Optimal Allowance with Limited Auditing Capacity

Thomas Brzustowski and Albin Erlanson*

November 21, 2024

PLEASE CLICK HERE FOR THE LATEST VERSION

Abstract

We analyze the mechanism-design problem of a principal allocating amounts of a perfectly divisible good to n agents, each of whom desires as much of the good as possible. The principal has an ideal allocation for each agent, which is private information held by that agent. The principal has access to an auditing technology that allows her to perfectly uncover the private information of any k ($< n$) of the agents. We present a tractable approach to solve for the principal's optimal mechanism, which combines targeted random audits with allocative distortions to ensure compliance. Agents whose reported type falls above a cutoff enter a pool for random audits. The allocation to audited agents coincides with the principal's ideal. For unaudited agents, upward distortions reward conservative reporting, while downward distortions discipline over-reporting.

1 Introduction

Across regulatory, fiscal, and organizational settings, auditing serves as a fundamental tool for ensuring compliance, whether to uphold public trust, enforce laws, safeguard financial stability, or maintain internal accountability. Audits are crucial for verifying information, deterring misconduct, and reinforcing regulatory or organizational policies. Yet, across public agencies, tax authorities, and corporations, auditing capacity is often constrained by tight budgets and limited resources. These constraints necessitate selective auditing strategies, raising the question of how best to deploy limited audits to encourage truthful reporting and optimal resource use.

This paper analyzes the mechanism-design problem of a principal who must allocate resources to agents with private information under a fixed auditing capacity. The goal of our analysis is to shed light on how limited auditing capacity can be optimally managed and on the allocative distortions that arise as agents strategically respond to the constraints imposed by limited audits. Our findings underscore the trade-offs inherent in designing effective, resource-constrained compliance policies.

In our model, the principal (she) is tasked with allocating perfectly divisible units of a good to n agents (he/they), each of whom desires as large an allocation as possible. The principal has an ideal

*Brzustowski: Department of Economics, University of Essex, Colchester CO4 3SQ, UK, t.brzustowski@essex.ac.uk.
Erlanson: Department of Economics, University of Essex, Colchester CO4 3SQ, UK, albin.erlanson@essex.ac.uk.
We have benefited from inspiring conversations with Francesco Caselli, Daniele Condorelli, Dan Garrett, Alkis Georgiadis-Harris, Deniz Kattwinkel, Andreas Kleiner, Jan Knoepfle, Ilan Kremer, Francesco Nava, and Balázs Szentes.

target allocation for each agent, and incurs a loss for deviating from this target. The difficulty is that the ideal allocation to an agent is private information held by that agent. In other words, each agent has a privately known type, which corresponds to the principal's bliss allocation for that agent. The principal has access to an auditing technology allowing her to perfectly certify the type of any k of the agents, where $k < n$. Under these constraints, the principal commits to a mechanism that specifies reports that the agents can submit, which agents to audit based on these reports, and how much to allocate to each agent based on the information gathered through reporting and audits.

For concreteness, let us map the model to the following regulatory situation. Consider an agency allocating carbon emissions permits among a number of firms. These permits, often allocated at no cost to firms,[1] represent production flexibility. Thus, each firm prefers a larger allocation to operate with less constraint. The regulator has an objective to curb emissions, but is also concerned with minimizing economic efficiency losses associated with disrupting the firms' production processes. As a consequence, the ideal permit allocation would depend on each firm's unique abatement costs, which are private information held internally. Although audits can reveal these abatement costs, the regulator is limited in the number of audits they can conduct due to budget and staffing constraints.[2]

Our assumption that audits are costless but subject to a hard constraint merits further discussion. Most of the literature on auditing (reviewed below) adopts the *costly state verification* paradigm, where the number of audits that can be conducted is flexible, with each additional audit being costly. In contrast, we study situations where investment in auditing capacity may have been decided ex-ante, but it is prohibitively costly to adjust it ex-post, conditional on the agents' reports. We believe this scenario is perfectly realistic in many environments, for instance in cases where the principal is a bureaucratic entity operating with a fixed number of auditors available on the payroll and limited time to implement an allocation. This approach allows us to focus on the optimal use of a pre-established auditing capacity. Furthermore, as we show in Section 5, the aforementioned distinction between ex-ante and ex-post flexibility disappears as the number of agents grows large, with the costly state verification problem nested as a limit case of our model. In other words, our fixed-auditing-capacity assumption is more general and allows for a precise assessment of the benefits of ex-post flexibility.[3]

We solve for the principal's optimal mechanism, which effectively determines an auditing strategy and a resulting allocation rule. Let us now describe its main features. The optimal mechanism adopts

[1] For instance, in the EU, home to the largest and longest-operating emissions trading system (ETS), 43 percent of carbon permits are allocated to firms for free (Matthes, Mehling, & Duan, 2021).

[2] This issue is highlighted in implementation guidelines as a crucial concern in practice. Matthes et al. (2021) devote an entire chapter (Chapter 7) to challenges posed by limited auditing resources.

[3] We also discuss the robustness of our results to other alternatives in modeling the cost of auditing at the end of Section 5.

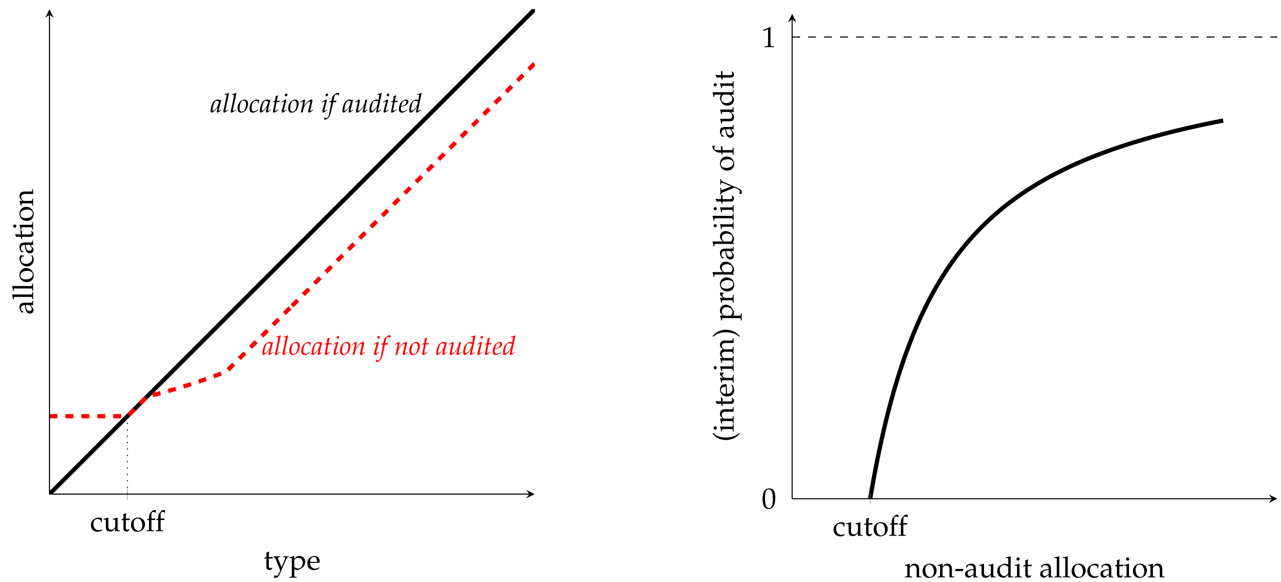
a *targeted random sampling* approach to auditing. Each agent is asked to report his private type, with a pre-specified cutoff in place for each agent. If an agent reports a type below his cutoff, he sidesteps any chance of audit and receives a fixed basic allocation, which is equal to his cutoff. However, agents reporting above their cutoff enter a random audit pool, from which up to k agents are selected for verification.[4] For agents reporting above their cutoff, the allocation depends on whether an audit occurs and, if it does, on the audit's outcome. Unsurprisingly, in order to encourage honest reporting, audited agents who are found to have misrepresented their type are maximally punished with an allocation of zero. Conversely, agents who are audited and verified as truthful receive an allocation exactly equal to their reported type, which, recall, corresponds to the principal's ideal allocation. Finally, agents in the audit pool who are not selected for auditing receive an allocation that is deliberately distorted away from the principal's ideal.

Let us examine more closely some key implications from the structure of the optimal mechanism. First, allocative distortions are specifically used to supplement auditing as a tool to discourage agents from misreporting. In fact, two kinds of distortions arise within the mechanism. Agents whose report falls below their cutoff receive more than their type (an upward distortion), while agents reporting above the cutoff face a downward distortion in case they are not selected for an audit.[5] Notably, it is unnecessary to distort the allocation to an agent who has been audited (and truthful). Perhaps interestingly, this feature can be interpreted as a reward for honesty: from the perspective of a truthful agent, getting selected for an audit is good news as it implies an increase in his allocation. As an illustration, on the left panel (a) of Figure 1, we plot the allocation rule that is part of the principal's optimal mechanism under specific assumptions covered by our model.

Second, in equilibrium, the principal's two instruments — audits and distortions — are optimally combined so that agents may strictly prefer to report their true type but remain indifferent among all possible lies. The intuition is simple. If an agent has a worst possible lie, it must be that either this lie is highly likely to be detected or that the distortion when undetected is severe. Therefore, the principal could (i) economize on scarce auditing resources by lowering the audit probability for that report, or (ii) improve allocative efficiency by reducing the distortion when no audit occurs. The resulting indifference among all lies implies a simple relationship between the interim probability that an agent is audited and the allocation that he would receive if an audit does not occur, illustrated on the right

[4]It is possible that fewer than k agents may report above their cutoff on the equilibrium path. In this event, the principal's auditing capacity is not exhausted, despite the fact that it would be costless to audit up to k agents. The reason is that there is no particular benefit to auditing agents reporting below their cutoff.

[5]Since upward distortions benefit the agent, they do not constitute an efficiency loss relative to the principal's first-best allocation, whereas downward distortions do.



(a) Amount allocated to each type, depending on whether the agent is audited (and truthful) or not. Non-truthful agents are allocated 0, if they are audited.

(b) Interim probability that an agent is audited, as a function of the allocation he would receive if not audited.

Figure 1: Solution to the principal's mechanism-design problem in a special case with $n = 2$ symmetric agents and $k = 1$ audit available. Types are independent and, for this special case, assumed to be uniformly distributed on $[0, 1]$. The principal's payoff for allocating amount z_i to agent $i = 1, 2$ with type t_i is $(z_1 - t_1) - z_1 \ln(z_1/t_1) + (z_2 - t_2) - z_2 \ln(z_2/t_2)$. The optimal mechanism treats agents symmetrically. The cutoff type is ≈ 0.1607.

panel (b) of Figure 1.

In the context of carbon permit allocation mentioned previously, the optimal mechanism derived in this paper suggests the following approach. Each firm has the option to accept a basic permit determined by observable characteristics, which ensures a minimum level of compliance with minimal scrutiny. Firms seeking a larger allocation are free to request any amount, but targeted random audits are such that their likelihood of being audited increases with the size of their request. Audited firms receive efficient allocations if their claims are truthful, while misrepresentation leads to severe penalties. Firms that are not audited, by contrast, receive permits intentionally skewed from their requested amounts, with under-provision reinforcing compliance incentives.

This last feature may appear blatantly unfair at first glance. In effect, firms that report equally honestly face different outcomes due to the random nature of audits, with truthful but unaudited firms receiving less than they request. Nevertheless, the optimal mechanism can be viewed in a different light. Firms' allocations are determined by a graduated system: the more they request, the more they receive, but full alignment with their request can only be granted after an audit. Given the limited auditing capacity, the inclusion of an element of chance reflects a practical necessity. Whether this randomness constitutes unfairness is a matter of debate, as similar approaches, such as randomized selection in research grant allocations, have often been defended as a step toward greater fairness.[6]

Literature Review.— As mentioned previously, the costly state verification framework has been the dominant approach to study auditing, with foundational contributions by Townsend (1979), Gale and Hellwig (1985), Border and Sobel (1987), and Mookherjee and Png (1989). This literature has experienced a revival in mechanism design environments without transfers, following the seminal contribution of Ben-Porath, Dekel, and Lipman (2014). Significant recent contributions to the topic include Ben-Porath, Dekel, and Lipman (2019), Erlanson and Kleiner (2020), Li (2020), Li (2021), Patel and Urgun (2022), and Kattwinkel and Knoepfle (2023). Within this broad literature, our model is most closely related to the problems studied by Harris and Raviv (1996) and Halac and Yared (2020).

Harris and Raviv (1996) apply costly state verification to the problem of budget allocation in organizations. Their environment is a special case of ours with a ternary type space and a particular payoff structure for the principal, both of which are encompassed by our framework.[7] Halac and Yared (2020) generalize the model of Harris and Raviv (1996) in several key directions. However, motivated by applications, they restrict attention to deterministic auditing rules. In our model, randomization in auditing is unavoidable due to the fixed capacity constraint. In the limit when the number of agents grows large, the limited-auditing-capacity and costly-state-verification frameworks coincide, and we obtain the optimal (random) mechanism in a special case of Halac and Yared (2020)'s model. This result allows for a detailed qualitative and quantitative assessment of the value of randomization in their problem, which we delve into in Section 5.

In contrast to the costly state verification framework, some papers have considered alternative assumptions where verification is constrained rather than costly. Notable examples include Glazer and Rubinstein (2004) and Carroll and Egorov (2019), who study environments where a single agent possesses multidimensional private information, and the principal is constrained in the number of dimensions she can audit. While this constraint is similar in spirit to our limited auditing capacity, the strategic structure of the problem is different since, in our setting, each dimension corresponds to the type of a distinct agent. The specific constraint that we impose on auditing is inspired by Erlanson and Kleiner (2024), who study capacity-constrained verification in a setting involving a fixed number of

[6] See Fang and Casadevall (2016).

[7] The assumptions that we impose on the principal's payoff are discussed in detail in Section 2. Regarding the type space, we should emphasize that we allow for any compact Borel set (where the upper bound will be normalized to 1), including in particular sets with three elements.

goods to allocate across multiple agents. Relative to their framework, our model provides the principal with more flexibility in designing agent-specific allocative distortions, clarifying the role of limited auditing capacity in shaping those distortions. Mylovanov and Zapechelnyuk (2017) also investigate a form of constrained verification, where the principal's ability to audit an agent is inherently tied to the implemented allocation ("ex-post verification").

Our paper focuses on auditing, which is related yet distinct from monitoring, where inspections are employed to deter fraudulent behavior rather than to verify private information.[8] Nevertheless, there are meaningful parallels between our analysis and those of Mookherjee and Png (1994) and Bond and Hagerty (2010), who emphasize the joint role of inspection frequency and the severity of penalties in models of deterrence.[9]

2 Model

Environment.— We consider the mechanism-design problem of a principal (she) who must allocate resources to n (≥ 2) agents (he, they). Agents are labeled by $i \in \{1, ..., n\}$. Throughout we use index $-i$ to refer to all agents except i. The set of feasible allocations is $[0, \infty)^n$, with typical element $z = (z_1, ..., z_n)$, where for each i, z_i is the non-negative quantity assigned to agent i. Each agent i has a type t_i ($\in [0, 1]$), which is his private information. Types are drawn independently. We denote F_i the cumulative distribution function of type t_i. Given the realized type profile $t = (t_1, ..., t_n)$ and an allocation z, the payoff to each agent i is z_i, while the principal's payoff is:

$$-\sum_{i=1}^{n} \ell_i(t_i, z_i),$$

where for all i and t_i, $\ell_i(t_i, \cdot) : [0, \infty) \to \mathbb{R}$ is uniquely minimized at t_i. Without loss, we normalize $\ell_i(t_i, t_i) = 0$. Additional technical assumptions on the functions ℓ_i are discussed later.

