FEDERAL RESERVE BANK of ST. LOUIS

RESEARCH DIVISION

Working Paper Series

Unemployment Insurance Fraud and Optimal Monitoring

David L. Fuller,
B. Ravikumar
and
Yuzhe Zhang

Working Paper 2012-024D
https://doi.org/10.20955/wp.2012.024

June 2014

FEDERAL RESERVE BANK OF ST. LOUIS
Research Division
P.O. Box 442
St. Louis, MO 63166

Unemployment Insurance Fraud and Optimal Monitoring*

David L. Fuller†, B. Ravikumar‡, and Yuzhe Zhang§

June 2014

Abstract

An important incentive problem for the design of unemployment insurance is the fraudulent collection of unemployment benefits by workers who are gainfully employed. We show how to efficiently use a combination of tax/subsidy and monitoring to prevent such fraud. The optimal policy monitors the unemployed at fixed intervals. Employment tax is nonmonotonic: it increases between verifications but decreases after a verification. Unemployment benefits are relatively flat between verifications but decrease sharply after a verification. Our quantitative analysis suggests that the optimal monitoring cost is 60 percent of the cost in the current U.S. system.

JEL Classification Numbers: D82, D86, J65.
Keywords and Phrases: Unemployment Insurance, Fraud, Concealed Earnings, Costly State Verification.

*We are grateful to the editor, Richard Rogerson, and an anonymous referee for comments that greatly improved the paper. We are also grateful to Árpád Ábrahám, Nicola Pavoni, seminar participants at the Federal Reserve Bank of St. Louis, University of Missouri, and Toulouse School of Economics, and participants at the Workshop on Macroeconomic Applications of Dynamic Games and Contracts, Midwest Macroeconomics Meeting, Midwest Theory Meeting, Asia Meeting of the Econometric Society, Society for the Advancement of Economic Theory Conference, and Tsinghua Workshop in Macroeconomics for their helpful comments. We would also like to thank George Fortier for editorial assistance. The views expressed in this article are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of St. Louis or the Federal Reserve System.

†Department of Economics, Concordia University, and CIREQ. Email: david.fuller@concordia.ca
‡Research Division, Federal Reserve Bank of St. Louis. Email: b.ravikumar@wustl.edu
§Department of Economics, Texas A&M University. Email: yuzhe-zhang@econmail.tamu.edu

1 Introduction

Unemployment insurance programs insure workers against the risk of losing their jobs through no fault of their own. Such insurance, however, has many potential incentive problems. In this paper, we study the incentive problem associated with fraudulent collection of unemployment benefits. The U.S. Department of Labor finds that more than 60 percent of unemployment insurance fraud overpayments are attributed to concealed earnings fraud—when a worker collecting unemployment benefits finds a job but continues collecting the benefits. Motivated by this fact, we study optimal unemployment insurance in an environment where workers can conceal earnings *and* collect unemployment benefits.

We study an infinitely lived worker in continuous time who has CARA preferences, is initially unemployed, and faces a stochastic arrival of employment opportunities. Employment is assumed to be an absorbing state. An employed worker can conceal his employment status and continue to claim unemployment benefits. The worker's employment status can be detected using a costly monitoring technology. In order to focus on the issue of hidden employment, we abstract from moral hazard issues by assuming that there is no search effort decision and that the wage offer distribution is degenerate.[1]

In our model, there are two instruments to deter fraudulent collection of unemployment benefits: tax/subsidy and monitoring. Both instruments are costly: The first distorts consumption relative to full insurance, and the second has a direct cost. We deliver a pre-commitment mechanism that optimally trades off between the two instruments. Our mechanism allows both instruments to be fully history dependent. As a result, the unemployed worker's consumption (i.e., the unemployment benefits) and the employed worker's consumption vary over time.

Since employment is an absorbing state in our model, the treatment of the worker who reports transitioning to employment is straightforward: constant consumption forever and

[1]The literature on the optimal provision of unemployment insurance concentrates on moral hazard and examines incentives for optimal search effort (e.g., Baily (1978), Shavell and Weiss (1979), and Hopenhayn and Nicolini (1997)). Hopenhayn and Nicolini (1997) and Wang and Williamson (2002) show that the search effort margin is quantitatively insignificant: The unemployed worker's *optimal* search effort almost equals what the current U.S. system implies.

no monitoring. Since employment status is private information, the worker who reports being unemployed is not fully insured and is monitored.

We consider two monitoring mechanisms: deterministic verification and stochastic verification. Under deterministic verification, the worker is either verified with probability one or not verified at all. We focus on this case for most of the paper since it is simpler and makes the results more transparent. We show later that our results remain the same under stochastic verification, where the worker is verified with a probability between zero and one. That is, even though our deterministic mechanism appears restrictive, the general mechanism of stochastic verification does not offer any additional economic insights on unemployment insurance and monitoring.

Under deterministic verification the optimal contract has three key features. First, monitoring occurs at fixed intervals and is independent of history. Second, the unemployment benefits decrease with the duration of unemployment between monitoring dates and jump downward at every monitoring date. Third, there is a nonmonotonic tax on employment.

The periodicity of monitoring follows from that fact that with CARA preferences the worker's utility flows in a new cycle are proportional to those in the previous cycle. Hence, his incentive to commit fraud remains the same and he is monitored in the same manner as in the previous cycle. Unemployment benefits decreasing with duration is a familiar feature from the previous literature. Unemployment benefits jump downward at the monitoring date because the unemployed worker's pre-monitoring consumption is distorted upward. In our model, increasing the unemployed worker's pre-monitoring consumption benefits the truth-teller more than it benefits the liar.[2] Within a monitoring cycle, the employment tax increases with duration of unemployment: the consumption for the worker who transitions to employment earlier exceeds that of the worker who transitions later. However, the employment tax decreases after the monitoring date. This is because the unemployed worker who transitions to employment shortly after the monitoring date can conceal earnings until the next monitoring date, while the worker who transitions to employment at the monitoring date cannot.

[2]For the same reason, in Mirrleesian taxation models with hidden ability, the labor supply of a low-ability worker is distorted downward.

Our optimal mechanism also deters fraud from quits. This occurs when workers quit their jobs, become unemployed, and start collecting unemployment benefits. The incentives in our optimal contract ensure that the employed workers do not engage in such behavior.[3]

To assess the empirical relevance of our theoretical analysis, we conduct a partial equilibrium quantitative exercise similar to Hopenhayn and Nicolini (1997). We find that the optimal monitoring cost is 60 percent of the cost incurred by the U.S. unemployment insurance system. Furthermore, using the same resources as the U.S. system, the optimal contract delivers higher utility to the average worker: 1.55 percent higher consumption at every date. This gain arises from two sources: (i) improved consumption smoothing between employed and unemployed states and (ii) reduced monitoring costs (or higher average consumption). Almost all of the gain in our optimal contract comes from (i). This is similar to the quantitative finding in Hopenhayn and Nicolini (1997) and Wang and Williamson (2002). The cost saving in their optimal contracts is due to improved consumption smoothing and not due to faster transitions from unemployment to employment.

