# Optimally Auditing Adversarial Agents

**Sanmay Das**[1], **Fang-Yi Yu**[2], **Yuang Zhang**[2]
[1]Virginia Polytechnic Institute and State University
[2]George Mason University
sanmay@vt.edu, fangyiyu@gmu.edu, yzhang78@gmu.edu

**Abstract**

Fraud can pose a challenge in many resource allocation domains, including social service delivery and credit provision. For example, agents may misreport private information in order to gain benefits or access to credit. To mitigate this, a principal can design strategic audits to verify claims and penalize misreporting. In this paper, we introduce a general model of audit policy design as a principal-agent game with multiple agents, where the principal commits to an audit policy, and agents collectively choose an equilibrium that minimizes the principal's utility. We examine both adaptive and non-adaptive settings, depending on whether the principal's policy can be responsive to the distribution of agent reports. Our work provides efficient algorithms for computing optimal audit policies in both settings and extends these results to a setting with limited audit budgets.

**Code** — https://github.com/dasddassad/Optimally-Auditing-Adversarial-Agents

## 1 Introduction

AI is increasingly used in making high-stakes societal decisions. One example that has recently gained considerable attention is the use of AI to decide whether to approve or deny the receipt of social benefits, with worries about how the scale of AI might systematically cut off thousands of people from benefits they are eligible for because of suspicion of fraud (Eubanks 2018). The reason to use AI in these domains, is because human time and resources are limited. However, an alternative method is the use of AI to flag a limited number of applicants for *audits* that can then be conducted by humans. How should one design such audit policies?

The problem of optimal audit design is relevant not just in the case of benefit receipt, but also in many other scenarios where agents must report their types to a principal in order for the principal to decide whether or not an agent is qualified to receive a benefit or service from the principal. In addition to qualification to receive social services or government benefits, other examples include credit or loan applications and tax relief. In all of these, the principal has the ability to, at some cost, audit agents to determine whether or not they are revealing their true types. The principal may also be able to impose a penalty on those caught misreporting (e.g., prosecution of tax fraud, ineligibility for future government services). The goals of the principal can vary. In some cases, they may want to minimize misreporting. For example, a social services agency can be thought of as a benevolent principal trying to maximize the social welfare of recipients, and it needs truthful elicitation in order to achieve this goal. In others, the principal may have their own utility function – for example, a bank making lending decisions.

There have been a number of papers that look at specific versions of the general auditing problem described above. In this paper, we systematically analyze the problem along three different dimensions, and present a number of new results. The first dimension is the goal, distinguishing between maximizing social welfare over the agents and maximizing the principal's utility. The second is whether or not the principal's strategy can be responsive to the actual distribution of agent reports (we call these the adaptive versus non-adaptive settings). Finally, the third dimension is whether audits are limited by a budget or whether the principal can undertake marginal-cost auditing by paying a specific cost for each additional audit.

To give a concrete example, consider the Social Security Administration's Supplemental Security Income program, which has a strict upper limit on assets for eligibility. The U.S. Supplemental Security Income program provides monthly cash only to applicants who (among other requirements) claim less than $ 2000 in so-called "deemed" resources (essentially countable assets). As verifying every claim is costly, the agency can only audit a fraction of applicants. Deciding which brackets to inspect, whether to adapt those fractions after seeing the week's claims, and how to weigh extra recoveries against audit costs mirrors the three axes we analyze: objective (social welfare vs. agency pay-off), adaptivity (fixed vs. responsive rules), and resources (budget vs. cost). Our model captures this core strategic tension: applicants choose whether to misreport their private asset levels, anticipating the mixed audit policy chosen by the principal.

**Our contributions** This paper studies audit mechanism problems when agents adversarially choose the *worst equilibrium*. A principal wants truthful reports from $n$ strategic

agents with private types in $[m]$. She may commit in advance to a probability of auditing different types (audit vector) or to an adaptive policy that decides the audit vector after seeing the agents' reports. Each audit costs $\lambda$ (in the costly setting) or counts against a budget of $B$ (in the budgeted setting); detected misreports incur a penalty. After the principal commits, agents adversarially select the worst-case equilibrium.

1. We fully characterize the equilibrium structure in the non-adaptive costly setting where the principal commits to an audit vector and each audit costs $\lambda$. This structure yields an $\epsilon$-approximation algorithm in $O(m^2)$ time for the principal's utility (Theorem 1). We further prove that exact optimality is impossible in Proposition 1.
2. When the prior is unknown and varies in each round, we develop an online learning algorithm in § 3.2 that has regret $O(n\sqrt{Tm^2 \log m})$ in $T$ rounds (Theorem 2). Interestingly, although exact optimality in the one-shot setting is impossible, careful choice of arms allows our online learning algorithm to satisfy the no regret property.
3. Beyond the principal's utility, §3.3 adapts both the efficient algorithm and the online learner to maximize social welfare, and §3.4 shows that increasing the penalty function or decreasing audit cost can only benefit the principal's utility and social welfare (Proposition 2).
4. For adaptive audits, although the principal has a larger action space, we show that they offer no advantage over nonadaptive audits under the insensitivity assumption Eq. (12) and the Wardrop equilibrium Eq. (6). The same algorithm applies in the adaptive costly-audit setting (Theorem 4). A similar algorithm for the budgeted case appears in the full version.

### 1.1 Related Work

At a high level, our setting is a principal–(multi-)agent Stackelberg game. In this section, we survey relevant techniques and highlight connections to three special cases—audit games, security games, and toll pricing in congestion games.

Our Stackelberg game features one leader and multiple heterogeneous followers. Computing the leader's optimal commitment is known to be hard even with two followers (Conitzer and Sandholm 2006). Our robustness notion relates to standard pessimistic equilibria. (Coniglio, Gatti, and Marchesi 2017), but as our model has a larger action space (real-valued probability of auditing each report) with a non-convex structure, the standard bi-level optimization technique is not feasible. Recent work also considers solving multi-follower games under no externality assumptions (Personnat et al. 2025).

Classic costly-state-verification work (Townsend 1979) and subsequent allocation papers (Mookherjee and Png 1989; Ben-Porath, Dekel, and Lipman 2014) study auditing or verification in resource-allocation settings, typically targeting truthful outcomes. Alm, Bahl, and Murray (1993); Ben abdelaziz, Neifar, and de Bourmont (2015); Coates, Florence, and Kral (2002) analyze equilibrium of audit game, whereas our work address the mechanism design problem. Lundy et al. (2019) study penalty design with an exogenous audit process, whereas in our setting the principal designs the audit strategy. In multi-agent settings, Estornell, Das, and Vorobeychik (2021); Estornell, Chen, and Das (2023) use audits to discourage misreporting and promote beneficial recourse. A key difference of our work from the above is equilibrium selection: rather than targeting truth-telling equilibria, we guarantee the principal's performance at the worst (pessimistic) equilibrium, which may be non-truthful. Recently, Jalota, Tsao, and Pavone (2024) connect information design to audit mechanisms when the agents can commit to a misreporting strategy; in our work, the principal is the one who can commit.

Less directly related are security games, where a defender allocates inspection or patrol resources to deter attackers and inspection costs do not scale with the number of attackers. (Pita et al. 2008; Tambe 2012) Closer to us are audit games that allow the leader to tune punishment for a single agent (Blocki et al. 2013, 2015).

