{w}_2 \cdot M\mathbf{x} - b_2 \in [-\ell, 0])\Big]\right| \\
&\leq \Pr_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x} \notin \Omega] + \ell \Pr_{\mathbf{x}\sim\mathcal{D}}\Big[\mathbf{w}_1 \cdot \mathbf{x} - b_1 \in [-\ell, 0] \text{ but } \mathbf{w}_2 \cdot M\mathbf{x} - b_2 \notin [-\ell, 0]\Big] \\
&\quad + \ell \Pr_{M\mathbf{x}\sim\mathcal{D}}\Big[\mathbf{w}_1 \cdot \mathbf{x} - b_1 \notin [-\ell, 0] \text{ but } \mathbf{w}_2 \cdot M\mathbf{x} - b_2 \in [-\ell, 0]\Big] \\
&\quad + \left|\int_{\Omega}\Big(p(\mathbf{x})(b_1 - \mathbf{w}_1 \cdot \mathbf{x}) - p(M\mathbf{x})(b_2 - \mathbf{w}_2 \cdot M\mathbf{x})\Big)\, d\mathbf{x}\right| \\
&\leq \nu + \frac{2\ell\epsilon}{\sqrt{2\pi}} + 2\int_{\Omega}\max\{|p(\mathbf{x}) - p(M\mathbf{x})|, |(b_1 - \mathbf{w}_1 \cdot \mathbf{x}) - (b_2 - \mathbf{w}_2 \cdot M\mathbf{x})|\} \\
&\leq \nu + \frac{2\epsilon\ell}{\sqrt{2\pi}} + 6R^2\epsilon + \sigma\epsilon \\
&\leq O(\nu + \epsilon R^2),
\end{aligned}$$

where the last transition is by the assumption that $\sigma \in O(R)$ and $\ell \in O(R)$.

C.4 Proof of Lemma 4.7

We suppress $\mathcal{D}$ and ℓ in the notation of margin density and soft-margin density below. Let $\mathbf{w}^*, b^*$ be the optimal solution to the weighted soft-margin density problem, i.e.,

$$(\mathbf{w}^* \cdot \mathbf{w}_f) \cdot \text{S-Den}^{\ell}_{\mathcal{D}}(\mathbf{w}^*, b^*) = \max_{\eta\in N}\eta \cdot \max_{\substack{\mathbf{w}\cdot\mathbf{w}_f=\eta \\ b\in\mathbb{R}}} (\mathbf{w} \cdot \mathbf{w}_f) \cdot \text{S-Den}(\mathbf{w}, b).$$

Note that for any $\mathbf{x} \sim \mathcal{D}$ whose contribution to S-Den$(\mathbf{w}^*, b^*)$ is non-zero, it must be that $\mathbf{w}^* \cdot \mathbf{x} - b \in [-\ell, 0]$. By definitions of margin and soft margin densities, we have

$$\alpha\left(\mathbf{w}^* \cdot \mathbf{w}_f\right) \cdot \text{S-Den}(\mathbf{w}^*, b^*) \leq \alpha\ell\left(\mathbf{w}^* \cdot \mathbf{w}_f\right) \cdot \text{Den}(\mathbf{w}^*, b^*) \leq \ell\left(\hat{\mathbf{w}} \cdot \mathbf{w}_f\right) \cdot \text{Den}(\hat{\mathbf{w}}, \hat{b}). \tag{9}$$

Consider the margin density of $(\hat{\mathbf{w}}, \hat{b})$. There are two cases:

1. At least half of the points, $\mathbf{x}$, in the ℓ-margin are at distance at least $\ell/2$ from the boundary, i.e., $\hat{\mathbf{w}} \cdot \mathbf{x} - \hat{b} \in [-\ell, -\ell/2]$. In this case, we have

$$\left(\frac{1}{2}\right)\left(\frac{\ell}{2}\right) \text{Den}(\hat{\mathbf{w}}, \hat{b}) \leq \text{S-Den}(\hat{\mathbf{w}}, \hat{b}). \tag{10}$$