The remainder of the paper proceeds as follows. In Section 2, we present the key facts on unemployment insurance fraud. We also provide evidence that deterring concealed earnings fraud involves a case-by-case investigation and, thus, a per-case cost, as in our model. Section 3 describes the model. In Section 4 we establish two properties of the optimal mechanism: scaling and periodic monitoring. In Section 5 we use these properties to analyze the optimal unemployment insurance scheme with exogenously given monitoring dates. Then, we characterize the optimal monitoring dates in Section 6. In Section 7 we show that our mechanism prevents employed workers from quitting. In Section 8 we examine the stochastic monitoring case. In this section, we also describe the similarities and differences between the insights from the deterministic mechanism and the insights from the stochastic mechanism. We conclude in Section 9.

[3] Hansen and Imrohoroglu (1992) study a model where unemployed workers can reject job offers and an exogenous fraction of such workers are denied benefits. In our optimal mechanism, the unemployed worker who receives a job offer has no incentive to refuse the offer.

2 Unemployment Insurance Fraud Data

In this section, we first briefly describe the program in place for determining the accuracy of payments in the U.S. unemployment insurance system. Second, we provide details on the nature of "fraud" overpayments by category for 2007 (Appendix A provides information for more years). Third, we present data on how these payments were detected. Finally, we discuss "off-the-books" employment.

Accuracy of Benefit Payments Unemployment insurance benefits in the U.S. are paid out by the states, with each state deciding its benefit levels and how to finance the benefits. The U.S. Department of Labor's BAM (Benefit Accuracy Measurement) program determines the accuracy of these expenditures by choosing a random sample of weekly unemployment insurance claims and determining whether there were any overpayments. The investigators also interview some claimants if necessary. Some overpayments are simple errors in calculating benefits, while some represent *fraud* overpayments.

The goal of the program is different from the goal of unemployment insurance fraud investigators. While the latter look to *recapture* overpayments, BAM investigators calculate statistics of the unemployment insurance program (see BAM State Operations Handbook ET No. 495, 4th edition). We use these statistics throughout the paper.

Overpayments due to Fraud There are several types of unemployment insurance fraud. Examples include collecting unemployment benefits while being employed, after quitting a job, or after refusing a suitable job offer. Table 1 categorizes the overpayments by type of fraud.

"Concealed Earnings" refers to cases where payments are made to individuals who are simultaneously earning wages and collecting unemployment benefits. "Insufficient Job Search" refers to cases where individuals did not meet the mandatory work search requirement (e.g., a minimum number of job applications must be filed each week). "Refused Suitable Offer" refers to cases where individuals were offered a job deemed suitable, but rejected it. "Quits" and "Fired," respectively, refer to cases where payments are made to individuals who voluntarily left their jobs or who were fired from their jobs for a valid

Table 1: Unemployment Insurance Overpayments in the U.S., 2007

Category	Percent of Fraud Overpayments
Concealed Earnings	60.06
Insufficient Job Search	4.95
Refused Suitable Offer	0.80
Quits	7.06
Fired	13.29
Unavailable for Work	4.17
Other	9.67
Total	100.00

Source: BAM program, U.S. Department of Labor. Note that these are our calculations. Our definitions of each type of fraud differ slightly from those used in the BAM reports available online.

reason (e.g., poor performance or missing work). "Unavailable for Work" refers to cases where payments are made to individuals who cannot work (e.g., disability).

Overpayments due to concealed earnings fraud in 2007 were *ten* times overpayments due to unemployed agents not actively searching or refusing suitable work (see Table 1). While the data indicate that concealed earnings fraud is the dominant source of overpayments, it does not imply that moral hazard from reduced search effort is unimportant for the design of unemployment insurance. It might be the case that the current unemployment insurance system provides adequate incentives to search but does not deter concealed earnings fraud.

Detection Technologies The detection technologies used by BAM are shown in Table 2. For example, "Verification of search contact" refers to cases when the BAM investigator verifies the potential job contact reported by the unemployed person; "Claimant interview" is an interview with the person collecting benefits.

Since 2003, states have used a cross-matching technology, comparing unemployment insurance records with employment records. One might think concealed earnings fraud could be automatically detected this way; however, only 7.5 percent of the fraud cases are detected by cross-matching with the state's directory of new hires (see Table 2). For in-

Table 2: Detection Technologies, 2007

Detection Method	*Percent of concealed earnings fraud overpayments detected by method*
Verification of search contact	1.31
Verification of wages and/or separation	62.02
Claimant interview	10.41
Verification of eligibility with 3rd parties	1.38
Unemployment insurance records	14.61
Job/employment service records	0.17
Verification with union	0.71
Crossmatch with state directory of new hires	7.52
Crossmatch with state wage record files	1.86

Source: Benefit Accuracy Measurement Program, U.S. Department of Labor

stance, cross-matching technology would not automatically catch a worker who is collecting unemployment benefits in one state while employed in another state. Furthermore, the directory of new hires is updated monthly, so even within individual states some workers who truthfully report unemployment in a specific week may show up in a cross-match of employment records and be mistakenly flagged for fraud. In most cases when a worker appears in both unemployment insurance records and employment records, further investigation is necessary to determine if fraud has actually occurred.

In addition, the worker could commit a more nuanced form of concealed earnings fraud by truthfully reporting the transition to employment but underreporting the earnings. (The worker is entitled to collect some unemployment benefits as long as the reported earnings are sufficiently low.) In 2007, roughly 40 percent of those committing concealed earnings fraud reported positive earnings. Less than 2 percent of these cases were detected by cross-matching the unemployment insurance records with wage records (updated quarterly) in each state (see Table 2). In fact, employees working in a sector not covered by the unemployment insurance system will never show up in the state wage records (e.g., federal employees and self-employed).

These data suggest that more than 90 percent of the overpayments due to concealed earnings fraud were not detectable under the automatic procedures available to the state authorities. Instead, detection involves a case-by-case investigation and, thus, a per-case cost of verification.

Working "Off-the-Books" A worker could collect unemployment benefits while working "off-the-books" and being paid in cash. In such cases, verifying the true employment status might be prohibitively expensive. However, the evidence suggests that concealed earnings fraud is committed by workers in "official" employment. While the worker is committing concealed earnings fraud, his weekly earnings are similar to the weekly earnings in the pre-unemployment job (which, by design, has to be official for the worker to collect unemployment benefits). In 2007, those committing concealed earnings fraud were earning 82 percent of their previous job's wages, on average. One-fourth of those committing this fraud were earning more while collecting benefits than before they became unemployed. Such relatively high earnings while committing fraud suggest official or "on-the-books" employment rather than "off-the-books" employment.[4]

3 Model

The Unemployment Insurance authority is a risk-neutral principal with a discount rate $r > 0$. She provides insurance to a risk-averse worker, whose preferences are given by

$$E\left[\int_0^\infty e^{-rt} r v(c(t)) dt\right],$$

where $c(t)$ is consumption at time t, $v(c) = -e^{-\rho c}$ is a CARA utility function with risk aversion ρ, r is the discount rate, and E is the expectation operator. Note that the flow utility is $rv(c)$ and that the agent's subjective discount rate is the same as the principal's.