Our audit probabilities play a role analogous to tolls in congestion games: they modify followers' payoffs to steer equilibrium flows. However in our model an agent's cost depends not only on its reported type but also the true type. Foundational work showed that marginal-cost tolls can implement the system optimum in nonatomic traffic (Roughgarden and Tardos 2002; Cole, Dodis, and Roughgarden 2003).

## 2 Audit Mechanism Problem

We study an audit mechanism design problem where a principal interacts with a continuum of (non-atomic) agents with total mass $n$. Each agent has a private type drawn from a known prior and may choose to misreport this type. The principal aims to incentivize truthful reporting by combining costly audits and penalties. We extend our analysis to a setting with a hard audit-budget constraint in the full version.

**Basic Setup** There are $m \geq 2$ ordered types, denoted by $[m] = \{0, 1, \ldots, m-1\}$. Each agent has a private type $i \in [m]$ which is sampled independently from a prior $\boldsymbol{q} \in \Delta_m$ with full support on $[m]$, and reports $k \in [m]$ which may differ from $i$. We use $i, j$ for true types and $k, l$ for reported types.

When a type $i$ agent reports $k$, the principal assigns a payment $\mathrm{pay}(k)$, and receives $\mathrm{val}(i, k)$. The principal detects misreporting through audits by choosing an *audit vector* $\boldsymbol{p} \in [0,1]^m$ where $p_k$ is the probability of auditing an agent reporting type $k$. Once audited, the principal gets $\mathrm{pen}(i, k)$ from the agent. We consider a type-independent penalty of the form $\mathrm{pen}(k)$ so that for all $i, k$ $\mathrm{pen}(i, k) = \mathbf{1}[i \neq k] \cdot \mathrm{pen}(k)$.[1] One example is an affine penalty where $\mathrm{pen}(k) = a\,\mathrm{pay}(k) + b$ with $a, b \geq 0$.[2] This includes the

[1] A type-independent penalty can admit a weaker notion of auditing—one that can detect inconsistencies between the reported and true types but cannot identify the true type itself.

[2] Here are two real-world examples of affine penalties. China's Export Control authority levies fines between five and ten times the illicit turnover from an unlicensed export, i.e., $\mathrm{pen}(k) = 10\,\mathrm{pay}(k)$ or $\mathrm{pen}(k) = 5\,\mathrm{pay}(k)$. Virginia requires an evading driver to pay the unpaid toll and an administrative fee of up to $100,

formulation in (Estornell, Das, and Vorobeychik 2021) as a special case when $a = 1$.

Without loss of generality, we order the indices so that $0 < \text{pay}(k) < \text{pay}(l)$ for all $k < l$ and $\text{pay}(-1) := 0$. Additionally, we assume misreporting higher can only decrease the value to the principal,

$$\text{val}(i,k) \geq \text{val}(i,l) \text{ for all } i \leq k \leq l, \tag{1}$$

and an agent that knows it will be audited would have no incentive to misreport:

$$\text{pen}(k) \geq \text{pay}(k) \text{ for all } k. \tag{2}$$

**Agents' Utilities and Strategies** Given an audit vector $\boldsymbol{p}$, a type $i$ agent reporting $k$ has expected utility

$$U_{i,k}(\boldsymbol{p}) := \text{pay}(k) - p_k \,\text{pen}(i,k) \tag{3}$$

Agents use a randomized *report strategy* represented by a matrix $\boldsymbol{Q}$ where $Q_{i,k}$ is the probability of a type $i$ agent reporting type $k$. The induced *report distribution*[3] is $\hat{\boldsymbol{q}} \in \Delta_m$ where $\hat{q}_k = \sum_i q_i Q_{i,k}$ for all $k$.

**Definition 1.** Given $\boldsymbol{p}$, a report strategy $\boldsymbol{Q}$ is a *Bayes-Nash equilibrium* if for all $i$ and $k, l \in [m]$ with $Q_{i,k} > 0$, $U_{i,k}(\boldsymbol{p}) \geq U_{i,l}(\boldsymbol{p})$. Let $Eqi(\boldsymbol{p})$ be the set of all equilibria.

**Principal's Utility** Given $\boldsymbol{p}$ and $\boldsymbol{Q}$, let $C(\boldsymbol{p}, \boldsymbol{Q}) := n \sum_{i,k} q_i Q_{i,k} p_k$ be the expected number of audit, and the principal's utility without audit costs $V(\boldsymbol{p}, \boldsymbol{Q}) = n \sum_{i,k \in [m]} q_i Q_{i,k} \left(\text{val}(i,k) - \text{pay}(k) + p_k \,\text{pen}(i,k)\right)$ where the final term is the gain from auditing.

We consider the *costly setting* where the principal can audit any number of agents, but incurs a cost $\lambda \geq 0$ per audit. The principal's utility is

$$V_\lambda(\boldsymbol{p}, \boldsymbol{Q}) = V(\boldsymbol{p}, \boldsymbol{Q}) - \lambda C(\boldsymbol{p}, \boldsymbol{Q}) \tag{4}$$

We assume that the cost of audits is less than the penalty

$$\lambda \leq \text{pen}(k) \text{ for all } k. \tag{5}$$

We defer the budgeted setting to the full version.

**Principal Strategies: Non-adaptive and Adaptive** The principal's audit vector may be fixed or adaptively chosen based on agents' reports, $\hat{\boldsymbol{q}}$.

In the *non-adaptive* setting, the principal commits to an audit vector $\boldsymbol{p}$. After observing $\boldsymbol{p}$, the agents collectively choose an equilibrium $\boldsymbol{Q} \in Eqi(\boldsymbol{p})$ that is worst for the principal. In the *adaptive* setting, the principal commits to an *audit strategy* $\pi$, which maps a reported distribution $\hat{\boldsymbol{q}}$ to an audit vector $\boldsymbol{p} = \pi(\hat{\boldsymbol{q}})$. After observing $\pi$, all agents collectively choose a worst equilibrium $\boldsymbol{Q}$ under $\pi$ so that

$$U_{i,k}(\pi(\hat{\boldsymbol{q}})) \geq U_{i,l}(\pi(\hat{\boldsymbol{q}})), \forall i,k,l \in [m] \text{ with } Q_{i,k} > 0 \tag{6}$$

The above is a *Wardrop equilibrium*; a single agent's deviation does not change the report distribution $\hat{\boldsymbol{q}}$. We denote $Eqi(\pi)$ as the set of equilibria among agents under strategy $\pi$, and $V_\lambda(\pi, \boldsymbol{Q}) = V_\lambda(\pi(\hat{\boldsymbol{q}}), \boldsymbol{Q})$ where $\hat{\boldsymbol{q}}$ is the report distribution of $\boldsymbol{Q}$.

i.e., $\text{pen} = \text{pay} + 100$.

[3] As agents are non-atomic, the observed report distribution equals the expectation. In particular, if all are truthful, the report distribution equals $\boldsymbol{q}$.

# 3 Optimal Non-Adaptive Audits with Costs

We study non-adaptive costly audit games with $(n, m, \boldsymbol{q}, \text{val}, \text{pay}, \text{pen})$ and $\lambda$. Most proofs are deferred to the full version.

