Optimal Auditing in a Dynamic Model of Tax Compliance

Ravikumar, B. and Zhang, Yuzhe

University of Iowa

2010

Optimal Auditing in a Dynamic Model of Tax Compliance*

B. Ravikumar†and Yuzhe Zhang‡

February 24, 2010

Abstract

We study the optimal auditing of a taxpayer's income in a dynamic principal-agent model of hidden income. Taxpayers in our model initially have low income and stochastically transit to high income that is an absorbing state. A low-income taxpayer who transits to high income can under-report his true income and evade his taxes. With a constant absolute risk-aversion utility function and a costly auditing technology, we show that the optimal auditing mechanism in our model consists of cycles. Within each cycle, a low-income taxpayer is initially unaudited, but if the duration of low-income report exceeds a threshold, then the auditing probability becomes positive. That is, the tax authority guarantees that the taxpayer will not be audited until the threshold duration is reached. We also find that auditing becomes less frequent if the auditing cost is higher or if the variance of income is lower.

JEL Classification Numbers: D82, D86, J65.
Keywords and Phrases: Tax Compliance, Tax Auditing, Stochastic Costly State Verification.

*We are grateful to V.V. Chari, April Franco, David Fuller, Hari Govindan, Larry Jones, Marek Kapicka, Ayca Kaya, Narayana Kocherlakota, Igor Livshits, Matt Mitchell, Chris Phelan, Shouyong Shi, Galina Vereschagina, Gustavo Ventura, Cheng Wang, and seminar participants of Canadian Macroeconomics Study Group 2009 for their helpful comments.

†Department of Economics, University of Iowa. Email: Ravikumar@uiowa.edu
‡Department of Economics, University of Iowa. Email: Yuzhe-Zhang@uiowa.edu

1 Introduction

There is a large literature on tax compliance following the approach to crime and punishment developed in Becker (1968) and Stigler (1970). For instance, Reinganum and Wilde (1985, 1986) examine a static model where taxpayers' incomes are private information. Using the costly state verification framework developed by Townsend (1979), they study optimal verification schemes when the tax and the penalty are exogenously specified. In this paper, we characterize the optimal auditing and taxation scheme in a dynamic stochastic costly-state-verification environment.

We develop a model where the tax authority (principal) is risk neutral and taxpayers (agents) have constant absolute risk-averse preferences. Each agent knows his own income but it is unobserved by the principal. The principal may audit an agent to verify his income, but this is costly. The tax authority designs an optimal taxation scheme as well as an optimal auditing scheme to maximize the present value of revenue net of audit cost. Taxpayers in our model initially have low income and receive stochastic opportunities each period to transit to high income. For convenience, we assume that high income is an absorbing state. Since income is private information, the taxpayer could conceal the fact that he has transited to high income and evade taxes. The punishment, if the taxpayer is audited and caught cheating, is assumed to be a constant. We use a dynamic mechanism-design approach to search for the best tax system and auditing system within a large family of state-contingent contracts.

Our model contains persistent private information and, as demonstrated by Fernandes and Phelan (2000), the principal's problem contains two state variables: the continuation utility for an agent who just transited to high income and the continuation utility for a low-income agent. We follow Zhang (2009) and set up the principal's problem in continuous time. We then formulate the Hamiltonian and apply the Pontryagin maximum principle to study the dynamic behavior of continuation utilities.

Since high income is an absorbing state in our model, the treatment of the agent who just transited to high income is straightforward – constant consumption forever and, hence, constant continuation utility. Furthermore, he is never audited. However, the principal has to provide incentives for the low-income agent to truthfully report the transition to high income. Since income is private information, the principal would not fully insure the low-income agent. The distortion in the consumption path for a low-income agent is a key object of interest. We measure this distortion as the difference between the cost of providing the continuation utility to the low-income agent and the cost of providing the same utility using a perfectly smooth consumption path. We show that the distortion is determined by the ratio of the two state variables noted above.

Our main result is that it is optimal for the principal to audit the agent *periodically*. The auditing mechanism in our model consists of cycles. The low-income agent could be in one of two states: (i) not audited or (ii) randomly drawn to be audited. Within each cycle, a low-income agent is initially in the not-audited state. He will be moved into the random audit state if the duration of his low-income report exceeds a threshold N, where N is pinned down by the primitives of the model. If he is randomly drawn to be audited, then he will be moved to the not-audited state after being audited, and a new cycle begins. While auditing is stochastic, the threshold duration N is not. Put differently, within each cycle the principal guarantees that the agent will not be audited until the duration N is reached. The intuition for the periodicity is that the benefit of auditing is increasing with the number of non-audited periods, while the cost of auditing is constant. Auditing occurs when the benefit exceeds the cost.

In our model, there are two instruments for providing incentives. One instrument is dynamic taxation that distorts the consumption path and makes future payoffs contingent on past history of reported incomes. This is the standard instrument used in dynamic mechanism-design. For instance, Green (1987) uses this instrument to provide incentives for truthful reporting of income by designing taxes and subsidies that are history-dependent. The second instrument is auditing; the principal has to pay a cost to use the instrument. The U.S. Internal Revenue Service uses the second instrument to provide incentives for taxpayers to pay their true share of taxes – those who are caught cheating will be penalized. The principal in our model has access not only to the past history of reported incomes but also to the history of auditing outcomes and, hence, can provide better incentives by using both instruments.

To understand the interaction between the two instruments, we study two versions of the model: one with only the dynamic taxation instrument and the other with both taxation and audit instruments. The model with only the dynamic taxation instrument implies that the consumption path is increasingly distorted with the duration of low-income report. That is, the ratio of continuation utilities approaches one with the duration of low-income report and the principal has to deliver a consumption stream to the low-income agent that yields almost the same utility as that to the high-income agent. In order to ensure that the high-income agent does not have the incentive to deviate, this consumption stream is such that the static gain to the high-income agent from deviation is small whereas the future losses are large. We show that this path is highly distorted since it implies a steeply declining consumption profile. We also show that the ratio of continuation utilities is close to one in the long run, meaning that the distortion in the consumption path converges to infinity. In contrast, if the income process was i.i.d., as in Green (1987), the distortion is constant.

When the auditing instrument is also available, we show that the principal uses this instrument to alleviate the distortion in the consumption path implied by the first instrument. Auditing reduces the distortion, because when the agent's true income is observed during the audit, the principal rewards the truth-teller relative to the cheater. In particular, the principal removes the distortion (accumulated up to the auditing date) in consumption and increases the continuation utility for the truth-teller. This is not possible when there is no technology to ever verify who is the truth-teller and who is the cheater. We show that the optimal mechanism implies a discrete upward jump in the continuation utility for the truth-teller after the audit. We also show that no matter how high the cost of auditing is, there always exists a threshold N at which the auditing probability becomes positive.

If the agent's absolute risk aversion is not constant, then he is audited minimally when risk aversion is either extremely high or extremely low. When risk aversion is extremely low, the distortion in consumption incurs little welfare loss, thus there is no need to use the costly auditing instrument to reduce the distortion in consumption. When risk aversion is large, then a small distortion in consumption will generate large incentive effects, hence there is again no need to use the auditing instrument. The model also implies that, as the variance in income increases, auditing occurs more frequently.

In related literature on dynamic costly-state-verification, Wang (2005) studies deterministic verification with i.i.d. hidden incomes. He finds that there is a critical level of verification cost, below which there is verification and above which there is no verification. That is, if it is optimal to verify in one period then it is optimal to verify in every period. Thus, the verification in Wang (2005) is a static decision: it only depends on the cost of verification and is independent of the continuation utility that summarizes the past history. In our model, the auditing decision is dynamic and depends on past history via the duration of low-income reports. Popov (2007) studies stochastic verification with i.i.d. hidden incomes. He specifies an exogenous lower bound for the agent's continuation utility and every cheater is moved to the lower bound if caught during the audit. He obtains a nonstatic verification probability: agents with high continuation utility are verified less frequently. In his model, these agents are induced to tell the truth because the lower bound implies a harsher punishment for them if they are caught lying. In our model, punishment is the same across all levels of continuation utilities and there is no lower bound on continuation utility. Our periodic auditing result comes from the persistent income shock and the resulting distorted consumption path. Monnet and Quintin (2005) study stochastic verification with linear utility and i.i.d. hidden incomes. They find that the continuation utility is increasing and verification will eventually not be used. We study the risk-averse case, thus consumption distortion plays a central role in our model.

Although we focus on tax compliance in the paper, the issue of fraud and optimal auditing is applicable to other areas in economics. For instance, a venture capitalist provides start-up funds to an entrepreneur to invent a new product. In the experimental stage, the entrepreneur receives outside funding but after the product is invented, he might have to share the profits with the venture capitalist. If the outcome of the experiment is private information, then the entrepreneur can delay the report of being successful, keep the profit by selling the product privately and continue to receive funding from the venture capitalist. In the problem of infant industry protection, domestic firms are subsidized for a certain period to help them increase their productivity and compete with foreign firms. If productivity is private information, the firms have strong incentives to cheat because they can earn monopoly rents and receive subsidies simultaneously. In the context of unemployment insurance, an unemployed worker might find a job at a random rate. The exact date when he finds the job might not be observable. By delaying the report of employment, the worker can receive both wage income and unemployment benefits.

The rest of the paper is organized as follows. Section 2 describes the basic model without auditing, and shows that the distortion in consumption increases with the duration of the low-income report. In Section 3, we introduce the auditing technology and show that it is optimal to audit the low-income agent periodically. Then we study the dependence of auditing frequency on the primitives of the model. Section 4 concludes. We provide the proofs of all the results in an appendix.

2 Model: No Auditing Technology

In this section we study a hidden income model in which the principal does not have access to an auditing technology. The characterization of the optimal contract in this section will help us examine the optimal auditing in Section 3 when an auditing technology is available.