In words, the principal's problem is to allocate amounts of a good to each of n agents. Agents care only about their own allocation, which we identify with their payoff.[10] Those are a priori unbounded at the top. However, we impose a non-negativity constraint on allocations, reflecting limited liability. The principal has an ideal allocation for each agent, and incurs a loss for deviating from this target. The

[8] Strausz (2006) highlights this distinction in moral hazard environments.

[9] Enforcement resources are fixed in the short run in Bond and Hagerty (2010), analogous to our assumption of a fixed auditing capacity.

[10] Since each agent's payoff coincides with his allocation, one interpretation of the model is that the principal directly allocates utils to each agent. In light of this interpretation, the model can be seen as a reduced-form representation of a variety of settings, e.g. involving richer risk attitudes.

challenge is that the ideal allocation for an agent is unknown to the principal, as it is private information held by the agent.

We endow the principal with limited auditing capacity. Specifically, we assume that the principal can perfectly verify, at no cost, the type of any k ($< n$) of the agents. Observe that, in the absence of additional tools, the principal would have no means to elicit useful information from any of the agents. We should also emphasize that the constraint on auditing is the only reason the agents are linked in this model.

Technical Assumptions.— We introduce four technical assumptions regarding the principal's payoff.

Assumption 1. *For each i and z_i, $\ell_i(\cdot, z_i)$ is F_i-measurable.*

Assumption 1 is a minimal requirement for the principal to be able to evaluate the likelihood of achieving any payoff.

Assumption 2. *For all i and t_i, $\ell_i(t_i, \cdot)$ is convex.*

Assumption 2 is not critical but is maintained throughout for clarity of exposition.[11] It implies that the principal is risk-averse, so she would not seek to implement a random mechanism just because she likes randomness.

A consequence of Assumption 2 is that, for each i and t_i, $\ell_i(t_i, \cdot)$ has well-defined left- and right-derivatives at each $z_i > 0$. Throughout, we write $(\partial \ell_i/\partial z_i)(t_i, z_i)$ to refer to an element of the interval $[(\partial \ell_i/\partial z_i)_-(t_i, z_i), (\partial \ell_i/\partial z_i)_+(t_i, z_i)]$. The following assumption applies to any such selection.

Assumption 3. *For each i and $z_i \in (0, 1)$, $(\partial \ell_i/\partial z_i)(\cdot, z_i)$ is F_i-integrable.*

Note that the assumption is only imposed for $z_i \in (0, 1)$. Since $\ell_i(t_i, z_i)$ is minimized when $z_i = t_i$, $(\partial \ell_i/\partial z_i)(\cdot, z_i)$ does not change sign on $[0, 1]$ when $z_i \notin (0, 1)$. Thus, even if the integral diverges, its sign remains unambiguous.

In order to state our next assumption, we define $\mathcal{D} = \{(t, z) \in [0, 1] \times [0, \infty) : 0 < z < t\}$, the set of type-allocation pairs such that the allocation is below the type but non-zero. Observe that $\mathcal{D}$ is a lattice.

Assumption 4. *For each i, the average loss, defined by $(t_i, z_i) \mapsto \ell_i(t_i, z_i)/z_i$, is submodular on $\mathcal{D}$.*

[11] If one wishes to relax the convexity assumption, they may define $\underline{\ell}_i(t_i, \cdot)$, the biconjugate (convexification) of $\ell_i(t_i, \cdot)$ (Hiriart-Urruty & Lemaréchal, 2001). As long as $\underline{\ell}_i(t_i, \cdot)$ remains uniquely minimized at 0, the whole analysis that follows applies to the model where we replace $\ell_i(t_i, \cdot)$ by $\underline{\ell}_i(t_i, \cdot)$. Then, in order to recover a solution to the original problem, it is sufficient to replace those allocations that fall where $\ell_i(t_i, \cdot)$ and $\underline{\ell}_i(t_i, \cdot)$ do not coincide by lotteries, which are constructed to achieve the convexified payoff.

A large part of our analysis remains valid if Assumption 4 is violated: it is only used to establish monotonicity properties of solutions (see Lemma 3 and Proposition 2). Therefore, we defer our discussion of the interpretation of this Assumption to Section 3, where those results are established.

For the moment, we emphasize that those four assumptions are satisfied by many natural modeling choices to capture the loss associated to misallocation relative to an ideal allocation.

Examples.[12]

(i) For any i, suppose that $\ell_i(t_i, z_i) = c_i(t_i - z_i)$, where c_i is a convex function uniquely minimized at 0 with $c_i(0) = 0$. Then Assumptions 1, 2, 3, and 4 are satisfied. Examples in this class include the standard quadratic loss function with $c_i(x) = x^2$, the usual distance with $c_i(x) = |x|$, or possibly asymmetric losses between upward and downward misallocations, e.g. with $c_i(x) = e^x + (x - 1/2)^2 - 5/4$.

(ii) [Bregman divergences] For any i, suppose that $\ell_i(t_i, z_i) = c_i(z_i) - c_i(t_i) - (z_i - t_i)c_i'(t_i)$, where c_i is a differentiable convex function. Then Assumptions 1, 2, 3, and 4 are satisfied. Examples in this class include the standard quadratic loss function with $c_i(x) = x^2$, or losses related to relative entropy with $c_i(x) = x\ln(x)$, as for the introductory example of Figure 1.

Mechanisms.— Next, we formulate the principal's mechanism-design problem. A standard revelation principle applies and implies that it is without the loss of generality to restrict attention to direct and truthful mechanisms. Therefore, a mechanism elicits simultaneous type reports from the agents. Conditional on the profile of reports, the mechanism selects (possibly randomly) up to k agents to audit. Conditional on the profile of reports and the information learned through auditing, the mechanism selects (possibly randomly) a feasible allocation $z \in [0, \infty)^n$. Furthermore, it is an equilibrium for all agents to report their type truthfully.

First, let us consider the auditing rule, which maps profiles of type reports into probability distributions over subsets of size at most k of the set of n agents, specifying whom are the audited agents. It will be convenient (and equivalent[13]) to specify directly the probability that each agent gets audited. That is, an auditing rule is given by a $(F_1, ..., F_n)$-measurable function $q : [0,1]^n \to [0,1]^n$, where for each $(t_1, ..., t_n) \in [0,1]^n$ and each $i \in \{1, ..., n\}$, $q_i(t_1, ..., t_n)$ is the probability that agent i gets audited when the profile of type reports is $(t_1, ..., t_n)$. Since at most k agents are audited, the following constraint must be satisfied.

$$\sum_{i=1}^{n} q_i(t_1, ..., t_n) \le k, \quad \forall (t_1, ..., t_n) \in [0,1]^n. \tag{1}$$

[12] Formal proofs of the claims associated to those examples are presented in Appendix B.

[13] See Appendix C.

In what follows, we refer to q as the ex-post auditing rule. It will also be useful to define the interim auditing rule $\{Q_i\}_{1\leq i\leq n}$, where for each i, $Q_i : [0,1] \to [0,1]$ maps agent i's type report into his probability of being audited if all other agents are truthful. That is:

$$Q_i(t_i) = \int_{[0,1]^{n-1}} q_i(t_i, t_{-i}) dF_{-i}(t_{-i}). \tag{2}$$

Let us turn to the allocation rule, specifying which allocation is implemented conditional on the reports and the information learned through auditing. Notice that, without the loss of optimality, the allocation implemented by the mechanism can be assumed to be deterministic. This follows from the fact that the agents are risk-neutral in the allocation they receive, and Assumption 2.[14] Furthermore, it can be assumed that, for each i, the amount z_i allocated to agent i depends only on (a) his own type report, (b) whether he is audited or not, and (c) in case he is audited, whether his report is truthful or not. The reason is that, at the reporting stage, only interim evaluations of agent i's probability of being audited and expected allocations are payoff relevant. There is no resource constraint on allocations, so interim allocations can always be trivially implemented ex-post. As a result, we can simply identify each agent's interim allocation with his ex-post allocation. However, due to the capacity constraint on auditing, the dependence of each agent's allocation on whether he is audited or not must remain explicit. Let us denote $x_i(t_i)$ the amount allocated to agent i if he is audited and found to have reported his type t_i truthfully. If agent i is audited and found to have misreported his type, it can be assumed that he is allocated 0. This way, deviations from truthful reporting are maximally punished. Finally, we denote $y_i(t_i)$ the amount allocated to agent i if his report is t_i and he is not audited by the mechanism.

It remains to express the incentive-compatibility constraints, which guarantee that it is an equilibrium for all agents to report their type truthfully.

$$Q_i(t_i)x_i(t_i) + [1 - Q_i(t_i)]y_i(t_i) \quad \geq \quad [1 - Q_i(\hat{t}_i)]y_i(\hat{t}_i), \qquad \forall i \forall t_i \forall \hat{t}_i. \tag{3}$$

To summarize, a mechanism is of the form $(q, \{Q_i\}_{1\leq i\leq n}, \{x_i\}_{1\leq i\leq n}, \{y_i\}_{1\leq i\leq n})$, where $q : [0,1]^n \to [0,1]^n$ is $(F_1, ..., F_n)$-measurable and, for each i, $Q_i : [0,1] \to [0,1]$, $x_i : [0,1] \to [0,\infty)$, and $y_i : [0,1] \to [0,\infty)$ are F_i-measurable. A mechanism is feasible if constraints (1), (2), and (3) are satisfied. Let us denote $\mathcal{M}$ the set of all feasible mechanisms.[15] The principal chooses a mechanism in $\mathcal{M}$ to minimize

[14] See also footnote 11.

[15] The set of feasible mechanisms is non-empty. For example, setting, for each i, $q_i(t_1, ..., t_n) = k/n$ for all $(t_1, ..., t_n)$, $Q_i(t_i) = k/n$ for all t_i, and $x_i(t_i) = y_i(t_i) = 0$ for all t_i, defines a feasible mechanism, which audits agents uniformly at random and allocates 0 in any case.

her expected loss:

$$\sum_{i=1}^{n}\int_{[0,1]}\Big[Q_i(t_i)\ell_i\big(t_i,x_i(t_i)\big)+\big[1-Q_i(t_i)\big]\ell_i\big(t_i,y_i(t_i)\big)\Big]dF_i(t_i).$$

3 Optimal Mechanism

This section lays out our approach to solve the principal's problem. As a first step, we fix an arbitrary auditing rule and study the problem of finding optimal allocations given this auditing rule. We show that, in an optimal mechanism, for each i, the functions x_i and y_i can be assumed to take a simple form, parameterized by agent i's best deviation payoff θ_i ($\in[0,1]$). As a second step, we show that the structure of those optimal allocations helps narrow our search for an optimal auditing rule. Indeed, we can, without loss, restrict attention to *economical auditing rules*, which minimize the expected number of audits conducted without affecting the equilibrium allocations. As a result, the principal's problem reduces to a well-behaved optimization problem over the auditing rule and the vector $\theta\in[0,1]^n$ of best-deviation payoffs, which we analyze in a third step. A generalized first-order approach applies and characterizes a solution. In general, the characterization remains difficult to work with. However, using Assumption 4, we show that an optimal mechanism treats each agent monotonically in his type. Therefore, in a fourth and final step, we develop an operational characterization of this monotonic solution.

Optimal Allocations for a Given Auditing Rule.— In what follows, we take as given a feasible auditing rule $(q,\{Q_i\}_{1\leq i\leq n})$. Let us first introduce a notation. Given a vector $\theta\in[0,1]^n$, we denote m_θ ($\in\mathcal{M}$) the mechanism with auditing rule $(q,\{Q_i\}_{1\leq i\leq n})$ and allocation rule defined by:[16]

$$x_i^*(t_i)=\max\{\theta_i,t_i\},\quad y_i^*(t_i)=\max\Big\{\theta_i,\min\big\{t_i,\theta_i/[1-Q_i(t_i)]\big\}\Big\},\quad\forall i\forall t_i. \tag{4}$$

Observe that m_θ is indeed a feasible mechanism. Measurability restrictions are inherited from Q_i and $t_i\mapsto t_i$. Moreover, the incentive compatibility constraints (3) hold. To see this, note that, for each i and t_i, $x_i^*(t_i)\geq\theta_i$ and $y_i^*(t_i)\geq\theta_i$, hence $Q_i(t_i)x_i^*(t_i)+[1-Q_i(t_i)]y_i^*(t_i)\geq\theta_i$. That is, truthful reporting guarantees an expected payoff of at least θ_i. In addition, for any deviation $\hat{t}_i$, either $y_i^*(\hat{t}_i)=\theta_i$ or $y_i^*(\hat{t}_i)=\min\big\{\hat{t}_i,\theta_i/[1-Q_i(\hat{t}_i)]\big\}\leq\theta_i/[1-Q_i(\hat{t}_i)]$, thus in any case $[1-Q_i(\hat{t}_i)]y_i^*(\hat{t}_i)\leq\theta_i$. That is, agent i's best deviation expected payoff is bounded above by θ_i.

[16] In the definition of $y_i^*(t_i)$, if it happens to be the case that $Q_i(t_i)=1$, we interpret the fraction with 0 in the denominator $\theta_i/[1-Q_i(t_i)]$ as ∞, so that $\min\big\{t_i,\theta_i/[1-Q_i(t_i)]\big\}=t_i\in[0,\infty)$.

Our first result establishes that an optimal mechanism is of the form m_θ, for some $\theta \in [0,1]^n$.

Lemma 1. *For any feasible mechanism* $m = (q, \{Q_i\}_{1\le i\le n}, \{x_i\}_{1\le i\le n}, \{y_i\}_{1\le i\le n}) \in \mathcal{M}$*, there exists* $\theta \in [0,1]^n$ *such that the principal prefers* m_θ *to* m.

Proof. See Appendix A.1 □

Let us briefly sketch the logic of the arguments leading to Lemma 1. A key observation is that an agent's type does not directly affect his own payoff. As a result, for each agent i and type $\hat{t}_i$, all types $t_i \neq \hat{t}_i$ would have the same expected payoff from deviating to reporting $\hat{t}_i$. It follows that we can define a single best deviation payoff θ_i for every type of agent i. Now, fixing the auditing rule and the vector $\theta = (\theta_1, ..., \theta_n)$ of best deviation payoffs, we can characterize allocations that are optimal point-wise for the principal, subject to the constraints that each agent i (1) has expected payoff at least θ_i from truthful reporting, and (2) has expected payoff at most θ_i from misreporting his type. Those point-wise optimal allocations turn out to be x_i^* and y_i^* defined previously. Moreover, the proof of Lemma 1 also establishes that the vector of best deviation payoffs θ can be taken in $[0,1]^n$.

Before proceeding, let us highlight some notable features of the optimal allocations. Recall that the principal aims to allocate his type to each agent. Effectively, each agent i has a cutoff type θ_i. All types of agent i below this cutoff are allocated the fixed amount θ_i, irrespective of whether they are audited or not. In particular, allocations to the lowest types are distorted upwards in order to meet incentive-compatibility constraints. Without this distortion, the lowest types, having little to lose due to limited liability, would be the most prone to misreport. Allocations to types above the cutoff depend on whether the agent is audited. If the agent is audited and found to be truthful, the allocation is always weakly greater, serving as a reward for truth-telling. In fact, in the case of auditing, the allocation is exactly the agent's type, thus coinciding with the principal-optimal allocation. However, if the agent is not audited, his allocation may be distorted downwards. This downward distortion to a type t_i ensures that it is not attractive for any other type to deviate to reporting t_i.

Economical Auditing Rules.— The shape of the optimal allocations characterized in (4) suggests that, in some cases, the allocation to a type t_i of agent i does not depend on whether the agent is audited or not (as long as he is truthful). Namely, if $Q_i(t_i) > 1 - \theta_i/t_i$,[17] marginal changes in the probability of auditing type t_i do not affect the principal's payoff. Since the principal's auditing capacity is limited, it is convenient and without the loss of optimality to restrict attention to auditing rules such that such excessive auditing is excluded. This observation is formalized in the following Lemma.