## 3.1 Optimizing the Principal's Utility

The principal wants to maximize her utility under the worst-case Bayes-Nash equilibrium, defined below

$$V_\lambda(\boldsymbol{p}) := \min_{\boldsymbol{Q} \in Eqi(\boldsymbol{p})} V_\lambda(\boldsymbol{p}, \boldsymbol{Q}). \tag{7}$$

An audit vector $\boldsymbol{p}$ *$\epsilon$-approximates* $\boldsymbol{p}'$ if $V_\lambda(\boldsymbol{p}) \geq V_\lambda(\boldsymbol{p}') - \epsilon$, and $\boldsymbol{p}$ is *$\epsilon$-optimal* if it $\epsilon$-approximates any $\boldsymbol{p}'$, i.e. $V_\lambda(\boldsymbol{p}) \geq \sup_{\boldsymbol{p}'} V_\lambda(\boldsymbol{p}') - \epsilon$.

Theorem 1 shows that there exists an algorithm that computes an $\epsilon$-optimal audit vector in time $O(m^2)$. This runtime is tight, as reading all entries of val already requires $\Omega(m^2)$ times. Moreover, Proposition 1 shows that computing an exactly optimal audit vector is impossible.

**Theorem 1.** *For any small enough $\epsilon > 0$, $(n, m, \boldsymbol{q}, \text{val}, \text{pay}, \text{pen})$ and $\lambda$, Algorithm 1 computes a $2n\epsilon$-optimal audit vector for Eq. (7) in $O(m^4)$ time.*

*Moreover, the time complexity can be improved to $O(m^2)$.*

The idea of Algorithm 1 is to search over a finite set of audit vectors, called critical audit vectors (Definition 2). We also show that any audit vector can be approximated by one from this set.

The remainder of this section is organized as follows. We begin by defining equalized and critical audit vectors and presenting the algorithm. Next, we characterize agents' best responses and equilibrium behavior in Lemma 1, a result that underpins both Theorem 1 and later analyses. We then show that exact optimization in Eq. (7) may be impossible, justifying our approximation approach. Finally, we prove Theorem 1.

Let $\rho_k(u)$ be the probability that a type $k$ report is audited when $u$ is the utility of misreporting.

$$\rho_k(u) = \frac{\text{pay}(k) - u}{\text{pen}(k)} \tag{8}$$

This is a valid probability when $0 \leq u \leq \text{pay}(k)$ by Eq. (2). Note that $\rho_k(u)$ is decreasing in $u$, and $p_k = \rho_k(U_{i,k}(\boldsymbol{p}))$, for all $\boldsymbol{p}$ and $i \neq k$. Hence, $\rho_k$ is a bijection between misreport utility and audit probability of type $k$.

**Definition 2** (Equalized and critical audit vectors)**.** Given $0 < u \leq \max_k \text{pay}(k)$ with $\iota = \min\{i : \text{pay}(i) \geq u\}$, $A \subseteq \{i \in [m] : i \geq \iota\}$, and $0 < \epsilon < u$, we define the *equalized audit vector* $\boldsymbol{p} = \text{equa}(u, A, \epsilon)$ such that for all $k \in [m]$

$$p_k = \begin{cases} 0 & k < \iota, \\ \rho_k(u) & k \in A, \\ \rho_k(u - \varepsilon) & \text{otherwise.} \end{cases}$$

If $\hat{A} = \{\kappa\}$, we write $\text{equa}(u, A, \epsilon)$ as $\text{equa}(u, \kappa, \epsilon)$. Given $0 < \epsilon < \gamma := \frac{1}{2} \min_k(\text{pay}(k) - \text{pay}(k-1))$ and $\iota \leq \kappa$ we define the *critical audit vectors* as $\text{equa}^+(\iota, \kappa, \epsilon) = \text{equa}(\text{pay}(\iota - 1) + \epsilon, \kappa, \epsilon)$ and $\text{equa}^-(\iota, \kappa, \epsilon) = \text{equa}(\text{pay}(\iota) - \epsilon, \kappa, \epsilon)$.

Algorithm 1: SuccinctSearch

**Require:** $\epsilon > 0$, $(n, m, \boldsymbol{q}, \text{val}, \text{pay}, \text{pen})$, and $\lambda \geq 0$
**Ensure:** Audit vector $\boldsymbol{p}^*$

Initialize $V_{\max} \leftarrow -\infty$ and $\boldsymbol{p}^* \leftarrow \mathbf{1}$
**for** $i \in [m]$ **do**
  **for** $k = i$ to $m-1$ **do**
    $\boldsymbol{p}^+ \leftarrow \text{equa}^+(i, k, \epsilon)$ ▷ critical audit vector
    $\boldsymbol{p}^- \leftarrow \text{equa}^-(i, k, \epsilon)$
    **if** $V_{\max} < \textsc{ComputeVal}(\boldsymbol{p}^+)$ **then**
      $\boldsymbol{p}^* \leftarrow \boldsymbol{p}^+$
      $V_{\max} \leftarrow \textsc{ComputeVal}(\boldsymbol{p}^+)$
    **if** $V_{\max} < \textsc{ComputeVal}(\boldsymbol{p}^-)$ **then**
      $\boldsymbol{p}^* \leftarrow \boldsymbol{p}^-$
      $V_{\max} \leftarrow \textsc{ComputeVal}(\boldsymbol{p}^-)$
**return** $\boldsymbol{p}^*$
**function** $\textsc{ComputeVal}(\boldsymbol{p})$
  $\hat{u} \leftarrow \max_k \{\text{pay}(k) - p_k \text{ pen}(k)\}$
  $\hat{A} \leftarrow \arg\max_k \{\text{pay}(k) - p_k \text{ pen}(k)\}$
  $v \leftarrow 0$
  **for** $i \in [m]$ **do**
    $v_i \leftarrow (\text{val}(i, i) - \text{pay}(i) - p_i \lambda)$
    $\hat{v}_i \leftarrow \min_{k \in \hat{A}} \text{val}(i, k) - \text{pay}(k) + p_k(\text{pen}(k) - \lambda)$
    **if** $\text{pay}(i) > \hat{u}$ **then** ▷ truthful
      $v \leftarrow v + q_i v_i$
    **else if** $\text{pay}(i) < \hat{u}$ **then** ▷ misreporting
      $v \leftarrow v + q_i \hat{v}_i$
    **else** ▷ indifferent
      $v \leftarrow v + q_i \min\{v_i, \hat{v}_i\}$
  **return** $v$

Intuitively, an equalized audit vector sets the misreport value of all types in $A$ to $u$, and minimizes audit probabilities for others so that agents either misreport as $A$ or report truthfully. Lemma 2 formalizes this property. Note that because the $\rho_k$ are decreasing, $\text{equa}^+(\iota, \kappa, \epsilon) \geq \text{equa}^-(\iota, \kappa, \epsilon)$ coordinate-wise.

**Characterizing Best Response and Equilibrium** Before proving the theorem, we show that the best response of each agent follows a threshold structure. There exists a minimal truthful type and a misreporting range such that all agents with lower types strictly prefer to misreport as a type within the misreport range, while higher types strictly prefer to report truthfully.

