Incentivizing Truthfulness Through Audits in Strategic Classification

Andrew Estornell
Computer Science & Engineering
Washington University in St. Louis
aestornell@wustl.edu

Sanmay Das
Computer Science
George Mason University
sanmay@gmu.edu

Yevgeniy Vorobeychik
Computer Science & Engineering
Washington University in St. Louis
yvorobeychik@wustl.edu

Abstract

In many societal resource allocation domains, machine learning methods are increasingly used to either score or rank agents in order to decide which ones should receive either resources (e.g., homeless services) or scrutiny (e.g., child welfare investigations) from social services agencies. An agency's scoring function typically operates on a feature vector that contains a combination of self-reported features and information available to the agency about individuals or households. This can create incentives for agents to misrepresent their self-reported features in order to receive resources or avoid scrutiny, but agencies may be able to selectively audit agents to verify the veracity of their reports.

We study the problem of optimal auditing of agents in such settings. When decisions are made using a threshold on an agent's score, the optimal audit policy has a surprisingly simple structure, uniformly auditing all agents who could benefit from lying. While this policy can, in general be hard to compute because of the difficulty of identifying the set of agents who could benefit from lying given a complete set of reported types, we also present necessary and sufficient conditions under which it is tractable. We show that the scarce resource setting is more difficult, and exhibit an approximately optimal audit policy in this case. In addition, we show that in either setting verifying whether it is possible to incentivize exact truthfulness is hard even to approximate. However, we also exhibit sufficient conditions for solving this problem optimally, and for obtaining good approximations.

1 Introduction

Algorithmic decision-making systems are increasingly used to make high-stakes resource allocation decisions by social services agencies. This includes both scarce resource settings, where the demand for a limited pool of resources exceeds supply (for example, housing for the homeless (Kube, Das, and Fowler, 2019)), as well as risk-scoring settings, where only those who fall above or below a certain threshold are either given a resource (for example, a loan (Agarwal, Skiba, and Tobacman, 2009)) or targeted for further scrutiny (for example, parents suspected of child maltreatment or neglect (Chouldechova et al., 2018)). As is standard in classification and ranking settings, each individual or household (henceforth *agent*) is associated with a feature vector. In many such settings, the feature vector will combine information submitted by the agents themselves with information about them available from other sources. For example, in prioritizing households for homeless services, agencies make decisions based on self-reported items (e.g., history of alcohol or drug use) as well as on information available to them in government records (e.g., child-support or welfare payments received) (Brown et al., 2018). Naturally, this creates incentives for agents to try and game the system by strategically choosing their self-reported features in order to maximize their chances of receiving the resource or avoiding scrutiny.

Prior work on *strategic* or *adversarial classification* has considered a closely related problem where agents subject to classification can modify feature values at some cost or subject to a constraint on the total magnitude of such modification, with the goal of inducing an incorrect prediction (Athalye, Carlini, and Wagner, 2018; Carlini and Wagner, 2017; Hardt et al., 2016; Milli et al., 2019; Papernot et al., 2018; Tong et al., 2019; Vorobeychik and Kantarcioglu, 2018). This research has typically focused on either assessing how vulnerable particular families of classifiers are to such attacks (often termed *adversarial examples*) (Athalye, Carlini, and Wagner, 2018; Carlini and Wagner, 2017; Lowd and Meek, 2005; Xu, Qi, and Evans, 2016), or on designing classifiers that are robust in the sense that the prediction remains unchanged even after budget-constrained feature modifications (Brückner and Scheffer, 2011, 2012; Hardt et al., 2016; Li and Vorobeychik, 2018; Madry et al., 2018; Tong et al., 2019; Wong and Kolter, 2018). In this literature, the interests of the agents are commonly viewed as opposed to those of the decision-maker (e.g., learner), often motivated by security considerations (Šrndic and Laskov, 2014; Xu, Qi, and Evans, 2016). Moreover, the typical models representing costs to agents of modifying features are at times not adequate at capturing realistic limits on what agents can do (Tong et al., 2019; Wu, Tong, and Vorobeychik, 2020). In contrast, in the kinds of social services settings we describe, and potentially numerous others (e.g., tax filing), the costs of misrepresenting one's self-reported features are better captured by the risk associated with being audited than, say, a hard constraint on how much the features are modified. Moreover, the agents' interests are not fundamentally opposed to the principal's; rather, this is a case of misaligned

incentives more akin to that studied in the incentive design literature (Haeringer, 2018; Nisan et al., 2007).

We consider a principal who has a limited budget of audits and can use these to determine whether an agent is telling the truth, with the cost of failing an audit the primary deleterious consequence to dishonest agents. For example, caseworkers can interview associates of the agent and ask about behavioral issues, alcohol or drug use, and the like, and impose restrictions or fines on the agent if the results reveal dishonesty. We suppose that the principal uses a score function f (for example, learned risk scores) that takes agent features as input in order to decide whether an agent is subject to further scrutiny whenever their score exceeds a predefined threshold (we term this the *threshold* setting), or to allocate resources to the agents with the top k values of f (we call this the *top-k* setting). We specifically focus on two problems: 1) designing an audit policy for the principal that minimizes incentives to lie, defined in terms of approximate Bayes-Nash incentive compatibility (ε-BNIC), and 2) verifying whether it is possible to ensure truthful reporting of features.

We show that in the threshold allocation setting an optimal policy audits uniformly at random all agents who are above the threshold, with special consideration for those who are either obviously lying or telling the truth. Although this policy is in general hard to compute, we present sufficient conditions under which it is tractable. In the top-k setting, we prove that auditing all agents who receive the scarce resource uniformly at random (again, modulo special treatment of agents who are either certainly truthful or dishonest) yields an additive approximation bound, although the problem is hard in general. Furthermore, we show that this audit policy is optimal if we consider dominant strategy incentive compatibility as a solution concept instead of ε-BNIC.