[17] Observe that the condition $Q_i(t_i) > 1 - \theta_i/t_i$ includes the case $t_i < \theta_i$ since probabilities are non-negative.

Lemma 2. *Without the loss of optimality, for any i, agent i is never audited when his type is below his best-deviation payoff. That is:*

$$t_i \leq \theta_i \quad \Rightarrow \quad q_i(t_i, t_{-i}) = 0, \quad \forall t_{-i}.$$

Moreover, for any type t_i of agent i, we have:

$$Q_i(t_i) \leq \max\{0, 1 - \theta_i/t_i\}.$$

Proof. First, observe that, for $t_i \leq \theta_i$, agent i receives the same allocation θ_i irrespective of whether he is audited or not. As a consequence, auditing type t_i does not benefit the principal. Setting the probability of audit to 0 in this case only slackens the auditing capacity constraint, without affecting the principal's payoff. In other words, we can assume that, for all t_{-i}, $q_i(t_i, t_{-i}) = Q_i(t_i) = 0 = \max\{0, 1 - \theta_i/t_i\}$.

Second, consider a type $t_i > \theta_i$. If $Q_i(t_i) > 1 - \theta_i/t_i$, then $x_i^*(t_i) = y_i^*(t_i) = t_i$, so again the allocation to that type does not depend whether he is audited or not. Therefore, a small reduction in the probability of audit slackens the auditing capacity constraint without affecting the principal's payoff. Specifically, denote $\varepsilon = (1 - \theta_i/t_i)/Q_i(t_i) \in (0, 1)$. For all t_{-i}, we replace the probability of auditing agent i by:

$$\hat{q}_i(t_i, t_{-i}) = \varepsilon q_i(t_i, t_{-i}).$$

This new auditing rule remains feasible and does not affect the induced allocations. Furthermore, the corresponding interim probability of auditing type t_i satisfies:

$$\hat{Q}_i(t_i) = \varepsilon Q_i(t_i) = 1 - \theta_i/t_i.$$

□

Optimal Audits.— Using Lemmata 1 and 2, we eliminate the allocations from the problem and simply optimize over the auditing rule and the vector of best deviation payoffs. Specifically, since types of agent i below the cutoff θ_i are audited with probability 0, we can replace the amount allocated conditional on a truthful audit by $x_i(t_i) = t_i$. Similarly, following Lemma 2, the amount allocated conditional on no audit can be replaced by $y_i(t_i) = \theta_i/[1 - Q_i(t_i)]$. As a result, the principal's choice set reduces to the set, which we shall denote $\widetilde{\mathcal{M}}$, of elements of the form $(\theta, q, \{Q_i\}_{1 \leq i \leq n})$, where $\theta \in [0, 1]^n$, $q : [0, 1]^n \to [0, 1]^n$ is $(F_1, ...F_n)$-measurable and, for all i, $Q_i : [0, 1] \to [0, 1]$ is F_i-measurable, such that constraints (1) and (2) are satisfied, as well as:

$$Q_i(t_i) \leq \max\{0, 1-\theta_i/t_i\}, \quad \forall i \forall t_i. \tag{5}$$

Given this definition, the principal's problem is:

$$\min_{(\theta,q,\{Q_i\}_i)\in\widetilde{\mathcal{M}}} \quad \sum_{i=1}^{n} \int_0^1 [1-Q_i(t_i)]\ell_i\big(t_i, \theta_i/[1-Q_i(t_i)]\big) dF_i(t_i). \tag{6}$$

We tackle this problem using a first-order approach. Necessary and sufficient optimality conditions are stated in the following result.[18]

Proposition 1. *$(\theta, q, \{Q_i\}_i) \in \widetilde{\mathcal{M}}$ is a solution to Problem (6) if and only if there exist $\psi : [0,1]^n \to [0,\infty)$ $(F_1, ..., F_n)$-measurable and, for each i, $\lambda_i : [0,1] \to \mathbb{R}$ F_i-measurable, $\alpha_i : [0,1]^n \to [0,\infty)$ $(F_1, ..., F_n)$-measurable, $\beta_i : [0,1]^n \to [0,\infty)$ $(F_1, ..., F_n)$-measurable, $\gamma_i \in [0,\infty)$, and $\delta_i \in [0,\infty)$, such that, for all i:*

$$\int_0^1 \frac{\partial \ell_i}{\partial z_i}\left(t_i, \frac{\theta_i}{1-Q_i(t_i)}\right) dF_i(t_i) - \gamma_i + \delta_i = 0, \tag{7}$$

$$\gamma_i \theta_i = 0, \tag{8}$$

$$\delta_i(1-\theta_i) = 0, \tag{9}$$

and for all i and $(t_1, ..., t_n)$:

$$\lambda_i(t_i) = \ell_i\left(t_i, \frac{\theta_i}{1-Q_i(t_i)}\right) - \frac{\theta_i}{1-Q_i(t_i)} \cdot \frac{\partial \ell_i}{\partial z_i}\left(t_i, \frac{\theta_i}{1-Q_i(t_i)}\right), \tag{10}$$

$$\beta_i(t_1, ..., t_n) - \alpha_i(t_1, ..., t_n) = \lambda_i(t_i) - \psi(t_1, ..., t_n), \tag{11}$$

$$\alpha_i(t_1, ..., t_n) q_i(t_1, ..., t_n) = 0, \tag{12}$$

$$\beta_i(t_1, ..., t_n)[1 - q_i(t_1, ..., t_n)] = 0, \tag{13}$$

$$\psi(t_1, ..., t_n)\left[\sum_{j=1}^{n} q_j(t_1, ..., t_n) - k\right] = 0. \tag{14}$$

Proof. See Appendix A.2. □

[18] As discussed in Section 2, the generalized partial derivatives of the loss functions ℓ_i with respect to its second argument are well-defined. At points where $\ell_i(t_i, \cdot)$ is not differentiable, the corresponding optimality condition should be understood as an inclusion. See Clarke (1990).

Let us discuss in turn each condition. Condition (7) is a first-order condition with respect to agent i's best-deviation payoff θ_i, where γ_i and δ_i are Lagrange-multipliers associated to the boundary conditions $0 \leq \theta_i \leq 1$. Conditions (8) and (9) are the corresponding complementary-slackness conditions. Recall that types $t_i < \theta_i$ of agent i receive an allocation that is distorted upward relative to the principal-optimal allocation, while the allocation to types $t_i > \theta_i$ is (weakly) distorted downward. Condition (7) implies that, across all types of agent i, the marginal loss to the principal from upward distortions and the marginal loss from downward distortions exactly compensate each other.

For each i and t_i, condition (10) is a first-order condition with respect to the interim probability $Q_i(t_i)$ of auditing type t_i of agent i. On the left-hand side, $\lambda_i(t_i)$, which is the Lagrange multiplier associated to constraint (2), can be interpreted as the interim evaluation of the marginal cost of auditing type t_i, to the extent that the principal's auditing resources are limited. The expression on the right-hand side captures the marginal benefit of auditing that type. Observe that there are two channels through which auditing is beneficial. First, if an audit is conducted, the mechanism implements the principal's optimal allocation, whereas the allocation to non-audited types is distorted. This effect is captured by the term $\ell_i(t_i, \theta_i/[1 - Q_i(t_i)])$ in equation (10). Second, even in the event of no audit, the increased likelihood that this type report would have been audited makes reporting type t_i a less attractive deviation to other types, implying that the allocative distortion for that type can be reduced. This second benefit is represented by the term $-(\theta_i/[1 - Q_i(t_i)]) \cdot (\partial \ell_i/\partial z_i)(t_i, \theta_i/[1 - Q_i(t_i)])$, which is non-negative when $t_i > \theta_i$.

Let us now consider the remaining conditions (11)-(14), which relate the interim optimality of the auditing rule from condition (10) to its ex-post feasibility. Condition (11) is a first-order condition with respect to the ex-post probability $q_i(t_1, ..., t_n)$ of auditing agent i, where $\alpha_i(t_1, ..., t_n)$ and $\beta_i(t_1, ..., t_n)$ are Lagrange-multipliers associated to the constraints that $0 \leq q_i(t_1, ..., t_n) \leq 1$, while $\psi(t_1, ..., t_n)$ is the Lagrange-multiplier associated to the auditing capacity constraint (1). Conditions (12), (13), and (14) are the corresponding complementary-slackness conditions. Those conditions yield useful insights about the structure of the ex-post auditing rule. In particular, if $\lambda_i(t_i) > \psi(t_1, ..., t_n)$, since $\alpha_i(\cdot)$ and $\beta_i(\cdot)$ are both non-negative, then it must be that $\beta_i(t_1, ..., t_n) > 0$, which, by complementary slackness, implies that $q_i(t_1, ..., t_n) = 1$. Similarly, $\lambda_i(t_i) < \psi(t_1, ..., t_n)$ implies that $\alpha_i(t_1, ..., t_n) > 0$ and $q_i(t_1, ..., t_n) = 0$. These observations suggest the following structure. Given a profile of type reports $(t_1, ..., t_n)$, agents are ranked by according to the profile $(\lambda_1(t_1), ..., \lambda_n(t_n))$ of marginal costs of audit. The multiplier $\psi(t_1, ..., t_n)$ can be assumed to be equal to the maximum between 0 and the k^{th} largest $\lambda_i(t_i)$. Those agents with $\lambda_i(t_i) > \psi(t_1, ..., t_n)$ are audited with probability 1, while those agents with

$\lambda_i(t_i) < \psi(t_1, ..., t_n)$ are audited with probability 0. The mechanism possibly randomizes among agents with $\lambda_i(t_i) = \psi(t_1, ..., t_n)$ in case of ties for the k^{th} position in the ranking. Now, for a fixed agent i, notice that $\lambda_i(t_i)$ depends only on that agent's own type t_i. Thus, for any profile of other reports t_{-i}, the higher $\lambda_i(t_i)$, the more likely it is that agent i gets audited. Furthermore, we have the following result.

Lemma 3. *Under the hypotheses of Proposition 1, for each i, $\lambda_i(\cdot)$ is monotonic on $[\theta_i, 1]$.*

Proof. See Appendix A.3. □

Building on Lemma 3, we obtain the following result.

Proposition 2. *Without loss, a solution to problem (6) features a monotonic interim auditing rule. That is, for all i, $Q_i(\cdot)$ is weakly increasing.*

Proof. See Appendix A.4. □

The proofs of Lemma 3 and Proposition 2 crucially rely on Assumption 4. To understand why, it is useful to reformulate the integrand in the principal's objective in Problem (6) in terms of the allocation rule:

$$[1 - Q_i(t_i)]\ell_i\left(t_i, \frac{\theta_i}{1 - Q_i(t_i)}\right) = \theta_i \frac{\ell_i\left(t_i, y_i(t_i)\right)}{y_i(t_i)},$$

where, recall, the identity $y_i(t_i) = \theta_i/[1 - Q_i(t_i)]$ follows from the restriction to *economical auditing rules*, which limit the use of auditing to ensuring incentive compatibility. In other words, since (i) distortions arise only when the agent is not audited, and (ii) the probability that the agent is not audited is inversely proportional to the allocation he would receive in that case, the principal's effective objective is, for fixed θ_i, to minimize the average loss per unit allocated to each type of the agent. Assumption 4 precisely imposes discipline on this effective objective. Indeed, it requires that, for a downward-distorted allocation, reducing the distortion by a fixed amount leads to a larger decrease in the average loss to the principal when the agent's type is higher. As a consequence of this assumption, an optimal mechanism treats each agent monotonically in his own type. By Proposition 2, higher type are more likely to be audited. Furthermore, it follows that both x_i and y_i are monotonic.

Reduced-Form Optimal Auditing.— Proposition 2 has important consequences in terms of tractability. Let us explain why. Notice that, in the formulation problem (6), only the best-deviation payoffs and the interim auditing rule directly enter the principal's objective, whereas the ex-post auditing rule is relevant exclusively for feasibility of the interim auditing rule. Therefore, the problem could be simplified if we could abstract away from the ex-post auditing rule and rely on a direct characterization of the feasible interim auditing rules. This very issue turns out to have been relevant in Auction Theory.[19] Following seminal contributions by Maskin and Riley (1984) and Border (1991), Che, Kim, and Mierendorff (2013) provide a general characterization of feasible interim rules, which can be applied to our context. This characterization is a priori hardly operational. However, having now established that we can restrict attention to monotonic interim auditing rules by Proposition 2, the formulation of the reduced-form problem drastically simplifies, as we show next. We rely on the following result by Che et al. (2013).

Lemma 4. *For each i, let $Q_i : [0,1] \to [0,1]$ F_i-measurable and non-decreasing. There exists a feasible ex-post auditing rule q consistent with $\{Q_i\}_{1 \le i \le n}$ if and only if:*

$$\sum_{i=1}^{n} \int_{s_i}^{1} Q_i(t_i) dF_i(t_i) \le \int_{[0,1]^n} \min\{k, \#\{i : t_i \ge s_i\}\} d\mathbf{F}(t), \quad \forall (s_1, ..., s_n). \tag{15}$$

Proof. Immediate application of Theorem 5 in Che et al. (2013). □

Following Lemma 4, we further reduce the principal's choice set to the set $\overline{\mathcal{M}}$ of elements of the form $(\theta, \{Q_i\}_{1 \le i \le n})$, where $\theta \in [0,1]^n$ and, for each i, $Q_i : [0,1] \to [0,1]$ is F_i-measurable and non-decreasing. Furthermore, constraints (5) and (15) must be satisfied by elements of $\overline{\mathcal{M}}$. The principal's problem becomes:

$$\min_{(\theta, \{Q_i\}_i) \in \overline{\mathcal{M}}} \quad \sum_{i=1}^{n} \int_{0}^{1} [1 - Q_i(t_i)] \ell_i\left(t_i, \theta_i / [1 - Q_i(t_i)]\right) dF_i(t_i). \tag{16}$$

In this reduced-form problem, the optimality conditions analogous to Proposition 1 can be expressed as follows.

Proposition 3. *$(\theta, \{Q_i\}_i) \in \overline{\mathcal{M}}$ is a solution to the principal's reduced-form problem (16) if and only if there exist vectors $\gamma \in [0,\infty)^n$ and $\delta \in [0,\infty)^n$, a positive measure μ on $[0,1]^n$,[20] and for each i, an F_i-measurable function $\nu_i : [0,1] \to [0,\infty)$, such that:*

$$\sum_{i=1}^{n} \int_{s_i}^{1} Q_i(t_i) dF_i(t_i) = \int_{[0,1]^n} \min\{k, \#\{i : t_i \ge s_i\}\} d\mathbf{F}(t), \quad \mu - a.s.,$$

[19]The formal connection is intuitive. In our setting, the principal has a limited number of audits to assign across multiple agents who, at the reporting stage, care about the probability that they will be audited. In an auction setting, the principal has a limited number of goods to allocate across multiple agents who, at the bidding stage, care about the probability that they will obtain a good.