Given $\boldsymbol{p}$, by Eq. (3), we can write the *best-response set* of type-$i$ agents as $A_i(\boldsymbol{p}) = \arg\max_{k \in [m]}(\text{pay}(k) - p_k \text{ pen}(i, k))$. To simplify the notation, we define the misreport utility of reporting $k$ as

$$\hat{U}_k(\boldsymbol{p}) := \text{pay}(k) - p_k \text{ pen}(k)$$

, which is independent of the misreporting agent's type, and the utility of being truthful as $U_k = \text{pay}(k)$. Finally, let $\hat{u}(\boldsymbol{p}) = \max_k \hat{U}_k(\boldsymbol{p})$ be the *highest misreport utility*, $i_{\text{truth}}(\boldsymbol{p}) = \min\{i \in [m] : U_i \geq \hat{u}(\boldsymbol{p})\}$ be the *minimal truthful type* (the lowest type that is willing to be truthful), and *misreporting range* $\hat{A}(\boldsymbol{p}) = \arg\max_k\{\hat{U}_k(\boldsymbol{p})\} \subseteq [m]$ be the set of types that have the highest misreport utility. We will omit $\boldsymbol{p}$ when it is clear in context.

**Lemma 1** (Threshold structure)**.** *Given $\boldsymbol{p}$ with $\hat{u}(\boldsymbol{p})$, $\hat{A}(\boldsymbol{p})$, and $i_{truth}(\boldsymbol{p})$ defined above, $\hat{A}(\boldsymbol{p}) \subseteq \{i \in [m] : i \geq i_{truth}(\boldsymbol{p})\}$ and*

$$A_i(\boldsymbol{p}) = \begin{cases} \{i\} & \text{if } i > i_{truth}, \\ \hat{A} & \text{if } i < i_{truth} \\ \{i_{truth}\} & \text{if } i = i_{truth} \text{ and } U_{i_{truth}} > \hat{u} \\ \hat{A} \cup \{i_{truth}\} & \text{if } i = i_{truth} \text{ and } U_{i_{truth}} = \hat{u} \end{cases}.$$

An audit vector is *strict* if $\hat{u} \notin \{U_i : i \in [m]\}$ so that every agent is either truthful or misreports as $\hat{A}$. Additionally, a report strategy $\boldsymbol{Q}$ is *single-minded* with $\iota \leq \kappa$ if all types $i \geq \iota$ are truthful and all types $i < \iota$ report as $\kappa$.[4]

By Lemma 1, $Eqi(\boldsymbol{p})$ is non-empty and closed, so the minimum in Eq. (7) is well-defined. However, Proposition 1 shows that the maximum of Eq. (7) does not always exist.

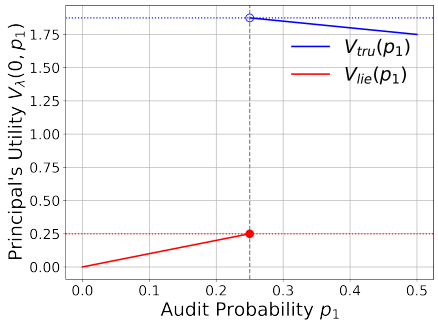



Figure 1: A non-adaptive audit game with unattainable optimum: As in Proposition 1, we consider binary types ($m = 2$) with $q_0 = q_1 = 1/2$, pay $= (1, 2)$, pen $= (3, 4)$, val $= \begin{pmatrix} 3 & 0 \\ 0 & 4 \end{pmatrix}$, and $\lambda = 1$. We vary the audit probability of the high type, using $\boldsymbol{p} = (0, p_1)$ since auditing the low type is useless. The principal's utility if all agents misreport as the high type is $V_{lie}(p_1) = p_1$ (red), and if all are truthful is $V_{tru}(p_1) = \frac{4-p_1}{2}$ (blue). When $p_1 < \frac{1}{4}$, misreporting as the highest type is the unique equilibrium, for $p_1 > \frac{1}{4}$, truth-telling is the unique equilibrium, and at the threshold $p_1 = \frac{1}{4}$, any mixture is an equilibrium. Therefore, the principal's worst-case equilibrium utility at $p_1 = \frac{1}{4}$ is $\frac{1}{4}$, but $\sup_{\boldsymbol{p}} V_\lambda(\boldsymbol{p}) = \lim_{p_1 \to 1/4^+} V_\lambda(0, p_1) = \frac{15}{8}$ which is not attained by any $\boldsymbol{p}$.

**Proposition 1.** *There exists a nonadaptive costly audit game so that $\sup_{\boldsymbol{p} \in [0,1]^m} V_\lambda(\boldsymbol{p}) < +\infty$ but for all $\boldsymbol{p}$, $V_\lambda(\boldsymbol{p}) < \sup_{\boldsymbol{p}' \in [0,1]^m} V_\lambda(\boldsymbol{p}')$.*

Lemma 2 establishes that the equalized audit vector is well-defined and corresponds to the $u$, $A$, and $\iota$ in Lemma 1.

[4] If $\iota = 0$ everyone is truthful.

**Lemma 2.** *Given* $\boldsymbol{p} = \text{equa}(u, A, \epsilon)$ *with* $\iota = \min\{i : \text{pay}(i) \geq u\}$*, if* $u \notin \{U_i : i \in [m]\}$*,* $\boldsymbol{p}$ *is a strict audit vector,* $\hat{u}(\boldsymbol{p}) = u$*,* $\hat{A}(\boldsymbol{p}) = A$*, and* $i_{truth}(\boldsymbol{p}) = \iota$*.*

$$A_i(\boldsymbol{p}) = \begin{cases} \{i\} & \text{if } i \geq i_{truth}, \\ A & \text{if } i < i_{truth} \end{cases}.$$

**Approximation by Equalized and Critical Audit Vectors** We now show that any audit vector can be approximated by some strict equalized vector (Lemma 3), and some critical vector (Lemma 4).

**Lemma 3.** *For any* $\boldsymbol{p} \in [0,1]^m$*, there is a strict equalized audit vector* $\boldsymbol{p}' = \text{equa}(u, \kappa, \epsilon)$ *with* $u > 0, \kappa \in [m]$*, and* $0 < \epsilon < u$ *so that* $Eqi(\boldsymbol{p}') \subseteq Eqi(\boldsymbol{p})$ *and for all* $\boldsymbol{Q} \in Eqi(\boldsymbol{p}')$

$$V_\lambda(\boldsymbol{p}, \boldsymbol{Q}) \leq V_\lambda(\boldsymbol{p}', \boldsymbol{Q}) + n\epsilon.$$

To prove Lemma 3, we use Lemma 2 to find a a strict equalized audit vector $\boldsymbol{p}'$ so that $Eqi(\boldsymbol{p}') = \{\boldsymbol{Q}\} \subseteq Eqi(\boldsymbol{p})$ consists of a single-minded equilibrium with $\iota$ and $\kappa$. Then we upper bound the difference of the principal's utilities with these two audit vectors, $V_\lambda(\boldsymbol{p}, \boldsymbol{Q}) - V_\lambda(\boldsymbol{p}', \boldsymbol{Q})$ to get

$$n \sum_{i,k \in [m]} q_i Q_{i,k} \left(\text{pen}(i,k) - \lambda\right)(p_k - p'_k). \tag{9}$$

To minimize Eq. (9), consider two cases. If $k \neq \kappa$, the smaller $p'_k$ yields a larger utility (which is the intuition of Eq. (8)). For type $\kappa$, the equalized audit vector ensures $p'_\kappa \approx p_\kappa$.