Surprisingly, the verification problem is even harder: determining if any audit policy can incentivize truthful reporting is #P-hard even for a uniform prior over features and only two agents. However, we give sufficient conditions under which verification becomes tractable in the threshold setting for both piecewise linear and logistic scoring functions. Our corresponding results are weaker for the top-k setting, where we require the distribution over features to be uniform to obtain a tractable algorithm for checking incentives to lie *assuming* that a uniform audit policy is used. Finally, we show that for distributions for which we can efficiently approximate integrals over intervals, we can also approximately verify incentive compatibility.

Our results are important for understanding the potential for audits to be useful in various social services settings. Of perhaps the most practical importance is the clear distinction we find between the threshold (modeling unlimited, but costly, deployment of resources) and top-k (modeling scarce resource allocation) settings in terms of the difficulty of finding a good audit policy, and the simplicity of the optimal audit policy in the threshold setting.

2 Preliminaries

We consider a setting with a collection of n agents in which either a scarce resource is distributed among k of them using a score function, or each agent is scored to determine whether they are selected to receive a resource. Each agent is associated with a vector of attributes (features) which are grouped into two categories: "known", denoted by $\mathbf{x}$, and "self-reported", denoted by $\mathbf{z}$. Throughout, we refer to $(\mathbf{x}, \mathbf{z})$ as an agent's *true type*, to contrast it with $(\mathbf{x}, \mathbf{z}')$ in which $\mathbf{z}'$ is self-reported and may be different from the true corresponding characteristics of the agent. For example, the agent may have a history of substance abuse, corresponding to "true" $z_j = 1$, but reports that they do not, with "reported" $z'_j = 0$. Let d be the number of known and s the number of self-reported features. We assume that each feature in either category either belongs to a continuous or discrete interval, i.e., each $x_j, z_k \in I = [a, b] \cap S$, where $S = \mathbb{R}$ (continuous interval) or $S = \mathbb{Z}$ (discrete interval). We further assume that the true types of each of the n agents are i.i.d. according to a (common knowledge) prior distribution D with PDF (or PMF, in the discrete case) denoted by $h : I^d \times I^s \to [0, 1]$. We will use $\mathbb{P}(\cdot)$ to denote the associated probability measure.

Let $\mathcal{A} = \{\mathbf{a}_1, ..., \mathbf{a}_n\}$ denote the collection of n agents, where $\mathbf{a}_i = (\mathbf{x}_i, \mathbf{z}_i) \in I^d \times I^s$ represents the agent's *true* type, and let $\mathcal{A}' = \{\mathbf{a}'_1, ..., \mathbf{a}'_n\}$ be the collection of reported types, $\mathbf{a}'_i = (\mathbf{x}_i, \mathbf{z}'_i)$. We assume that each agent knows their own type, but only knows the common prior h about the types of other agents.

The principal publishes a score function $f : I^d \times I^s \to \mathbb{R}$ that takes each agent's *reported* type $\mathbf{a}'_i$ as input, and returns a real-valued score. For example, f may represent the probability (learned from historical data) that a homeless person will be safely and stably housed in 1 year if allocated a housing resource. There are two common ways that f is used in resource allocation: (1) **Threshold allocation:** all agents scoring above a threshold θ are allocated a resource (e.g., not chosen for further scrutiny in a child neglect case), and (2) **Top-k allocation:** agents with the highest k scores based on reported types are allocated a resource (e.g., housing).

The principal can *audit* up to B agents and thereby verify whether their reported type matches their true type. Let ϕ denote the audit policy, which is a function of the full collection of n reported types $\mathcal{A}'$. We consider stochastic audit policies, where $\phi_i(\mathcal{A}') \in [0, 1]$ is the probability that agent i is audited. If an audit of agent i determines that the agent has lied, i.e., $\mathbf{z}'_i \neq \mathbf{z}_i$, there are two consequences: 1) the agent does not receive the resource, and 2) the agent pays a penalty (fine) $c \geq 0$. Let α denote the allocation policy with $\alpha_i(f, \mathcal{A}', \phi) = 1$ if agent i receives the resource, and 0 otherwise. Further, let $\mathcal{L}_i = 1$ if agent i is audited and $\mathbf{z}'_i \neq \mathbf{z}_i$ (the agent is caught lying) and 0 otherwise; note that since the audit policy is stochastic, $\mathcal{L}_i$ is a random variable. We assume that an agent obtains a value of 1 for receiving the resource and 0 otherwise. Consequently, the agent's utility is $u_i(\mathcal{A}') = \alpha_i(f, \mathcal{A}', \phi)(1 - \mathcal{L}_i) - c\mathcal{L}_i$.

This game between a principal and agents can be expressed as the following sequence of events:

1. The principal knows D, n, $I^d \times I^s$, α, c, and f, and announces an audit policy ϕ.
2. Realizations of n agents are drawn i.i.d. from D. Each agent knows its own type $(\mathbf{x}_i, \mathbf{z}_i)$, D, n, $I^d \times I^s$, α, c, ϕ,

and f, but does not know the types of other agents.

3. Agents simultaneously submit their reported type $(\mathbf{x}_i, \mathbf{z}'_i)$, where $\mathbf{z}'_i$ need not equal $\mathbf{z}_i$.
4. The principal audits up to B agents, according to ϕ. Any agent i found to have reported $\mathbf{z}'_i \neq \mathbf{z}_i$ is removed from consideration (not allocated the resource), and pays a fine of c.
5. The remaining agents are distributed a resource according to α.

Note that if an agent i is found to be dishonest through an audit in the top-k allocation setting, *another agent would receive the resource in place of* i.

The goal of the principal is to achieve truthful reporting of types by the agents in an (approximate) Bayes-Nash equilibrium, or (approximate) *Bayes-Nash incentive compatibility (BNIC)*. Formally:

Definition 1. *(ε-BNIC) An audit policy ϕ is ε-Bayes-Nash incentive compatible (ε-BNIC) if for all i and $\mathbf{a}_i$,*

$$\mathbb{E}_{\mathcal{A}_{-i}\sim D}[u_i(\mathbf{a}_i, \mathcal{A}_{-i})|f, \phi, \alpha] \geq \mathbb{E}_{\mathcal{A}_{-i}\sim D}[u_i(\mathbf{a}'_i, \mathcal{A}_{-i})|f, \phi, \alpha] - \varepsilon \quad \forall \mathbf{a}'_i : \mathbf{x}'_j = \mathbf{x}_j.$$

ϕ is BNIC if it is 0-BNIC.