[20]More precisely, μ must be defined on the sigma algebra of $(F_1, ..., F_n)$-measurable subsets of $[0,1]^n$.

for all i:

$$\int_0^1 \frac{\partial \ell_i}{\partial z_i}\left(t_i, \frac{\theta_i}{1-Q_i(t_i)}\right) dF_i(t_i) - \gamma_i + \delta_i = 0,$$

$$\gamma_i \theta_i = 0,$$

$$\delta_i(1-\theta_i) = 0,$$

and for all i and t_i:

$$\mu\left([0,t_i]\times[0,1]^{n-1}\right) - \nu_i(t_i) = \ell_i\left(t_i, \frac{\theta_i}{1-Q_i(t_i)}\right) - \frac{\theta_i}{1-Q_i(t_i)} \cdot \frac{\partial \ell_i}{\partial z_i}\left(t_i, \frac{\theta_i}{1-Q_i(t_i)}\right),$$

$$Q_i(t_i)\nu_i(t_i) = 0.$$

The proof is analogous to that of Proposition 1, so we do not repeat the argument. The measure μ acts as a Lagrange multiplier associated to constraint (15). The first condition is the corresponding complementary-slackness condition. The following three conditions are identical to conditions (7), (8), and (9). The next condition is a reformulation of condition (10), where $\lambda_i(t_i)$ is decomposed into the difference between a positive part $\mu\left([0,t_i]\times[0,1]^{n-1}\right)$ and a negative part $\nu_i(t_i)$. The interpretation remains the same: $\lambda_i(t_i)$ is an evaluation of the marginal cost of auditing type t_i of agent i. This term may be positive due to the principal's limited auditing resources, which are now captured by constraint (15). Thus, the positive part directly corresponds to the impact of type t_i on the Lagrange multiplier associated to (15). In addition, the negative part $\nu_i(t_i)$ is the Lagrange multiplier associated to the non-negativity constraint $Q_i(t_i) \geq 0$. By complementary slackness, expressed as the final optimality condition, this negative part plays a role only for those types that the principal finds optimal not to audit at all.

Let us also remark that it is not necessary to define Lagrange multipliers associated to the monotonicity constraint on Q_i nor constraint (5). This follows from arguments similar to those used in the proofs of Propositions 1 and 2.

Summary.— As we illustrate in the next section, Proposition 3 provides a practically useful characterization of solutions to the principal's problem. The optimality conditions can be used to construct candidate solutions and to certify their optimality. Before moving on, let us briefly take stock of our findings so far regarding the structure of an optimal mechanism.

From the interim perspective, each agent i is treated monotonically in his own type. In terms of the auditing rule, the best deviation payoff θ_i corresponds to a cutoff type. Types below the cutoff are not audited. Types above the cutoff are audited randomly, with a probability $Q_i(t_i)$ that is monotonic in the type t_i. The allocation to an audited agent (if truthful) matches the type $x_i(t_i) = t_i$. If the agent is not audited, his allocation satisfies $y_i(t_i)[1 - Q_i(t_i)] = \theta_i$. This allocation rule implies that the agent remains indifferent between any lie.

We can abstract away from the problem of constructing an optimal ex-post auditing rule. However, the optimality conditions of Proposition 1 still provide some insight into the structure of optimal ex-post auditing rules. Specifically, types of agent i above the cutoff θ_i are mapped into a non-decreasing index $\lambda_i(t_i)$. Now, given a profile $(t_1, .., t_n)$, agents are ranked according to their respective $\lambda_i(t_i)$, and the mechanism selects agents to audit in order of this ranking until possibly reaching the full capacity k. Ex-post randomization may remain needed when multiple agents tie for the k^{th} position.

4 Illustration of Explicit Constructions

In this section, we illustrate how to use the optimality conditions of Proposition 3 to construct an explicit solution to the principal's problem and certify its optimality. For clarity, we slightly simplify the problem and restrict attention to symmetric environments. That is, we assume that for all i, $\ell_i = \ell$ and $F_i = F$, where F is an atomless cumulative distribution function on $[0, 1]$. This last assumption simplifies the exposition by ensuring that agents almost surely have different types.

Since the agents are ex-ante symmetric, an optimal mechanism is symmetric.[21] Therefore, from now on, we omit the index i and denote by $\theta \in [0, 1]$ the cutoff type and $Q(\cdot)$ the interim auditing rule. In the symmetric case, the reduced-form auditing capacity constraint takes a particularly simple form:

$$\int_s^1 Q(t)dF(t) \leq \int_s^1 R_k^n(t)dF(t), \quad \forall s, \tag{17}$$

where R_k^n is the interim evaluation of the rank-based auditing rule, which audits agents with the k highest reported types. That is, since F has no atom:[22]

$$R_k^n(t) = \sum_{j=0}^{k-1} \binom{n-1}{j} F(t)^{n-1-j}[1 - F(t)]^j.$$

[21] To see this, observe that, given an optimal mechanism, one can create $n!$ optimal mechanisms by relabeling the agents. In each of these, agent's have the same optimal strategy to report their type truthfully, and the principal achieves the same expected payoff. The mechanism consisting of privately randomizing uniformly among the $n!$ possible relabelings is a symmetric direct mechanism in which it is an equilibrium for all agents to report their type truthfully. In addition, the principal again achieves the same expected payoff, so it is an optimal mechanism.

[22] It would be easy to adjust the notations in case F has atoms, observing that, due to symmetry, ties must be broken uniformly at random.

A first remark is that, having imposed symmetry and ruled out atoms, we can guarantee that the cutoff type θ is interior. Indeed, suppose by contradiction that $\theta = 0$. Then, the symmetric counterpart to condition (7) would imply:

$$\int_0^1 \frac{\partial \ell}{\partial z}(t,0)\, dF(t) - \gamma + \delta = 0,$$

with, for all $t > 0$, $(\partial\ell/\partial z)(t,0) < 0$, $\gamma \geq 0$, and $\delta = 0$ since $\theta = 0 < 1$. This obviously cannot hold. A similar argument rules out the possibility that $\theta = 1$.

With this remark in mind, the optimality conditions of Proposition 3 can be translated under our simplifying assumptions as follows:

$$\int_s^1 Q(t)dF(t) = \int_s^1 R_k^n(t)dF(t), \quad \mu\text{-a.s.,} \tag{18}$$

$$\int_0^1 \frac{\partial \ell}{\partial z}\left(t, \frac{\theta}{1-Q(t)}\right) dF(t) = 0, \tag{19}$$

$$\mu(t) - \nu(t) = \ell\left(t, \frac{\theta}{1-Q(t)}\right) - \frac{\theta}{1-Q(t)} \cdot \frac{\partial \ell}{\partial z}\left(t, \frac{\theta}{1-Q(t)}\right), \quad \forall t, \tag{20}$$

$$Q(t)\nu(t) = 0, \quad \forall t, \tag{21}$$

where $\nu : [0,1] \to [0,\infty)$ is F-measurable and μ is a positive measure on $[0,1]$. As in condition (20), we write $\mu(t)$ for $\mu([0,t])$.

Next, we investigate necessary implications of those conditions. In particular, we show that we can partition the type space $[0,1]$ into different regions. Moreover on each region, we can derive an explicit parameterized functional form for the interim auditing rule Q. Thus, finding a solution boils down to identifying the boundaries of those regions together with the corresponding parameters which satisfy the feasibility constraints and the optimality conditions (18)-(21).

As a first step, we write $[0,1] = \text{supp}\mu \cup ([0,1] \setminus \text{supp}\mu)$. On the support of μ, by condition (18), the reduced-form auditing capacity constraint (17) binds. As a consequence, we can set, without loss:

$$\forall t \in \text{supp}\mu, \quad Q(t) = R_k^n(t).$$

To see this, consider $t \in \text{supp}\mu$. We have, for all $t' \geq t$, $\int_{t'}^1 Q(u)dF(u) \leq \int_{t'}^1 R_k^n(u)dF(u)$, with equality at $t' = t$. Thus, for all $t' > t$, $\int_t^{t'} Q(u)dF(u) \geq \int_t^{t'} R_k^n(u)dF(u)$. Furthermore, recall that Q must be monotonic and that R_k^n inherits the continuity of F. Hence, it must be the case that: $\lim_{t' \to t, t' > t} Q(t') \geq R_k^n(t)$. Moreover, a similar argument implies that $\lim_{t' \to t, t' < t} Q(t') \leq R_k^n(t)$. If the two limits coincide, then it must be the case that $Q(t) = R^n_k(t)$ by monotonicity. Otherwise, it is without loss to redefine $Q(t) = R^n_k(t)$ at that point, since Q has a measure zero of such discontinuities, again by monotonicity (and using the fact that F is atomless).

The previous argument specifies an explicit solution for Q on suppμ. Next, we derive an expression for Q outside suppμ. As a special case, consider the region below the support of μ, that is for $t \in [0, \bar{t}_0)$, where $\bar{t}_0$ denotes the lower end of suppμ. Observe that, for $t \leq \theta$, $Q(t) = 0$, so constraint (17) must be slack below θ, implying that $\bar{t}_0 \geq \theta$. If there exists $t \in (\theta, \bar{t}_0)$, condition (20) at that point can only be satisfied with $\mu(t) = \nu(t) = 0$, thus $Q(t) = 1 - \theta/t$. Combining those observations, we conclude that:

$$\forall t < \bar{t}_0, \quad Q(t) = \max\{0, 1 - \theta/t\}.$$

By definition, $\bar{t}_0 \in \text{supp}(\mu)$. Hence $Q(\bar{t}_0) = R^n_k(\bar{t}_0) \geq 1 - \theta/\bar{t}_0$, where the inequality is by monotonicity of Q. In addition, with $R^n_k(\bar{t}_0) > 0$, condition (20) implies:

$$\ell\left(\bar{t}_0, \frac{\theta}{1 - R^n_k(\bar{t}_0)}\right) - \frac{\theta}{1 - R^n_k(\bar{t}_0)} \cdot \frac{\partial \ell}{\partial z}\left(\bar{t}_0, \frac{\theta}{1 - R^n_k(\bar{t}_0)}\right) = \mu(\bar{t}_0) \geq 0.$$

As a consequence, since $\ell(\bar{t}_0, \cdot)$ is convex and minimized at $\bar{t}_0$, it must be the case that $\theta/[1 - R^n_k(\bar{t}_0)] \leq \bar{t}_0$, or equivalently $R^n_k(\bar{t}_0) \leq 1 - \theta/\bar{t}_0$. As a result, we obtain the following condition, which relates $\bar{t}_0$ to the best deviation payoff θ:

$$R^n_k(\bar{t}_0) = 1 - \theta/\bar{t}_0.$$

Another special case concerns the region above the support of μ. Indeed, note that $R^n_k(t) \xrightarrow[t \to 1]{} 1$, since the rank-based auditing rule must audit with probability converging to 1 an agent whose type approaches 1. However, by constraint (5), $Q(t)$ is bounded above by $1 - \theta < 1$. As a result, constraint (17) must be slack for s sufficiently close to 1. Therefore, there exists $\underline{t}_1 < 1$, such that $\mu(t)$ is constant on $(\underline{t}_1, 1]$. Let us denote μ_1 the value of this constant. On this interval, $\nu(t) = 0$, so condition (20) implies:

$$\mu_1 = \ell\left(t, \frac{\theta}{1 - Q(t)}\right) - \frac{\theta}{1 - Q(t)} \cdot \frac{\partial \ell}{\partial z}\left(t, \frac{\theta}{1 - Q(t)}\right).$$

Equivalently, we can write:

$$\forall t > \underline{t}_1, \quad Q(t) = 1 - \frac{\theta}{\overline{y}_{\mu_1}(t)},$$

where $\overline{y}_{\mu_1}(t)$ is solution to $\ell(t, y) - y(\partial \ell/\partial z)(t, y) = \mu_1$.

Note that the expression $Q(t) = 1 - \theta/\overline{y}_{\mu_1}(t)$ remains valid at the top as long as constraint (17) remains slack. Thus, $\underline{t}_1$ can be defined as the upper boundary of the support of μ, where:

$$\int_{\underline{t}_1}^{1} \left(1 - \frac{\theta}{\overline{y}_{\mu_1}(t)}\right) dF(t) = \int_{\underline{t}_1}^{1} R_k^n(t) dF(t).$$

This is a condition relating μ_1, $\underline{t}_1$, and θ. A second condition is obtained by monotonicity of Q and μ. Indeed, by monotonicity of Q, we must have:

$$\lim_{t \to \underline{t}_1, t > \underline{t}_1} \left(1 - \frac{\theta}{\overline{y}_{\mu_1}(t)}\right) \geq R_k^n(\underline{t}_1).$$

On the other hand, monotonicity of μ requires:

$$\ell\left(\underline{t}_1, \frac{\theta}{1 - R_k^n(\underline{t}_1)}\right) - \frac{\theta}{1 - R_k^n(\underline{t}_1)} \cdot \frac{\partial \ell}{\partial z}\left(\underline{t}_1, \frac{\theta}{1 - R_k^n(\underline{t}_1)}\right) \leq \mu(\underline{t}_1) = \ell\left(\underline{t}_1, \overline{y}_{\mu_1}(\underline{t}_1)\right) - \overline{y}_{\mu_1}(\underline{t}_1) \cdot \frac{\partial \ell}{\partial z}\left(\underline{t}_1, \overline{y}_{\mu_1}(\underline{t}_1)\right).$$

Since $\ell(\underline{t}_1, \cdot)$ is convex, we obtain equivalently:

$$R_k^n(\underline{t}_1) \leq 1 - \frac{\theta}{\overline{y}_{\mu_1}(\underline{t}_1)}.$$

In particular, if $\overline{y}_{\mu_1}$ happens to be continuous at $\underline{t}_1$, we obtain an exact condition relating again the parameters μ_1, $\underline{t}_1$, and θ.

We now extend the logic applied to those two special cases in order to complete the construction. Observe that $[0, 1] \setminus \text{supp}\mu$ can be partitioned into a (at most) countable union of disjoint open intervals:

$$[0, 1] \setminus \text{supp}\mu = \bigcup_{p \in \mathbb{N}} I_p,$$

where $I_0 = [0, \bar{t}_0)$ and $I_1 = (\underline{t}_1, 1]$ correspond to the two special cases already discussed. For $p \geq 2$, let us denote $I_p = (\underline{t}_p, \bar{t}_p)$. On each of those intervals, $\mu(t)$ must be constant, and we denote μ_p its constant value. We have already discussed the case $p = 0$, with $\mu_0 = 0$. Thus, let us consider cases with $\mu_p > 0$. In this case, $\nu(t) = 0$ on I_p, so condition (20) implies:

$$\forall t \in (\underline{t}_p, \bar{t}_p), \quad Q(t) = 1 - \frac{\theta}{\overline{y}_{\mu_p}(t)},$$

where, again, $\overline{y}_{\mu_p}(t)$ is solution to $\ell(t, y) - y(\partial \ell / \partial z)(t, y) = \mu_p$.

As for the special cases, we obtain conditions that relate the parameters μ_p, $\underline{t}_p$, and $\bar{t}_p$ to the best-deviation payoff θ. First, note that constraint (17) binds at both boundaries, thus:

$$\int_{\underline{t}_p}^{\bar{t}_p} \left(1 - \frac{\theta}{\overline{y}_{\mu_p}(t)}\right) dF(t) = \int_{\underline{t}_p}^{\bar{t}_p} R_k^n(t) dF(t). \tag{22}$$

Furthermore, similar to the previous cases, we have boundary conditions:

$$\lim_{t \to \underline{t}_p, t > \underline{t}_p} \left(1 - \frac{\theta}{\overline{y}_{\mu_p}(t)}\right) \geq R_k^n(\underline{t}_p) \geq 1 - \frac{\theta}{\overline{y}_{\mu_p}(\underline{t}_p)}, \tag{23}$$

$$\lim_{t \to \bar{t}_p, t < \bar{t}_p} \left(1 - \frac{\theta}{\overline{y}_{\mu_p}(t)}\right) \leq R_k^n(\bar{t}_p) \leq 1 - \frac{\theta}{\overline{y}_{\mu_p}(\bar{t}_p)}. \tag{24}$$

To summarize, we have obtained a full description of a candidate solution in terms of the parameters θ and $\{\mu_p, \underline{t}_p, \bar{t}_p\}_p$. Those are restricted by conditions (22), (23), (24), and the optimality condition (19). Reciprocally, to verify optimality of the candidate solution, it is sufficient to verify that, on suppμ, the induced:

$$\mu(t) = \ell\left(t, \frac{\theta}{1 - R_k^n(t)}\right) - \frac{\theta}{1 - R_k^n(t)} \cdot \frac{\partial \ell}{\partial z}\left(t, \frac{\theta}{1 - R_k^n(t)}\right)$$

remains monotonic.