**Lemma 4.** *For any* $\epsilon'$ *and* $\boldsymbol{p} = \text{equa}(u, \kappa, \epsilon)$ *with* $\iota = i_{truth}(\boldsymbol{p})$ *and* $\epsilon, \epsilon' < \gamma$*, there are* $\boldsymbol{p}^- = \text{equa}(U_\iota - \epsilon', \kappa, \epsilon)$ *and* $\boldsymbol{p}^+ = \text{equa}(U_{\iota-1} + \epsilon', \kappa, \epsilon)$ *so that* $Eqi(\boldsymbol{p}^+) = Eqi(\boldsymbol{p}^-) = Eqi(\boldsymbol{p})$ *and* $V_\lambda(\boldsymbol{p}, \boldsymbol{Q}) \leq \max\{V_\lambda(\boldsymbol{p}^+, \boldsymbol{Q}), V_\lambda(\boldsymbol{p}^-, \boldsymbol{Q})\} + n\epsilon'$ *for all* $\boldsymbol{Q} \in Eqi(\boldsymbol{p})$*.*

To prove Lemma 4, we note that fixing $\kappa \in [m]$ and $\epsilon > 0$, an equalized audit vector $\boldsymbol{p} = \text{equa}(\kappa, u, \epsilon)$ is parameterized by a single parameter $u \in \mathbb{R}$. Moreover, the audit probabilities are affine in $u$, so the principal's utility is affine in $u$ by Eq. (9). Therefore, we can optimize the principal's utility using the extreme value of $u \in (U_{\iota-1}, U_\iota)$ by Lemma 2.

**Proof of Theorem 1** Algorithm 1 iterates over critical audit vectors with all combinations of $i \leq k$ and computes the principal's worst-case utility. By Lemma 1, `ComputeVal` computes $V_\lambda(\boldsymbol{p})$ by considering whether a type is truthful or misreporting as $\hat{A}$. Therefore, Algorithm 1 returns the optimal critical audit vector.

For the approximation guarantee, given any audit vector $\boldsymbol{p} \in [0,1]^m$, by Lemma 3, there exists a strict equalized audit vector $\boldsymbol{p}' = \text{equa}(u, \kappa, \epsilon)$ with $\iota = i_{\text{truth}}(\boldsymbol{p}')$ so that $V_\lambda(\boldsymbol{p}) \leq V_\lambda(\boldsymbol{p}') + n\epsilon$. By Lemma 4, there exists a critical audit vector $\boldsymbol{p}'' = \text{equa}^+(\iota, \kappa, \epsilon)$ or $\text{equa}^-(\iota, \kappa, \epsilon)$ with $\iota \leq \kappa$ so that $V_\lambda(\boldsymbol{p}') \leq V_\lambda(\boldsymbol{p}'') + n\epsilon$. Therefore, there exists some critical audit vector

$$V_\lambda(\boldsymbol{p}) \leq V_\lambda(\boldsymbol{p}'') + 2n\epsilon,$$

and the algorithm is $2n\epsilon$-optimal.

The algorithm searches through all $2(\binom{m}{2} + m) = O(m^2)$ critical vectors. By Lemma 1, `ComputeVal`computes $V_\lambda(\boldsymbol{p})$ by computing the worst report in $A_i(\boldsymbol{p})$ that minimizes the principal's utility for all $i$. This takes $O(m^2)$ for each audit vector. Therefore, the time complexity is in $O(m^4)$. We can improve the running time of Algorithm 1 to $O(m^2)$ using dynamic programming for `ComputeVal`.

## 3.2 No-Regret Auditing Without a Prior

One limitation of Algorithm 1 is assuming access to the prior $\boldsymbol{q}$. We provide a no-regret online learning algorithm when the prior $\boldsymbol{q}$ is unknown and can vary in each round.

Let $V_\lambda(\boldsymbol{p}, \boldsymbol{Q}; \boldsymbol{q})$ be the principal's (single-round) utility from Eq. (4) under prior $\boldsymbol{q}$, and $V_\lambda(\boldsymbol{p}; \boldsymbol{q}) := \min_{\boldsymbol{Q} \in Eqi(\boldsymbol{p})} V_\lambda(\boldsymbol{p}, \boldsymbol{Q}; \boldsymbol{q})$.

Consider the principal and agents interacting over $T$ rounds. The principal knows $(n, m, \text{val}, \text{pay}, \text{pen})$ and $\lambda$ while Nature secretly chooses $\vec{\boldsymbol{q}} := (\boldsymbol{q}^0, \ldots, \boldsymbol{q}^{T-1})$. For round $t = 0, \ldots, T-1$,

1. The principal with algorithm $\mathcal{A}$ samples an audit vector $\boldsymbol{p}^t$ from a distribution $P^t$ based on the history $(\boldsymbol{p}^0, v^0, \ldots, \boldsymbol{p}^{t-1}, v^{t-1})$.
2. After observing $\boldsymbol{p}^t$, agents collectively choose the *worst equilibrium* $\boldsymbol{Q}^t \in \arg\min_{\boldsymbol{Q} \in Eqi(\boldsymbol{p}^t)} V_\lambda(\boldsymbol{p}^t, \boldsymbol{Q}; \boldsymbol{q}^t)$
3. The principal gets $v^t = V_\lambda(\boldsymbol{p}^t, \boldsymbol{Q}^t; \boldsymbol{q}^t) = V_\lambda(\boldsymbol{p}^t; \boldsymbol{q}^t)$.

The principal designs an online learning algorithm $\mathcal{A}$ that maximizes her accumulative expected utility. Formally, the algorithm is evaluated by its (multi-agent) *Stackelberg regret* (Dong et al. 2018; Chen, Liu, and Podimata 2020)[5] against the optimal audit vector in hindsight which knows agents' prior $\vec{\boldsymbol{q}}$. We define

$$Reg_T(\mathcal{A}, \vec{\boldsymbol{q}}) = \sup_{\boldsymbol{p}} \sum_{t \in [T]} V_\lambda(\boldsymbol{p}; \boldsymbol{q}^t) - \mathbb{E}_{\mathcal{A}} \left[ \sum_{t \in [T]} V_\lambda(\boldsymbol{p}^t; \boldsymbol{q}^t) \right], \tag{10}$$

and $Reg_T(\mathcal{A}) = \sup_{\vec{\boldsymbol{q}}} Reg_T(\mathcal{A}, \vec{\boldsymbol{q}})$ where the randomness is over the choice of audit vectors.

**Theorem 2.** *Given any* $(n, m, \text{val}, \text{pay}, \text{pen})$ *and* $\lambda$*, the online learning algorithm* $\mathcal{A}$ *in Algorithm 2 has* $Reg_T(\mathcal{A}) = O(n\sqrt{Tm^2 \log m})$*.*

The key observation is that the equalized and critical audit vectors in Definition 2 are independent of prior $\boldsymbol{q}^t$. Additionally, we can reuse Lemmas 3 and 4 to show that the critical vectors are approximately optimal as in Lemma 5.