2. At least half of the points, $\mathbf{x}$, in the ℓ-margin are at distance at most $\ell/2$ from the boundary, i.e., $\hat{\mathbf{w}} \cdot \mathbf{x} - \hat{b} \in [-\ell/2, 0]$. Note that all such points are in the ℓ-margin of the halfspace defined by $(\hat{\mathbf{w}}, \hat{b} + \ell/2)$. Additionally, these points are at distance of at least $\ell/2$ from this halfspace. Therefore,

$$\left(\frac{1}{2}\right)\left(\frac{\ell}{2}\right) \text{Den}(\hat{\mathbf{w}}, \hat{b}) \leq \text{S-Den}(\hat{\mathbf{w}}, \hat{b} + \ell/2). \tag{11}$$

The claim is proved using Equations 9, 10, and 11.

C.5 Proof of Theorem 4.4

For ease of exposition, let $\epsilon' = \min\{\epsilon^4, \epsilon^2\sigma^4/r^4\}$, $\nu = 1/\exp(1/\epsilon)$, and $R = 2r + \sigma\sqrt{2\ln(1/\nu)}$. Let's start with the reformulating the optimization problem as follows.

$$\max_{g \in \mathcal{G}_{0\text{-}1}} \text{Gain}(g) = \max_{\substack{\mathbf{w} \in \text{Img}(P) \\ b \in \mathbb{R}}} \left(\mathbf{w} \cdot \mathbf{w}_f\right) \text{S-Den}_{\mathcal{D}}^{1/c}(\mathbf{w}, b) = \max_{\eta \in [0,1]} \eta \max_{\substack{\mathbf{w} \in \text{Img}(P) \\ \mathbf{w} \cdot \mathbf{w}_f = \eta \\ b \in \mathbb{R}}} \text{S-Den}_{\mathcal{D}}^{1/c}(\mathbf{w}, b). \tag{12}$$

Let $(\mathbf{w}^*, b^*)$ be the optimizer of Equation (12). Next, we show the value of solution $(\mathbf{w}^*, b^*)$ can be approximated well by a solution $(\hat{\mathbf{w}}, \hat{b})$ where $\hat{\mathbf{w}} \cdot \mathbf{w}_f$ and b belongs to a discrete set of values. Let $N := \{0, \epsilon'\|P\mathbf{w}_f\|, 2\epsilon'\|P\mathbf{w}_f\| \ldots, \|P\mathbf{w}_f\|\}$. Let $\eta^* = \mathbf{w}^* \cdot \mathbf{w}_f$ and define i such that $\eta^* \in (i\epsilon'\|P\mathbf{w}_f\|, (i+1)\epsilon'\|P\mathbf{w}_f\|]$. If $i > \lfloor \frac{1}{\epsilon'} \rfloor - 2$, then let $\hat{\mathbf{w}} = \frac{P\mathbf{w}_f}{\|P\mathbf{w}_f\|}$. If $i \geq \lfloor \frac{1}{\epsilon'} \rfloor - 2$, then let $\hat{\mathbf{w}}$ be the unit vector defined by Lemma 4.5 when $\mathbf{w}_1 = \mathbf{w}^*$ and $\mathbf{w}_2 = \frac{P\mathbf{w}_f}{\|P\mathbf{w}_f\|}$. Then,

$$\hat{\mathbf{w}} \cdot \mathbf{w}_f = j\epsilon'\|P\mathbf{w}_f\|, \text{where} j = \begin{cases} \lfloor \frac{1}{\epsilon'} \rfloor & \text{If } i \geq \lfloor \frac{1}{\epsilon'} \rfloor - 2 \\ i + 1 & \text{Otherwise} \end{cases}.$$

By these definitions and Lemma 4.5, we have that $\hat{\mathbf{w}} \cdot \mathbf{w}^* \geq 1 - 2\epsilon'\|P\mathbf{w}_f\|$, which implies that $\theta(\mathbf{w}^*, \hat{\mathbf{w}}) \leq \sqrt{2\epsilon'\|P\mathbf{w}_f\|}$.