For concreteness, let us apply this approach to the introductory example presented in Figure 1.

Example.— Suppose that there are $n = 2$ agents and only $k = 1$ audit available. Furthermore, the agents' types are independently and identically distributed uniformly on $[0, 1]$, that is $F(t) = t$. In this case, we have:

$$R_1^2(t) = t.$$

The principal's loss function is assumed to take the functional form $\ell(t, z) = z \ln(z/t) - (z - t)$. As a consequence, we have:

$$\ell(t, y) - y \frac{\partial \ell}{\partial z}(t, y) = t - y,$$

so that:

$$\overline{y}_{\mu_p}(t) = t - \mu_p.$$

In particular, note that $\overline{y}_{\mu_p}$ is continuous. As a result, conditions (23) and (24) simplify to exact equalities in this case.

Let us now follow the approach outlined previously using those simple functional forms. At the bottom of the type space, we have:

$$\forall t < \bar{t}_0, \quad Q(t) = \max\{0, 1 - \theta/t\},$$

with the condition:

$$\bar{t}_0(1 - \bar{t}_0) = \theta.$$

At the top, we have:

$$\forall t > \underline{t}_1, \quad Q(t) = 1 - \frac{\theta}{t - \mu_1},$$

with the condition:

$$\mu_1 = \underline{t}_1 - \frac{\theta}{1 - \underline{t}_1}.$$

Furthermore, the binding constraint (17) provides the additional condition:

$$\int_{\underline{t}_1}^{1} \left(1 - \frac{\theta}{t - \mu_1}\right) dt = \int_{\underline{t}_1}^{1} t dt.$$

Simple algebra, where we substitute $\mu_1 = \underline{t}_1 - \theta/(1 - \underline{t}_1)$, leads to the equivalent condition:

$$\frac{(1 - \underline{t}_1)^2}{\theta} = 2 \ln \left(1 + \frac{(1 - \underline{t}_1)^2}{\theta}\right).$$

For future reference, note that this condition implies that $(1 - \underline{t}_1)^2/\theta > e - 1 > 1$, thus $\underline{t}_1 < 1 - \sqrt{\theta}$.

On the support of μ, we have $Q(t) = t$, thus:

$$\mu(t) = t - \frac{\theta}{1 - t}.$$

Observe that the expression on the right-hand side is monotonic in t on $[0, 1 - \sqrt{\theta}] \supseteq [\bar{t}_0, \underline{t}_1]$. As a result, we can set supp$\mu = [\bar{t}_0, \underline{t}_1]$, and it is not necessary to introduce additional parameters.

To summarize, we propose the following parameterized candidate:

$$Q(t) = \begin{cases} 0 & \text{if } t \leq \theta, \\ 1 - \theta/t & \text{if } \theta \leq t \leq \bar{t}_0, \\ t & \text{if } \bar{t}_0 \leq t \leq \underline{t}_1, \\ 1 - \theta/(t - \mu_1) & \text{if } t \geq \underline{t}_1, \end{cases}$$

where:

$$\bar{t}_0(1-\bar{t}_0)=\theta, \quad \mu_1=\underline{t}_1-\frac{\theta}{1-\underline{t}_1}, \quad \frac{(1-\underline{t}_1)^2}{\theta}=2\ln\left(1+\frac{(1-\underline{t}_1)^2}{\theta}\right).$$

It remains to use condition (19) which expresses optimality with respect to the best-deviation payoff θ, which in the context of this example, takes the form:

$$\int_0^1 \ln\left(\frac{\theta}{t[1-Q(t)]}\right)dt=0.$$

Using previous conditions, we can compute the integral to obtain the equivalent condition:

$$\bar{t}_0(1-\bar{t}_0)+2(\underline{t}_1-\bar{t}_0)+\ln(\bar{t}_0)+\left(1-\underline{t}_1+\frac{\bar{t}_0(1-\bar{t}_0)}{1-\underline{t}_1}\right)\cdot\ln\left(1+\frac{(1-\underline{t}_1)^2}{\bar{t}_0(1-\bar{t}_0)}\right)=0.$$

Thus, it is sufficient to solve a system of 4 conditions, with 4 unknown parameters. Given a solution to this system, it is easy to verify that conditions (18)-(21) hold with:

$$\nu(t)=\max\left\{-\ell\left(t,\frac{\theta}{1-Q(t)}\right)+\frac{\theta}{1-Q(t)}\cdot\frac{\partial\ell}{\partial z}\left(t,\frac{\theta}{1-Q(t)}\right),0\right\}=\max\{\theta-t,0\},$$

and

$$\mu(t)=\max\left\{\ell\left(t,\frac{\theta}{1-Q(t)}\right)-\frac{\theta}{1-Q(t)}\cdot\frac{\partial\ell}{\partial z}\left(t,\frac{\theta}{1-Q(t)}\right),0\right\}=\max\left\{t-\frac{\theta}{1-Q(t)},0\right\}.$$

Furthermore, a solution can be computed numerically, and we obtain $\theta\approx 0.1607$, $\mu_1\approx 0.1116$, $\bar{t}_0\approx 0.2012$, $\underline{t}_1\approx 0.3645$.

It is insightful to plot the marginal benefit of auditing any type t at the optimum, which, by condition (20), is equal to $\mu(t)-\nu(t)$. Recall that, by Proposition 1, an optimal ex-post auditing rule ranks agents according to this marginal benefit and, if an agent is audited, his marginal benefit must be weakly higher than that of the other agent. The plot, which is presented in Figure 2, reveals that the type space can be decomposed into different categories.

A first category contains types below the cutoff θ, for whom the marginal benefit is negative. Indeed, those types are never audited by the optimal mechanism. A second category consists of the interval of types $(\bar{t}_0,\underline{t}_1)$, where the marginal benefit is strictly increasing and positive. As a result, for those types, a tie in marginal benefit is a zero-probability event, and from the ex-post perspective, they must be audited if and only if the other agent's report is below their type. This is exactly what the optimal interim auditing rule that we have derived suggests. Since this interval corresponds to (the interior of)

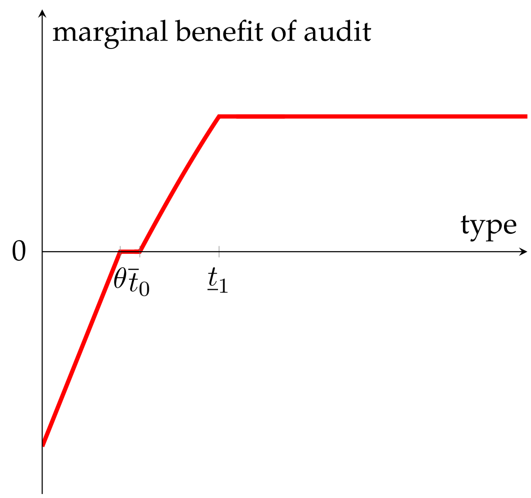


Figure 2: Marginal benefit of auditing any type, as implied by the optimal mechanism.

the support of μ, the reduced-form auditing capacity constraint binds, and they are audited according to the rank-based auditing rule R_1^2. Finally, a third category includes types in $[\theta, \bar{t}_0]$ and in $[\underline{t}_1, 1]$, where having the same marginal benefit of audit as the other agent is a positive-probability event. There, an ex-post implementation of the optimal interim auditing rule may rely on randomization.

Beyond the specific functional forms of this example, those insights carry over more generally. Given the structure of solutions that we have obtained previously, we can define a first category of types, below the cutoff θ, that are never audited. The interior of the support of the measure μ defines a second category, of types who are audited according to the rank-based auditing rule. Finally, the rest of the type space forms the third category. In a sense, this third category corresponds to a true pool of targeted random audits as we have described it in the Introduction. Instead the second category of types constitute a *reserve pool*, of types that are considered for auditing when the principal's capacity allows it, that is, if they make it to the top k.

5 Scaling Up the Problem and Costly Verification

In our model, the principal can conduct k audits for free, but the $(k+1)^{\text{th}}$ audit is infinitely costly. In this section, we address the sensitivity of our results to this assumption.

It is clear that allowing the principal to conduct additional audits can only benefit her, as it would expand her choice set. Slightly less obvious is the fact that the principal would also find it advantageous to scale up the problem. This is most vividly apparent when agents are ex-ante symmetric, as in Section

4. Indeed, for two natural integers $p \leq \hat{p}$, we have:

$$\int_s^1 R_{pk}^{pn} dF(t) \leq \int_s^1 R_{\hat{p}k}^{\hat{p}n} dF(t), \quad \forall s.$$

Thus, if k and n both increase in a way that maintains k/n constant, the reduced-form auditing capacity constraint (17) slackens, widening the set of feasible auditing rules.

We can push this argument to the limit, when both k and n grow towards infinity, with k/n remaining constant. Let us denote $\kappa \in (0,1)$ this constant. The principal's value in the limit problem constitutes a lower bound for her expected loss from an optimal mechanism in any finite environment with $k/n = \kappa$. Moreover, in its reduced form, the limit problem is a special case of the mechanism-design problem we have analyzed, with rank-based auditing rule:

$$R_\kappa(t) = \mathbb{1}\left(F(t) \geq 1 - \kappa\right).$$

Now, observe that we have the equivalence:

$$\int_s^1 Q(t) dF(t) \leq \int_s^1 R_\kappa(t) dF(t), \quad \forall s \qquad \Leftrightarrow \qquad \int_0^1 Q(t) dF(t) \leq \kappa.$$

Therefore, for the limit problem, the family of constraints (17) could be replaced by the single constraint limiting the ex-ante probability of auditing any agent not to exceed κ. The Lagrange multiplier associated with this constraint then acts as a constant marginal cost of each audit. In other words, the standard *costly state verification* problem emerges by duality as a special case of our model.

This connection is intuitive. As mentioned in the Introduction, our model applies to environments such that investment in auditing capacity would have to be decided ex-ante because it is prohibitively costly to adjust in the short run. In contrast, the costly state verification model assumes that auditing capacity can be adjusted flexibly ex-post. However, with infinitely many (i.i.d.) agents, population uncertainty vanishes and there is no difference between ex-ante and ex-post from the principal's perspective. As a result, the two models coincide in this case.

The limit problem can be solved using the approach of Section 4, where we use R_κ as the rank-based auditing rule. Again, there exist μ_1 and $\underline{t}_1$ such that:

$$\forall t > \underline{t}_1, \quad Q(t) = 1 - \frac{\theta}{\overline{y}_{\mu_1}(t)},$$

where $\overline{y}_{\mu_1}(t)$ is solution to $\ell(t, y) - y(\partial\ell/\partial z)(t, y) = \mu_1$. At $\underline{t}_1$, the reduced-form auditing capacity constraint binds, so:

$$\int_{\underline{t}_1}^{1} \left(1 - \frac{\theta}{\overline{y}_{\mu_1}(t)}\right) dF(t) = \int_{\underline{t}_1}^{1} R_\kappa(t) dF(t).$$

This condition can only be satisfied if $R_\kappa(\underline{t}_1) = 0$. To see this, note that $1 - \theta/\overline{y}_{\mu_1}(t) \leq 1 - \theta < 1$, whereas R_κ is a step function jumping once from 0 to 1. As a result, for any $t \leq \underline{t}_1$, it must be the case that $Q(t) = 0$. Then, the solution is characterized by the three parameters θ, $\underline{t}_1$, and μ_1, satisfying the three conditions:

$$\int_{\underline{t}_1}^{1} \left(1 - \frac{\theta}{\overline{y}_{\mu_1}(t)}\right) dF(t) = \kappa,$$

$$\overline{y}_{\mu_1}(\underline{t}_1) \leq \theta \leq \lim_{t \to \underline{t}_1, t > \underline{t}_1} \overline{y}_{\mu_1}(t),$$

$$\int_0^{\underline{t}_1} \frac{\partial \ell}{\partial z}(t, \theta) dF(t) + \int_{\underline{t}_1}^{1} \frac{\partial \ell}{\partial z}\left(t, \overline{y}_{\mu_1}(t)\right) dF(t) = 0.$$

The optimal mechanism conducts random audits for types $t > \underline{t}_1$. In this region, audits indeed have constant marginal cost equal to μ_1. Thus, as explained previously, the corresponding mechanism is optimal for the costly state verification problem with cost μ_1.

This limit optimal mechanism shares most of its qualitative features with the solution to the principal's problem in the finite model discussed in previous sections of this paper. One notable difference though, is that audits are conducted only for types above the threshold $\underline{t}_1$, which is typically strictly greater than the best deviation payoff θ.[23] In other words, in this case, the cutoff type marking entry into the audit pool and the basic allocation to those types below the cutoff are two distinct objects. Observe that, in contrast to types below θ, which she does not value auditing, the principal's marginal benefit from auditing those types in $(\theta, \underline{t}_1)$ is positive. However, although positive, it remains below the marginal cost μ_1.[24] In the finite model, those types correspond to the second category discussed at the end of Section 4, of types left in reserve in case auditing resources permit. With infinitely many agents, auditing resources never permit to audit them, so they completely exit the audit pool.

The difference between the optimal mechanisms in the finite and infinite-limit model is illustrated in Figure 3, where we reproduce the left panel of Figure 1, plotting allocations to each type in the introductory example with $n = 2$ agents and $k = 1$ audit available. We add to the figure the allocations to each type in the limit model with $\kappa = 1/2$, under the same assumptions regarding the principal's

[23] It is easy to verify that a sufficient condition for $\underline{t}_1 > \theta$ is that, for all t, $z \mapsto \ell(t, z)$ is differentiable at $z = t$.

[24] Indeed, in this case, the condition at the lower boundary of the support of μ yields $\overline{t}_0 = \theta$. Moreover, $\mu(t)$ increases from $\mu(\theta) = 0$ to $\mu(\underline{t}_1) = \mu_1$ on $[\theta, \underline{t}_1]$.

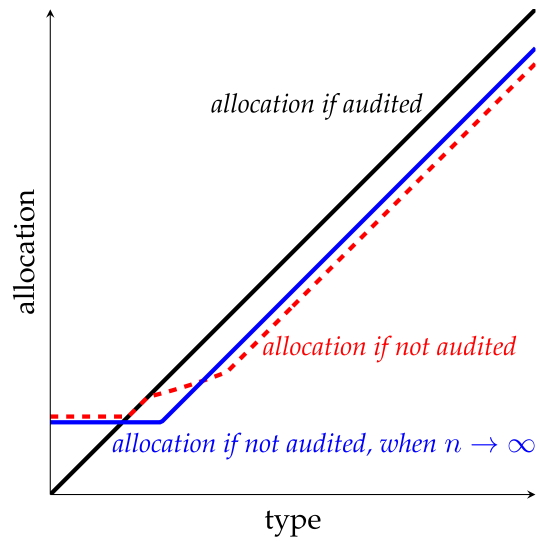


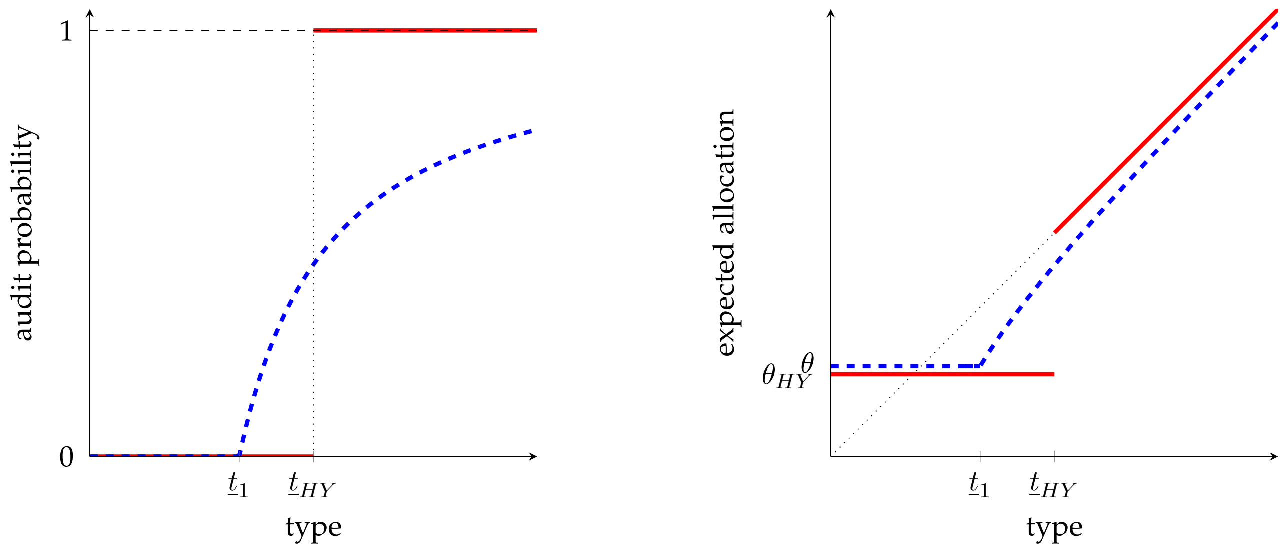
Figure 3: Amount allocated to each type by the optimal mechanism when $n = 2$ and $k = 1$ (dashed red) and when $k/n = \kappa = 1/2$ with $n \to \infty$ (solid blue). If the agent is audited, allocations coincide to the type in the two cases (solid black).

preferences and the distribution of types.