A worker can be either employed with wage $w > 0$ or unemployed with wage zero. The worker is unemployed at $t = 0$ and transitions to employment with Poisson rate $\pi > 0$. We assume that employment is permanent. (For similar assumptions, see the unemployment

[4]The BAM program detects 10.5 percent of the fraud overpayments by interviewing the claimants (see Table 2). Such interviews might reveal some cash earnings.

insurance model of Hopenhayn and Nicolini (1997) and the disability insurance model of Golosov and Tsyvinski (2006).)

The worker's employment status is private information, so an employed worker can claim to be unemployed and continue collecting the unemployment benefits. We refer to this as *fraud*. The principal can verify the worker's unemployment report at a cost of γ units of the consumption good. Verification reveals the worker's true employment status.

We study pre-commitment mechanisms that efficiently deliver unemployment benefits and deter fraud. In addition to the tax/subsidy instrument used by the unemployment insurance literature, our mechanism uses the monitoring instrument to provide incentives.[5]

We assume that the principal always collects the wage, so an unemployed worker can never claim to be employed. Hence, there is no need for verification when the worker reports a transition to employment. Furthermore, since employment is an absorbing state, verification is unnecessary forever if the worker reports to be employed just once in the past. The incentive problem then reduces to ensuring that an employed worker does not claim to be unemployed.

We focus on *deterministic* verification mechanisms: in each period the worker is either verified with probability one or not verified at all. This mechanism is sub-optimal; it is dominated by a stochastic verification mechanism in our environment. One may then ask why study the deterministic case? Our goal is to characterize the optimal combination of the two instruments: tax/subsidy and monitoring. In Section 8, we show that the key economic insights on these two instruments are nearly identical in both the deterministic and stochastic cases. In both cases, optimal monitoring and employment tax have the same pattern. The stochastic monitoring case requires cumbersome notation and provides less intuition so we start by analyzing the deterministic case.

In our deterministic mechanism, the verification in any period is based on the history of employment status reports and past verifications outcomes. Since verification is necessary only for agents who have been reporting unemployment in every period in the past, a

[5]See Setty (2011) for a model of optimal unemployment insurance where the agent's search effort is monitored. Empirically, as noted in Table 1, fraudulent behavior in search effort is not as costly as concealment of earnings.

sufficient statistic for past history is the duration of unemployment reports. In other words, at $t = 0$ the principal commits to all future verification periods, mapping durations of unemployment reports to $\{0, 1\}$. In a verification period, clearly no worker would misreport. (Any penalty $\epsilon > 0$ induces truth telling in the verification period.) Thus, the principal does not have to keep track of the outcomes of past verifications. We represent the set of verification periods as $\{m_i; i = 1, 2, ...\}$, where m_i is the date of the ith verification.[6]

The timing is as follows. In the initial period, the worker is unemployed. Then the stochastic job opportunity arrives. The worker either remains unemployed or transitions to employment. He then chooses to report either employment or unemployment to the principal. Conditional on the unemployment report, the principal verifies the true employment status if the period is a verification period. Then, conditional on the report and the outcome of the verification, the principal assigns current and future consumptions. In subsequent periods, if the worker reported employment in the past, he is in an absorbing state and no further reports are necessary. If the worker reported unemployment in every period in the past, then the sequence of events is the same as in the initial period.

If an unemployed worker transitions to employment at t, let $c^E(t, s)$ denote his consumption at time $s \geq t$. Because the principal and the worker have the same discount rate and employment is an absorbing state, efficiency requires that the worker's consumption remain constant after t for all s. We therefore suppress s in $c^E(t, s)$ and denote this constant level of consumption as $c^E(t)$. The flow utility from this level of consumption then is $rv\left(c^E(t)\right)$. We denote the discounted sum of utilities to a worker who accepts a job offer for the first time at t as $E(t)$, i.e., $E(t) = \int_t^\infty e^{-r(s-t)} rv(c^E(t))ds = v(c^E(t))$. Since employment status is private information, $E(t)$ is also the continuation utility to a worker who accepted an offer before t, but reports employment for the first time at t.

An unemployed worker's consumption at t is denoted by $c^U(t)$ and his flow utility is $rv(c^U(t))$. His continuation utility,

$$U(t) \equiv \int_t^\infty e^{-r(x-t)} e^{-\pi(x-t)} rv(c^U(x))dx + \int_t^\infty e^{-r(x-t)} e^{-\pi(x-t)} \pi E(x)dx,$$

[6]There is no loss of generality in assuming a countable collection of verification periods. Since each verification costs $\gamma > 0$, the principal would not want to verify infinitely many times in any finite time interval.

is the sum of expected utilities before and after the transition ($e^{-\pi(x-t)}$ in the first integral is the conditional probability of remaining unemployed at date x and $e^{-\pi(x-t)}\pi$ in the second integral is the density function of the transition time). Hence,

$$\begin{aligned} U(t) &= \int_t^{\infty} e^{-(r+\pi)(x-t)} \left(\pi E(x) + ru(x)\right) dx \\ &= \int_t^{s} e^{-(r+\pi)(x-t)} \left(\pi E(x) + ru(x)\right) dx + e^{-(r+\pi)(s-t)} U(s), \text{ for all } t < s, \end{aligned} \tag{1}$$

where $u(x) \equiv v(c^U(x))$. We will refer to (1) as *promise-keeping* constraints.

The principal commits at $t = 0$ to verification periods $\{m_i; i = 1, 2, ...\}$ and consumptions $\left\{(c^E(t), c^U(t)); t \geq 0\right\}$. The verification periods and consumptions are history dependent. We denote this pre-commitment contract as σ.

Incentive compatibility requires that a worker who transitioned to employment at $t \in (m_i, m_{i+1})$ does not have the incentive to delay the report of the transition to a later time $s \in (t, m_{i+1})$, i.e., report unemployment and commit fraud from t to s, and then report employment from s onward:

$$E(t) \geq \int_t^{s} e^{-r(x-t)} rv(c^U(x) + w)dx + e^{-r(s-t)} E(s), \forall s \in (t, m_{i+1}). \tag{2}$$

Note that the worker cannot delay the report beyond the next verification period m_{i+1}.

We restrict contract allocations to

$$E(t) \geq U(t), \text{ for all } t. \tag{3}$$

Restriction (3) rules out the fraud due to refusal of offers noted in Table 1 (0.8 percent of total fraud overpayments). This restriction can be derived by adding a job-refusal option to our model. For ease of exposition we have imposed the restriction on the mechanism; Appendix B describes the job-refusal option and derives this restriction.