The tax authority is a risk-neutral principal with a discount rate $r > 0$. The taxpayer is a risk-averse agent, whose preferences are given by

$$E\left[\int_0^\infty re^{-rt}u(c_t)dt\right],$$

where c_t is consumption at time t, $u(c) = -\exp(-\rho c)$ is a constant absolute risk-aversion (CARA) utility function with risk aversion ρ, r is the discount rate (same as that of the principal) and E is the expectations operator. Let $c : (-\infty, 0) \to \mathbb{R}$ denote the inverse of the utility function:

$$c(u) = \frac{-\log(-u)}{\rho}. \tag{1}$$

To simplify our presentation, we first describe a discrete-time analogue of the model, and then consider the continuous-time model as the limit of a sequence of discrete-time models when the period length shrinks to zero. In a discrete-time model, period n ($n = 0, 1, 2, ...$) represents the time interval $[ndt, (n+1)dt)$ where $dt > 0$ is the length of one period.

Agents have either high income, w_H, or low income, w_L, where $w_L < w_H$. The high-income state is permanent.[1] All agents start with low income. In each period, a low-income agent transits to w_H with probability $\pi.dt$, where $\pi > 0$ is the Poisson arrival rate of w_H.

True income is not observable by the principal, so a high-income agent can underreport his income and pose as a low-income agent. We assume that the principal always asks the agent to show his reported income, so the low-income agent can never pretend to have w_H. Hence, there are no incentive constraints when the agent reports w_H.

The timing is as follows. In the initial period, the agent receives an income, either w_H or w_L. He chooses to report either w_H or w_L to the principal. The principal assigns current and future consumptions based on the report. In subsequent periods, if an agent had reported w_H in the past, he is in an absorbing state and no further reports are necessary. If an agent had reported w_L in every period in the past, then he receives an income, either w_H or w_L. The sequence of events then is the same as in the initial period.

The principal commits to delivering two sequences of consumptions, $\left\{ \left(c^H(n), c^L(n)\right); n = 0, 1, 2, ...\right\}$. We will denote this pre-commitment contract as σ. If an agent transits to w_H for the first time in period j, efficiency requires that the agent's consumption remains constant afterwards. This is because the principal and the agent have the same discount rate and w_H is an absorbing state. We denote this constant level of consumption by $c^H(j)$. The flow utility in each period from this level of consumption then is $ru\left(c^H(j)\right)$. Let $H(j) \equiv u(c^H(j))$ denote the discounted sum of utilities to an agent who transits to w_H for the first time in period j. Note that $H(j)$ is also the continuation utility to an

[1]That high income is an absorbing state allows us to focus on one spell of transiting to high income. If high income is not permanent, then the agent might experience multiple spells; nevertheless, the analysis within each spell would be similar to what we carry out here.

agent who transited to w_H before j but reports w_H for the first time in period j, since true income is not observable.

For a low-income agent, the consumption sequence $\left\{c^L(n)\right\}$ has to provide incentives to truthfully report the transition to the high income state. The continuation utility $L(n)$ to an agent who has low income until $n-1$ is

$$L(n) = \sum_{i=n}^{\infty} e^{-(i-n)rdt}(1-\pi dt)^{i-n}\left(\pi dt H(i) + (1-\pi dt)\,(rdt)\,u(c^L(i))\right).$$

The *temporary incentive compatibility* constraint requires that an agent who transited to high income in the current period does not have the incentive to delay the report of the transition to the next period, i.e., report w_L in the current period and w_H in future periods:

$$H(n) \geq (rdt)u(c^L(n) + w_H - w_L) + (1-rdt)H(n+1), n = 0, 1, 2, ...$$

The above constraint can be simplified as follows. CARA utility implies that $u(c^L(n) + w_H - w_L) = u(c^L(n))|u(w_H - w_L)|$. Define

$$b \equiv |u(w_H - w_L)| \in (0, 1),$$

so the temporary incentive compatibility constraint can be written as

$$H(n) \geq (rdt)bu(c^L(n)) + (1-rdt)H(n+1). \tag{2}$$

The expected cost for the principal is

$$C(\sigma) = \sum_{n=0}^{\infty} e^{-nrdt}(1-\pi dt)^n\left(\pi dt c^H(n) + (1-\pi dt)(rdt)c^L(n)\right).$$

There should, in fact, be an additional term in $C(\sigma)$: the discounted income obtained by the principal, $w_L + \frac{\pi(w_H - w_L)}{(r+\pi)}$. However, unlike the unemployment insurance literature that endogenizes search efforts and job-finding probabilities, the discounted income in our model is a constant, so it does not affect the optimal σ.

The principal's problem is to find an *incentive compatible* (I.C.) σ that delivers a level of initial utility L_0 to a low-income agent and minimizes $C(\sigma)$, i.e.,

$$\begin{aligned} \min_{\sigma} \quad & C(\sigma) \\ s.t. \quad & L_0 = \sum_{n=0}^{\infty} e^{-nrdt}(1-\pi dt)^n \left(\pi dt H(n) + (1-\pi dt)\,(rdt)\, u(c^L(n))\right), \\ & \text{and (2) for } n = 0, 1, 2, ... \end{aligned}$$

Next we will obtain a continuous-time representation of the above problem. First, denote $u(c^L(t))$ as $u^L(t)$ and write the promise-keeping constraints and incentive constraints recursively as,

$$\begin{aligned} L(t) &= \pi dt H(t) + (1-\pi dt)\left[(rdt)u^L(t) + (1-rdt)L(t+dt)\right], \\ H(t) &\geq (rdt)bu^L(t) + (1-rdt)H(t+dt). \end{aligned}$$

Second, transform these into differential equations and inequalities. For example, the inequality above can be rewritten as $\frac{H(t+dt)-H(t)}{dt} \leq rH(t+dt) - rbu^L(t)$. Taking limit $dt \to 0$ yields the differential inequality below.

$$\begin{aligned} \frac{dL(t)}{dt} &= (r+\pi)L(t) - \pi H(t) - ru^L(t), \\ \frac{dH(t)}{dt} &\leq rH(t) - rbu^L(t). \end{aligned}$$

Introducing a slack variable $\mu(t) \geq 0$ in the above differential inequality and rewriting the cost in continuous time, we get

$$\min_{\sigma} \quad C(\sigma) = \int_0^{\infty} e^{-(r+\pi)t}\left(\pi c(H(t)) + rc(u^L(t))\right) dt \tag{3}$$

$$s.t. \quad \frac{dL(t)}{dt} = (r+\pi)L(t) - \pi H(t) - ru^L(t), \tag{4}$$

$$\frac{dH(t)}{dt} = rH(t) - rbu^L(t) - \mu(t). \tag{5}$$

Following Fernandes and Phelan (2000) and Zhang (2009), we write the principal's problem as a dynamic programming problem, with L and H as the state variables and u^L and μ as the control variables. With a slight abuse of notation, denote the principal's cost function as $C(L, H)$.

Remark 1 *We include H in the state variable for incentive reasons. The principal chooses $H(0)$ freely to minimize cost (i.e., $\frac{\partial C}{\partial H} = 0$). In any continuation contract, however, H is no longer a free variable, because H acts as a threat utility. Raising $H(t)$ might induce an agent who transited to w_H in earlier periods to postpone the high-income report until t.*

Remark 2 *The domain of the cost function $C(L, H)$ in the dynamic programming problem is $\{(L, H) : L < H < 0\}$. If L is not strictly below H, then a high-income agent would pose as a low-income agent and consume more than a low-income agent.*

In the rest of this section, we study the solution to problem (3) in three steps. In subsection 2.1, we prove a homogeneity property of the cost function $C(L, H)$ and use it to introduce a measure of consumption distortion. In subsection 2.2, we formulate the Hamiltonian of (3) and derive a system of ordinary differential equations (ODE) that fully characterizes the optimal contract. In subsection 2.3, we obtain properties of the cost function and the dynamics of the state variables. In particular, we show that the distortion is increasing with the duration of low-income report. We summarize these results in Lemma 3 and will use them in Section 3.

2.1 A Measure of Distortion

Recall that the agent's utility function belongs to the CARA class. A property of the utility function is that

$$-\exp\left(-\rho\left(-\frac{\log(\alpha)}{\rho} + c\right)\right) = -\alpha \exp(-\rho c), \text{ for all } \alpha > 0.$$

Suppose that a contract $\sigma = \left\{\left(c^L(t), c^H(t)\right); t \geq 0\right\}$ delivers the continuation utility pair (L, H). Then, a contract

$$\sigma_\alpha = \left\{\left(-\frac{\log(\alpha)}{\rho} + c^L(t), -\frac{\log(\alpha)}{\rho} + c^H(t)\right); t \geq 0\right\}$$

delivers the pair $(\alpha L, \alpha H)$. The reverse is also true. Further, σ is I.C. if and only if σ_α is I.C. Therefore, $\left\{\left(c^{L*}(t), c^{H*}(t)\right); t \geq 0\right\}$ is the optimal contract to deliver (L, H) if and only if $\left\{\left(-\frac{\log(\alpha)}{\rho} + c^{L*}(t), -\frac{\log(\alpha)}{\rho} + c^{H*}(t)\right); t \geq 0\right\}$ is the optimal contract to deliver $(\alpha L, \alpha H)$. The next lemma states this homogeneity property and will be used to establish other properties of the cost function.[2]

LEMMA 1 *The cost function C has the following properties:*

(i) (Homogeneity) For any $\alpha > 0$,

$$C(\alpha L, \alpha H) = C(L, H) + \frac{-\log(\alpha)}{\rho}, \tag{6}$$

$$C_L L + C_H H = -\frac{1}{\rho}; \tag{7}$$

(ii) (Monotonicity) $C_L > 0$, $C_H \leq 0$.

Recall that c is the inverse of the utility function, so equation (6) is the same as

$$C(\alpha L, \alpha H) = C(L, H) + c(-\alpha).$$

We can thus decompose the cost $C(L, H)$ as

$$C(L, H) = C\left(-1, -\frac{H}{L}\right) + c(L).$$

Under full information, the principal will deliver L to the low-income agent via a stream of constant consumption and the cost of delivering L is $c(L)$. The *distortion* of consumption to the low-income agent in our contract can be measured by the difference between the cost $C(L, H)$ and the full information cost $c(L)$:

$$C(L, H) - c(L) = C\left(-1, -\frac{H}{L}\right). \tag{8}$$

It is helpful to compare the distortion in our model to that in Green (1987). With i.i.d. incomes, private information and CARA utility, Green (1987) shows that the cost function implied by the optimal contract differs from the full information cost function only by a constant. Thus, the distortion in any continuation contract in the i.i.d. case is constant. In particular, the distortion is independent of the history, or the level of evolving continuation utilities. With persistent shocks, the distortion is independent of the level of the continuation utility L, but depends on the ratio $\frac{H}{L}$, as noted in equation (8). Part (ii) of LEMMA 1 implies that the higher the ratio is, the higher the distortion will be.