Next, we use the lipschitzness of the soft-margin density to show that the soft-margin density of $\mathbf{w}^*$ and $\hat{\mathbf{w}}$ are close. Let $\hat{b}$ be the closest multiple of $\epsilon'\|P\mathbf{w}_f\|$ to b. Using Lemma 4.6, we have that

$$\left|\text{S-Den}_{\mathcal{D}}(\mathbf{w}^*, b^*) - \text{S-Den}_{\mathcal{D}}(\hat{\mathbf{w}}, \hat{b})\right| \leq O\left(\nu + \frac{\epsilon'\|P\mathbf{w}_f\|}{\sigma^2} R^2\right) \in O\left(\nu + \frac{R^2}{\sigma^2}\sqrt{2\epsilon'\|P\mathbf{w}_f\|}\right).$$

Note that for any $|a_1 - a_2| \leq \epsilon'$ and $|b_1 - b_2| \leq \epsilon'$, $|a_1 b_1 - a_2 b_2| \leq (a_1 + b_1)\epsilon' + \epsilon'^2$, therefore,

$$\left(\hat{\mathbf{w}} \cdot \mathbf{w}_f\right) \text{S-Den}_{\mathcal{D}}^{\ell}(\hat{\mathbf{w}}, \hat{b}) \geq \left(\mathbf{w}^* \cdot \mathbf{w}_f\right) \text{S-Den}_{\mathcal{D}}^{\ell}(\mathbf{w}^*, b^*) - O\left(\nu + \frac{R^2}{\sigma^2}\sqrt{2\epsilon'\|P\mathbf{w}_f\|}\right). \tag{13}$$

By Lemma 4.7, the outcome of the algorithm has the property that

$$\text{Gain}(\bar{\mathbf{w}}, \bar{b}) \geq \frac{1}{4}\text{Gain}(\hat{\mathbf{w}}, \hat{b})$$

Putting this all together, we have

$$\text{Gain}(\bar{\mathbf{w}}, \bar{b}) \geq \max_{g \in \mathcal{G}_{0\text{-}1}} \text{Gain}(g) - O\left(\nu + \frac{R^2}{\sigma^2}\sqrt{2\epsilon' \|P\mathbf{w}_f\|}\right)$$

Now, replacing back $\epsilon' = \min\{\epsilon^4, \epsilon^2\sigma^4/r^4\}$, $\nu = 1/\exp(1/\epsilon)$, and $R = 2r + \sigma\sqrt{2\ln(1/\nu)}$, we have

$$\nu + \frac{R^2}{\sigma^2}\sqrt{2\epsilon' \|P\mathbf{w}_f\|} \leq O\left(\epsilon + \frac{r^2}{\sigma^2}\sqrt{\epsilon'} + \frac{\sigma^2}{\sigma^2\epsilon}\sqrt{\epsilon'}\right) \leq O(\epsilon)$$

D Proof of Theorem 5.3

The first claim directly follows from Lemma 5.2, Theorem 5.1 and the facts that $\text{Gain}_{\mathbf{x}}(g) \in [0, \ell]$ and $\text{Gain}(g) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\text{Gain}_{\mathbf{x}}(g)]$.

For the second claim, note that Algorithm 1 accesses distribution $\mathcal{D}$ only in steps that maximize the margin density of a classifier (the calls to oracle $\text{O}(\mathcal{D}, \frac{1}{c}, \mathcal{K}_\eta)$) or maximize the gain of a classifier. By Theorem 5.1, Lemma 5.2, and the fact that $\text{Den}(g) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\text{Den}_{\mathbf{x}}(g)]$, we have that the density of any outcome of the oracle $\text{O}(S, \frac{1}{c}, \mathcal{K}_\eta)$ is optimal up to $O(\epsilon/c)$ for cost unit c. Lemma 4.7 shows that the soft-margin density and the gain is also within a factor $O(\epsilon)$ from the 1/4 approximation.