The expected cost for the principal is

$$C(\sigma) = \int_0^{\infty} e^{-(r+\pi)t} \left(\pi c^E(t) + rc^U(t)\right) dt + \sum_i e^{-(r+\pi)m_i} \gamma.$$

There should, in fact, be an additional term in $C(\sigma)$: the discounted income obtained by the principal, $\frac{\pi w}{r+\pi}$. However, unlike the unemployment insurance literature that endogenizes

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* σ that minimizes $C(\sigma)$ and delivers the initial promised utility $U(0)$, i.e.,

$$\begin{aligned} \min_{\sigma} \quad & C(\sigma) \\ \text{subject to} \quad & U(0) = \int_0^\infty e^{-(r+\pi)t} \left(\pi E(t) + ru(t)\right) dt, \\ & \text{and constraints (2), (3).} \end{aligned} \tag{4}$$

With a slight abuse of notation, denote the principal's cost function as $C(U(0))$.[7]

4 A Simplification of the Optimal Contract

We begin our analysis by presenting two features of the optimal contract. In Section 4.1 we establish a "scaling" property. Then, in Section 4.2 we show that the optimal monitoring is periodic. These properties simplify our analysis of the optimal contract by narrowing the search of a solution to problem (4) to a smaller space.

To help us simplify, we rewrite problem (4) in terms of continuation utilities $E(\cdot), U(\cdot)$ and flow variable $u(\cdot)$, instead of consumptions. The objective becomes

$$C(\sigma) = \int_0^\infty e^{-(r+\pi)t} \left(\pi c(E(t)) + rc(u(t))\right) dt + \sum_i e^{-(r+\pi)m_i} \gamma,$$

where $c : (-\infty, 0) \to \mathbb{R}$ denotes the inverse of the utility function:

$$c(v) = -\log(-v)/\rho. \tag{5}$$

The incentive constraint (2) becomes

$$E(t) \geq \int_t^s e^{-r(x-t)} e^{-\rho w} ru(x) dx + e^{-r(s-t)} E(s), \forall s \in (t, m_{i+1}), \tag{6}$$

[7] Ravikumar and Zhang (2012) analyze the problem of tax compliance in a costly state verification model where the verification technology is imperfect (a low-income agent might be mistakenly labeled as high income). They solve for the principal's cost function using the Hamilton-Jacobi-Bellman equation. In contrast, we study optimal unemployment insurance in an environment with a perfect verification technology. We characterize the path of unemployment benefits by formulating the optimal control problem and using the Pontryagin minimum principle.

since CARA utility implies that $v(c^U(x) + w) = e^{-\rho w} v(c^U(x)) = e^{-\rho w} u(x)$.

4.1 Scaling

Our mechanism exhibits a scaling property: if the initial promise $U(0)$ is scaled by $\alpha > 0$, then the optimal contract is also scaled by α. More formally,

LEMMA 1 *If* $\{(U(t), E(t), u(t)) ; t \geq 0\}$ *are optimal utilities for initial promise* $U(0)$*, then the optimal utilities for initial promise* $\alpha U(0)$ *are*

$$\{(\alpha U(t), \alpha E(t), \alpha u(t)) ; t \geq 0\}.$$

Alternatively, Lemma 1 states that the consumption of the worker with initial promise $\alpha U(0)$ differs from that of the worker with promise $U(0)$ by a constant, $-\log(\alpha)/\rho$, at all dates and states.

The scaling property in Lemma 1 is related to the fact that CARA utility has no wealth effect. Although a worker with high promised utility consumes (permanently) more than a worker with low promised utility, the level of promised utility does not have an effect on the worker's incentives to conceal earnings. In other words, the incentive constraint (6) holds when all of the utilities are scaled by the same factor.

Since the incentives to conceal earnings are the same for workers with different promised utilities, the optimal sequence of monitoring dates, $\{m_i; i \geq 1\}$, is independent of the initial promised utility. Again, no wealth effect implies that the level of promised utility does not change how the worker is monitored, even if it does change the worker's consumption.

4.2 Periodicity

At time 0, the principal knows the true employment status of the agent. After the verification at m_1, the principal again knows the true employment status. Hence, the continuation problem at m_1 is the same as the problem at time 0, except for the "initial" promised utility. The scaling property implies that, if $U(m_1) = \alpha U(0)$, then the optimal utilities from m_1 forward are scaled by α. Thus, starting with a promise $U(0)$, if the principal finds it optimal to monitor the unemployed agent at m_1, then it must be the case

that starting with the promise $\alpha U(0)$ the principal would again find it optimal to monitor at m_1. Put differently, having monitored the agent at m_1, the next optimal monitoring period is $2m_1$. We immediately conclude that

PROPOSITION 1 *The optimal monitoring is periodic, i.e., $m_i = im_1$ for all $i \geq 1$.*

To understand the intuition for the periodic monitoring, consider policies where the interval between verifications is either increasing or decreasing over time. First, it is suboptimal for the planner to verify more frequently at the beginning. Since the worker starts out unemployed, he stays unemployed for some duration initially. Frequent verifications early on merely incur unnecessary verification cost. Second, one might think that it is optimal to verify more frequently later since the probability of a long duration of unemployment is small. However, this policy is also suboptimal. The worker's conditional probability of transitioning to employment is independent of how long he has been unemployed. Moreover, because the principal knows the true employment status after each verification, the scaling property implies that from the principal's perspective the worker who was just verified to be unemployed is no different from the worker at time zero. Thus the interval between consecutive monitoring periods is a constant.

While we have established that the optimal monitoring is periodic, finding the optimal periodicity is difficult. To determine the optimal m_1 we must first determine the optimal utilities in the intervals $[0, m_1]$, $[m_1, 2m_1]$, etc. Toward this end, we break the principal's problem into two steps. First, assume that m_1 is exogenous and the principal learns the agent's employment status at dates m_1, $2m_1$, etc. Given m_1, the principal solves for the endogenous utility paths in $[0, m_1]$, $[m_1, 2m_1]$, etc. Second, the principal chooses m_1 optimally. We analyze the first step in the next section and the second step in section 6.

5 Optimal Unemployment Insurance with Exogenous Monitoring

Given the simplification in Section 4, we now present the features of the optimal unemployment insurance scheme. For a given m_1, we first formulate the optimal control problem

in Section 5.1. This allows us to analyze the time paths of the variables of interest. We then describe some features of the continuation utilities $E(\cdot)$ and $U(\cdot)$ in Section 5.2 and use these features to illustrate the employment tax in Section 5.3 and unemployment benefits in Section 5.4. Finally, in Section 5.5 we use the Pontryagin Minimum Principle to explicitly characterize $E(\cdot)$ and $U(\cdot)$.

5.1 Optimal Control

Following Zhang (2009), we formulate the principal's problem for interval $[0, m_1]$ as one of optimal control. Our analysis for $[0, m_1]$ applies to other intervals as well.

First, we rewrite the constraints recursively. The promise-keeping constraint (1) is equivalent to the differential equation:

$$U'(t) = r\left(U(t) - u(t)\right) + \pi\left(U(t) - E(t)\right).$$

On the right side of the differential equation, the first term is the rate of change of U when there is no uncertainty (i.e., when there is no transition to employment), and the second term captures the additional rate of change due to uncertainty.