[2] We use C_L, C_H, C_{LL}, C_{LH}, and C_{HH} to denote partial derivatives $\frac{\partial C}{\partial L}, \frac{\partial C}{\partial H}, \frac{\partial^2 C}{\partial L^2}, \frac{\partial^2 C}{\partial L \partial H}$, and $\frac{\partial^2 C}{\partial H^2}$, respectively.

2.2 The Hamiltonian

The problem faced by the principal is to choose a time path $(u^L(t), \mu(t))$ to minimize the cost in (3). Given the path $(u^L(t), \mu(t))$ and an initial state $(L(0), H(0))$, the promise-keeping and incentive constraints (4) and (5) imply a time path $(L(t), H(t))$ for continuation utilities. One way to think about this problem is to think of choosing $(u^L(t), \mu(t))$ at each date, given the values of $(L(t), H(t))$ that have been attained by that date. The principal faces a tradeoff between the current-period cost, $c(u^L(t))$, and the cost of delivering continuation utility, and hence needs to set "prices", Φ and $-\lambda$, on increments to the continuation utilities L and H:

$$\Phi = C_L, -\lambda = C_H \leq 0.$$

A central construct in the study of *optimal* allocation is the *current value Hamiltonian* $\mathscr{H}$ defined by

$$\mathscr{H}(L, H, \Phi, \lambda, u^L, \mu) = (\pi c(H) + rc(u^L)) + \Phi((r+\pi)L - ru^L - \pi H) - \lambda(rH - rbu^L - \mu),$$

which is just the sum of current-period cost and the rate of increase in continuation utilities (see (4) and (5)), the latter valued at $\Phi(t)$ and $-\lambda(t)$. An optimal allocation must minimize $\mathscr{H}$ at each date t.

The first-order condition for minimizing $\mathscr{H}$ with respect to u^L is

$$c'(u^L) = \Phi - b\lambda. \tag{9}$$

The left hand is the marginal cost of today's utility, while the right hand is the marginal cost of starting with higher continuation utility tomorrow, offset by the benefit of a slacker incentive constraint. The utility u^L must be chosen to equalize the costs at each date. Note that equation (9) implies $\lambda \in [0, \Phi/b]$ along the optimal path.

The prices $\Phi(t)$ and $-\lambda(t)$ must satisfy

$$\frac{d\Phi(t)}{dt} = (r+\pi)\Phi(t) - \frac{\partial \mathscr{H}(L(t), H(t), \Phi(t), \lambda(t), u^L(t), \mu(t))}{\partial L} = 0, \tag{10}$$

$$\begin{aligned} \frac{d\lambda(t)}{dt} &= (r+\pi)\lambda(t) + \frac{\partial \mathscr{H}(L(t), H(t), \Phi(t), \lambda(t), u^L(t), \mu(t))}{\partial H} \\ &= \pi(\lambda(t) + c'(H(t)) - \Phi(t)), \end{aligned} \tag{11}$$

at each date t if $(u^L(t), \mu(t))$ is an optimal path. Equation (10) implies that Φ is a constant. Furthermore, since the principal chooses the optimal $H(0)$ to satisfy $C_H(L(0), H(0))$=0, we can pin down Φ by the level of initial utility L_0 and equation (7) in Lemma 1. Hence,

$$\Phi = \frac{-1}{\rho L_0}. \tag{12}$$

Using the above differential equations, the following lemma shows that the incentive constraints are always binding.

Lemma 2 $\lambda(0) = 0$ *and* $\lambda(t) > 0$, *for all* $t > 0$; $\mu(t) = 0$ *for all* t.

To summarize, the dynamics are given by

$$\begin{aligned} \frac{d\lambda}{dt} &= \pi(\lambda + c'(H) - \Phi), \\ \frac{dL}{dt} &= (r + \pi)L - \pi H - ru^L, \\ \frac{dH}{dt} &= rH - rbu^L. \end{aligned}$$

From equation (1), the inverse of the utility function implies that $c'(H) = -\frac{1}{\rho H}$. From (1) and (9), we can determine $u^L = -\frac{1}{\rho(\Phi - b\lambda)}$.

Substituting for u^L and $c'(H)$, the dynamics are described by an ODE system:

$$\frac{d\lambda}{dt} = \pi\left(\lambda - \frac{1}{\rho H} - \Phi\right), \tag{13}$$

$$\frac{dL}{dt} = (r + \pi)L - \pi H + \frac{r}{\rho\left(\Phi - b\lambda\right)}, \tag{14}$$

$$\frac{dH}{dt} = rH + \frac{rb}{\rho\left(\Phi - b\lambda\right)}. \tag{15}$$

2.3 The Analysis of the ODE System

In the ODE system, equations (14) and (15) are nonlinear in λ, while (13) is nonlinear in H. To facilitate the analysis of the ODE system, in this subsection, we will first reduce the above system to a system with only two variables, λ and H. Second, we will eliminate the nonlinearity in one of the differential equations in the reduced system by a simple transformation of variables. Finally, we will obtain some properties of the cost function and dynamics of the state variable in the reduced and transformed ODE system. These properties will be used in Section 3.

To reduce the ODE system, we will show that equation (14) is redundant if equations (13), (15), and LEMMA 1 hold. Rewrite equation (7) in LEMMA 1 as $\Phi L - \lambda H = -\frac{1}{\rho}$ and differentiate both sides with respect to time:

$$\Phi \frac{dL}{dt} - \frac{d\lambda}{dt} H - \lambda \frac{dH}{dt} = 0.$$

After the substitution of (13) and (15), the above equation becomes

$$\begin{aligned}
\Phi \frac{dL}{dt} &= \pi \left(\lambda - \frac{1}{\rho H} - \Phi \right) H + \lambda \left(rH + \frac{rb}{\rho\,(\Phi - b\lambda)} \right) \\
&= (r+\pi)\lambda H - \frac{\pi}{\rho} + \frac{rb\lambda}{\rho\,(\Phi - b\lambda)} - \pi \Phi H \\
&= (r+\pi) \left(\lambda H - \frac{1}{\rho} \right) + \frac{r\Phi}{\rho\,(\Phi - b\lambda)} - \pi \Phi H \\
&= (r+\pi)\Phi L - \pi \Phi H + \frac{r\Phi}{\rho\,(\Phi - b\lambda)},
\end{aligned}$$

which is equation (14) multiplied by a positive constant Φ.

The reduced ODE system now consists of only equations (13) and (15). Since we will show later that $\lim_{t\to\infty} H(t) = -\infty$, it is convenient to eliminate the nonlinearity in H in the reduced system. To this end, we introduce a new variable

$$y \equiv c'(H) = -\frac{1}{\rho H}$$

to replace H. We then have

$$\frac{dy}{dt} = \frac{1}{\rho H^2} \frac{dH}{dt} = \frac{ry^2}{\Phi/b - \lambda} - ry.$$

Thus, the ODE system in the previous section can now be written as

$$\frac{d\lambda}{dt} = \pi(\lambda + y - \Phi), \tag{16}$$

$$\frac{dy}{dt} = \frac{ry^2}{\Phi/b - \lambda} - ry, \tag{17}$$

where $\lambda \geq 0$ and $y > 0$.

In what follows, we will characterize the solution to (16) and (17). It is easy to see that $\frac{d\lambda}{dt} > 0$ if and only if $y > \Phi - \lambda$, and $\frac{dy}{dt} > 0$ if and only if $y > \frac{\Phi}{b} - \lambda$. Figure 1 is the

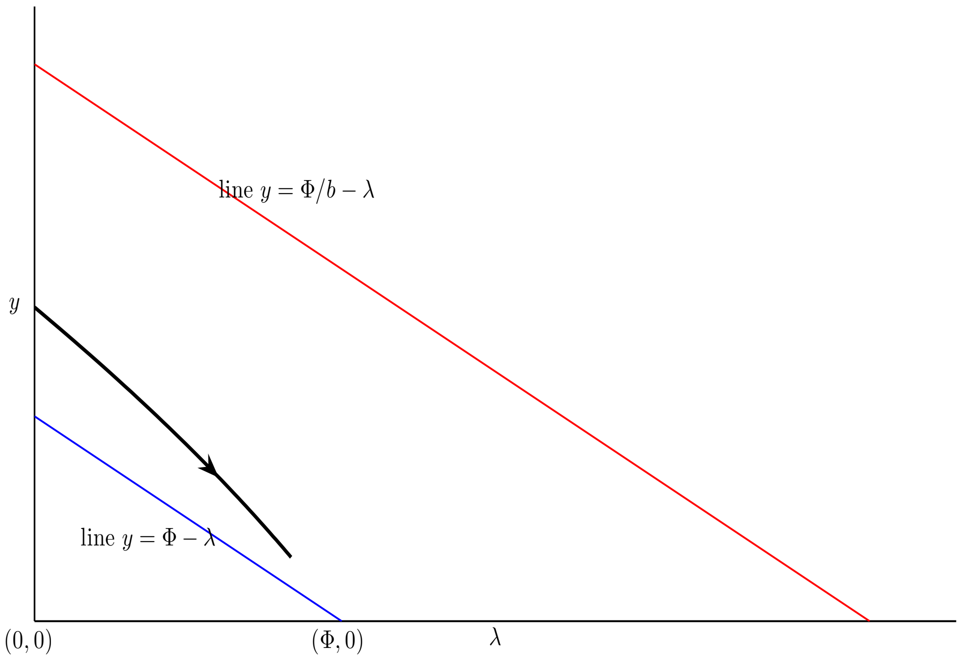

Figure 1: Phase diagram for the system in equations (16) and (17).