We consider two problems in this setting. First, since it is in general impossible to induce BNIC, as we show below, we aim to identify an *optimal* audit policy, defined as follows.

Definition 2. *(Optimal) An audit policy ϕ is optimal if ϕ induces an ε^*-BNIC, and there does not exist another policy ϕ' for which truthful reporting is an ε-BNIC with $\varepsilon < \varepsilon^*$.*

In other words, the optimal ϕ induces the least incentive to lie among all policies.[1] As a consequence, if we find that an optimal policy is not BNIC, then no policy can be. Our second problem is to determine the smallest ϵ that can be induced by an audit policy. We show that in general, these problems have differing complexity.

Before proceeding with a general analysis, we make three observations about our model: 1) if $B = n$, any score function f can be made BNIC; 2) if $k \in \{0, n\}$, the top-k case is trivially BNIC; and 3) if $(1 + c)(B/k) \geq 1$, the top-k case is again trivially BNIC.

We begin by showing that without auditing the self-reported features (equivalently, when the audit budget $B = 0$), ensuring BNIC amounts to ignoring $\mathbf{z}$ altogether whenever we use a deterministic scoring function f. Since self-reported features may be important in determining priority of individuals for resources, this impossibility motivates a careful treatment of optimal auditing, which follows.

Proposition 1. *Suppose $B = 0$. Then, both the top-k and threshold mechanism are incentive compatible iff $f(\mathbf{x}, \mathbf{z}) = f(\mathbf{x})$. Moreover, in the threshold setting, BNIC can be achieved only if $c > 0$.*

Due to space constraints, this and other full proofs are deferred to the supplement.

[1] To avoid confusion, note that the principal could have other objectives, and our definition of optimality is specific to inducing the "best" approximation of BNIC.

3 Design of Optimal Audit Policies

The problem of incentivizing truthfulness via auditing can be broken into two primary components: design and verification. The first component, design, is the construction of *optimal* or approximately *optimal* audit policies. The second, verification, focuses on computing the maximum incentive to lie under an *optimal* audit policy, denoted as ε^*. Although both problems are in general hard, we show that verification is intrinsically "harder" in the sense that in a wide range of settings optimally auditing agents is tractable, but computing ε^* remains hard. The focus of this section is on design. In particular, we exhibit a simple audit policy which is guaranteed to be optimal under the threshold allocation setting, and approximately optimal under the top-k allocation setting.

We begin with some remarks and notation that will be subsequently used in characterizing the optimal audit policies. When selecting which agents to audit, the principal is unaware of each agent's true type $\mathbf{a}_i = (\mathbf{x}_i, \mathbf{z}_i)$, and sees only the reported type $\mathbf{a}'_i = (\mathbf{x}_i, \mathbf{z}'_i)$. Since the principal is interested in minimizing the marginal gain that *any* agent can achieve from lying, agents' true types must be considered through the lens of worst-case analysis. Note that the type with the largest incentive to report $(\mathbf{x}_i, \mathbf{z}'_i)$ is the type with the lowest scoring $\mathbf{z}$, given *known* type $\mathbf{x}_i$ (denoted as $\mathbf{a}^*_i = (\mathbf{x}, \mathbf{z}^*_i)$. From the principal's perspective, any agent reporting $(\mathbf{x}_i, \mathbf{z}'_i)$ must be assumed to have true type $(\mathbf{x}_i, \mathbf{z}^*_i)$.

With this in mind, agent reports can be classified as one of the following: a sure-truth, a sure-lie, or suspicious. Sure-truths are reports which are guaranteed to be honest (e.g. $\mathbf{z}'_i = \mathbf{z}^*_i$). Sure-lies are reports which are guaranteed to be false (these are only of the form $h(\mathbf{x}_i, \mathbf{z}'_i) = 0$). Suspicious reports are those with an unknown truth value. The following two definitions formalize these observations.

Definition 3. *(Minimum Type) For any known partial type $\mathbf{x}_i$, we say the* minimum type *of $\mathbf{x}_i$ is* $\mathbf{a}^*_i = (\mathbf{x}_i, \mathbf{z}^*_i) = \arg\min_{\mathbf{z}\in I^s : h(\mathbf{x}_i, \mathbf{z}) > 0} f(\mathbf{x}_i, \mathbf{z})$.

Definition 4. *(Suspicious) We say a type $\mathbf{a}'_i$ is* suspicious *if the* minimum type $\mathbf{a}^*_i$ *has a strictly lower chance of being allocated a resource barring auditing, i.e.,* $\mathbb{E}_{\mathcal{A}_{-i}}\left[\alpha_i\left(f, \mathcal{A}_{-i} \cup \{\mathbf{a}'_i\}\right)\right] > \mathbb{E}_{\mathcal{A}_{-i}}\left[\alpha_i\left(f, \mathcal{A}_{-i} \cup \{\mathbf{a}^*_i\}\right)\right]$

The key point here is that the principle should never waste an audit on a sure-truth, and when looking at incentive compatibility (i.e. single deviations from collective truth-telling), there is at most one sure-lie in any set of reports, which should be audited with probability 1. The more interesting question regarding audit polices is; what to do with *suspicious* reports.

3.1 Threshold Allocation

Recall that in the threshold allocation setting, an agent receives a resource if $f(\mathbf{x}, \mathbf{z}') \geq \theta$, where $(\mathbf{x}, \mathbf{z}')$ is the agent's reported type. We first show that, in general, optimal auditing under threshold allocation is NP-hard in general, but is tractable if and only if identifying sure-truths is tractable. The hardness of auditing stems from the possibly arbitrary relationship between the distribution D and the score function f.