The incentive constraint (6) is equivalent to the following differential inequality:

$$r(v(c^U(t) + w) - v(c^E(t))) + E'(t) \leq 0. \tag{7}$$

That is, the short term benefit that the agent gets from fraud, $r(v(c^U(t) + w) - v(c^E(t)))$, is offset by lower continuation utility he receives after he delays the employment report. Note that $E(\cdot)$ could have downward jumps: when $E(t) > \lim_{s \downarrow t} E(s)$, we interpret the discontinuity as $E'(t) = -\infty$, and the differential inequality (7) still holds under this interpretation. Introducing a slack variable $\mu(t) \geq 0$, we may rewrite (7) as

$$E'(t) = rE(t) - e^{-\rho w} r u(t) - \mu(t).$$

In Lemma C.1 in Appendix C, we show that the above differential equation and inequality are equivalent to (1) and (6).

Second, the scaling property implies that the cost function $C(\cdot)$ satisfies

$$C(\alpha U) \;=\; C(U) - \log(\alpha)/\rho.$$

Recalling the definition of $c(\cdot)$ in (5), we rewrite $C(U)$ as

$$C(U) = C\left(|U|(-1)\right) = C\left(-1\right) - \log(-U)/\rho \equiv \psi + c(U), \tag{8}$$

where $\psi \equiv C\left(-1\right)$ is the cost of private information: it is the one-time cost that the principal is willing to pay to permanently remove private information from the model.

With $\psi + c(U(m_1))$ as the continuation cost at m_1, we rewrite the principal's problem as one of optimal control with a convex objective and linear constraints.

$$\min_{\substack{u(t),U(t),E(t),\\ 0\le t\le m_1}} \int_0^{m_1} e^{-(r+\pi)t}\left(\pi c(E(t)) + rc(u(t))\right)dt + e^{-(r+\pi)m_1}(\gamma + \psi + c(U(m_1))) \tag{9}$$

$$\text{subject to} \quad U'(t) = (r+\pi)U(t) - \pi E(t) - ru(t), \tag{10}$$

$$E'(t) = rE(t) - e^{-\rho w}ru(t) - \mu(t), \tag{11}$$

$$E(t) \ge U(t), \tag{12}$$

$$U(0) \text{ is given.}$$

5.2 Continuation Utilities

The continuation utilities $E(\cdot)$ and $U(\cdot)$ help us uncover the consumption paths for the employed and the unemployed. We focus on the properties of $E(\cdot)$ and $U(\cdot)$ in $[0, m_1]$; those in other monitoring cycles can be obtained by scaling (see Lemma 1).

We demonstrate five properties:

(i) $E(t) > E(s)$ for $t < s \le m_1$.

(ii) $E(t) > U(t)$ for all $t < m_1$.

(iii) $E(m_1) = U(m_1)$.

(iv) $E(\cdot)$ jumps up immediately after m_1.

(v) $U(\cdot)$ declines over time.

Property (i) states that the payoff to a worker who reports the transition to employment earlier is higher than the payoff to one who reports the transition later. The worker who transitions to employment at t but commits fraud consumes $c^U(t) + w$ at t, whereas the

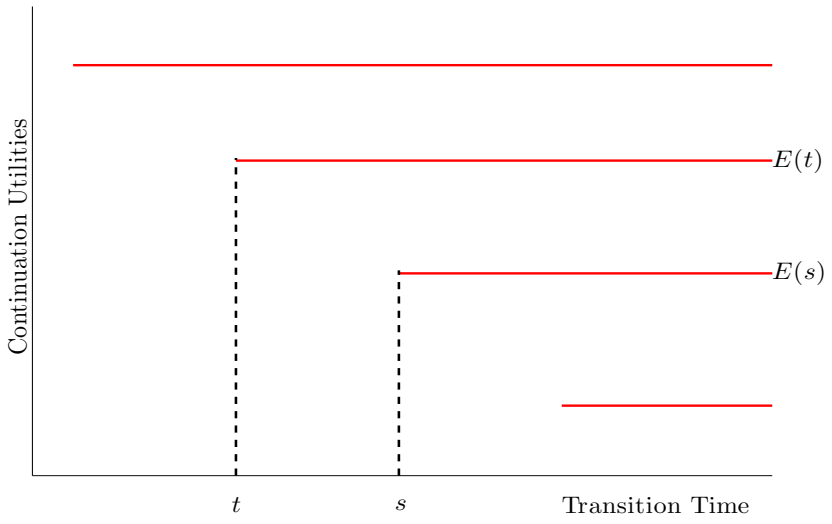

Figure 1: Lower payoff for late reporters ($E(t) > E(s)$ for $t < s$)

worker who tells the truth consumes $c^E(t)$. It is intuitive that $c^E(t) < c^U(t) + w$; otherwise deterring fraud would not be an issue. In terms of utilities, $E(t) < e^{-\rho w} u(t)$. Incentive compatibility (11) requires that delaying the report yields a lower payoff (see Figure 1). Thus, $E(t) > E(s)$ within a monitoring cycle.

For property (ii), recall that restriction (12) imposes $E(t)$ must be greater than or equal to $U(t)$. If the agent who transitions to employment before m_1 is offered the same payoff as the agent who remains unemployed, then the employed agent will claim to be unemployed and consume more than the unemployed agent. He can continue cheating until the verification period m_1 (see Figure 2). Thus, within a monitoring cycle, $E(t)$ must be greater than $U(t)$.

To understand (iii), note that the true employment status is revealed at m_1, so the principal does not face an incentive problem at that instant. Hence, there is no reason to reward the (lucky) agent who transitioned to employment at m_1 relative to the (unlucky) agent who remains unemployed i.e., no reason to set $E(m_1) > U(m_1)$. Thus, $E(m_1) = U(m_1)$. (Again, recall restriction (12): $E(t) \geq U(t)$ for all t.)

Property (iv) states that $U(m_1) = E(m_1) < E(m_1+)$, where $E(m_1+)$ is the utility for a worker who is unemployed at m_1 but transitions to employment immediately after m_1, i.e.,

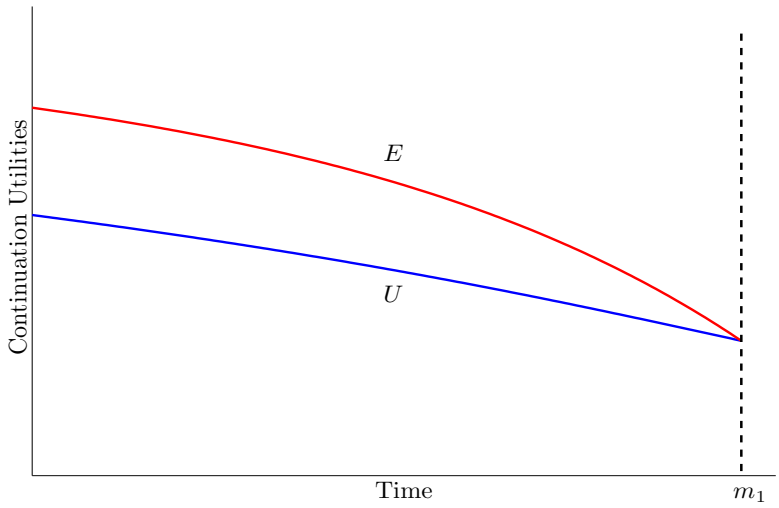

Figure 2: Continuation utilities $E(\cdot)$ and $U(\cdot)$ in $[0, m_1]$.