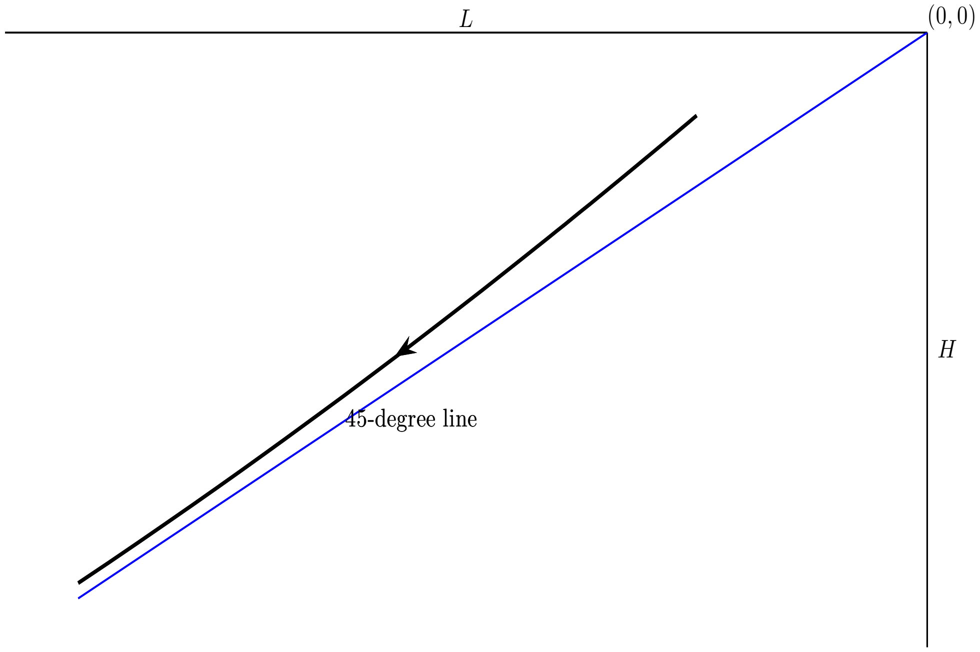

Figure 2: Time path of $(L(t), H(t))$.

phase diagram for the system in equations (16) and (17). We summarize the dynamics of the ODE system in LEMMA 3.

LEMMA 3 *The time path* $(\lambda(t), y(t))$ *and cost function* $C(L, H)$ *have the following properties:*

(i) The initial condition $(\lambda(0), y(0))$ *satisfies* $\lambda(0) = 0$. *The path always stays below the straight line* $y = \frac{\Phi}{b} - \lambda$ *and above the straight line* $y = \Phi - \lambda$, *i.e.,* $\Phi - \lambda(t) < y(t) < \frac{\Phi}{b} - \lambda(t)$ *for all* $t \geq 0$; *(see Figure 1)*

(ii) The path $(\lambda(t), y(t))$ *moves southeast, converges to* $(\Phi, 0)$ *and also approaches (but never reaches) the line* $y = \Phi - \lambda$, *i.e.,* $\frac{d\lambda}{dt} > 0$, $\frac{dy}{dt} < 0$, $\lim_{t\to\infty}(\lambda(t), y(t)) = (\Phi, 0)$ *and* $\frac{d\lambda}{dt} + \frac{dy}{dt} < 0$, *for all* $t \geq 0$; *(see Figure 1)*

(iii) The continuation utilities $L(t)$ *and* $H(t)$ *decline with low-income report, and the path* $(L(t), H(t))$ *approaches (but never reaches) the 45-degree line, i.e.,* $\frac{dL}{dt} < 0$, $\frac{dH}{dt} < 0$, $\lim_{t\to\infty}(L(t), H(t)) = (-\infty, -\infty)$ *and* $\frac{d\left(\frac{L}{H}\right)}{dt} < 0$ *for all* $t \geq 0$; *(see Figure 2)*

(iv) The cost function $C(L, H)$ *is strictly convex;*

(v) For a fixed L, $\lim_{H\downarrow L} C(L, H) = \infty$.

Part (iii) states that the time path $(L(t), H(t))$ moves toward $(-\infty, -\infty)$ and approaches the 45-degree line, i.e., $\lim_{t\to\infty} \frac{H(t)}{L(t)} = 1$ (see Figure 2). That L and H decrease with low-income report is for incentive reasons. In order to prevent an agent who transits to high income from postponing the high-income report, the principal punishes late reporters by reducing H with the duration of low-income report. Moveover, H declines faster than L, so the distortion (8) in the continuation contract increases with the duration of low-income report.

Part (v) states that for a given $L(t)$, the distortion approaches infinity when $H(t)$ is sufficiently close to $L(t)$. Suppose $H(t) = L(t) + \epsilon$, where $\epsilon > 0$ is a small number. Note that the high-income agent's consumption path is not distorted. For the low-income agent, consider three possibilities for the consumption path: flat, increasing, or decreasing. When the path is flat, a high-income agent may underreport his income and obtain utility of at least $[u(c_t^L + w_H - w_L) - u(c_t^L)]dt + L$. Because $[u(c_t^L + w_H - w_L) - u(c_t^L)]dt + L \approx [u(c_t^L + w_H - w_L) - u(c_t^L)]dt + H > H$, the incentive constraint is violated. When the path is increasing, c_t^L must be below c_t^H, which leads to a similar violation of the incentive constraint. Hence when ϵ is sufficiently small, c_t^L must be above c_t^H and must approach infinity. Put differently,

$$
\begin{aligned}
H(t) &\geq (rdt)bu^L(t) + (1 - rdt)H(t + dt), \\
L(t) &= \pi dt H(t) + (1 - \pi dt)\left[(rdt)u^L(t) + (1 - rdt)L(t + dt)\right],
\end{aligned}
$$

imply that

$$
\begin{aligned}
H(t) - L(t) &= (1 - \pi dt)\left[(rdt)(b - 1)u^L(t) + (1 - rdt)(H(t + dt) - L(t + dt))\right] \\
&> (1 - \pi dt)(rdt)(b - 1)u^L(t),
\end{aligned}
$$

which implies $u^L(t) > \frac{-\epsilon}{(1-\pi dt)(rdt)(1-b)}$. When ϵ is sufficiently small, $u^L(t)$ needs to be sufficiently close to 0, which requires a large consumption at t. The large consumption at t needs to be offset by much lower levels of consumption in the future, so as to deliver a given level of continuation utility $L(t)$. Hence the consumption path is very distorted when H is just slightly above L.

3 Model: Costly Auditing

Besides distorting the consumption path to provide incentives, now the principal can deter cheating by auditing the agent's report. Auditing reveals the agent's true income but costs γ units of consumption good. Since high income is an absorbing state, it is easy to see that auditing is unnecessary forever if the agent reports w_H just once in the past. In each period, conditional on low-income report, the principal chooses to audit according to a Poisson arrival rate $p(t) \geq 0$. That is, over a period of length dt, she audits with probability $p(t)dt$ and she does not audit with probability $1 - p(t)dt$. Note that, since our model is in continuous time, $p(t)$ is the (endogenous) arrival rate of an audit, not the auditing probability itself. If $p(t) = 0$, no auditing arrives, while if $p(t) = \infty$, the auditing probability has an atom at t. We assume that if an agent is audited and caught cheating, he needs to pay a finite penalty of $\psi > 0$ forever. We model finite penalty, because if infinite penalty ($\psi = \infty$) is allowed, then an arbitrarily small auditing probability would deliver the full information constant consumption.

The principal pre-commits not only to the two sequences of consumption, as in the previous section, but also to the sequence of arrival rates of audit. We can again represent the principal's cost minimization problem as a dynamic program with L and H as state variables. We continue to exclude the discounted income from the cost function $C(L, H)$ for the same reason as in the case without auditing; however, $C(L, H)$ now includes both the cost of delivering consumption and the cost of auditing.

Remark 3 *When the principal audits and observes the true income, it is feasible for her to deliver any continuation utility pair (L, H), $(L < 0, H < 0)$. But when she delivers less utility to a high-income agent than to a low-income agent (i.e., $H < L < 0$), it induces the high-income agent to quit his high-income job and become a low-income agent. Hence, we exclude the region $\{(L, H) : H < L < 0\}$ from the domain of the cost function. This is equivalent to assuming that a high-income agent can secretly become a low-income agent whenever he wants to. Therefore, the domain of the cost function is $\{(L, H) : L \leq H < 0\}$.*

Remark 4 *The utility $L = H < 0$ can be delivered because the true income is observable with an auditing technology.*

The timing is as follows. In the initial period, the agent receives an income, either w_H or w_L. He chooses to report either w_H or w_L to the principal. Then conditional on the report, the principal chooses the auditing probability. Conditional on the report and the outcome of the audit, the principal assigns current and future consumptions. (Recall that auditing probability is zero if the report is w_H.) In subsequent periods, if an agent had reported w_H in the past, he is in an absorbing state and no further reports or auditing are necessary. If an agent had reported w_L in every period in the past, then he receives an income, either w_H or w_L. The sequence of events then is the same as in the initial period.

For now we impose a restriction that atomic auditing is not allowed. In subsection 3.3, we verify that the principal will not use atomic auditing even if it is allowed. When there is no atomic auditing (i.e., $p(t) < \infty$), the promise-keeping and incentive constraints are

$$\begin{aligned} L(t) &= \pi dt H(t) + (1-\pi dt)\left(p(t)dt\tilde{L}(t) + (1-p(t)dt)\left[(rdt)u^L(t) + (1-rdt)L(t+dt)\right]\right), \\ H(t) &\geq p(t)dte^{\rho\psi}H(t) + (1-p(t)dt)\left[(rdt)bu^L(t) + (1-rdt)H(t+dt)\right], \end{aligned}$$

where $e^{\rho\psi}H(t)$ is the agent's continuation utility if he is audited and found to be a liar, and $\tilde{L}(t)$ denotes the low-income agent's continuation utility if he is audited and found to be a truth-teller. Thus, the differential equations for the state variables are

$$\frac{dL}{dt} = (r+\pi)L - \pi H - ru^L - p(\tilde{L} - L), \tag{18}$$

$$\frac{dH}{dt} = rH - rbu^L - p(e^{\rho\psi} - 1)H - \mu, \tag{19}$$

where μ is, again, a slack variable. Note that if p is exogenously set to zero, then the above differential equations are identical to (4) and (5) in Section 2.