Theorem 1. *For a given set of n reports $\mathcal{A}'$ and a budget B, computing an optimal audit policy is NP-hard.*

Proof Sketch. This result stems from the observation that the principal would never want to "waste" an audit on an agent whose report is guaranteed to be truthful. For example, suppose agent $\mathbf{a}_1 = (\mathbf{x}_1, \mathbf{z}_1)$ reports type $\mathbf{a}'_1 = (\mathbf{x}_1, \mathbf{z}'_1)$ with $f(\mathbf{x}_1, \mathbf{z}'_1) \geq \theta$. Suppose further that for all $\mathbf{z}$ with $f(\mathbf{x}_1, \mathbf{z}) < \theta$, we have $h(\mathbf{x}_1, \mathbf{z}) = 0$. Then the principal is certain that agent 1 is truthful since this agent's true type could not have scored below the threshold. Due to this dependency on the underlying distribution, one can encode a SAT formula into the distribution such that determining if there exists a $\mathbf{z}$ such that $h(\mathbf{x}_1, \mathbf{z}) > 0$ and $f(\mathbf{x}_1, \mathbf{z}) < \theta$ is equivalent to determining the satisfiability of the SAT instance. □

To better understand the nature of the problem of characterizing an optimal audit policy, consider the following simple example.

Example 1. Suppose there are two agents with one *known* and one *self-reported* binary feature, and suppose that $z = 1$ if $x = 1$, and can be either 0 or 1 according to some prior distribution if $x = 0$. Further, suppose that $f(x, z) = z$ and $\theta = 1/2$, which means that an agent receives the resource iff $z = 1$. Now, suppose that $B = 1$ and the principal observes two types: $(1, 1)$ and $(0, 1)$. Clearly, the principal would not audit the former, since $x = 1$ already implies that the agent is honest, but would audit the latter. This simple example suggests that one could expect an optimal audit policy to depend in rather complex ways on the observed types $\mathcal{A}'$.

However, we show that a simple policy of uniformly auditing all *suspicious* agents (Definition 4), is optimal. We call this policy UNIFORM, and define it formally next.

Definition 5. *(*UNIFORM*) For a given set of reports $\mathcal{A}'$, let $G(\mathcal{A}')$ be the set of all agent's whose reports are* suspicious*. Given budget B, the* UNIFORM *audit policy audits each $\mathbf{a}' \in \mathcal{A}'$ with probability*

$$\phi_i(\mathcal{A}') = \begin{cases} 1 & \text{if } h(\mathbf{a}'_i) = 0 \\ \min\left(\frac{B}{|G(\mathcal{A}')|}, 1\right) & \text{if } \mathbf{a}'_i \in G(\mathcal{A}'), \\ 0 & \text{otherwise} \end{cases}$$

Next, we show that in the threshold allocation setting, this UNIFORM audit policy is optimal.

The intuition for the optimality of UNIFORM comes from the fact that any type $\mathbf{z}$ can report any other type $\mathbf{z}'$ at no cost. This means that any lie that gets an agent above the threshold is equivalent, modulo auditing. Thus, if an audit is non-uniform (as long as the reported type is above the threshold), some lies become more valuable than others, and we should shift auditing to those lies (more precisely, to agents who feature such lies). The discontinuity arises by observing a sure-lie (i.e., $h(x, z) = 0$); only in this case do we know which agent was dishonest, and can thus place higher audit weight on this agent without increasing the value of lying for any other agent.

Note that this implies optimal auditing is equivalent to identifying sure-truths.

Theorem 2. *In the threshold allocation setting, for any score function f,* UNIFORM *is an* optimal *audit policy.*

Proof Sketch. For the sake of illustration, we demonstrate how this result holds in the cases of a discrete distribution over agent types. An identical idea holds for continuous features, although the technical details differ.

When analyzing ε-BNIC, we are considering the value that any agent gains when deviating from a truthful reporting, while all other agents remain truthful, i.e. we consider this case when at most one report is dishonest. In any set of reports $\mathcal{A}'$, if the principal sees a sure-lie, they are immediately aware of the dishonest agent's identity and should exclusively audit that agent, since all other agents are guaranteed to be truthful.

The principal's objective is to minimize the expected gain of any type $\mathbf{a}_i$ misreporting their type as $\mathbf{a}'_i$, when all other agents are truthful. Note that when all agents, aside from agent i are honest, the set of reported types $\mathcal{A}' = \mathcal{A}_{-i} \cup \{\mathbf{a}'_i\}$ (where $\mathcal{A}_{-i}$ is the set true types for all other agents). As such, we can express the minimum expected gain of misreporting, achievable by any audit policy ϕ, as

$$\varepsilon = \min_{\phi} \max_{\mathbf{a}'_i, \mathbf{a}_i} \Bigg(\mathop{\mathbb{E}}_{\mathcal{A}_{-i}} \big[\alpha_i(f, \mathcal{A}') - \alpha_i(f, \mathcal{A})\big] \tag{1}$$

$$- \mathop{\mathbb{E}}_{\mathcal{A}_{-i}} \big[\big(\alpha_i(f, \mathcal{A}') + c\big)\phi_i(\mathcal{A}')\big] \Bigg) \tag{2}$$

Where term (1) the expected difference in the allocation decision between agent i falsely reporting $\mathbf{a}'_i$ or truthfully reporting $\mathbf{a}_i$, and term (2) represents the expected cost of being caught lying when reporting $\mathbf{a}'_i$. Making use of two simple observations, we can simplify this equation. First, in the threshold setting, agents know both their own type and the threshold θ, thus agent i knows the allocation decision on both the true type $\mathbf{a}_i$, and reported type $\mathbf{a}'_i$, meaning that the expectations on α can be dropped. Second, we need only consider this term for *suspicious* agents, so we may assume that $\alpha_i(f, \mathcal{A}') = 1$ and $\alpha_i(f, \mathcal{A}) = 0$. With this, the equation can be simplified to

$$\varepsilon = \min_{\phi} \max_{\substack{\mathbf{a}'_i, \mathbf{a}_i \\ f(\mathbf{a}'_i) \geq \theta > f(\mathbf{a}_i)}} 1 - (1 + c)\mathbb{E}_{\mathcal{A}_{-i}}\big[\phi_i(\mathcal{A}')\big]$$