$E(m_1+) = \lim_{t \downarrow m_1} E(t)$ (see Figure 3). Suppose, to the contrary, that $U(m_1) = E(m_1+)$. Then incentive compatibility in $[m_1, 2m_1]$ would be violated because the worker employed immediately after m_1 can claim to be unemployed and consume more than the employed until the next verification period, $2m_1$. Note that if there is no verification at date t, then an upward jump in $E(\cdot)$ violates the incentive constraint: a worker who transitions to employment prior to t would benefit from delaying the employment report. At the moment of verification, however, the worker cannot delay the employment report since the true employment status is revealed.

To understand why $U(\cdot)$ declines, suppose $U(m_1) > U(0)$. Then lowering $U(m_1)$ has two benefits. First, the unemployed agent's continuation utility path is flatter, which implies better insurance for the unemployed. Second, lower $U(m_1)$ (and $E(m_1)$) reduces $E'(\cdot)$, generating stronger incentives to deter fraud. In addition, $U(\cdot)$ can never jump. Because $U(\cdot)$ is the promised utility to the unemployed agent, any jump in $U(\cdot)$ would violate the promise-keeping constraint.

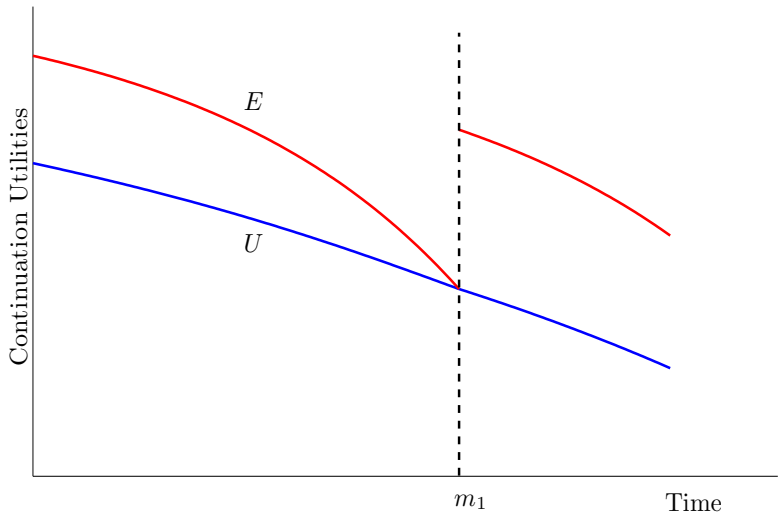

Figure 3: Continuation utility $E(\cdot)$ is nonmonotonic.

5.3 Employment Tax

Here we examine the consumption allocated to the agent who reports employment earlier relative to the consumption for the agent who reports it later. Recall that $E(t) > E(s)$ within a monitoring cycle and the continuation utility $E(\cdot)$ jumps up after verification. Since employment is an absorbing state, any agent who reports a transition to employment at t is allocated constant consumption $c^E(t)$ forever and is not monitored. Thus, $E(t)$ maps into $c^E(t)$ instant by instant and, hence, $c^E(t) > c^E(s)$ within a monitoring cycle. Furthermore, the consumption for the agent who reports the transition to employment immediately after m_1 is higher than that for the employed agent at m_1 (see Figure 4).

The nonmonotonicity is closely related to the way incentives are provided in our model. Within a cycle, the principal does not monitor, and relies exclusively on consumption distortions to induce truth-telling: c^E must fall sufficiently fast for the worker not to postpone his report of employment. At m_1, c^E falls to a level such that the agent is indifferent between transitioning to employment and remaining unemployed. The principal can perfectly insure the agent against the unemployment shock at m_1 because the true employment status is revealed. Immediately after m_1, the principal treats the worker

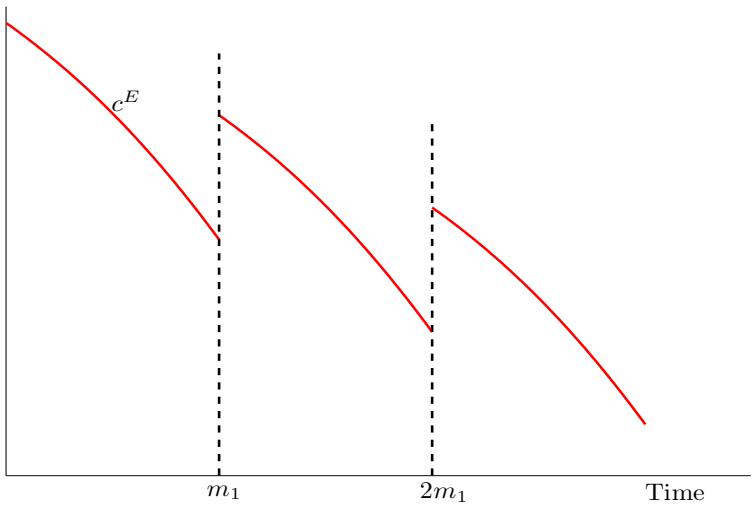

Figure 4: Permanent consumption for workers who transition to employment in different periods

employed right after m_1 better than the worker employed at m_1. This is because the worker who transitions to employment after m_1 can commit fraud until the next monitoring period, while the worker who transitions to employment at m_1 cannot commit fraud. Hence, the principal must offer the former a higher permanent consumption to induce truth-telling.

The difference between wage w and consumption c^E can be interpreted as an employment tax. Our contract implies that within a verification cycle, the employment tax for late reporters is higher than that for the early reporters. However, unlike the existing unemployment insurance literature, the employment tax is nonmonotonic: it decreases immediately following verification.

5.4 Unemployment Benefits

Unlike the case where $c^E(t)$ maps into $E(t)$ at every instant, $c^U(t)$ is not pinned down at every instant by $U(t)$, since the unemployed agent is not fully insured. Instead, the path of $c^U(\cdot)$ in $[0, m_1]$ requires knowledge of the entire path of $U(\cdot)$ in the interval. We obtain the entire trajectories of $c^U(\cdot)$ and $U(\cdot)$ after solving (9) in Section 5.5. However, monotonicity

of $U(\cdot)$ in Section 5.2 suggests that $c^U(\cdot)$ declines with unemployment duration. As in Hopenhayn and Nicolini (1997), our contract implies that the unemployment benefit c^U eventually reaches an arbitrarily low level with positive probability.[8]