We can write the *Hamilton-Jacobi-Bellman (HJB)* equation satisfied by the cost function $C(L, H)$ as:

$$\begin{aligned} (r+\pi)C(L,H) = \min_{u^L,p,\tilde{L},\tilde{H},\mu} \quad & rc(u^L) + \pi c(H) + p\left(C(\tilde{L},\tilde{H}) + \gamma - C(L,H)\right) \\ & + C_L(L,H)\left((r+\pi)L - \pi H - ru^L - p(\tilde{L} - L)\right) \\ & + C_H(L,H)\left(rH - rbu^L - p(e^{\rho\psi} - 1)H - \mu\right). \end{aligned} \tag{20}$$

In the HJB equation, $\tilde{H}$ denotes the continuation utility to a low-income agent who transited to high income immediately after he was audited.

The presence of the control variable p in the HJB equation (20) makes it difficult to study the optimal contract directly. To simplify the analysis, in subsection 3.1, we first study a restricted problem where the principal is able to audit only when $L = H$ (i.e., $p(t) = 0$ if $L(t) < H(t)$). We show the properties of the cost function under this restriction. Then in subsection 3.2, we show that the principal would, in fact, not audit when $L < H$, even if she is allowed to do so. In subsection 3.3, we show that the principal would not use atomic auditing. Finally, in subsection 3.4, we study implications of the optimal contract.

3.1 A Restricted Problem

When $L < H$ and $p = 0$, the HJB equation (20) reduces to

$$(21)\ (r+\pi)C(L,H) = \min_{u^L,\mu} \quad rc(u^L) + \pi c(H) + C_L(L,H)\left((r+\pi)L - \pi H - ru^L\right) + C_H(L,H)\left(rH - rbu^L - \mu\right).$$

When $L = H$ and auditing is allowed, the HJB equation (20) applies. Recall that the domain of the cost function in the unrestricted problem is $L \leq H < 0$ (see Remark 3). In the restricted problem, the domain remains the same. However, the feasible values of p are restricted in parts of the domain: when $L < H$, the feasible set for p is a singleton $\{0\}$ and when $L = H$, the feasible set is $[0, \infty)$.

The restricted problem and the problem in Section 2 (where auditing is completely shut down and the domain of the cost function is $L < H < 0$) share a lot of similarities. When $L < H$, Lemma 1 and Lemma 2 continue to hold, i.e., the cost function is homogeneous, $\mu \equiv 0$ in equation (19) and the ODE system in (16) and (17) characterize the dynamics.

But there are two important differences between the restricted problem and the problem in Section 2. The first difference is that when $L = H$, $C(L,H)$ is defined and finite in the restricted problem, but it is not defined in Section 2 (or, the cost is infinity, see part (v) in Lemma 3). The second difference is that the time path of $(L(t), H(t))$ in the restricted problem reaches the 45-degree line in finite time (see Figure 3), but it does not reach the 45-degree line in Section 2 (see Figure 2).

To see the first difference, note that on the 45-degree line, perfectly smooth consumption can be achieved as long as the arrival rate p of auditing is sufficiently large. For example, when $L = H$, set $u^L = L = H$, $\tilde{L} = \tilde{H} = L$, $\mu = 0$ and $p = \frac{r(1-b)}{e^{\rho\psi}-1}$. Then, equations (18) and (19) imply that

$$\begin{aligned} \frac{dL}{dt} &= (r+\pi)L - \pi H - ru^L - p(\tilde{L} - L) = 0, \\ \frac{dH}{dt} &= rH - rbu^L - p(e^{\rho\psi} - 1)H - \mu = 0. \end{aligned}$$

Thus, both L and H are constant and the cost of implementing (L, H) cannot be infinite when $L = H$, if auditing is allowed.

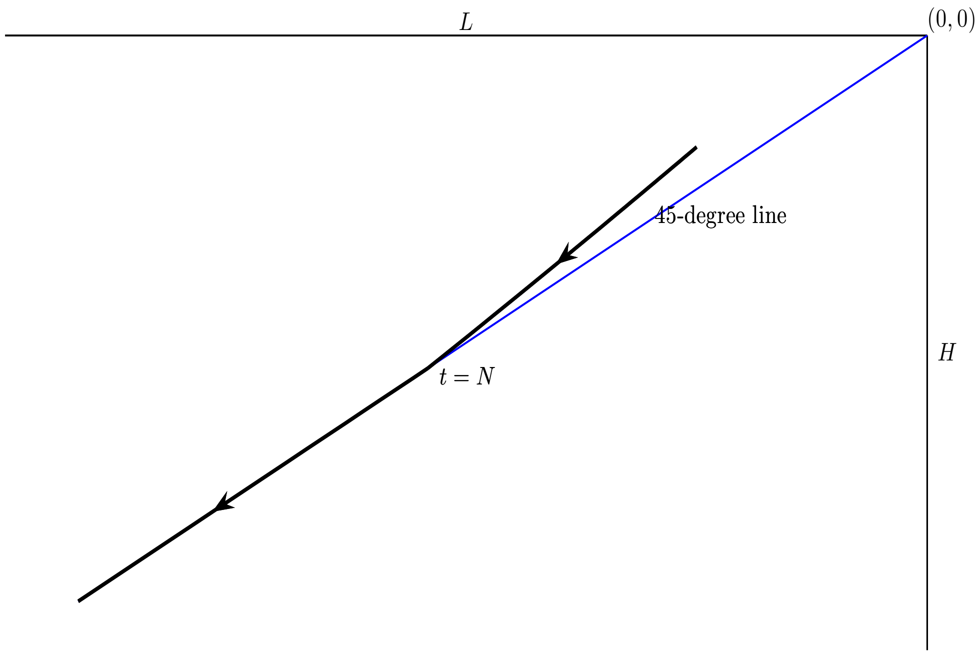

Figure 3: Time path of $(L(t), H(t))$ in the restricted problem.

To see the second difference, suppose to the contrary that the dynamic path of $(L(t), H(t))$ never reaches the 45-degree line. Then in the region $\{(L, H) : L < H < 0\}$, the restricted problem is no different from the problem in Section 2 and, hence, $\lim_{H \downarrow L} C(L, H) = \infty$, as in part (v) in Lemma 3. But this contradicts the property in Lemma 1 that $C(L, H)$ is decreasing in H and $C(L, L)$ is finite.

Despite the two differences, we can characterize the dynamics in the restricted problem using the results from subsection 2.3. To begin, since the path of $(L(t), H(t))$ in the restricted problem reaches the 45-degree line in finite time (Figure 3), let N denote the first time that the path reaches the 45-degree line, i.e., $L(t) = H(t)$ for the first time at $t = N$. Similar to subsection 2.3 we can describe the path of $(L(t), H(t))$ in terms of $(\lambda(t), y(t))$. Equations (16) and (17) completely describe the dynamic path before N in the restricted problem. At N, $H(N) = L(N)$, so using the prices in subsection 2.2 equation (7) implies that

$$\Phi H(N) - \lambda H(N) = \Phi L(N) - \lambda H(N) \;\; = \;\; -\frac{1}{\rho}.$$

Thus $y(N) = \frac{-1}{\rho H(N)} = \Phi - \lambda$ and the path of $(\lambda(t), y(t))$ reaches the straight line $y = \Phi - \lambda$ at time N. Lemma 4 below demonstrates that the auditing probability is positive at $t = N$ (when $L(t) = H(t)$ for the first time). The evolution of L and H is then pinned down by (18) and (19).

Lemma 4 *When $t = N$, $p > 0$.*

3.2 Optimal Auditing

The remaining issue is the auditing probability before N. In the restricted problem, we forced the auditing probability to be zero for $t < N$ (when $L < H$). In this subsection, we show that the constraint $p = 0$ when $L < H$ is not binding, i.e., the principal would choose $p = 0$ even if the feasibility set for p is $[0, \infty)$.

Lemma 5 *Let $t < N$. Suppose $p \in [0, \infty)$. Then the principal chooses $p = 0$.*

When $t < N$, the path of $(L(t), H(t))$ is above the 45-degree line and the principal does not audit. That is, in the pre-commitment contract, the principal guarantees that she would not audit until N, despite the fact that income is private information.

Our main result is that the auditing pattern is periodic. The optimal mechanism consists of cycles. A low-income agent begins each cycle with (L, H), $L < H$, and is initially not audited. When the duration of his low-income reports reaches N, he will be audited randomly according to an endogenous arrival rate $p > 0$. The actual instant of audit depends on the realization of the audit random variable, so the actual audit could be at any $t \geq N$. The new cycle starts the moment after he is audited. In the new cycle, the low-income agent begins with updated continuation utilities $(\tilde{L}, \tilde{H})$. Note that while auditing is stochastic, N is deterministic and is completely pinned down by the differential equations below:

$$\begin{aligned} \frac{dL}{dt} &= (r + \pi)L - \pi H - ru^L, \\ \frac{dH}{dt} &= rH - rbu^L. \end{aligned}$$

(These equations follow directly from (18) and (19), since $p = 0$ and $\mu = 0$ when $L < H$.) Starting from (L, H), N is the time taken to reach the 45-degree line along the path implied by these equations.

The threshold duration N depends on primitives of the model, but does not depend on the initial promised utility L_0. The homogeneity property implies that, if $\{(L(t), H(t)); t \geq 0\}$ is the optimal time path when the initial promise is L_0, then $\{(\frac{\bar{L}_0}{L_0}L(t), \frac{\bar{L}_0}{L_0}H(t)); t \geq 0\}$ is the optimal path when the initial promise is $\bar{L}_0 \neq L_0$. The two paths reach the 45-degree line at the same time.