Thus ε is solely determined by the value of $\mathbb{E}_{\mathcal{A}_{-i}}\big[\phi_i(\mathcal{A}')\big]$ for any *suspicious* type $\mathbf{a}'_i$. Let

$$G(\mathcal{A}') = \{(\mathbf{x}, \mathbf{z}') \in \mathcal{A}' : f(\mathbf{x}, \mathbf{z}') \geq \theta \text{ and } \exists \mathbf{z}^* \text{s.t. } f(\mathbf{x}, \mathbf{z}^*) < \theta \text{ and } h\big((\mathbf{x}, \mathbf{z}^*)\big) > 0 \text{ and } h\big((\mathbf{x}, \mathbf{z}')\big) > 0\}$$

be the set of *suspicious* types in $\mathcal{A}'$. In the case when agent features are distributed according to a discrete distribution, this expectation can be expressed as

$$\begin{aligned} \mathbb{E}_{\mathcal{A}_{-i}}\big[\phi_i(\mathcal{A}')\big] &= \sum_{\mathcal{A}_{-i}} \phi_i(\mathcal{A}') Q(\mathcal{A}_{-i}) \\ &= \sum_{\mathcal{A}_{-i}} \min\left(1, \frac{B}{G(\mathcal{A}')}\right) Q(\mathcal{A}_{-i}), \end{aligned}$$

where $Q(\mathcal{A}_{-i})$ is the probability of any realization of the specified types of agents other than i induced by D. The probability which sure-lies are audited has no effect on the value of other lies, and thus sure-liescan be audited with probability 1. Moreover, the sum is equal for any two *suspicious* agents with $h(\mathbf{a}') > 0$. In each set of reports $\mathcal{A}'$, the principle fully spends their budget (or audits all *suspicious* types with probability 1) and $Q(\mathcal{A}_{-i})$ is independent of the type agent i reports. Thus, for any policy different from UNIFORM, at least one audit weight must be changed, i.e., $\phi_i(\mathcal{A}') \neq \min\left(1, {}^{B}/_{G(A')}\right)$, for some i and some $\mathcal{A}'$. As a result of the tightness and independence of $Q(\mathcal{A}_{-i})$, this change of audit weight could only result in a (not necessarily strict) increase in the expected gain of misreporting for any agent type. □

Note that while optimal, UNIFORM is in general intractable because of the combinatorial structure of such policies that may be induced by $h(\cdot)$. However, we now show that for sufficiently well-behaved h and f we can compute UNIFORM efficiently.

Theorem 3. *The audit policy* UNIFORM *can be computed in polynomial time if for any report $(\mathbf{x}, \mathbf{z}')$, it can be efficiently determined if $(\mathbf{x}, \mathbf{z}')$ is a sure-truth, (i.e. there exists a self reported type $\mathbf{z}^*$, such that $h(\mathbf{x}, \mathbf{z}^*) > 0$ and $f(\mathbf{x}, \mathbf{z}^*) < \theta$).*

3.2 Top-k Allocation

We now turn our attention to selecting the *optimal* audit policy when resources are given to the k highest scoring agents. In this case, the optimal policy no longer admits a clean characterization. The main challenge is that now there are far more complex interdependencies among agents' benefits from lying, other agents' reports, and the audit policy. For example, if an agent in the top-k is caught lying, another agent would now receive the resource. Instead, we study a natural adaptation of UNIFORM to this setting, and exhibit an additive approximation bound for its optimality. We then show that if we use dominant strategy incentive compatibility (defined formally below) as a solution concept in place of BNIC, uniform auditing is optimal even in this setting.

We begin by showing that optimal auditing in the top-k setting is NP-hard even when sure-truths are identifiable in constant time.

Theorem 4. *In the top-k allocation setting determining which agents should be audited is NP-hard, even for $n = 4$ agents, monotone f, uniform D, and even if sure-truth can be identified in constant time.*

Proof Sketch. We can encode an instance of Vertex Cover into f such that agents with a self-reported type, which constitutes a vertex cover, ranks in the top-k with extremely low probability, while all other types have score proportional to number of vertices that their self-reported type "covers". For a sufficiently small budget and penalty for lying, there will be agents whose expected value of lying (even if never audited) is smaller than agents who receive the highest probability weight. As such, the principal must determine which agents should receive zero audit weight, which is NP hard due to the encoding of VC. □

Now, consider a variant of the UNIFORM policy in the top-k setting where we uniformly at random audit agents who have scores in the top k. We first define this policy formally.

Definition 6. *(*UNIFORM-K*) For any set of reported types $\mathcal{A}'$, let* UNIFORM-K *denote the policy of auditing each of the top-k agents (refereed to as the set $T_k \subset \mathcal{A}'$) with probability* $\min(1, {}^{B}/_{k})$.

Next, we show that UNIFORM-K admits an additive approximation of an optimal audit policy in the top-k setting. Recall that multiplicative approximations are in general NP-hard to achieve.

Theorem 5. *Let ϕ denote the audit policy* UNIFORM-K. *Then the maximum utility gained by lying under ϕ is no more than $\max\left(0, 1 - \frac{(1+c)B}{k}\right)$ greater than that of the optimal audit policy, and this bound is tight.*

Proof. This is the result of simple worst case analysis on the expected value of lying, which can be expressed as

$$\begin{aligned}&\mathbb{E}_{\mathcal{A}_{-i}}\left[\alpha_i(\mathcal{A}')\big(1-\phi_i(\mathcal{A}')\big) - c\phi_i(\mathcal{A}') - \alpha_i(\mathcal{A})\right]\\ =&\mathbb{P}(\mathbf{a}'_i \in T_k)\mathbb{E}[\phi_i(\mathcal{A}')|\mathbf{a}'_i \in T_k] - c\mathbb{E}[\phi_i(\mathcal{A}')] + \mathbb{P}(\mathbf{a}_i \in T_k)\end{aligned}$$

In the worst case, the expected value of lying could be 0 for all agents. However, the uniform audit policy will have expected value of lying equal to $\mathbb{P}(\mathbf{a}'_i \in T_k)(1 - (1+c)\frac{B}{k}) - \mathbb{P}(\mathbf{a}_i \in T_k)$. Which again in the worst case is equal to $((1 - (1+c)\frac{B}{k})$