**Lemma 5.** *Given any* $0 < \epsilon < \gamma$*, there exist* $\boldsymbol{p}^+ = \text{equa}^+(\iota, \kappa, \epsilon)$ *or* $\boldsymbol{p}^- = \text{equa}^-(\iota, \kappa, \epsilon)$ *with* $\iota \leq \kappa$ *so that for all* $\vec{\boldsymbol{q}} = (\boldsymbol{q}^0, \ldots, \boldsymbol{q}^{T-1})$ *and* $\boldsymbol{p}$*,*

$$V_\lambda(\boldsymbol{p}; \vec{\boldsymbol{q}}) \leq \max\{V_\lambda(\boldsymbol{p}^+; \vec{\boldsymbol{q}}), V_\lambda(\boldsymbol{p}^-; \vec{\boldsymbol{q}})\} + 2n\epsilon T$$

*where* $V_\lambda(\boldsymbol{p}; \vec{\boldsymbol{q}}) := \sum_t V_\lambda(\boldsymbol{p}; \boldsymbol{q}^t)$*.*

With Lemma 5, given $\epsilon > 0$ we run a no regret algorithm for adversarial bandits (e.g., EXP3) on $O(m^2)$ critical audit vectors in order to achieve regret bounded by

[5] Classical online-Stackelberg work assumes a single agent (follower) who best-responds to the leader's action. In our model the follower is a population of $n$ agents who play the worst equilibrium under the $n$-player game induced by the audit vector.

Algorithm 2: EXP3 algorithm on critical audit vectors

**Require:** Game parameters $(n, m, \text{val}, \text{pay}, \text{pen})$ with $L = n \max_{i,k} (\text{val}(i,k) + \text{pay}(k) + \text{pen}(k))$, cost $\lambda \geq 0$, horizon $T$, and learning rate $\eta = \sqrt{\frac{\log(2m^2)}{2m^2 T}}$

Initialize $\epsilon_0 \leftarrow \frac{2}{3}\gamma$ and $s_\sigma^0 \leftarrow 0$ for all $\sigma \in \Sigma$.
**for** $t = 1$ to $T$ **do**
  Compute $P_t$ with $P_{t,\sigma} \propto \exp(\eta s_\sigma^t)$ for all $\sigma$
  Sample $\sigma^t \sim P_t$ and set $\boldsymbol{p}^t = \text{equa}(\sigma^t, \epsilon_t)$
  Observe reward $v^t = V_\lambda(\boldsymbol{p}^t; \boldsymbol{q}^t)$,
  Update $\epsilon_{t+1} \leftarrow \frac{1}{2}\epsilon_t$ and for all $\sigma$

$$s_\sigma^{t+1} \leftarrow s_\sigma^t + 1 - \frac{L - v^t}{2L} \cdot \frac{\mathbb{I}[\sigma = \sigma^t]}{P_{t,\sigma}}.$$

$O(\sqrt{Tm^2 \log m} + n\epsilon T)$. However, to achieve no-regret, Algorithm 2 considers the set of critical vectors in Definition 2 with $\epsilon_t = 2^{-t}\epsilon_0$. Specifically, we consider the set of all critical audit vectors $\sigma \in \Sigma := \{(i,k,+), (i,k,-) : i,k \in [m]\}$, and at round $t$ we use $\text{equa}(\sigma, \epsilon_t) = \text{equa}^+(i,k,\epsilon_t)$ if $\sigma = (i,k,+)$ and $\text{equa}^-(i,k,\epsilon_t)$ if $\sigma = (i,k,-)$ as the set of arms.[6]

### 3.3 Optimizing Social Welfare

Now we show how to maximize social welfare, the sum of utility between the principal and all agents,

$$\begin{aligned} W_\lambda(\boldsymbol{p}, \boldsymbol{Q}) :=& V_\lambda(\boldsymbol{p}, \boldsymbol{Q}) + \sum_{i,k \in [m]} q_i Q_{i,k} U_{i,k}(\boldsymbol{p}) \\ =& n \sum_{i,k \in [m]} q_i Q_{i,k} \left(\text{val}(i,k) - p_k \lambda\right). \end{aligned} \tag{11}$$

For instance, if $\text{val}(i,k) = \mathbf{1}[i = k]$, the social welfare is the number of truthful agents minus the cost of audits.

As agents are strategic, we need to design an audit vector $\boldsymbol{p}$ so that $W_\lambda(\boldsymbol{p}) := \min_{\boldsymbol{Q} \in Eqi(\boldsymbol{p})} W_\lambda(\boldsymbol{p}, \boldsymbol{Q})$ is large, and say $\boldsymbol{p}$ is $\epsilon$-optimal if $W_\lambda(\boldsymbol{p}) \geq W_\lambda(\boldsymbol{p}') - \epsilon$ for all $\boldsymbol{p}'$.

**Theorem 3.** *There is an algorithm that computes a $2n\epsilon$-optimal audit vector for Eq.* (11) *in time $O(m^2)$ for any $\epsilon > 0$ and nonadaptive audit game with $(n, m, \boldsymbol{q}, \text{val}, \text{pay}, \text{pen})$ and cost $\lambda$.*

The algorithm is nearly identical to Algorithm 1. Since the agent's best-response still follows from Lemma 1, we can reuse Lemmas 3 and 4 and search through all critical audit vectors as in Algorithm 1 and return the one that maximizes the worst-case social welfare. We omit the proof. Similarly, we can adopt Algorithm 2 to have a no-regret algorithm for social welfare maximization.

### 3.4 Monotonicity in Penalty and Audit Cost

Besides designing the audit vector, the principal may also adjust the penalty function or face a different audit cost $\lambda$. We show that increasing the penalty or decreasing the audit cost $\lambda$ can only improve the principal's utility and social welfare.

Let $V_\lambda(\boldsymbol{p}; \text{pen})$ and $W_\lambda(\boldsymbol{p}; \text{pen})$ be the principal's worst case utility (Eq. (7)) and worst-case social welfare respectively under penalty function pen and cost $\lambda$.

**Proposition 2.** *If $\lambda \geq \lambda'$ and $\text{pen}(k) \leq \text{pen}'(k)$ for all $k \in [m]$, for any $\boldsymbol{p}$ there exists $\boldsymbol{p}'$ so that $V_\lambda(\boldsymbol{p}; \text{pen}) \leq V_{\lambda'}(\boldsymbol{p}'; \text{pen}')$ and $W_\lambda(\boldsymbol{p}; \text{pen}) \leq W_{\lambda'}(\boldsymbol{p}', \text{pen}')$.*

The idea of Proposition 2 is that if the penalty increases, we can decrease the audit probability $p'_k = \frac{\text{pen}(k)}{\text{pen}'(k)} p_k$, which preserves the same equilibria and expected penalty gain, but lowers the audit cost.

## 4 Optimal Adaptive Audits With Costs

We now explore adaptive audit games, where the principal's strategy depends on both the agents' prior distribution $\boldsymbol{q}$ and report distribution $\hat{\boldsymbol{q}}$. We discuss the costly setting and defer the budgeted setting to the full version.

In this section, we assume that the penalty is less sensitive than the payment: for all $k \leq l$ in $[m]$

$$\frac{\text{pay}(l)}{\text{pay}(k)} \geq \frac{\text{pen}(l)}{\text{pen}(k)}. \tag{12}$$

In particular, any positive affine function $\text{pen}(k) = a\,\text{pay}(k) + b$ with $a, b \geq 0$ for all $k$ satisfies Eq. (12).