PROPOSITION 1 *When $t \geq N$ and conditional on low-income report, the principal audits with an arrival rate $p > 0$. The time path $(L(t), H(t))$ stays on the 45-degree line and moves along it toward $(-\infty, -\infty)$ until the agent is randomly drawn to be audited. After the audit, (L, H) jumps to a new state $(\tilde{L}, \tilde{H})$. Then the optimal contract enters a new cycle.*

In the unrestricted problem, there are two instruments to provide incentives for truthfully reporting the transition to high income. The first instrument is dynamic taxation that distorts the consumption path. The principal always uses this instrument. The second instrument of auditing, however, is not used by the principal if $L < H$. Since $\frac{H}{L}$ measures the distortion in the continuation contract (see the discussion after LEMMA 1), the closer $\frac{H}{L}$ is to 1, the higher is the distortion. PROPOSITION 1 shows that the principal uses the auditing instrument only when the distortion is the highest, i.e., when $L = H$.

The reason that the principal uses the two instruments asymmetrically is because the marginal cost of the first instrument is increasing with the distortion, while that of the second is constant. Starting with the full insurance consumption path, a first-order distortion in consumption generates only a second-order welfare loss. Thus when consumption distortion is small, it is nearly costless to use the first instrument and the principal will avoid the second instrument. The principal uses the second instrument of auditing only when the benefit of correcting the distortion is larger than the cost, namely when (L, H) reaches the 45-degree line.

The principal audits periodically, no matter how high the auditing cost γ is. This is because the distortion in consumption converges to infinity and the benefit of using the auditing instrument will eventually surpass any finite cost. This result contrasts with that in Wang (2005), where the principal does not audit when γ is large, since the income is i.i.d. in his environment and the distortion is constant. It also contrasts with Monnet and Quintin (2005), where auditing is not used eventually, since the agent is risk neutral.

3.3 No Atomic Auditing

In the previous two subsections, we have shown that the principal would set $p = 0$ when $L < H$ and would set $p > 0$ but finite when $L = H$. In this subsection we will show that the principal would never set $p = \infty$, i.e., she will never use atomic auditing. For the moment, denote P as the size of the atom if there is an atom in the auditing probability when the state is (L, H). With probability $P > 0$ the principal audits and with probability $1 - P$ she does not. Thus, the cost minimization problem for the principal is

$$M(P) = \min_{\tilde{L},\tilde{H},\bar{L},\bar{H}} \quad P\left(C(\tilde{L},\tilde{H}) + \gamma\right) + (1-P)C\left(\bar{L},\bar{H}\right) \tag{22}$$

$$s.t. \quad L = P\tilde{L} + (1-P)\bar{L}, \tag{23}$$

$$H \geq Pe^{\rho\psi}H + (1-P)\bar{H}, \tag{24}$$

where $(\tilde{L}, \tilde{H})$ denotes the state if the agent is audited and $(\bar{L}, \bar{H})$ denotes the state if the agent is not audited. Note that $M(0) = C(L, H)$. Lemma 6 below states that atomic auditing is not optimal.

Lemma 6 *At any (L, H), $L \leq H$, there is no atomic auditing, i.e., $M(P) > M(0)$ for all $P > 0$.*

3.4 Implications of the Optimal Contract

1. *Reducing the cost of auditing (smaller γ) increases the auditing frequency.*

A smaller γ makes the auditing instrument cheaper than the dynamic taxation instrument. As a result, the principal is willing to audit more frequently.

2. *Increasing the variance of income (i.e., larger $w_H - w_L$) increases the auditing frequency.*

With a larger $w_H - w_L$, a high-income agent benefits more from underreporting income. In the absence of auditing, the low-income agent's consumption path needs to be distorted more to provide dynamic incentives for truth telling. As a result, the auditing instrument is used more frequently to reduce the distortion.

For instance, the rich might have more volatile income relative to the poor since capital income might be a larger component of the income for the rich. Our model implies that the rich would be audited more frequently, which is roughly consistent with the IRS practice.

3. *Agents with an intermediate level of risk aversion, ρ, are audited more frequently relative to agents with either low or high risk aversion.*

When ρ is small, distortion in the consumption path incurs little welfare loss, thus there is no need to use auditing to reduce the distortion.

When ρ is large, a small distortion in consumption is able to generate large incentive effects. Again, there is no need to audit. More specifically, let $L_0 = -1$ and consider a no-auditing contract in which consumptions decline linearly, i.e., $c^H(t) = c^L(t) = \frac{\log(\frac{\pi}{r+\pi}) - rt}{\rho}$. This contract delivers the promised utility L_0 because

$$\int_0^\infty e^{-(r+\pi)t}\left(\pi H(t) + ru^L(t)\right) dt = -\int_0^\infty e^{-(r+\pi)t}\left((\pi + r)e^{rt}\frac{\pi}{r+\pi}\right) dt = -1.$$

It is I.C. because

$$\frac{dH(t)}{dt} = rH(t) \quad < \quad rH(t) - rbu^L(t).$$

When $\rho \to \infty$, the cost of the contract converges to the full information cost $c(L_0)$, which is zero. The optimal no-auditing contract has very little distortion in consumption, and there is no need to correct it frequently by auditing.

4 Conclusion

We have studied a repeated hidden income environment with persistent incomes. A principal, with imperfect ability to audit, designs an optimal taxation scheme as well as an optimal auditing scheme. When the agent's absolute risk aversion is constant, we have shown that it is optimal to audit the low-income agent periodically. The optimal mechanism consists of cycles. Within each cycle, an agent reporting low income is guaranteed that he will not be audited until the duration of the low-income reports exceeds a threshold. After the threshold is reached, the agent is audited randomly.

Unlike repeated hidden income model with i.i.d. incomes, the distortion in the consumption path increases with the duration of low-income reports in our model. Auditing helps the principal detect who is the truth-teller and who is the cheater. She can thus correct the distortion in the consumption path after an audit. The benefit of auditing increases with the duration of low-income reports whereas the cost of auditing is constant. Consequently, the principal would use the auditing instrument no matter how high the auditing cost is.

Our model is limited in several respects. First, our results are valid for the case of CARA utility. With more general utility specifications, the length of each auditing cycle would not be constant. However, we expect the periodic feature to remain.

Second, the binary nature of income levels – income is either w_H or w_L – is restrictive. One implication of this assumption is that auditing occurs only at the lowest income level (i.e., w_L). When there are more than two income levels, the principal might audit any income level below the maximum. If the only binding incentive constraint when the report is w_i is for the agent at the next higher income level w_{i+1}, then our measure of distortion remains useful in the more general setup, namely, the distortion is the ratio between the continuation utility for the agent who just transited to w_{i+1} and that for the agent who remains at w_i. We conjecture that the auditing in the optimal contract would still contain cycles.

Appendix

Proof of Lemma 1:

(i) Equation (6) holds because $\left\{\left(c^{L*}(t), c^{H*}(t)\right); t \geq 0\right\}$ is the optimal contract to implement (L, H) if and only if $\left\{\left(-\frac{\log(\alpha)}{\rho} + c^{L*}(t), -\frac{\log(\alpha)}{\rho} + c^{H*}(t)\right); t \geq 0\right\}$ is the optimal contract to implement $(\alpha L, \alpha H)$. Differentiating (6) with respect to α and then setting $\alpha = 1$ yield (7).

(ii) We show $C_H \leq 0$ first. It is equivalent to show that $C(\bar{L}, \bar{H}) \leq C(L, H)$ for any (L, H) and $(\bar{L}, \bar{H})$ with $L = \bar{L}, H < \bar{H}$. The evolution of $(\bar{L}, \bar{H})$ is

$$\begin{aligned}\frac{d\bar{L}(t)}{dt} &= (r+\pi)\bar{L}(t) - \pi\bar{H}(t) - r\bar{u}^L(t),\\ \frac{d\bar{H}(t)}{dt} &= r\bar{H}(t) - rb\bar{u}^L(t) - \bar{\mu}(t).\end{aligned}$$

By picking $\bar{\mu}(0) = \infty$, $\bar{H}(t)$ could jump to $H(t)$ immediately after time 0. Thus $C(\bar{L}, \bar{H}) \leq C(L, H)$. It follows from (7) and $C_H \leq 0$ that

$$C_L = \frac{-\frac{1}{\rho} - C_H H}{L} > 0.$$

Q.E.D.

Proof of Lemma 2: That $\lambda(0) = 0$ is because the principal chooses $H(0)$ to minimize the cost $C(L(0), H(0))$.[3] To prove the second statement by contradiction, suppose for some $t^* > 0$, $\lambda(t^*) = 0$. Since $\lambda(t)$ is non-negative, it achieves a minimum at t^*, and the first- and second-order conditions are

$$\frac{d\lambda(t^*)}{dt} = 0, \tag{25}$$

$$\frac{d^2\lambda(t^*)}{dt^2} \geq 0. \tag{26}$$

Equations (11) and (25) imply that $c'(H(t^*)) = \Phi$. The first-order condition (9) implies that $c'(u^L(t^*)) = \Phi$, hence $u^L(t^*) = H(t^*)$. Therefore equation (5) and $b \in (0, 1)$ imply that $\frac{dH(t^*)}{dt} \leq rH(t^*) - rbu^L(t^*) < 0$. Thus differentiating (11) with respect to t yields

$$\frac{d^2\lambda(t^*)}{dt^2} = \pi\left(\frac{d\lambda(t^*)}{dt} + c''(H(t^*))\frac{dH(t^*)}{dt}\right) < 0,$$

which contradicts (26). *Q.E.D.*

Proof of Lemma 3:

[3]By contradiction, suppose $\lambda(0) > 0$, then the principal can further lower the cost by increasing $H(0)$. Increasing $H(0)$ does not violate any incentive constraints and the principal has complete freedom in picking $H(0)$.

(i) We proved $\lambda(0)=0$ in LEMMA 2. To show that $y(t) < \frac{\Phi}{b} - \lambda(t)$ for all t, suppose to the contrary that $y(t^*) \geq \frac{\Phi}{b} - \lambda(t^*)$ for some t^*. Then the phase diagram in Figure 1 shows that the path will remain above the line $y = \frac{\Phi}{b} - \lambda$ forever, because $\frac{dy(t)}{dt} > 0$ and $\frac{d\lambda(t)}{dt} > 0$ for $t > t^*$. Eventually y and λ become unbounded, which contradicts the fact that $\lambda \in [0, \Phi/b]$.