This bound is tight to within any small $\beta > 0$. To see this, construct an instance with 3 agents of types $x \in \{0, 1\}, z \in \{0, 1\}$. Where $f(x, z) = x \wedge z$. Let $\mathbb{P}(x = 1, z = 1) = \beta$ and the rest have probability $\frac{1-\beta}{3}$. Let $B = 1$ and $k = 2$. Then an optimal audit policy yields $\varepsilon^* = 0$, but uniformly auditing the top-k yields $\varepsilon = ((1 - \beta^2)(1 - (1+c)\frac{B}{k})$. □

A major part of what makes auditing difficult is the dependence on the distribution. We now consider an alternative solution concept which eliminates this dependence: ε-Dominant Strategy Incentive Compatibility (ε-DSIC). Specifically, under ε-DSIC the principal aims to design a policy under which truthful reporting is (approximately) optimal for agents regardless of other agents' types.

Definition 7. *An audit policy ϕ is ε-DSIC if for all i and $\mathbf{a}_i$,*

$$\begin{aligned}&\mathbb{E}[u_i(\mathbf{a}_i, \mathcal{A}_{-i})|f, \phi, \alpha, \mathcal{A}'_{-i}]\\ &\geq \mathbb{E}[u_i(\mathbf{a}'_i, \mathcal{A}_{-i})|f, \phi, \alpha, \mathcal{A}'_{-i}] - \varepsilon \;\forall \mathbf{a}'_i : \mathbf{x}'_j = \mathbf{x}_j \text{ and } \forall \mathcal{A}'_{-i}.\end{aligned}$$

Theorem 6. *In the top-k setting,* UNIFORM-K *yields ε^*-DSIC with an optimal ε^*.*

Proof Sketch. In the top-k setting the key difference from ε-BNIC is that for any realization $\mathcal{A}_{-i}$ and any set of corresponding reports $\mathcal{A}'_{-i}$, agent i knows the allocation decision on both their true type $\mathbf{a}_i$ and any reported type $\mathbf{a}'_i$. This certainty of outcomes it precisely what made all *suspicious* reports equivalent in the threshold case. Using a similar argument for the optimality of UNIFORM in the threshold case, we can see that UNIFORM-K is optimal in the top-k case. □

4 Verification of Policy Effectiveness

In the previous section we showed that in many circumstances we can fully characterize the optimal audit policy, and it can be efficiently computed for a broad range of settings. We now consider the problem of *verification*, that is, computing the smallest ϵ^* that we can achieve for an optimal audit policy. We show that this problem is hard even when auditing is easy. Subsequently, we first show that we can often effectively approximate this problem, and then exhibit special cases in which we can even compute this ϵ^* efficiently.

4.1 Complexity of Verification

In the threshold setting, we will show that computing the minimum ε^* inducible by any policy is #P-hard, even in cases when optimal auditing is tractable. This complexity stems from *both* the score function f and distribution D. Intuitively, these uniquely define both the set of agent types which are considered *suspicious* and the probability that a *suspicious* type will occur. As *suspicious* types are more likely to occur, the probability that any particular agent is audited decreases. Thus, we can encode "hard" problems into f or D where agent types (binary vectors) correspond to satisfying assignments of the encoded problem. We can also observe that if the number of possible agent types is polynomial, then the problem is trivially tractable through brute force search.

We show here hardness in terms of f; a similar construction works to show the hardness in terms of D. In this construction, optimal auditing is easy even in the top-k case.

Theorem 7. *In both the threshold and top-k setting, computing the minimum ε inducible by any audit policy is #P-hard, for both continuous and discrete features, even when the feature distribution is uniform, there are only 2 agents, and f is both monotone and binary.*

Proof Sketch. For this proof sketch we will work in the setting of threshold allocation and discrete features, similar logic holds in the other cases. We reduce from #VC. For a graph $G = (V, E)$, let D be uniform and agents be $\mathbf{a} = \langle x_1, ..., x_{|V|}, z_1 \rangle$, for $x, z \in \{0, 1\}$. Let $\theta = \frac{1}{2}$ and set

$$f(\mathbf{x}, z) = \Big(\bigwedge_{(v_r, v_t) \in E} (x_r \vee x_t) \Big) \wedge z_1.$$

Under this construction of f we see that an agent scores $f(\mathbf{a}) = 1$ if and only if $\mathbf{x}$ constitutes a vertex cover and $z = 1$. Thus when $B = 1$ and $n = 2$, if agent 1 scores below $\frac{1}{2}$ and is considering misreporting their type, they are audited with lower probability if $f(\mathbf{a}_2) = 1$. Since D is uniform, the probability of this occurring is equivalent to the number of vertex covers of G. □

In addition to hardness of checking BNIC, we can show that it is even hard to multiplicatively approximate an ε-BNIC in the threshold and top-k settings.

Theorem 8. *Multiplicatively approximating to any constant factor the smallest ε such that there is an ε-BNIC audit policy, in both threshold and top-k allocation is NP-hard even for $\Theta(1)$ agents.*

Proof. This result is a straightforward consequence of the construction in the proof of Theorem 7. In that proof we encode an NP-hard problem into an instance of our problem, and show that determining if truthful reporting is BNIC is equivalent to counting the number of satisfying assignments of vertex covers. If we reduce instead from an Unambiguous-SAT instance (f is no longer monotone), then the mechanism is BNIC if and only if the formula has exactly one satisfying assignment. This would imply that $\varepsilon = 0$ if and only if the U-SAT instance has no satisfying assignment, and any multiplicative factor ε would likewise be zero, immediately indicating the satisfiability of the U-SAT instance. □

Note that UNIFORM-K is the optimal audit policy in these cases, implying that not only is verification of an optimal policy hard, but also verification of UNIFORM-K is also in general hard.

In summary, the problem of *checking* whether a particular setting is ε-BNIC is hard, even in instances when auditing is tractable. To further outline the relation of the complexity of both problems we make the following observation.