On the Value of Random Verification.— As we have explained, the optimal mechanism in the limit model is also optimal in a model with costly state verification. The specific model in question happens to be a special case of the model studied by Halac and Yared (2020), who analyze optimal delegation to a biased agent with costly state verification. In particular, given our assumption that each agent's payoff is always increasing in his allocation, our environment corresponds to the special case of an agent with extreme bias. In their analysis, Halac and Yared (2020) restrict attention to deterministic verification, while suggesting: *"the study of random verification could be an interesting extension of our work."* Since our results allow to address this suggestion directly, we take this opportunity to compare the optimal random-auditing mechanism we have obtained to the optimal deterministic-auditing mechanism derived by Halac and Yared (2020).

Let us first describe the optimal deterministic-auditing mechanism. In the case of extreme bias, it falls in the class of *threshold with escape clause* (TEC) mechanisms. Specifically, there is a threshold $\underline{t}_{HY}$, such that types below the threshold are never audited and receive a fixed basic allocation $\theta_{HY} < \underline{t}_{HY}$. Types t above the threshold are always audited and receive an allocation equal to t. Denoting ϕ the cost of auditing, the parameters θ_{HY} and $\underline{t}_{HY}$ are characterized by the following optimality conditions:

$$\int_0^{\underline{t}_{HY}} \frac{\partial \ell}{\partial z}(t, \theta_{HY}) dF(t) = 0,$$



(a) Optimal auditing rule with deterministic audits (solid, red) and with random audits (dashed, blue).

(b) Expected allocation for each type with deterministic audits (solid, red) and with random audits (dashed, blue).

Figure 4: Comparison between the principal's optimal deterministic mechanism and her optimal random mechanism with loss $\ell(t,z) = z\ln(z/t) - (z-t)$, uniformly distributed type on $[0,1]$, and cost of audit $\phi = 1/2 - 1/e$.

$$\ell(\underline{t}_{HY}, \theta_{HY}) - \phi = 0.$$

As attested by the optimal mechanism in the limit model we have described previously, the following properties of the TEC mechanism are robust to allowing random auditing. First, both the auditing rule and the allocation rule are monotonic in the agent's type. Second, there is a cutoff type, below which the agent is not audited and receives a constrained basic allocation. Third, after an audit has been conducted, the allocation features no distortion relative to the principal's ideal allocation for that type.

However, the inability to randomize considerably restricts the principal. In fact, if possible, the probability of auditing any type would remain bounded away from 1. Starting from the cutoff type, the probability of audit slowly increases from 0 to a value bounded above by $1-\theta$. Those random audits are combined with allocative distortions that are fine-tuned to the agent's type report, in case he is not audited. We illustrate those differences in Figure 4.

For a more quantitative evaluation of the value of random auditing, we also plot in Figure 5 the evolution of the principal's expected loss from the two mechanisms as the cost of audits varies.

General Auditing Cost.— We end this section with a discussion of the robustness of our approach and

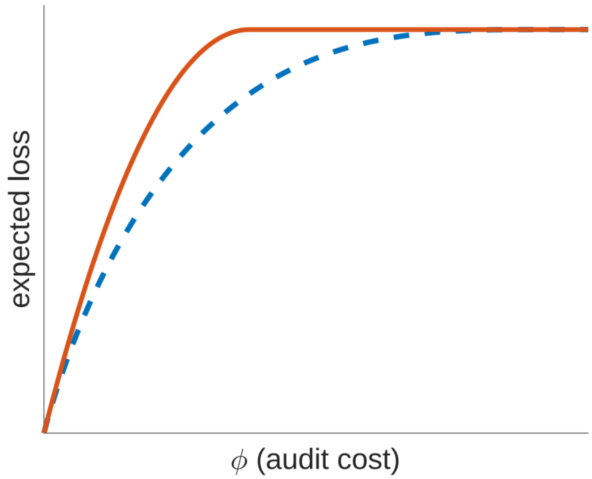


Figure 5: Expected loss to the principal as a function of the audit cost with deterministic audits (solid, red) and random audits (dashed, blue), under the same functional assumptions as for Figure 4.

results to modeling the cost of auditing via a more general function $C : \{0, 1, ..., n\} \to [0, \infty]$, which depends on the number of audited agents. In our main model, we assumed $C(m) = 0$ if $m \leq k$ and $C(m) = \infty$ if $m > k$. In this section, we have also considered the case where the cost is of the form $C(m) = \phi m$, for $\phi > 0$, and saw that the optimal mechanism in this case is a limit of optimal mechanisms in our main model, thus sharing most of their properties.

More generally, as long as C (i) is monotonic, and (ii) features increasing marginal differences (that is, $C(m + 1) - C(m)$ is monotonic), the problem can be analyzed following an approach similar to the one we have discussed in Section 3. Specifically, Lemmata 1 and 2 apply verbatim. Furthermore, solutions to the principal's problem can be characterized by optimality conditions analogous to those of Proposition 1, where the Lagrange multiplier $\psi(t_1, ..., t_n)$ is to be replaced by the marginal cost of an additional audit. Thus, the ex-post auditing rule still has the same structure, where agents are ranked by their $\lambda_i(t_i)$ and are audited not until the capacity is exhausted but until the marginal cost of an additional audit exceeds the marginal benefit. Moreover, the monotonicity results of Lemma 3 and Proposition 2 remain valid.

However, the reduced-form characterization of interim auditing rules does not apply in this case. We do not mean to suggest that the problem would necessarily remain untractable, but that the best approach to get to an explicit solution would likely depend on the specific cost function C.

That said, the two cases of costs that we have discussed are two polar cases in the aforementioned class of cost functions. Yet, they lead to fairly similar results. Therefore, we find it reasonable to speculate that any cost 'in between' would also lead to fairly similar results.

6 Conclusion

This paper has explored the problem of designing optimal mechanisms for resource allocation when auditing capacity is limited. We have laid out a tractable approach to solving the principal's mechanism design problem. This approach provides insights into the use of targeted random auditing and allocative distortions as complementary instruments to achieve compliance and mitigate efficiency losses despite restricted auditing resources. Our approach applies under fairly weak assumptions and has been demonstrated to be operational in Section 4. The simplicity of the model offers a clear benchmark to understand the nature and sources of allocative distortions. In particular, the optimal mechanism relies on upward distortions to reward conservative reporting and downward distortions to discipline unverified over-reporting.

While our analysis is rooted in a stylized environment, the techniques and findings have potential relevance for more applied contexts, including taxation, regulation, and organizational design. In these settings, the efficient use of scarce auditing resources is often a pressing concern, and the insights provided here may serve as a foundation for future investigations into policy design.

That said, several limitations and directions for future research warrant attention. The model abstracts away from global constraints on resource allocation, focusing instead on the limited capacity for auditing. In some applied settings, such allocation constraints as well as payoff externalities are likely to play a significant role. Extending the framework to incorporate these elements would broaden its applicability. Furthermore, our analysis relies on a strong form of commitment from the principal, especially regarding severe punishments for dishonesty and distortions for unaudited agents. While these features are essential for the mechanism's effectiveness, they may pose practical challenges in implementation. A valuable extension would study the mechanism design problem under partial commitment.

References

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

Ben-Porath, E., Dekel, E., & Lipman, B. L. (2019). Mechanisms with evidence: Commitment and robustness. *Econometrica*, *87*, 529–566.

Bond, P., & Hagerty, K. (2010). Preventing crime waves. *American Economic Journal: Microeconomics*, *2*, 138–159.

Border, K. C. (1991). Implementation of reduced form auctions: A geometric approach. *Econometrica*, *59*, 1175–1187.

Border, K. C., & Sobel, J. (1987). Samurai accountant: A theory of auditing and plunder. *The Review of Economic Studies*, *54*(4), 525–540.

Carroll, G., & Egorov, G. (2019). Strategic communication with minimal verification. *Econometrica*, *87*(6), 1867-1892. doi: https://doi.org/10.3982/ECTA15712

Che, Y.-K., Kim, J., & Mierendorff, K. (2013). Generalized reduced-form auctions: A network-flow approach. *Econometrica*, *81*(6), 2487-2520.

Clarke, F. H. (1990). *Optimization and nonsmooth analysis*. SIAM.

Erlanson, A., & Kleiner, A. (2020). Costly verification in collective decisions. *Theoretical Economics*, *15*, 923–954.

Erlanson, A., & Kleiner, A. (2024). Optimal allocations with capacity constrained verification. *Working Paper*.

Fang, F. C., & Casadevall, A. (2016). Research funding: the case for a modified lottery. *mBio*, *7*(2), 10.1128/mbio.00422-16.

Gale, D., & Hellwig, M. (1985). Incentive-compatible debt contracts: the one-period problem. *Review of Economic Studies*, *52 (4)*, 647–663.

Glazer, J., & Rubinstein, A. (2004). On optimal rules of persuasion. *Econometrica*, *72 No. 6*, 1715–1736.

Halac, M., & Yared, P. (2020). Commitment versus flexibility with costly verification. *Journal of Political Economy*, *128*(12), 4523–4573.

Harris, M., & Raviv, A. (1996). The capital budgeting process: Incentives and information. *The Journal of Finance*, *51*(4), 1139–1174.

Hiriart-Urruty, J.-B., & Lemaréchal, C. (2001). *Fundamentals of convex analysis* (1st ed.). Berlin, Heidelberg: Springer Berlin, Heidelberg. (Based on volume 305 and 306 in the series: Grundlehren der mathematischen Wissenschaften, 1993. Published as part of the Springer Book Archive. Copyright: Springer-Verlag Berlin Heidelberg 2001.) doi: 10.1007/978-3-642-56468-0

Kattwinkel, D., & Knoepfle, J. (2023). Costless information and costly verification: A case for transparency. *Journal of Political Economy*, *131*(2), 504–548.

Li, Y. (2020). Mechanism design with costly verification and limited punishments. *Journal of Economic Theory*, *186*, 105000.

Li, Y. (2021). Mechanism design with financially constrained agents and costly verification. *Theoretical Economics*, *16*, 1139–1194.

Luenberger, D. G. (1997). *Optimization by vector space methods*. John Wiley & Sons.

Maskin, E., & Riley, J. (1984). Optimal auctions with risk averse buyers. *Econometrica*, *52 No.6*, 1473–1518.

Matthes, F., Mehling, M., & Duan, M. (2021). *Emissions trading in practice: A handbook on design and implementation*. World Bank Group.

Mookherjee, D., & Png, I. (1989). Optimal auditing, insurance and redistribution. *The Quarterly Journal of Economics*, *104*(4), 399–415.

Mookherjee, D., & Png, I. (1994). Marginal deterrence in enforcement of law. *The Journal of Political Economy*, *102*(5), 1039–1066.

Mylovanov, T., & Zapechelnyuk, A. (2017, September). Optimal allocation with ex post verification and limited penalties. *American Economic Review*, *107*(9), 2666-94.

Patel, R., & Urgun, C. (2022). Costly verification and money burning. *Working Paper*.

Strausz, R. (2006). Timing of verification procedures: Monitoring versus auditing. *Journal of Economic Behavior & Organization*, *59*(1), 89-107.

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

APPENDIX

A Omitted Proofs

A.1 Proof of Lemma 1

Let $m = (q, \{Q_i\}_{1\le i\le n}, \{x_i\}_{1\le i\le n}, \{y_i\}_{1\le i\le n}) \in \mathcal{M}$. For each i, define:

$$\tilde{\theta}_i \quad = \quad \sup_{t_i} \quad [1 - Q_i(t_i)]y_i(t_i).$$

Combining this definition with the incentive-compatibility constraints (3), we obtain that, for any $t_i \in [0, 1]$, it must be the case that:

$$Q_i(t_i)x_i(t_i) + [1 - Q_i(t_i)]y_i(t_i) \quad \ge \quad \tilde{\theta}_i \quad \ge \quad [1 - Q_i(t_i)]y_i(t_i). \tag{25}$$

Observe that those inequalities imply that $\tilde{\theta}_i \in [0, \infty)$.

Now, for each agent i and type t_i, let us define the value:

$$\begin{aligned} \omega_i(t_i, \tilde{\theta}_i) = \min_{x,y\ge 0} \quad & Q_i(t_i)\ell_i(t_i, x) + [1 - Q_i(t_i)]\ell_i(t_i, y), \\ & \text{subject to } Q_i(t_i)x + [1 - Q_i(t_i)]y \ge \tilde{\theta}_i \ge [1 - Q_i(t_i)]y. \end{aligned}$$

Due to the inequalities established in (25), type t_i of agent i's allocations $x_i(t_i)$ and $y_i(t_i)$ are feasible for this minimization problem. As a result, it must be the case that type t_i of agent i's contribution to the principal's expected loss satisfies:

$$Q_i(t_i)\ell_i\big(t_i, x_i(t_i)\big) + \big[1 - Q_i(t_i)\big]\ell_i\big(t_i, y_i(t_i)\big) \ge \omega_i(t_i, \tilde{\theta}_i).$$

Furthermore, we establish below that $\omega_i(t_i, \tilde{\theta}_i) \ge \omega_i(t_i, \theta_i)$, where $\theta_i = \min\{\tilde{\theta}_i, 1\} \in [0, 1]$. As a result, the principal would be weakly better off replacing the allocations $x_i(t_i)$ and $y_i(t_i)$ by solutions to the minimization problem defining the value $\omega_i(t_i, \theta_i)$. We show next that those turn out to be exactly the allocations $x_i^*(t_i)$ and $y_i^*(t_i)$ implemented by m_θ, thus establishing the result of the lemma.

In the remainder of this proof, we analyze the minimization problem defining $\omega_i(t_i, \tilde{\theta}_i)$ in order to prove those results that we promised. We consider in turn two cases.