To show that $y(t) > \Phi - \lambda(t)$ for all t, suppose to the contrary that $y(t^*) \leq \Phi - \lambda(t^*)$ for some t^*. Then the phase diagram in Figure 1 shows that the path will remain below the line $y = \Phi - \lambda$ afterwards, because $\frac{dy(t)}{dt} < 0$ and $\frac{d\lambda(t)}{dt} < 0$ for $t > t^*$. Eventually y becomes negative. This contradicts the fact that y is positive.

(ii) Part (i) states that $\Phi - \lambda(t) < y(t) < \frac{\Phi}{b} - \lambda(t)$. It then follows from equations (16) and (17) that $\frac{d\lambda}{dt} > 0$ and $\frac{dy}{dt} < 0$. To show that $\lim_{t\to\infty}(\lambda(t), y(t)) = (\Phi, 0)$, note that if a path between the two straight lines does not converge to $(\Phi, 0)$, then it will eventually hit either one of the straight lines, or the horizontal axis $y = 0$. Hitting a straight line contradicts part (i) and hitting the horizontal axis contradicts that $y > 0$.

To prove $\frac{d\lambda}{dt} + \frac{dy}{dt} < 0$, suppose to the contrary that $\frac{d(\lambda+y)(t^*)}{dt} \geq 0$ for some t^*. We argue that $\frac{d(\lambda+y)(t)}{dt} \geq 0$ for all $t \geq t^*$. By contradiction, suppose $t^{**} \equiv \inf_s\{s > t^* : \frac{d(\lambda+y)(s)}{dt} < 0\}$ exists. It is easily seen that $\frac{d(\lambda+y)(t^{**})}{dt} = 0$ and $\frac{d^2(\lambda+y)(t^{**})}{dt^2} \leq 0$. Adding equations (16) and (17) yields $\frac{d(\lambda+y)}{dt} = \frac{ry^2}{\Phi/b-\lambda} - ry + \pi(\lambda + y - \Phi)$. It then follows from $\frac{d\lambda(t^{**})}{dt} = -\frac{dy(t^{**})}{dt}$ that

$$\begin{aligned} \frac{d^2(\lambda+y)(t^{**})}{dt^2} &= \left(\frac{2ry(\Phi/b-\lambda) - ry^2}{(\Phi/b-\lambda)^2} - r\right)\frac{dy(t^{**})}{dt} \\ &= -r\frac{(\Phi/b-\lambda-y)^2}{(\Phi/b-\lambda)^2}\frac{dy(t^{**})}{dt} > 0, \end{aligned}$$

which contradicts that $\frac{d^2(\lambda+y)(t^{**})}{dt^2} \leq 0$. Since $\frac{d(\lambda+y)(t)}{dt} \geq 0$ for all $t \geq t^*$, equation (16) and $(\lambda + y) > \Phi$ imply that $\lambda(t)$ grows unboundedly after t^*, and contradicts the fact that $\lambda(t)$ is bounded.

(iii) That $\frac{dH}{dt} < 0$ and $\lim_{t\to\infty} H(t) = -\infty$ follow from that $\frac{dy}{dt} < 0$ and $\lim_{t\to\infty} y(t) = 0$, because $y \equiv \frac{-1}{\rho H}$. To see that $\frac{dL}{dt} < 0$, recall equation (7),

$$\Phi L = H\lambda - 1/\rho.$$

Since $\lambda \geq 0$ is increasing and $H < 0$ is decreasing, $H\lambda$ and L both decrease with time. It follows from $L(t) < H(t)$ and $\lim_{t\to\infty} H(t) = -\infty$ that $\lim_{t\to\infty} L(t) = -\infty$. Dividing both sides of equation (7) by ΦH yields $\frac{L}{H} = (\lambda + y)/\Phi$. Thus it follows from part (ii) that $d\left(\frac{L(t)}{H(t)}\right)/dt < 0$ and $\lim_{t\to\infty} \frac{L(t)}{H(t)} = 1$.

(iv) Differentiating equation (7) with respect to H yields $C_{LH}L + C_{HH}H = -C_H$, which is

$$C_{LH} = -C_{HH}\frac{H}{L} - \frac{C_H}{L}. \tag{27}$$

Substituting equation (27) into

$$\frac{dC_H}{dt} = C_{LH}\frac{dL}{dt} + C_{HH}\frac{dH}{dt},$$

we get

$$C_{HH}\left(\frac{dH/dt}{dL/dt} - \frac{H}{L}\right) = \frac{dC_H/dt}{dL/dt} + \frac{C_H}{L}.$$

Then $C_{HH} > 0$ follows from $dC_H/dt = d(-\lambda)/dt < 0$, $dL/dt < 0$, $C_H \leq 0$, $L < 0$, and $\frac{dH/dt}{dL/dt} - \frac{H}{L} > 0$.

Equation (27) and $C_{HH} > 0$ imply that $C_{LH} < 0$, which, together with equation $C_{LL}\frac{dL}{dt} + C_{LH}\frac{dH}{dt} = \frac{dC_L}{dt} = 0$, imply that $C_{LL} > 0$. To finish the proof, note that

$$\begin{aligned} C_{LH}\frac{dL}{dt} + C_{HH}\frac{dH}{dt} &= \frac{dC_H}{dt} < 0, \\ C_{LL}\frac{dL}{dt} + C_{LH}\frac{dH}{dt} &= \frac{dC_L}{dt} = 0, \end{aligned}$$

imply that

$$\begin{aligned} \left|C_{LH}\frac{dL}{dt}\right| &< \left|C_{HH}\frac{dH}{dt}\right|, \\ \left|C_{LH}\frac{dH}{dt}\right| &= \left|C_{LL}\frac{dL}{dt}\right|. \end{aligned}$$

Therefore, $C_{LH}^2 < C_{LL}C_{HH}$.

(v) For a fixed L, we first show that $\lim_{\epsilon\downarrow 0} C_L(L, L+\epsilon) = \infty$. Consider a path $(L(t), H(t))$ with $L(0) = L$ and $H(0)$ chosen optimally so that $C_H(L(0), H(0)) = 0$. Homogeneity (6) implies that

$$C_L\left(L, \frac{H(t)}{L(t)}L\right) = \frac{L(t)}{L}C_L(L(t), H(t)) = \frac{L(t)}{L}C_L(L(0), H(0)).$$

Since $\lim_{t\to\infty} L(t) = -\infty$, it follows that $\lim_{t\to\infty} C_L\left(L, \frac{H(t)}{L(t)}L\right) = \infty$. Since $\lim_{t\to\infty}\frac{H(t)}{L(t)} = 1$, we have

$$\lim_{\epsilon\downarrow 0} C_L(L, L+\epsilon) = \lim_{t\to\infty} C_L\left(L, \frac{H(t)}{L(t)}L\right) = \infty.$$

Second we show $\lim_{H\downarrow L} C(L, H) = \infty$. Consider a path $(L(t), H(t))$ with $L(0) = L$ and $H(0) = L+\epsilon$, where $\epsilon > 0$ is a small number. Equation (9) implies $c'(u^L(t)) = \Phi - b\lambda \geq (1-b)\Phi = (1-b)C_L(L(0), L(0)+\epsilon)$. If ϵ is sufficiently small, $u^L(t)$ and c_t^L become sufficiently large, which implies that $\lim_{H\downarrow L} C(L, H) = \infty$.

Q.E.D.

Proof of Lemma 4: To show that $p > 0$ at $t = N$, suppose to the contrary that $p = 0$, then equations (18) and (19) imply

$$\frac{dH}{dt} = rH - rbu^L < r(L - u^L) = \frac{dL}{dt},$$

which contradicts that $H(t) \geq L(t)$, for all $t \geq N$. *Q.E.D.*

Proof of Lemma 5: To prove that $p = 0$ is optimal when $L < H$, it is sufficient to show the first-order condition

$$C(\tilde{L}, \tilde{H}) + \gamma - C(L, H) - C_L(L, H)(\tilde{L} - L) - C_H(L, H)(e^{\rho\psi} - 1)H > 0, \text{ when } L < H.$$

Firstly, we show that when $L = H$,

$$C(\tilde{L}, \tilde{H}) + \gamma - C(L, H) - C_L(L, H)(\tilde{L} - L) - C_H(L, H)(e^{\rho\psi} - 1)H = 0. \tag{28}$$

By contradiction, suppose $C(\tilde{L}, \tilde{H}) + \gamma - C(L, H) - C_L(L, H)(\tilde{L} - L) - C_H(L, H)(e^{\rho\psi} - 1)H \neq 0$. If

$$C(\tilde{L}, \tilde{H}) + \gamma - C(L, H) - C_L(L, H)(\tilde{L} - L) - C_H(L, H)(e^{\rho\psi} - 1)H < 0,$$

then picking a large enough $p > 0$ makes the right side of the HJB (20) less than the left side, which is a contradiction. If

$$C(\tilde{L}, \tilde{H}) + \gamma - C(L, H) - C_L(L, H)(\tilde{L} - L) - C_H(L, H)(e^{\rho\psi} - 1)H > 0,$$

then since $p > 0$ when $L = H$,

$$\begin{aligned}
(r + \pi)C(L, H) &= \min_{u^L, p, \tilde{L}, \tilde{H}} \quad rc(u^L) + \pi c(H) + p(C(\tilde{L}, \tilde{H}) + \gamma - C(L, H)) \\
&\qquad + C_L(L, H)\left((r + \pi)L - \pi H - ru^L - p(\tilde{L} - L)\right) \\
&\qquad + C_H(L, H)\left(rH - rbu^L - p(e^{\rho\psi} - 1)H\right) \\
&> \min_{u^L} \quad rc(u^L) + \pi c(H) + C_L(L, H)\left(r(L - u^L) - \pi(H - L)\right) \\
&\qquad + C_H(L, H)\left(rH - rbu^L\right).
\end{aligned}$$

By continuity, the above strict inequality at $t = N$ continues to hold when t is sufficiently close to N, which violates the HJB equation (21) when $L < H$.