As multiple equilibria may exist, the principal optimizes for the worst-case utility by solving the following optimization problem:

$$\sup_{\pi : \Delta_m \to [0,1]^m} \min_{\boldsymbol{Q} \in Eqi(\pi)} V_\lambda(\pi, \boldsymbol{Q}). \tag{13}$$

We define $V_\lambda(\pi) = \min_{\boldsymbol{Q} \in Eqi(\pi)} V_\lambda(\pi, \boldsymbol{Q})$ as the principal's worst case utility and set to $-\infty$ if $Eqi(\pi) = \emptyset$ following the standard convention in pessimistic Stackelberg games. (Coniglio, Gatti, and Marchesi 2017) We say that $\pi$ *$\epsilon$-approximates* $\pi'$ if $V_\lambda(\pi) \geq V_\lambda(\pi') - \epsilon$, and $\pi$ is *$\epsilon$-optimal* if it $\epsilon$-approximates any $\pi'$.

**Theorem 4.** *There is an algorithm that computes an $\epsilon$-optimal audit vector for Eq.* (13) *with Eq.* (4) *in $O(m^2)$ time for all small enough $\epsilon > 0$ and adaptive audit game with cost $\lambda \geq 0$ and parameters $(n, m, \boldsymbol{q}, \text{val}, \text{pay}, \text{pen})$ satisfying Eq.* (12)*. Moreover,*

$$\sup_{\pi} \min_{\boldsymbol{Q} \in Eqi(\pi)} V_\lambda(\pi, \boldsymbol{Q}) = \sup_{\boldsymbol{p} \in [0,1]^m, \boldsymbol{Q} \in Eqi(\boldsymbol{p})} V_\lambda(\boldsymbol{p}, \boldsymbol{Q}).$$

**Proof Sketch.** To prove Theorem 4, we use three key observations. First, due to Lemma 1 equilibria depend only on the output vector $\boldsymbol{p}$. Adaptive strategies cannot yield new equilibria beyond those already attainable by some fixed $\boldsymbol{p}$. Consequently,

$$V_\lambda(\pi) \leq \sup_{\boldsymbol{p}, \boldsymbol{Q} \in Eqi(\boldsymbol{p})} V_\lambda(\boldsymbol{p}, \boldsymbol{Q}) \tag{14}$$

and we will show that Eq. (14) holds with equality.

[6] We treat each tuple $(i,k,+)$ or $(i,k,-)$ as a template arm. EXP3 maintains weights over these templates, while the audit vector played in round $t$ depends on the template and $\epsilon_t$

Second, let a *dictator audit strategy* with $\boldsymbol{p}^*$ and $\hat{\boldsymbol{q}}^* \in \Delta_m$

$$\pi_{dict}(\hat{\boldsymbol{q}}) = \begin{cases} \boldsymbol{p}^* & \text{if } \hat{\boldsymbol{q}} = \hat{\boldsymbol{q}}^* \\ \mathbf{1} & \text{if } \hat{\boldsymbol{q}} \neq \boldsymbol{q} \text{ and } \hat{\boldsymbol{q}} \neq \hat{\boldsymbol{q}}^* \\ \mathbf{0} & \text{if } \hat{\boldsymbol{q}} = \boldsymbol{q} \text{ and } \hat{\boldsymbol{q}} \neq \hat{\boldsymbol{q}}^* \end{cases}. \quad (15)$$

Intuitively, if the observed reports differ from $\boldsymbol{q}^*$, we either audit everyone (making any misreporting strictly unprofitable) or audit no one (agents strictly prefer to over-report as the highest type). Lemma 6 shows that a dictator audit strategy can eliminate any report strategy with $\hat{\boldsymbol{q}} \neq \hat{\boldsymbol{q}}^*$, while ensuring the existence of an equilibrium with $\hat{\boldsymbol{q}} = \hat{\boldsymbol{q}}^*$ by choosing $\boldsymbol{p}^*$ appropriately.

**Lemma 6** (Dictator strategies)**.** *For any dictator audit strategy $\pi_{dict}$ in Eq.* (15) *with $\boldsymbol{p}^*$ and $\hat{\boldsymbol{q}}^*$, $Eqi(\pi_{dict}) = \{\boldsymbol{Q} \in Eqi(\boldsymbol{p}^*) : \hat{\boldsymbol{q}} = \hat{\boldsymbol{q}}^*\}$.*

Finally, Lemma 7 shows that for any audit vector $\boldsymbol{p}$, the best equilibrium can be single-minded. Therefore, it is sufficient to iterate all single-minded strategies $\boldsymbol{Q}$ and search the optimal audit vector $\boldsymbol{p}$ with $\boldsymbol{Q} \in Eqi(\boldsymbol{p})$. Moreover, by a similar argument as in Lemma 4, we show that the optimal audit vector is critical. This reduces the search to $O(m^2)$ candidates, yielding the claimed $O(m^2)$ running time.

**Lemma 7.** *For any audit vector $\boldsymbol{p}$, if Eq.* (12) *holds, there exists a single-minded equilibrium $\boldsymbol{Q}' \in Eqi(\boldsymbol{p})$ so that for all $\boldsymbol{Q} \in Eqi(\boldsymbol{p})$ $V_\lambda(\boldsymbol{p}, \boldsymbol{Q}) \leq V_\lambda(\boldsymbol{p}, \boldsymbol{Q}')$.*

**Remark 1.** Note that the argument to prove Lemma 7 also applies to social welfare, so Theorem 4 also holds for optimizing social welfare. Additionally, by Lemma 7, if Eq. (12) holds, Algorithm 1 also finds an approximately optimal audit vector in the non-adaptive setting, and the worst-case utility coincides with the best-case utility

$$\sup_{\boldsymbol{p}} \min_{\boldsymbol{Q} \in Eqi(\boldsymbol{p})} V_\lambda(\boldsymbol{p}, \boldsymbol{Q}) = \sup_{\boldsymbol{p}} \max_{\boldsymbol{Q} \in Eqi(\boldsymbol{p})} V_\lambda(\boldsymbol{p}, \boldsymbol{Q}).$$

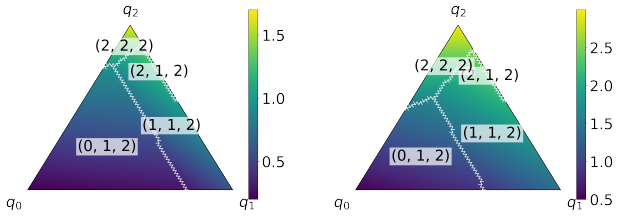
Conversely, if Eq. (12) is not satisfied, the optimal equilibrium may not be single-minded, and this equivalence no longer applies.

## 5 Simulations

Thus far, we have analyzed the optimal audit policy theoretically. We now provide simple simulations to illustrate how the optimal policy and the resulting equilibria depend on the prior and the payment function in small three-type examples.

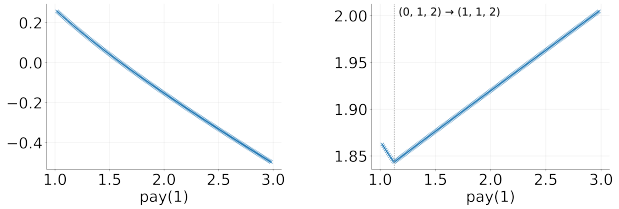
Figure 2 illustrates the effect of the prior $\boldsymbol{q}$. In the lower-left corner, most agents have the lowest type (type 0), which admits the truthful equilibrium $(0, 1, 2)$. At the top corner, most agents have type 2, and it becomes preferable to allow everyone to report the highest type $(2, 2, 2)$ rather than impose huge audit costs to enforce truth-telling. Similarly, in the lower-right corner, it is optimal to allow type 0 to misreport as type 1. Finally, we note that the principal-optimal policy in Fig. 2a is stricter than the welfare-optimal one in Fig. 2b, and yields a larger truth-telling region. This is because misreports impose greater costs on the principal than on overall welfare.