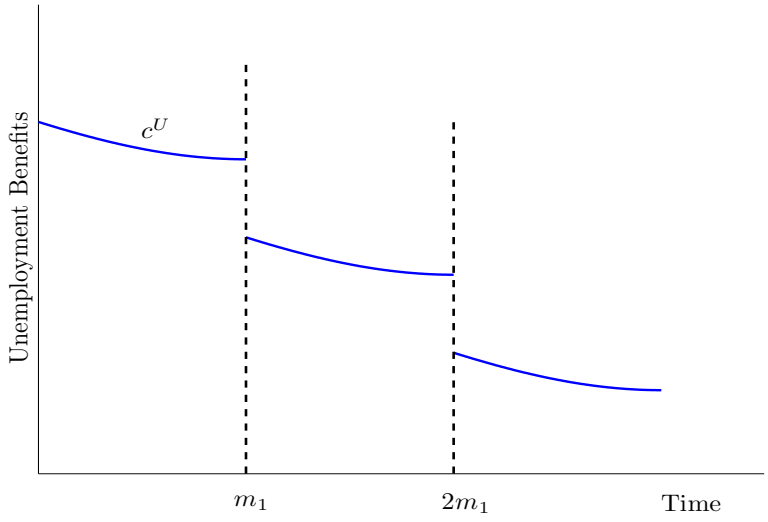

Figure 5: Consumption for the Unemployed

Figure 5 shows that the unemployment benefits jump down at the verification period. To understand the jump, we argue that it is optimal for the principal to set $u(t)$ above $u(m_1)$ when $m_1 - t > 0$ is small. Doing this relaxes the incentive constraint at time t, as the following variational argument shows. The promise-keeping constraint at $m_1 - \delta$, for a small positive δ, is

$$\begin{aligned} U(m_1 - \delta) &= r\delta u(m_1 - \delta) + e^{-r\delta}[(\pi\delta)E(m_1) + (1 - \pi\delta)U(m_1)] \\ &= r\delta u(m_1 - \delta) + e^{-r\delta}U(m_1), \end{aligned}$$

where the second equality uses the aforementioned property $E(m_1) = U(m_1)$. The incentive

[8] In contrast to Hopenhayn and Nicolini (1997) and our paper, Pavoni (2007) imposes an exogenous lower bound on promised utility and shows that the optimal benefits decrease with the duration of unemployment, but remain constant after the promised utility reaches the lower bound. Alvarez-Parra and Sanchez (2009) show a similar result in a model with an endogenous lower bound on promised utility.

constraint at $m_1 - \delta$ is

$$E(m_1 - \delta) \geq r\delta e^{-\rho w} u(m_1 - \delta) + e^{-r\delta} E(m_1).$$

Suppose $u(m_1 - \delta) = u(m_1)$. Then the principal can maintain the promise-keeping constraint but relax the incentive constraint by increasing $u(m_1 - \delta)$ and decreasing $u(m_1)$. Specifically, consider the variation

$$\tilde{u}(m_1 - \delta) = u(m_1 - \delta) + e^{-r\delta}\epsilon, \quad \tilde{u}(m_1) = u(m_1) - \epsilon, \quad \tilde{E}(m_1) = E(m_1) - r\delta\epsilon.$$

Because the unemployed worker's consumption after m_1 remains unchanged, his continuation utility at m_1 is $\tilde{U}(m_1) = U(m_1) - r\delta\epsilon$, which is equal to $\tilde{E}(m_1)$. Therefore, the promise-keeping constraint $U(m_1 - \delta) = r\delta\tilde{u}(m_1 - \delta) + e^{-r\delta}\tilde{U}(m_1)$ still holds, and the incentive constraint is relaxed:

$$\begin{aligned} r\delta e^{-\rho w}\tilde{u}(m_1 - \delta) + e^{-r\delta}\tilde{E}(m_1) &= r\delta e^{-\rho w} u(m_1 - \delta) + e^{-r\delta} E(m_1) - (1 - e^{-\rho w}) r\delta\epsilon \\ &< r\delta e^{-\rho w} u(m_1 - \delta) + e^{-r\delta} E(m_1). \end{aligned}$$

Starting from $u(m_1 - \delta) = u(m_1)$, the additional cost of consumption incurred by this variation is second order, but the effect on incentive constraint is first order. Hence the principal always chooses $u(t)$ above $u(m_1)$ when t is close to (but below) m_1.

We summarize these findings in the following proposition. The proof is in Appendix C.

PROPOSITION 2 *The unemployment benefit, $c^U(\cdot)$ is monotonically decreasing with unemployment duration, with downward jumps at verification, while $c^E(\cdot)$ is nonmonotonic: it decreases between verifications with upward jumps immediately after verification.*

Unemployment insurance systems in many countries feature benefits schemes similar to the one in Proposition 2. For example in Spain, workers receive a replacement rate of 70 percent for the first 6 months of unemployment, 60 percent for the next 18 months, and a minimum payment thereafter.

5.5 Pontryagin Minimum Principle

We construct a solution to the optimal control problem (9) in which the incentive constraint (11) binds (i.e., $\mu(t) = 0$) for all $t < m_1$. The problem faced by the principal

is to choose an initial state $E(0)$ and a time path $u(\cdot)$ to minimize the cost in (9), given $U(0)$. The promise-keeping and incentive constraints (10) and (11) then imply a time path $(U(\cdot), E(\cdot))$ for continuation utilities. One way to think about this problem is to think of choosing $u(t)$ at each date, given the values of $U(t)$ and $E(t)$ that have been attained by that date. The principal faces a tradeoff between the current-period cost and the cost of delivering continuation utilities. Hence, she needs to set "prices", Φ and λ, on increments to the continuation utilities U and E. Because it is costly for the principal to maintain a low E as a threat, it must be the case that $\lambda \leq 0$. Moreover, we have argued in Section 5.2 that $E(t) \geq U(t)$ is slack except at m_1, so we impose only the constraint $E(m_1) = U(m_1)$.

A central construct in the optimal control problem is the *current value Hamiltonian* $\mathscr{H}$ defined by

$$\mathscr{H} = \pi c(E(t)) + rc(u(t)) + \Phi(t)((r+\pi)U(t) - \pi E(t) - ru(t)) + \lambda(t)(rE(t) - e^{-\rho w} ru(t)),$$

which is just the sum of current-period cost and the rate of increase in continuation utilities 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 is

$$c'(u) = \Phi + e^{-\rho w}\lambda. \tag{13}$$

The left-hand side is the marginal cost of today's utility, while the right-hand side is the marginal cost of starting with higher continuation utility U tomorrow, offset by the benefit of a slacker incentive constraint (it is a benefit because $\lambda \leq 0$). The utility u must be chosen to equalize the costs at each date.

The prices Φ and λ must satisfy

$$\Phi'(t) = (r+\pi)\Phi - \frac{\partial \mathscr{H}}{\partial U} = 0, \tag{14}$$

$$\lambda'(t) = (r+\pi)\lambda - \frac{\partial \mathscr{H}}{\partial E} = \pi(\Phi - c'(E) + \lambda), \tag{15}$$

at each date t if $(u(\cdot), U(\cdot), E(\cdot))$ is an optimal path. Equation (14) implies that $\Phi(t)$ is a constant. Moreover, since multiplier $\Phi(0)$ is the marginal cost of $U(0)$, we have

$$\Phi = C'(U(0)) = -(\rho U(0))^{-1} > 0.$$

Since the planner can choose $E(0)$ freely,

$$\lambda(0) = 0. \tag{16}$$

At m_1, the shadow prices Φ and $\lambda(m_1)$ must satisfy

$$\Phi = -\kappa + c'(U(m_1)), \tag{17}$$

$$\lambda(m_1) = \kappa, \tag{18}$$

where $e^{-(r+\pi)m_1}\kappa$ is the multiplier on the constraint $E(m_1) = U(m_1)$. Since the principal's problem is convex, these conditions (13–18) are both necessary and sufficient for a minimum.