Suppose first that $\tilde{\theta}_i \geq t_i$. For feasible x and y, the convexity of $\ell_i(t_i, \cdot)$ implies that:

$$Q_i(t_i)\ell_i(t_i, x) + [1 - Q_i(t_i)]\ell_i(t_i, y) \geq \ell_i\Big(t_i, Q_i(t_i)x + [1 - Q_i(t_i)]y\Big).$$

In addition, since $\ell_i(t_i, \cdot)$ is uniquely minimized at t_i and convex, it must be increasing above t_i. Using the fact that $Q_i(t_i)x + [1 - Q_i(t_i)]y \geq \tilde{\theta}_i \geq t_i$, we obtain:

$$\ell_i\Big(t_i, Q_i(t_i)x + [1 - Q_i(t_i)]y\Big) \geq \ell_i(t_i, \tilde{\theta}_i).$$

As a result, since $x = y = \tilde{\theta}_i$ is feasible for this minimization problem, it must be the solution. We conclude that, for $\tilde{\theta}_i \geq t_i$, we have:

$$\omega_i(t_i, \tilde{\theta}_i) = \ell_i(t_i, \tilde{\theta}_i).$$

In particular, as explained previously, this expression is decreasing in $\tilde{\theta}_i$ on the domain $\tilde{\theta}_i \geq t_i$. It follows that, if $\tilde{\theta}_i > 1$, since all types satisfy $t_i \leq 1$, then:

$$\omega_i(t_i, \tilde{\theta}_i) \geq \omega_i(t_i, 1) = \omega_i(t_i, \theta_i).$$

Since the other case $\tilde{\theta}_i \leq 1$ is trivial, this establishes that indeed, in general, $\omega_i(t_i, \tilde{\theta}_i) \geq \omega_i(t_i, \theta_i)$. Furthermore, for $\theta_i \geq t_i$, we have $x_i^*(t_i) = y_i^*(t_i) = \theta_i$, thus $x_i^*(t_i)$ and $y_i^*(t_i)$ indeed achieve the value $\omega_i(t_i, \theta_i)$.

Next, we consider the case $t_i > \tilde{\theta}_i = \theta_i$. Let us show that $x_i^*(t_i)$ and $y_i^*(t_i)$ constitute a solution to the minimization problem defining $\omega_i(t_i, \theta_i)$. First, note that a solution (x, y) to this problem satisfies $y \geq \theta_i$. To see this, note that if (x, y) is feasible with $y < \theta_i$, then (x, θ_i) remains feasible. Furthermore, since $\ell_i(t_i, \cdot)$ is uniquely minimized at t_i and convex, it must be decreasing below t_i. Using the fact that $y < \theta_i < t_i$, we conclude that, in addition to being feasible, (x, θ_i) also achieves a lower value than (x, y). From now on, we restrict attention to those (x, y) that are feasible with $y \geq \theta_i$. In particular, for any such y, observe that $x_i^*(t_i) = t_i > \theta_i$ satisfies: $Q_i(t_i)x_i^*(t_i) + [1 - Q_i(t_i)]y \geq \theta_i$. That is, it is feasible to set $x = x_i^*(t_i)$, where $\ell_i(t_i, x)$ is minimized. Finally, consider the optimal choice of y. If feasible, the solution must be at $y = t_i$. Feasibility is satisfied if $[1 - Q_i(t_i)]t_i \leq \theta_i$, that is exactly when $y_i^*(t_i) = t_i$.

Otherwise, y must be chosen such that $y \leq \theta_i/[1 - Q_i(t_i)] < t_i$. As explained previously, $\ell_i(t_i, y)$ is decreasing in y below t_i. As a result, it is optimal to set $y = \theta_i/[1 - Q_i(t_i)]$, which corresponds again to $y_i^*(t_i)$ in this case. Thus, in any case, we find that it is indeed optimal to use allocations $x_i^*(t_i)$ and $y_i^*(t_i)$.

A.2 Proof of Proposition 1

The result follows from standard arguments. We refer to Luenberger (1997), chapters 8 and 9 for details. The principal's choice set $\widetilde{\mathcal{M}}$ is a subset of the normed vector space $X = \mathbb{R} \times \mathcal{B}_{\mathbf{F}}([0,1]^n) \times \prod_{i=1}^n \mathcal{B}_{F_i}([0,1])$ equipped with the sup norm, where $\mathcal{B}_{\mathbf{F}}([0,1]^n)$ refers to the set of bounded and $(F_1, ..., F_n)$-measurable functions on $[0,1]^n$ and, for each i, $\mathcal{B}_{F_i}([0,1])$ is the set of bounded and F_i-measurable functions on $[0,1]$. All constraints are linear and define a convex subset of X. Note also that an immediate consequence of Lemma 5 in Appendix A.5 is that the principal's objective function is convex.

To guarantee that the Slater condition is satisfied, we relax constraints (5). We verify later that it is satisfied by a solution to conditions (7)-(14).

The dual variables associated to the constraints $\theta_i \geq 0$ and $\theta_i \leq 1$, for all i, can be identified with non-negative real numbers, denoted γ_i and δ_i in the statement of the proposition. The dual variables associated to the auditing capacity constraint (1), as well as $q_i(t_1, ..., t_n) \geq 0$ and $q_i(t_1, ..., t_n) \leq 1$, for all i and $(t_1, .., t_n)$, are measures that are absolutely continuous with respect to $(F_1, ..., F_n)$. In the statement of the proposition, we formulate optimality conditions immediately in terms of their Radon-Nikodym derivatives ψ, α_i, and β_i respectively. Similarly, for each i, λ_i is the Radon-Nikodym derivative of the signed measure which acts as a Lagrange multiplier associated to constraint (2).

For necessity, conditions (7)-(9) are valid by Assumption 3. Condition (10) can be taken as a point-wise definition of λ_i. Then, conditions (11) - (14) are required to hold $(F_1, ..., F_n)$-almost surely. However, we can always select representatives for ψ, α_i, and β_i such that they are valid point-wise. Sufficiency can be established equally well in either case.

It remains to verify that a solution to conditions (7)-(14) satisfies:

$$\forall i \forall t_i, \quad Q_i(t_i) \leq \max\{0, 1 - \theta_i/t_i\}.$$

For some i and $t_i > 0$, suppose that $Q_i(t_i) > 1 - \theta_i/t_i$, that is $\theta_i/[1 - Q_i(t_i)] > t_i$. With reference to condition (10), we have:

$$\begin{aligned}\lambda_i(t_i) &= \ell_i\left(t_i, \frac{\theta_i}{1-Q_i(t_i)}\right) + \left[t_i - \frac{\theta_i}{1-Q_i(t_i)}\right]\frac{\partial \ell_i}{\partial z_i}\left(t_i, \frac{\theta_i}{1-Q_i(t_i)}\right) - t_i\frac{\partial \ell_i}{\partial z_i}\left(t_i, \frac{\theta_i}{1-Q_i(t_i)}\right)\\ &< \ell_i\left(t_i, \frac{\theta_i}{1-Q_i(t_i)}\right) + \left[t_i - \frac{\theta_i}{1-Q_i(t_i)}\right]\frac{\partial \ell_i}{\partial z_i}\left(t_i, \frac{\theta_i}{1-Q_i(t_i)}\right),\end{aligned}$$

where the inequality follows from the fact that $\ell_i(t_i, \cdot)$ is single-dipped at t_i, with $\theta_i/[1-Q_i(t_i)] > t_i$, so $(\partial \ell_i/\partial z_i)\,(t_i, \theta_i/[1-Q_i(t_i)]) > 0$. Observe that on the right-hand side of the inequality is the expression of a tangent to $\ell_i(t_i, \cdot)$ evaluated at t_i. By convexity, we have:

$$\ell_i\left(t_i, \frac{\theta_i}{1-Q_i(t_i)}\right) + \left[t_i - \frac{\theta_i}{1-Q_i(t_i)}\right]\frac{\partial \ell_i}{\partial z_i}\left(t_i, \frac{\theta_i}{1-Q_i(t_i)}\right) \leq \ell_i(t_i, t_i) = 0,$$

implying that $\lambda_i(t_i) < 0$. Now, with reference to condition (11), since for any t_{-i}, $\mu(t_i, t_{-i}) \geq 0$, it must be the case that $\alpha_i(t_i, t_{-i}) > 0$, which implies, by complementary slackness, that $q_i(t_i, t_{-i}) = 0$. Integrating this result, it follows that $Q_i(t_i) = 0$. In other words, we have shown that:

$$Q_i(t_i) > 1 - \theta_i/t_i \quad \Rightarrow \quad Q_i(t_i) = 0,$$

which is equivalent to:

$$Q_i(t_i) \leq \max\{0, 1 - \theta_i/t_i\}.$$

For $t_i = 0$, note that the same reasoning implies that $\lambda_i(t_i) \leq 0$. If $Q_i(t_i) > 0$, we can define $\hat{q}_i(0, t_{-i}) = 0$ for all t_{-i}, and $\hat{q}_i = q_i$ on the rest of the domain. Similarly, let $\hat{Q}_i(0) = 0$ and $\hat{Q}_i = Q_i$ on $(0, 1]$. Note that $(\theta, \hat{q}, \{\hat{Q}_i\}_i)$ satisfy the optimality conditions (7)-(14). Applying this construction for all i yields the result.

A.3 Proof of Lemma 3

The proof relies on the fact that, the higher λ_i, the more likely it is that agent i gets audited, as explained in the paragraph preceding the statement of the Lemma.

To establish monotonicity above θ_i, let us suppose, by contradiction, that $t_i' > t_i \geq \theta_i$ with $\lambda_i(t_i') < \lambda_i(t_i)$. Then, type t_i must be more likely to get audited than type t_i'. As a result, it must be the case that $Q_i(t_i) \geq Q_i(t_i')$, thus $t_i \geq \theta_i/[1-Q_i(t_i)] \geq \theta_i/[1-Q_i(t_i')] \geq 0$. Using condition (10) and the convexity of

$\ell_i(t_i, \cdot)$, we write:

$$\begin{aligned}\lambda_i(t_i) = \ell_i\left(t_i, \frac{\theta_i}{1-Q_i(t_i)}\right) + \left[\frac{\theta_i}{1-Q_i(t_i')} - \frac{\theta_i}{1-Q_i(t_i)}\right]\frac{\partial \ell_i}{\partial z_i}\left(t_i, \frac{\theta_i}{1-Q_i(t_i)}\right) \\ - \frac{\theta_i}{1-Q_i(t_i')}\cdot\frac{\partial \ell_i}{\partial z_i}\left(t_i, \frac{\theta_i}{1-Q_i(t_i)}\right) \\ \leq \ell_i\left(t_i, \frac{\theta_i}{1-Q_i(t_i')}\right) - \frac{\theta_i}{1-Q_i(t_i')}\cdot\frac{\partial \ell_i}{\partial z_i}\left(t_i, \frac{\theta_i}{1-Q_i(t_i)}\right) \\ \leq \ell_i\left(t_i, \frac{\theta_i}{1-Q_i(t_i')}\right) - \frac{\theta_i}{1-Q_i(t_i')}\cdot\frac{\partial \ell_i}{\partial z_i}\left(t_i, \frac{\theta_i}{1-Q_i(t_i')}\right),\end{aligned}$$

where the first inequality uses the fact that $\ell_i(t_i, \cdot)$ remains above its tangents, and the second the fact that $(\partial \ell_i/\partial z_i)(t_i, \cdot)$ is non-decreasing, both consequences of convexity. Next, we use Assumption 4 to show that the term on the right-hand side of the last inequality is bounded above by $\lambda_i(t_i')$, thereby establishing a contradiction. Indeed, by Assumption 4, it must be that:

$$-\frac{\partial}{\partial z_i}\left(\frac{\ell_i(\tilde{t}_i, z_i)}{z_i}\right) = \frac{1}{z_i^2}\left[\ell_i(\tilde{t}_i, z_i) - z_i\frac{\partial \ell_i}{\partial z_i}(\tilde{t}_i, z_i)\right]$$

is non-decreasing in $\tilde{t}_i$. Applying this fact to $z_i = \theta_i/[1-Q_i(t_i')]$, with $t_i' > t_i$, we obtain:

$$\begin{aligned}\ell_i\left(t_i, \frac{\theta_i}{1-Q_i(t_i')}\right) - \frac{\theta_i}{1-Q_i(t_i')}\cdot\frac{\partial \ell_i}{\partial z_i}\left(t_i, \frac{\theta_i}{1-Q_i(t_i')}\right) \\ \leq \ell_i\left(t_i', \frac{\theta_i}{1-Q_i(t_i')}\right) - \frac{\theta_i}{1-Q_i(t_i')}\cdot\frac{\partial \ell_i}{\partial z_i}\left(t_i', \frac{\theta_i}{1-Q_i(t_i')}\right).\end{aligned}$$

Recognizing on the right-hand side the value of $\lambda_i(t_i')$, from condition (10), we conclude that $\lambda_i(t_i) \leq \lambda_i(t_i')$, the desired contradiction.

A.4 Proof of Proposition 2

Consider a solution $(\theta, q, \{Q_i\}_i)$ to problem (6). If, for some i, there exist $t_i' > t_i$ with $Q_i(t_i) > Q_i(t_i')$, we will show that we can construct another solution $(\theta, \hat{q}, \{\hat{Q}_j\}_j)$ such that $\hat{Q}_i(t_i) \leq \hat{Q}_i(t_i')$, and, for all $\tilde{t}_i \neq t_i$, $\hat{Q}_i(\tilde{t}_i) = Q_i(\tilde{t}_i)$, as well as, for all $j \neq i$ and all t_j, $\hat{Q}_j(t_j) = Q_j(t_j)$. Applying this construction to any such pair (t_i, t_i') yields a solution with a monotonic interim auditing rule.

As a solution to problem (6), $(\theta, q, \{Q_i\}_i)$ satisfies the optimality conditions of Proposition 1, with appropriately defined Lagrange multipliers. In particular, for each i, types with a strictly higher λ_i are more likely to be audited. Thus, if $Q_i(t_i) > Q_i(t_i')$, it must be the case that $\lambda_i(t_i) \geq \lambda_i(t_i')$. However, if in addition $t_i' > t_i$, then, by Lemma 3, we have $\lambda_i(t_i') = \lambda_i(t_i)$. As a result, with reference to Appendix A.3, the chain of inequalities obtained in the proof of Lemma 3 is a chain of equalities. It follows that $\ell_i(t_i, \cdot)$ must be linear on $[\theta_i/[1 - Q_i(t_i')], \theta_i/[1 - Q_i(t_i)]]$.

Given this linearity, we will be able to establish the optimality of a mechanism $(\theta, \hat{q}, \{\hat{Q}_j\}_j)$ which differs from $(\theta, q, \{Q_j\}_j)$ only in terms of the probability of auditing type t_i of agent i. Specifically, set $\hat{Q}_i(t_i) = Q_i(t_i')$. This can be implemented at the ex-post stage by setting:

$$\hat{q}_i(t_i, t_{-i}) = \frac{Q_i(t_i')}{Q_i(t_i)} q_i(t_i, t_{-i}),$$

for all t_{-i}. Note that the auditing capacity constraint is then relaxed, so this transformation does not jeopardize feasibility.

Optimality follows from verifying that all the conditions of Proposition 1 are satisfied. With respect to the Lagrange multipliers, the only required changes are:

$$\hat{\alpha}_i(t_i, t_{-i}) = \alpha_i(t_i', t_{-i}), \quad \hat{\beta}_i(t_i, t_{-i}) = \beta_i(t_i', t_{-i}), \quad \hat{\psi}(t_i, t_{-i}) = \psi(t_i', t_{-i}).$$

Due to the aforementioned linearity, condition (7) holds with:

$$\frac{\partial \ell_i}{\partial z_i}\left(t_i, \frac{\theta_i}{1 - Q_i(t_i)}\right) = \frac{\partial \ell_i}{\partial z_i}\left(t_i, \frac{\theta_i}{1 - \hat{Q}_i(t_i)}\right).$$

Conditions (8) and (9) are clearly unaffected. Since $\lambda_i(t_i) = \lambda_i(t_i')$, condition (10) remains unchanged. Finally, the Lagrange multipliers have been appropriately modified to guarantee that conditions (11)-(14) continue to hold.