E Threshold Mechanisms when All Features are Visible

Consider the special case where the projection matrix P is the identity matrix (or, more generally, has full rank), and the mechanism is tasked with implementing a desired threshold rule. In this case, we want to maximize the number of individuals whose features satisfy some arbitrary condition, and the features are fully visible to the mechanism. Suppose that $f : \mathbb{R}^n \to \{0, 1\}$ is an arbitrary binary quality metric, and the mechanism design space is over all functions $g : \mathbb{R}^n \to [0, 1]$ (even non-binary functions). Then in particular it is possible to set $g = f$, and we observe that this choice of g is always optimal.

Proposition E.1. *For any $f : \mathbb{R}^n \to \{0, 1\}$, if $\mathcal{G}$ is the set of all functions $g : \mathbb{R}^n \to [0, 1]$, then $f \in \operatorname{argmax}_{g\in\mathcal{G}} \text{Val}(g)$.*

Proof. Fix f, and write $S_f = \{\mathbf{x} \in \mathbb{R}^n \mid f(\mathbf{x}) = 1\}$ for the set of feature vectors selected by f. Let $\bar{S}_f = \{\mathbf{x} \in \mathbb{R}^n \mid \exists \mathbf{y} \text{ s.t. } f(\mathbf{y}) = 1, \text{cost}(x, y) < 1/c\}$. That is, $\bar{S}_f$ contains all points of S_f, plus all points that lie within distance $1/c$ of any point in S_f. Write $\Phi_f = \Pr_{\mathbf{x}\sim\mathcal{D}}\left[\mathbf{x} \in \bar{S}_f\right]$. Then note that Φ_f is an upper bound on $\text{Val}(g) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[f(\delta_g(\mathbf{x}))\right]$ for any g, since Φ_f is the probability mass of all individuals who could possibly move their features to lie in S_f at a total cost of at most 1, even if compelled to make such a move by a social planner, and 1 is the maximum utility that can be gained by any move.

We now argue that setting $g = f$ achieves this bound of Φ_f, which completes the proof. Indeed, by definition, if $\mathbf{x} \in \bar{S}_f$ then there exists some $\mathbf{y} \in S_f$ such that $\text{cost}(x, y) < 1$. So in particular $U_f(\mathbf{x}, \mathbf{y}) > 0$, and hence $U_f(\mathbf{x}, \delta_f(\mathbf{x})) > 0$. We must therefore have $f(\delta_f(\mathbf{x})) = 1$ for all $\mathbf{x} \in \bar{S}_f$, since $U_f(\mathbf{x}, \mathbf{x}') \leq 0$ whenever $f(\mathbf{x}') = 0$. We conclude that $\text{Val}(f) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[f(\delta_f(\mathbf{x}))\right] = \Phi_f$, as required. □

This proposition demonstrates that our problem is made interesting in scenarios where the projection matrix P is not full-rank, so that some aspects of the feature space is hidden from the mechanism.

We note that Proposition E.1 does not hold if we allow f to be an arbitrary non-binary quality metric $f : \mathbb{R}^n \to [0, 1]$. For example, suppose $f(\mathbf{x}) = 0.1$ when $x_1 \geq 0$ and $f(\mathbf{x}) = 0$ otherwise, $c = 1$, and $\mathcal{D}$ is the uniform distribution over $[-1, 1]^n$. Then setting $g = f$ results in all agents with $x_1 \geq -0.1$ contributing positively to the objective, since agents with $x_1 \in (-0.1, 0)$ will gain positive utility by shifting to $x_1' = 0$. However, if we instead set g such that $g(\mathbf{x}) = 1$ when $x_1 \geq 0$ and $g(\mathbf{x}) = 0$ otherwise, then all agents will contribute positively to the objective, for a strictly greater aggregate quality.