Figure 3 shows effect of the payment function is non-monotone when all other parameters are fixed. In Fig. 3a,




(a) Principal's utility (b) Social Welfare

Figure 2: Effect of prior $\boldsymbol{q}$: There are three types $m = 3$ with $n = 1$, $\text{val} = \text{diag}(0.5, 1.4, 3.0)$, $\text{pay} = (0.3, 0.8, 1.3)$, $\text{pen} = (1.0, 1.2, 1.4)$, and $\lambda = 0.7$. Each point corresponds to a prior vector $\boldsymbol{q} = (q_0, q_1, q_2)$, and the color encodes the principal's optimal utility by Theorem 1 with $\epsilon = 10^{-3}$ in Fig. 2a, and the optimal social welfare by Theorem 3 in Fig. 2b. We also indicate the region of the worst equilibrium.




(a) Principal's utility (b) Social Welfare

Figure 3: Effect of pay: There are three types $m = 3$ and change the payment of type 1 with the following parameters $n = 1$, $\boldsymbol{q} = (0.4, 0.3, 0.3)$, $\text{val} = \begin{pmatrix} 0.99 & 0.90 & 0.50 \\ 0 & 1.50 & 1.40 \\ 0 & 0 & 4.00 \end{pmatrix}$, $\text{pay}(0) = 1, \text{pay}(2) = 3$, $\text{pen} = \text{pay} + 0.5$, and $\lambda = 1$.

the worst equilibrium is always truth-telling, and increasing the payments monotonically decreases the principal's utility. In contrast, in Fig. 3b, when the type-1 payment is small, the equilibrium is still truth-telling and welfare decreases. However, for large type-1 payment, type 0 agents begin to misreport as type 1, and increasing $\text{pay}(1)$ reduces audit probability $p_2$ and increases welfare.

## 6 Discussion and Future Work

We provide several optimal and efficient audit policies for utility- and welfare-maximizing under pessimistic equilibrium selection. At the same time, extending our model suggests fruitful directions for future work. First, extending our guarantees to finite agents, noisy or partial verification, and richer penalty structures remains open. Second, we take the classifier or allocation rule as exogenous; jointly designing the predictive model and the audit policy could yield better performance. Finally, it would be interesting to extend the incentive-minimization framework of Estornell, Chen, and Das (2023) to non-binary payment outcomes.


## Acknowledgments

SD is grateful for support from NSF Award 2533162.

## References


Alm, J.; Bahl, R.; and Murray, M. N. 1993. Audit selection and income tax underreporting in the tax compliance game. *Journal of development Economics*, 42(1): 1–33.

Ben abdelaziz, F.; Neifar, S.; and de Bourmont, M. 2015. Auditing and game theory: A survey. In *Multiple Criteria Decision Making in Finance, Insurance and Investment*, 249–272. Springer.

Ben-Porath, E.; Dekel, E.; and Lipman, B. L. 2014. Optimal allocation with costly verification. *American Economic Review*, 104(12): 3779–3813.

Blocki, J.; Christin, N.; Datta, A.; Procaccia, A. D.; and Sinha, A. 2013. Audit Games. In Rossi, F., ed., *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, 41–47. IJCAI/AAAI.

Blocki, J.; Christin, N.; Datta, A.; Procaccia, A. D.; and Sinha, A. 2015. Audit Games with Multiple Defender Resources. In Bonet, B.; and Koenig, S., eds., *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, 791–797. AAAI Press.

Chen, Y.; Liu, Y.; and Podimata, C. 2020. Learning Strategy-Aware Linear Classifiers. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Coates, C. J.; Florence, R. E.; and Kral, K. L. 2002. Financial statement audits, a game of chicken? *Journal of Business Ethics*, 41(1): 1–11.

Cole, R.; Dodis, Y.; and Roughgarden, T. 2003. Pricing network edges for heterogeneous selfish users. In *Proceedings of the Thirty-Fifth Annual ACM Symposium on Theory of Computing*, STOC '03, 521–530. New York, NY, USA: Association for Computing Machinery. ISBN 1581136749.

Coniglio, S.; Gatti, N.; and Marchesi, A. 2017. Pessimistic Leader-Follower Equilibria with Multiple Followers. In Sierra, C., ed., *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, 171–177. ijcai.org.

Conitzer, V.; and Sandholm, T. 2006. Computing the optimal strategy to commit to. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, EC '06, 82–90. New York, NY, USA: Association for Computing Machinery. ISBN 1595932364.

Dong, J.; Roth, A.; Schutzman, Z.; Waggoner, B.; and Wu, Z. S. 2018. Strategic Classification from Revealed Preferences. In Tardos, É.; Elkind, E.; and Vohra, R., eds., *Proceedings of the 2018 ACM Conference on Economics and Computation, Ithaca, NY, USA, June 18-22, 2018*, 55–70. ACM.

Estornell, A.; Chen, Y.; and Das, S. 2023. Incentivizing Recourse through Auditing in Strategic Classification. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI)*, 400–408.

Estornell, A.; Das, S.; and Vorobeychik, Y. 2021. Incentivizing Truthfulness Through Audits in Strategic Classification. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 5347–5354. AAAI Press.

Eubanks, V. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY: St. Martin's Press. ISBN 1250074312.

Jalota, D.; Tsao, M.; and Pavone, M. 2024. Catch Me If You Can: Combatting Fraud in Artificial Currency-Based Government Benefits Programs. arXiv:2402.16162.

Lundy, T.; Wei, A.; Fu, H.; Kominers, S. D.; and Leyton-Brown, K. 2019. Allocation for Social Good: Auditing Mechanisms for Utility Maximization. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, EC '19, 785–803. New York, NY, USA: Association for Computing Machinery. ISBN 9781450367929.

Mookherjee, D.; and Png, I. 1989. Optimal auditing, insurance, and redistribution. *The Quarterly Journal of Economics*, 104(2): 399–415.

Personnat, G.; Lin, T.; Hossain, S.; and Parkes, D. C. 2025. Learning to Play Multi-Follower Bayesian Stackelberg Games. arXiv:2510.01387.

Pita, J.; Jain, M.; Marecki, J.; Ordóñez, F.; Portway, C.; Tambe, M.; Western, C.; Paruchuri, P.; and Kraus, S. 2008. Deployed ARMOR protection: the application of a game theoretic model for security at the Los Angeles International Airport. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems: Industrial Track*, AAMAS '08, 125–132. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Roughgarden, T.; and Tardos, E. 2002. How bad is selfish routing? *J. ACM*, 49(2): 236–259.

Tambe, M. 2012. *Security and Game Theory - Algorithms, Deployed Systems, Lessons Learned*. Cambridge University Press. ISBN 978-1-10-709642-4.

Townsend, R. M. 1979. Optimal contracts and competitive markets with costly state verification. *Journal of Economic theory*, 21(2): 265–293.