Theorem 9. *In the threshold allocation setting, verification being in P implies optimal auditing is also in P.*

Next, we turn to positive results. To begin, we now show that when agents' *minimum* type can be efficiently computed we can achieve a probabilistic bound on the value of lying in polynomial time, via Monte Carlo simulations.

Theorem 10. *Suppose that ε^* is the minimum value for which* UNIFORM *is ε^*-BNIC. Then, for any $\gamma \in \Theta(1)$, performing n^γ rounds of Monte-Carlo sampling will yield a value ε', such that $\varepsilon' = \varepsilon^* \pm \Theta\big(1/\sqrt{n^\gamma - 3}\big)$ with probability at least $1 - 1/n^2$. This can be done in time $\Theta(n^{\gamma+1})$.*

Observe from Theorem 10 that as n increases, the error of approximation tends towards 0 with probability tending towards 1. Next, we consider special cases in which verification is tractable.

4.2 Tractable Special Cases

Thus far, our results are negative when it comes to checking incentive compatibility, and mixed in terms of devising an optimal audit policy. We now proceed to identify a number of special cases in which we can check incentive compatibility in polynomial time. In the threshold setting, we focus on checking ε-BNIC for a UNIFORM audit policy, which we showed earlier is optimal, while in the top-k setting we focus on the UNIFORM-K audit policy. We consider, in particular, three common machine learning models for f: linear, piecewise linear, and logistic (sigmoid) functions. Throughout, we assume that distributions over types are sufficiently well behaved, in that it is tractable to compute probabilities of intervals.

We begin by showing that verification is tractable in any instance in which the CDF (CMF) of h can be computed over the set of *suspicious* agent types. As can be surmised from the complexity results regarding verification, the

"hardness" of the problem stems from determining the probability that an agent's true type is suspicious. However, when this can be computed efficiently, so can ε^*.

Theorem 11. *Let $U = \{(\mathbf{x}, \mathbf{z}') \in I^d \times I^s : f(\mathbf{x}, \mathbf{z}') \geq \theta, h(\mathbf{x}, \mathbf{z}) \neq 0$, and $\exists (\mathbf{x}, \mathbf{z}^*)$ with $f(\mathbf{x}, \mathbf{z}^*) < \theta$ and $h(\mathbf{x}, \mathbf{z}^*) \neq 0\}$. If $\mathbb{P}_{\mathbf{a}\sim D}(\mathbf{a} \in U)$ can be efficiently computed, then so can ε^*.*

Proof Sketch. Let $p_U = \mathbb{P}_{\mathbf{a}\sim D}(\mathbf{a} \in U)$. Suppose an agent initially scores below the threshold, then this agent's only means for allocation is to report a type in U. Moreover, UNIFORM only audits agents in U and does so uniformly. Thus, for a given realization, the more agents with true types in U, the lower the probability that the dishonest agent is audited. More specifically, suppose that some agent $\mathbf{a}_i = (\mathbf{x}_i, \mathbf{z}_i)$, with $f(\mathbf{a}_i) < \theta$, is able to falsely submit $\mathbf{a}'_i = (\mathbf{x}_i, \mathbf{z}'_i)$ with $f(\mathbf{a}'_i) \geq \theta$. Then, this agent's expected marginal gain is,

$$\mathbb{E}_{\mathcal{A}_{-i}}[u_i(\mathbf{a}_i, \mathbf{a}'_i)|f, \alpha, \phi] = \mathbb{E}_{\mathcal{A}_{-i}}[1 - (1+c)\phi_i(\mathcal{A}')]$$

$$= 1 - (1+c)\sum_{\ell=0}^{n-1} \binom{n-1}{\ell} p_U^{\ell}(1-p_U)^{n-\ell-1} \min\left(1, {}^{B}/_{\ell+1}\right)$$

Since, under UNIFORM all dishonest reports have either value 0 or value $\mathbb{E}_{\mathcal{A}_{-i}}[u_i(\mathbf{a}_i, \mathbf{a}'_i)|f, \alpha, \phi]$, we need only compute this single sum, for any agent type, and have found ε. Moreover, UNIFORMis optimal and thus $\varepsilon = \varepsilon^*$. □

In both the discrete and continuous case, when $\mathbb{P}(\mathbf{a} \in U)$ can be computed exactly, verification is tractable. Next, we give a sufficient condition on this, and present several tractable special cases.

Definition 8. *We say a PDF h is* well-behaved *if h is zero on a polynomial number of $s + d-$dimensional maximal intervals, and over any any interval $[a, b] \subset \mathbb{R}$, the value of $\int_a^b h(\mathbf{x}, \mathbf{z})dz_r$ and $\int_a^b h(\mathbf{x}, \mathbf{z})dx_t$ for observed features r and unobserved features s have closed-form solutions derivable in polynomial-time w.r.t. $(n, B, s, d, \log(c))$.*

Remark 1. *Many commonly used distributions, such as uniform and exponential, are* well-behaved. *In many other common cases, such as Gaussian, we can obtain a good numerical approximation, so that the approaches below can approximately apply in these also. We formalize this below.*

As we show next, in the threshold case, checking ε-BNIC is easy for piecewise linear and logistic score functions as long as the distribution over types is *well-behaved*. For top-k, we need a much stronger assumption that types are distributed uniformly to obtain comparable positive results. Each proof proceeds as follows (see the supplement for details). The score function f partitions $I^d \times I^s$ into two disjoint regions, one of which is U (the set of suspicious types). We then show that one of the two partitions it is easy to compute the CDF, as long as h is *well-behaved*. Once this is done, we have computed either p_U or $1 - p_U$ and from here Theorem 11 directly implies that ε^* is computable in polynomial time.