Secondly, we show that,

$$C(\tilde{L}, \tilde{H}) + \gamma - C(L, H) - C_L(L, H)(\tilde{L} - L) - C_H(L, H)(e^{\rho\psi} - 1)H > 0, \text{ when } L < H.$$

The first-order conditions for $(\tilde{L}, \tilde{H})$ in (20) are

$$\begin{aligned}
C_L(\tilde{L}, \tilde{H}) &= C_L(L, H), \\
C_H(\tilde{L}, \tilde{H}) &= 0.
\end{aligned}$$

Because $C_L(L,H)$ remains constant when $t<N$, these first-order conditions imply that $\tilde{L}=L(0)$ and $\tilde{H}=H(0)$. Since $\tilde{L}$ and $\tilde{H}$ are constants for all $t<N$, the derivative of $(C(\tilde{L},\tilde{H})+\gamma-C(L,H))-C_L(L,H)(\tilde{L}-L)-C_H(L,H)(e^{\rho\psi}-1)H$ with respect to t is

$$\begin{aligned}
& -C_L(L,H)\frac{dL}{dt}-C_H(L,H)\frac{dH}{dt}+C_L(L,H)\frac{dL}{dt} \\
& -C_H(L,H)(e^{\rho\psi}-1)\frac{dH}{dt}-\frac{dC_H(L,H)}{dt}(e^{\rho\psi}-1)H \\
= \; & -C_H(L,H)e^{\rho\psi}\frac{dH}{dt}-\frac{dC_H(L,H)}{dt}(e^{\rho\psi}-1)H<0,
\end{aligned}$$

where the inequality follows from $\frac{dH}{dt}<0$, $-\frac{dC_H(L,H)}{dt}=\frac{d\lambda}{dt}>0$ and $H<0$, as shown in parts (ii) and (iii) of LEMMA 3. Because $(C(\tilde{L},\tilde{H})+\gamma-C(L,H))-C_L(L,H)(\tilde{L}-L)-C_H(L,H)(e^{\rho\psi}-1)H$ decreases over time and it reaches zero at time N, as shown in (28), $(C(\tilde{L},\tilde{H})+\gamma-C(L,H))-C_L(L,H)(\tilde{L}-L)-C_H(L,H)(e^{\rho\psi}-1)H$ must be positive when $t<N$. *Q.E.D.*

PROOF OF PROPOSITION 1: The time path $(L(t),H(t))$ stays on the 45-degree when $t\geq N$ because all the states above the 45-degree line will eventually converge to it, as can be seen in Figure 3. To see that the state is not constant after it reaches the 45-degree line, we show that $\frac{dL}{dt}=\frac{dH}{dt}<0$. Since $L=H=-\frac{1}{\rho y}$ and $u^L=-\frac{1}{\rho(\Phi-b\lambda)}$, equation (18) is

$$\begin{aligned}
\frac{dL}{dt} &= (r+\pi)L-\pi H-ru^L-p(\tilde{L}-L) \\
&= r(L-u^L)-p(\tilde{L}-L) \\
&= r\left(\frac{1}{\rho(\Phi-b\lambda)}-\frac{1}{\rho y}\right)-p(\tilde{L}-L).
\end{aligned}$$

Because $y=\Phi-\lambda$ on the 45-degree line, $\frac{1}{\rho(\Phi-b\lambda)}<\frac{1}{\rho y}$. It follows from the first-order conditions for $(\tilde{L},\tilde{H})$ that

$$\tilde{L}=\frac{-1/\rho}{C_L(\tilde{L},\tilde{H})} \geq \frac{-1/\rho-C_H(L,H)H}{C_L(\tilde{L},\tilde{H})}=\frac{-1/\rho-C_H(L,H)H}{C_L(L,H)}=L.$$

Hence,

$$\frac{dL}{dt} = r\left(\frac{1}{\rho(\Phi-b\lambda)}-\frac{1}{\rho y}\right)-p(\tilde{L}-L)<0.$$

The auditing arrival rate p can be solved by $\frac{dL}{dt}=\frac{dH}{dt}$. Equations (18), (19), and $L=H$ yield

$$(r+\pi)L-\pi H-ru^L-p(\tilde{L}-L) = rH-rbu^L-p(e^{\rho\psi}-1)H,$$

and

$$p = \frac{r(b-1)u^L}{\tilde{L}-e^{\rho\psi}L}>0.$$

Q.E.D.

PROOF OF LEMMA 6: First, we show that the objective $M(P)$ in (22) is a strictly convex function of P. Denote the optimal solution by $\tilde{L}(P)$, $\tilde{H}(P)$, $\bar{L}(P)$, and $\bar{H}(P)$. Suppose $\theta \in (0,1)$ and $P = \theta P_1 + (1-\theta)P_2$, $P_1 \neq P_2$. Then construct a solution for the problem $M(P)$ as follows,

$$
\begin{aligned}
(\tilde{L}, \tilde{H}) &= \frac{\theta P_1}{\theta P_1 + (1-\theta)P_2}\left(\tilde{L}(P_1), \tilde{H}(P_1)\right) + \frac{(1-\theta)P_2}{\theta P_1 + (1-\theta)P_2}\left(\tilde{L}(P_2), \tilde{H}(P_2)\right), \\
(\bar{L}, \bar{H}) &= \frac{\theta(1-P_1)}{\theta(1-P_1) + (1-\theta)(1-P_2)}\left(\bar{L}(P_1), \bar{H}(P_1)\right) \\
&\quad + \frac{(1-\theta)(1-P_2)}{\theta(1-P_1) + (1-\theta)(1-P_2)}\left(\bar{L}(P_2), \bar{H}(P_2)\right).
\end{aligned}
$$

Using the same proof as that in part (iv) in LEMMA 3, we can prove that the cost function $C(L, H)$ is strictly convex. Therefore,

$$
\begin{aligned}
& M(P) \\
\geq\ & P\left(C(\tilde{L}, \tilde{H}) + \gamma\right) + (1-P)C(\bar{L}, \bar{H}) \\
>\ & P\left(\frac{\theta P_1}{\theta P_1 + (1-\theta)P_2} C\left(\tilde{L}(P_1), \tilde{H}(P_1)\right) + \frac{(1-\theta)P_2}{\theta P_1 + (1-\theta)P_2} C\left(\tilde{L}(P_2), \tilde{H}(P_2)\right)\right) + P\gamma \\
& + (1-P)\Big(\frac{\theta(1-P_1)}{\theta(1-P_1) + (1-\theta)(1-P_2)} C(\bar{L}(P_1), \bar{H}(P_1)) \\
& + \frac{(1-\theta)(1-P_2)}{\theta(1-P_1) + (1-\theta)(1-P_2)} C(\bar{L}(P_2), \bar{H}(P_2))\Big) \\
=\ & \theta\left(P_1\left(C\left(\tilde{L}(P_1), \tilde{H}(P_1)\right) + \gamma\right) + (1-P_1)C(\bar{L}(P_1), \bar{H}(P_1))\right) \\
& + (1-\theta)\left(P_2\left(C\left(\tilde{L}(P_2), \tilde{H}(P_2)\right) + \gamma\right) + (1-P_2)C(\bar{L}(P_2), \bar{H}(P_2))\right) \\
=\ & \theta M(P_1) + (1-\theta)M(P_2).
\end{aligned}
$$

Second, we show that, for all $P > 0$, $M(P) > C(L, H) = M(0)$. Because $M(P)$ is strictly convex, it is sufficient to prove that $M'(0) \geq 0$. To finish the proof, we will show that

$$
M'(0) = \left(C(\tilde{L}, \tilde{H}) + \gamma - C(L, H)\right) - \left(C_L(L, H)(\tilde{L} - L) + C_H(L, H)(e^{\rho\psi} - 1)H\right).
$$

Denote the Lagrangian multipliers on constraints (23) and (24) by $\xi_L(P)$ and $\xi_H(P)$, respectively. Then

$$
\begin{aligned}
M'(P) &= \left(C(\tilde{L}(P), \tilde{H}(P)) + \gamma - C(\bar{L}(P), \bar{H}(P))\right) \\
&\quad - \left(\xi_L(P)(\tilde{L}(P) - \bar{L}(P)) + \xi_H(P)(e^{\rho\psi}H - \bar{H}(P))\right), \\
\xi_L(P) &= C_L\left(\tilde{L}(P), \tilde{H}(P)\right) = C_L\left(\bar{L}(P), \bar{H}(P)\right), \\
\xi_H(P) &= C_H(\bar{L}(P), \bar{H}(P)).
\end{aligned}
$$

Since $\lim_{P\to 0} \bar{L}(P) = L$, $\lim_{P\to 0} \bar{H}(P) = H$, we have

$$\lim_{P\downarrow 0} M'(P) = \Big(C(\tilde{L},\tilde{H}) + \gamma - C(L,H)\Big) - \Big(C_L(L,H)(\tilde{L}-L) + C_H(L,H)(e^{\rho\psi}-1)H\Big) \geq 0.$$

Q.E.D.

References

Becker, G. (1968): "Crime and Punishment: An Economic Approach," *Journal of Political Economy*, 76, 169–217.

Fernandes, A., and C. Phelan (2000): "A Recursive Formulation for Repeated Agency with History Dependence," *Journal of Economic Theory*, 91, 223–247.

Green, E. (1987): *Contractual Agreement for Intertemporal Trade*, vol. Lending and the Smoothing of Uninsurable Income. University of Minnesota Press.

Monnet, C., and E. Quintin (2005): "Optimal Contracts in a Dynamic Costly State Verification Model," *Economic Theory*, 26, 867–885.

Popov, L. (2007): "Stochastic Costly State Verification and Dynamic Contracts," *working paper, University of Iowa.*

Reinganum, J., and L. Wilde (1985): "Income Tax Compliance in a Principal-Agent Framework," *Journal of Public Economics*, 26, 1–18.

——— (1986): "Equilibrium Verification and Reporting Policies in a Model of Tax Compliance," *International Economic Review*, 27, 739–760.

Stigler, G. (1970): "The Optimum Enforcement of Laws," *Journal of Political Economy*, 78, 526–536.

Townsend, R. (1979): "Optimal Contracts and Competitive Markets With Costly State Verification," *Journal of Economic Theory*, 21, 265–293.

Wang, C. (2005): "Dynamic Costly State Verification," *Economic Theory*, 25, 887–916.

Zhang, Y. (2009): "Dynamic Contracting with Persistent Shocks," *Journal of Economic Theory*, 144, 635–675.