A.5 Additional Results

Lemma 5. *Let $f : (0, \infty) \to \mathbb{R}$ convex. Then $g : (0, \infty) \times (0, \infty) \to \mathbb{R}$, defined by $g(x, y) = x f(y/x)$, is also convex.*

Proof. Let x, $\hat{x}$, y, $\hat{y}$ all in $(0, \infty)$, and $\lambda \in [0,1]$. Observe that:

$$\frac{\lambda y + (1-\lambda)\hat{y}}{\lambda x + (1-\lambda)\hat{x}} = \frac{\lambda x}{\lambda x + (1-\lambda)\hat{x}} \cdot \frac{y}{x} + \frac{(1-\lambda)\hat{x}}{\lambda x + (1-\lambda)\hat{x}} \cdot \frac{\hat{y}}{\hat{x}},$$

with $\lambda x/[\lambda x + (1-\lambda)\hat{x}] \in [0,1]$. As a result, given that f is convex, we have:

$$[\lambda x + (1-\lambda)\hat{x}] f\left(\frac{\lambda y + (1-\lambda)\hat{y}}{\lambda x + (1-\lambda)\hat{x}}\right) \leq \lambda x f\left(\frac{y}{x}\right) + (1-\lambda)\hat{x} f\left(\frac{\hat{y}}{\hat{x}}\right).$$

In other words:

$$g(\lambda x + (1-\lambda)\hat{x}, \lambda y + (1-\lambda)\hat{y}) \leq \lambda g(x,y) + (1-\lambda) g(\hat{x}, \hat{y}),$$

establishing the result. $\square$

B Examples of Loss Functions

In this appendix, we verify that the examples listed in Section 2 satisfy Assumptions 1, 2, 3, and 4.

(i) Suppose that, for all i, $\ell_i(t_i, z_i) = c_i(t_i - z_i)$, where c_i is convex, uniquely minimized at 0, with $c_i(0) = 0$. Assumptions 1 and 2 are obviously satisfied. Assumption 3 follows from noting that $(\partial \ell_i/\partial z_i)(t_i, z_i) = -c_i'(t_i - z_i)$, which is monotonically decreasing in t_i, and bounded for $t_i \in [0,1]$. Let us verify Assumption 4. To that end, let $(t_i, z_i) \in \mathcal{D}$ and $(\hat{t}_i, \hat{z}_i) \in \mathcal{D}$. Without loss, suppose that $\hat{t}_i \geq t_i$. It is sufficient to verify that, if $\hat{z}_i < z_i$, then:

$$\frac{\ell_i(t_i, z_i)}{z_i} + \frac{\ell_i(\hat{t}_i, \hat{z}_i)}{\hat{z}_i} \geq \frac{\ell_i(t_i, \hat{z}_i)}{\hat{z}_i} + \frac{\ell_i(\hat{t}_i, z_i)}{z_i},$$

that is:

$$\frac{c_i(t_i - z_i)}{z_i} + \frac{c_i(\hat{t}_i - \hat{z}_i)}{\hat{z}_i} \geq \frac{c_i(t_i - \hat{z}_i)}{\hat{z}_i} + \frac{c_i(\hat{t}_i - z_i)}{z_i}.$$

Observe that, in this case, we can write:

$$t_i - \hat{z}_i = \frac{z_i - \hat{z}_i}{\hat{t}_i - t_i + z_i - \hat{z}_i}(\hat{t}_i - \hat{z}_i) + \frac{\hat{t}_i - t_i}{\hat{t}_i - t_i + z_i - \hat{z}_i}(t_i - z_i),$$

where $(z_i - \hat{z}_i)/(\hat{t}_i - t_i + z_i - \hat{z}_i) \in [0,1]$, and similarly:

$$\hat{t}_i - z_i = \frac{z_i - \hat{z}_i}{\hat{t}_i - t_i + z_i - \hat{z}_i}(t_i - z_i) + \frac{\hat{t}_i - t_i}{\hat{t}_i - t_i + z_i - \hat{z}_i}(\hat{t}_i - \hat{z}_i).$$

Therefore, using the convexity of c_i, we have:

$$c_i(t_i - \hat{z}_i) \leq \frac{z_i - \hat{z}_i}{\hat{t}_i - t_i + z_i - \hat{z}_i} c_i(\hat{t}_i - \hat{z}_i) + \frac{\hat{t}_i - t_i}{\hat{t}_i - t_i + z_i - \hat{z}_i} c_i(t_i - z_i),$$

$$c_i(\hat{t}_i - z_i) \leq \frac{z_i - \hat{z}_i}{\hat{t}_i - t_i + z_i - \hat{z}_i} c_i(t_i - z_i) + \frac{\hat{t}_i - t_i}{\hat{t}_i - t_i + z_i - \hat{z}_i} c_i(\hat{t}_i - \hat{z}_i).$$

Those two inequalities, together with simple algebraic manipulations imply:

$$\frac{c_i(t_i - z_i)}{z_i} + \frac{c_i(\hat{t}_i - \hat{z}_i)}{\hat{z}_i} - \frac{c_i(t_i - \hat{z}_i)}{\hat{z}_i} - \frac{c_i(\hat{t}_i - z_i)}{z_i} \geq \frac{(\hat{t}_i - t_i)(z_i - \hat{z}_i)}{z_i \hat{z}_i (\hat{t}_i - t_i + z_i - \hat{z}_i)} [c_i(\hat{t}_i - \hat{z}_i) - c_i(t_i - z_i)].$$

Furthermore, by assumption $\hat{t}_i \geq t_i > z_i > \hat{z}_i$, with c_i increasing on $[0, \infty)$ (since it is convex and uniquely minimized at 0). It follows that:

$$\frac{c_i(t_i - z_i)}{z_i} + \frac{c_i(\hat{t}_i - \hat{z}_i)}{\hat{z}_i} - \frac{c_i(t_i - \hat{z}_i)}{\hat{z}_i} - \frac{c_i(\hat{t}_i - z_i)}{z_i} \geq 0,$$

which is what we wanted to show.

(ii) Suppose that, for all i, $\ell_i(t_i, z_i) = c_i(z_i) - c_i(t_i) - (z_i - t_i)c_i'(t_i)$, where c_i is a differentiable convex function. Assumptions 1 and 3 are easy to verify given c_i's convexity, which implies that $c_i'(t_i)$ is monotonic in t_i. Assumption 2 is obviously satisfied. To verify Assumption 4, note that $\ell_i(t_i, z_i)$ is differentiable in z_i, thus so is the average loss, with:

$$\frac{\partial}{\partial z_i}\left(\frac{\ell_i(t_i, z_i)}{z_i}\right) = \frac{[z_i c_i'(z_i) - c_i(z_i)] - [t_i c_i'(t_i) - c_i(t_i)]}{z_i^2}.$$

It is sufficient to show that this expression is non-increasing in t_i, which implies Assumption 3. Thus, consider $t_i < \hat{t}_i$. By convexity:

$$c_i(t_i) \geq c_i(\hat{t}_i) + (t_i - \hat{t}_i)c_i'(\hat{t}_i),$$

which can be rearranged as:

$$\hat{t}_i c_i'(\hat{t}_i) - c_i(\hat{t}_i) \geq t_i c_i'(t_i) - c_i(t_i) + t_i[c_i'(\hat{t}_i) - c_i'(t_i)],$$

which, by monotonicity of c_i' implies:

$$\hat{t}_i c_i'(\hat{t}_i) - c_i(\hat{t}_i) \geq t_i c_i'(t_i) - c_i(t_i),$$

which implies the result.

C Marginal Representation of Auditing Rules

This Appendix explains why it is sufficient to represent ex-post auditing rules by the marginal probability of auditing each agent.

The set of all subsets of $\{1, ..., n\}$ of size k is denoted $\mathcal{P}_k(n)$, while $\mathcal{P}_{\leq k}(n)$ denotes the set of all subsets of size at most k. For any finite set S, the set of probability distributions on S is denoted ΔS.

If $p \in \Delta\mathcal{P}_{\leq k}(n)$ and $i \in \{1, ..., n\}$, we can define:

$$q_i = \sum_{S \in \mathcal{P}_{\leq k}(n)} \mathbb{1}(i \in S) p_S.$$

Then, $q \in [0, 1]^n$ and:

$$\sum_{i=1}^{n} q_i \leq k.$$

Reciprocally, if $q \in [0, 1]^n$ with $\sum_{i=1}^{n} q_i \leq k$, we will show below that there exists $p \in \Delta\mathcal{P}_{\leq k}(n)$ such that:

$$\forall i, \quad q_i = \sum_{S \in \mathcal{P}_{\leq k}(n)} \mathbb{1}(i \in S) p_S.$$

For all n and $k \leq n$, we denote $[k, n]$ the claim:

$$\forall q \in [0, 1]^n, \quad \sum_{i=1}^{n} q_i = k \quad \Rightarrow \quad \exists p \in \Delta\mathcal{P}_k(n), \quad \forall i, \quad q_i = \sum_{S \in \mathcal{P}_k(n)} \mathbb{1}(i \in S) p_S.$$

Note that it is sufficient to prove the claim $[k, n]$ for all k and n. To see this, observe that, if $k - 1 <$

$\sum_{i=1}^{n} q_i < k$, we can define $q_{n+1} = k - \sum_{i=1}^{n} q_i \in (0,1)$ and apply $[k, n+1]$. Subsets of size k of $\{1, ..., n+1\}$ have size at most k if we remove the fictitious agent $n+1$.

Next, we will prove $[k, n]$ by induction. Observe that it is sufficient to prove the result for $k \leq n/2$. Indeed, defining $\tilde{q}_i = 1 - q_i$, it is easy to see that $[k, n]$ is implied by $[n-k, n]$.

$k = 0$.— For $k = 0$, we must have $q_1 = q_2 = ... = q_n = 0$, which is consistent with a probability distribution placing mass 1 on the empty subset of $\{1, ..., n\}$. So $[0, n]$ holds for all n.

Inductive step.— Fix q satisfying the premise. Without loss, assume that $0 < q_1 \leq q_2 \leq ... \leq q_n < 1$.[25] As a consequence $q_1 \leq k/n$. If $q_1 = k/n$, then $q_i = k/n$ for all i. In this case, we can choose p the uniform distribution on $\mathcal{P}_k(n)$. In what follows, assume $0 < q_1 < k/n$. As a result, it is also the case that $k/n < q_n < 1$. We consider two cases.

<u>Case 1</u>: $(n-k)q_1 + kq_n \leq k$. Let $a, b \in [0,1]^n$ such that:

$$\forall i, \quad a_i = k/n, \quad b_i = \frac{q_i - q_1}{1 - q_1 n/k}.$$

It is clear that $a_i \in [0,1]$ and $b_i \geq 0$. To verify that $b_i \leq 1$, note that b_i is non-decreasing in i, so it is sufficient to verify that $b_n \leq 1$. Using the assumption $(n-k)q_1 + kq_n \leq k$, observe that:

$$b_n \leq \frac{1 - q_1(n-k)/k - q_1}{1 - q_1 n/k} = 1.$$

In addition, it is easy to verify that:

$$\sum_{i=1}^{n} a_i = \sum_{i=1}^{n} b_i = k.$$

Finally, note that:

$$q = (n/k)q_1 a + (1 - q_1 n/k)b,$$

with $0 < q_1 n/k < 1$. a can be implemented by a uniform distribution on $\mathcal{P}_k(n)$. $b_1 = 0$, so b can be implemented by a solution to $[k, n-1]$. By mixing between the two, we can implement q.

[25]We can always re-order the agents. If $q_1 = 0$ we easily get back to the case $[k, n-1]$. If $q_n = 1$ we easily get back to the case $[k-1, n-1]$.

Case 2: $(n-k)q_1 + kq_n \geq k$. The argument is similar. Define a the same way. Let:

$$c_i = \frac{(1-k/n)q_i - (1-q_n)k/n}{q_n - k/n}.$$

We have $c_n = 1$. Similar arguments as in Case 1 establish that c can be implemented by a solution to $[k-1, n-1]$, and q is a convex combination between a and c.

D Details on the Figures of Section 5

Figures 3, 4, and 5 present explicit solutions in the case where the principal's loss takes the form $\ell(t,z) = z\ln(z/t) - (z-t)$ and types are uniformly distributed on $[0,1]$. Let us explain how the solutions are obtained in this case.

First, for Figure 3, we apply the optimality conditions for those functional forms. We have $\overline{y}_{\mu_1}(t) = t - \mu_1$, which is continuous, so we immediately obtain $\underline{t}_1 = \theta + \mu_1$. The other two optimality conditions are:

$$\int_{\theta+\mu_1}^{1} \left(1 - \frac{\theta}{t-\mu_1}\right) dt = \kappa,$$

$$\int_{0}^{\theta+\mu_1} \ln\left(\frac{\theta}{t}\right) dt + \int_{\theta+\mu_1}^{1} \ln\left(\frac{t-\mu_1}{t}\right) dt = 0.$$

With $\kappa = 1/2$ and simple algebra, the system simplifies to:

$$1 - (\theta + \mu_1) - \theta \ln\left(\frac{1-\mu_1}{\theta}\right) = \frac{1}{2},$$

$$\mu_1 \ln(\theta) + (1-\mu_1)\ln(1-\mu_1) + (\theta + \mu_1) = 0.$$

The system possesses a unique solution which can be evaluated numerically. We obtain $\theta \approx 0.1489$ and $\mu_1 \approx 0.0799$.

For Figure 4, we first solve for the parameters of the optimal deterministic mechanism. The optimality conditions are:

$$\int_{0}^{\underline{t}_{HY}} \ln\left(\frac{\theta_{HY}}{t}\right) dt = 0,$$

$$\theta_{HY} \ln(\theta_{HY}/\underline{t}_{HY}) - (\theta_{HY} - \underline{t}_{HY}) - \phi = 0.$$

The unique solution is given by $\theta_{HY} = \phi/(e-2)$ and $\underline{t}_{HY} = \phi e/(e-2)$, for $\phi \leq (e-2)/e$. The optimal random mechanism corresponds to the limit mechanism as $n \to \infty$ we have characterized previously, with marginal cost of audits exogenously fixed to $\mu_1 = \phi$. Therefore, we have $\underline{t} = \theta + \phi$, where θ is the unique solution to:

$$\phi \ln(\theta) + (1-\phi)\ln(1-\phi) + (\theta + \phi) = 0,$$

for $\phi \leq (e-1)/e$. For the figure, we use $\phi = 1/2 - 1/e$, so $\underline{t}_{HY} = 1/2$, $\theta_{HY} = 1/(2e)$, while the parameters of the optimal random mechanism can be evaluated numerically as $\theta \approx 0.2021$ and $\underline{t} \approx 0.3342$.

For Figure 5, we evaluate the principal's expected loss from the two mechanisms as a function of the cost of audits ϕ. In both cases, this loss can be expressed as:

$$L(\phi) = \int_0^1 \left[(1-Q(t))\,\ell\left(t, \frac{\theta}{1-Q(t)}\right) + Q(t)\phi\right] dt.$$

The deterministic mechanism has $Q_{HY}(t) = \mathbb{1}(t \geq \underline{t}_{HY})$. For $\phi \leq (e-2)/e$, simple algebra yields:

$$L_{HY}(\phi) = \phi(1 - \phi e/[2(e-2)]).$$

For $\phi > (e-2)/e$, $\underline{t}_{HY} = 1$, so the mechanism becomes constant as no type is ever audited. In that case, $L_{HY}(\phi) = (e-2)/(2e)$. Similarly, for the optimal random mechanism, we can compute the principal's expected loss in terms of the parameters of the mechanism and obtain:

$$L(\phi) = \phi - \underline{t}^2/2,$$

or equivalently, using the optimality conditions characterizing those parameters, $L(\phi)$ is solution to:

$$\phi \ln\left(\sqrt{2[\phi - L(\phi)]} - \phi\right) + (1-\phi)\ln(1-\phi) + \sqrt{2[\phi - L(\phi)]} = 0,$$

valid for $\phi \leq (e-1)/e$. For $\phi > (e-1)/e$, the mechanism again becomes constant and no type is audited. In this case, $L(\phi) = (e-2)/(2e)$.