Definition 9. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be* Piecewise Linear *if for some partition of $\mathbb{R}^{d+s}$ into disjoint rectangular regions, given by $P = \{L_1, ..., L_m\}$ the function $f\big|_{L_s} : \mathbb{R}^{d+s} \to \mathbb{R}$ is linear for each $L_s \in P$.*

Corollary 1. *Suppose the distribution of agent types is well-behaved and f is piecewise linear, or logistic, and $\alpha(f, \mathcal{A}')$ is threshold allocation. Then determining the minimum $\varepsilon \geq 0$ such that* UNIFORM *is ε-BNIC, can be done in polynomial time.*

Corollary 2. *Suppose the distribution of agent types is uniform. Suppose further that f is piecewise-linear, or logistic, and $\alpha(f, \mathcal{A}')$ is top-k allocation. Then determining the minimum $\varepsilon \geq 0$ such that* UNIFORM-K *is ε-BNIC, can be done in polynomial time.*

For many common continuous distributions, such as Gaussian, only a numerical approximation of $p_U = \mathbb{P}(\mathbf{a} \in U)$ can be computed. Our final result is to quantify the error in ε^*, in terms of the additive numerical error γ in p_U.

Theorem 12. *Suppose with error γ we have a numerical approximation $p'_U = p_U \pm \gamma$. Then we can compute $\varepsilon' = \varepsilon^* \pm (n - B)\binom{n-1}{B-1}\int_{p_U}^{p_U+\gamma} x^{B-1}(1-x)^{n-B}dx$.*

Although the error term looks messy, it is tight and in general small relative to γ, which itself is also in general a small value. As an illustration, when we have error $\gamma = 4.44\text{E}^{-4}$, a typical absolute error for a standard Gaussian, and $n = 1000, B = 250$, and $p_U = 0.6$, then $\epsilon' = \epsilon^* \pm 6\text{E}^{-60}$.

5 Conclusion

We study the problem of auditing self-reported attributes in resource allocation settings from two perspectives: 1) the complexity of checking whether a particular audit policy is incentive compatible, and 2) characterizing and computing an audit policy that minimizes incentives to lie. We find that checking incentive compatibility is, in general, hard. However, in settings where resources are assigned by thresholding the individual's computed score, a uniform audit policy, particularly appealing for its simplicity, is optimal. In addition, we show that in two important classes of score functions, piecewise linear and logistic, we can check incentive compatibility in polynomial time under some assumptions on the distribution of agent types.

A number of open questions remain. While we show that computing an optimal audit policy in the setting where resources are allocated to the top-k scoring agents is hard, it may be possible to achieve a better approximation of optimal than what we exhibit for the uniform policy. Moreover, our model presumes that agents incur no direct costs of misreported preferences besides the endogenous costs of being audited. In practice, there may be both cognitive and tangible costs involved, and these can be considered as an extension to our model. Finally, we assume that the distribution over agent types is known a priori, whereas it likely needs to be learned from data.

Acknowledgments

This research was partially supported by the National Science Foundation (IIS-1910392, IIS-1939677, IIS-1905558, IIS-1903207, IIS-1927422, and ECCS-2020289), Army Research Office (W911NF-19-1-0241), and Amazon.

References

Agarwal, S.; Skiba, P. M.; and Tobacman, J. 2009. Payday loans and credit cards: New liquidity and credit scoring puzzles? *American Economic Review Papers & Proceedings* 99(2): 412–17.

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. In *International Conference on Machine Learning*, 274–283.

Brown, M.; Cummings, C.; Lyons, J.; Carrión, A.; and Watson, D. P. 2018. Reliability and validity of the Vulnerability Index-Service Prioritization Decision Assistance Tool (VI-SPDAT) in real-world implementation. *Journal of Social Distress and the Homeless* 27(2): 110–117.

Brückner, M.; and Scheffer, T. 2011. Stackelberg games for adversarial prediction problems. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 547–555.

Brückner, M.; and Scheffer, T. 2012. Static Prediction Games for Adversarial Learning Problems. *Journal of Machine Learning Research* 13: 2617–2654.

Carlini, N.; and Wagner, D. 2017. Towards Evaluating the Robustness of Neural Networks. In *IEEE Symposium on Security and Privacy*, 39–57.

Chouldechova, A.; Benavides-Prado, D.; Fialko, O.; and Vaithianathan, R. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, 134–148.

Haeringer, G. 2018. *Market Design: Auctions and Matching*. The MIT Press.

Hardt, M.; Megiddo, N.; Papadimitriou, C.; and Wootters, M. 2016. Strategic classification. In *ACM Conference on Innovations in Theoretical Computer Science*, 111–122.

Kube, A.; Das, S.; and Fowler, P. J. 2019. Allocating Interventions Based on Predicted Outcomes: A Case Study on Homelessness Services. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 622–629.

Li, B.; and Vorobeychik, Y. 2018. Evasion-robust classification on binary domains. *ACM Transactions on Knowledge Discovery from Data* .

Lowd, D.; and Meek, C. 2005. Adversarial Learning. In *ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 641–647.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

Milli, S.; Miller, J.; Dragan, A.; and Hardt, M. 2019. The social costs of strategic classification. In *Conference on Fairness, Accountability, and Transparency*.

Nisan, N.; Roughgarden, T.; Tardos, E.; and Vazirani, V. V., eds. 2007. *Algorithmic Game Theory*. Cambridge University Press.

Papernot, N.; McDaniel, P.; Sinha, A.; and Wellman, M. 2018. Towards the Science of Security and Privacy in Machine Learning. In *IEEE European Symposium on Security and Privacy*.

Šrndic, N.; and Laskov, P. 2014. Practical Evasion of a Learning-Based Classifier: A Case Study. In *IEEE Symposium on Security and Privacy*, 197–211.

Tong, L.; Li, B.; Hajaj, C.; Xiao, C.; Zhang, N.; and Vorobeychik, Y. 2019. Improving Robustness of ML Classifiers against Realizable Evasion Attacks Using Conserved Features. In *USENIX Security Symposium*.

Vorobeychik, Y.; and Kantarcioglu, M. 2018. *Adversarial Machine Learning*. Morgan and Claypool.

Wong, E.; and Kolter, J. Z. 2018. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*.

Wu, T.; Tong, L.; and Vorobeychik, Y. 2020. Defending Against Physically Realizable Attacks on Image Classification. In *International Conference on Learning Representations*.

Xu, W.; Qi, Y.; and Evans, D. 2016. Automatically Evading Classifiers: A Case Study on PDF Malware Classifiers. In *Network and Distributed System Security Symposium*.