When (11) holds as equality, the states (U, E) and the costate λ satisfy differential equations:

$$U'(t) = (r+\pi)U - \pi E - ru, \tag{19}$$

$$E'(t) = rE - re^{-\rho w}u, \tag{20}$$

$$\lambda'(t) = \pi(\Phi - c'(E) + \lambda). \tag{21}$$

The ODE system contains three variables and would be difficult to analyze in a general context. However, we can solve (20) and (21) regardless of (19), because neither (20) nor (21) relies on U. Once (20) and (21) are solved, it is easy to solve (19). Formally,

LEMMA 2 *If* (20) *and* (21) *hold, then* (19) *holds if and only if*

$$\Phi U(t) + \lambda(t)E(t) + \rho^{-1} = 0, \quad \forall t \in [0, m_1]. \tag{22}$$

To solve the reduced ODE system, (20) and (21), we need two boundary conditions. The first is (16), $\lambda(0) = 0$. The second cannot be a value for $E(0)$, as $E(0)$ is endogenous and unknown a priori. We obtain the second boundary condition, $E(m_1) = -\rho^{-1}(\Phi+\lambda(m_1))^{-1}$, from $E(m_1) = U(m_1)$ and equation (22).

The following lemma shows that these two boundary conditions pin down a unique solution curve for the system (20) and (21). Figure 6 shows the phase diagram. That $\lambda < 0$ implies that the incentive constraint binds for all $t < m_1$.

LEMMA 3 *For any $m_1 > 0$, there is a unique initial condition $E(0)$ such that the solution starting at $(\lambda(0) = 0, E(0))$ satisfies $E(m_1) = -\rho^{-1}(\Phi + \lambda(m_1))^{-1}$.*

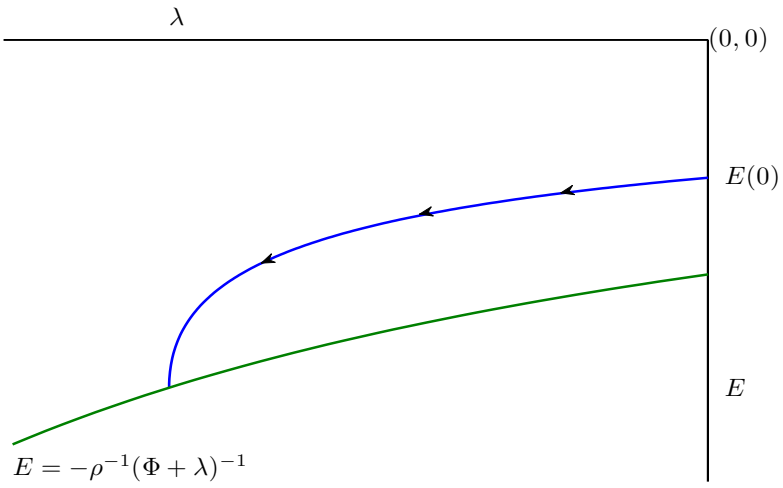

Figure 6: Phase Diagram for (λ, E).

6 Optimal Monitoring

Until this point, we have taken m_1 as exogenous. In this section, we characterize the optimal choice of m_1. The tradeoff in choosing m_1 is as follows. Monitoring more frequently implies higher verification cost, but the principal can provide better insurance: the consumption path for the unemployed is similar to that for the employed. Monitoring less frequently implies lower verification cost but worse insurance.

For any $m_1 > 0$, denote the minimized cost in (9) as $\mathscr{C}(m_1)$; that is,

$$\mathscr{C}(m_1) = \int_0^{m_1} e^{-(r+\pi)t} \left(\pi c(E(t)) + r c(u(t))\right) dt + e^{-(r+\pi)m_1} \left(\gamma + \psi + c(U(m_1))\right).$$

Intuitively, delaying monitoring (i.e., a small increase in m_1) saves the principal both the cost of monitoring and the cost of (after-monitoring) consumptions, because the payment of $\gamma + \psi + c(U(m_1))$ is postponed. By doing so, however, the principal must maintain the consumptions $c(E(\cdot))$ and $c(u(\cdot))$ for a longer duration. Subtracting the benefit from the cost (algebraic details in Appendix C) yields

$$\mathscr{C}'(m_1) = e^{-(r+\pi)m_1} \left(r\rho^{-1} \log\left(\frac{\Phi + e^{-\rho w}\lambda(m_1)}{\Phi + \lambda(m_1)} \right) - (r+\pi)(\gamma + \psi) \right).$$

Thus, the first-order condition for m_1 is

$$r\rho^{-1}\log\left(\frac{\Phi+e^{-\rho w}\lambda(m_1)}{\Phi+\lambda(m_1)}\right)=(r+\pi)(\gamma+\psi). \tag{23}$$

PROPOSITION 3 *The optimal m_1 is the unique solution to (23). That is, (23) is both necessary and sufficient for the minimum of $\mathscr{C}(m_1)$.*

REMARK 1 *Although our analysis relies on an undetermined parameter ψ, the parameter can be uniquely pinned down by a fixed-point condition that the actual cost function at time zero must equal the conjectured function $\psi + c(U(0))$. Further details are in Appendix C.*

REMARK 2 *Our analytical results rely on the assumption of CARA preferences. Unlike the CARA case where the length of the monitoring cycle is independent of history, the cycle length in the CRRA case depends on the worker's continuation utility. However, most of the main features of the optimal contract remain valid even if the worker has CRRA preferences. We demonstrate this through a numerical example in Fuller, Ravikumar, and Zhang (2013).*

7 Quits

Another type of fraud that could arise in our model is quits. An agent in our model could transition to employment in period t, claim to be unemployed until almost m_1, and then quit to become unemployed at m_1. The verification at m_1 would not reveal him to be a cheater. Thus, quitting is possible in our model.

Our mechanism guarantees that the agent does not commit such a fraud. The continuation utilities $E(\cdot)$ and $U(\cdot)$ are such that the agent is indifferent between reporting the transition immediately and delaying it to the next period. By following the path above and quitting at m_1, he becomes truly unemployed, is subject to the stochastic arrival rate of employment opportunity, and is worse off.

Hopenhayn and Nicolini (2009) examine a model where quits cannot be distinguished from layoffs and the only fraudulent behavior is quits. In their model, the employment status is observable and non-absorbing, and disutility from working is greater than that from searching for employment. Employed agents might want to opportunistically quit

their job, enjoy more leisure, and collect unemployment benefits. To discourage quits, the principal offers (i) higher consumption to the employed workers who stay on the job longer and (ii) more generous benefits to unemployed workers with longer employment spells, as quitters have shorter employment spells on average. In our model, the utility functions for the unemployed worker and the employed worker are the same, and employment status is private information. Since employment is an absorbing state, quitting as considered in Hopenhayn and Nicolini (2009) cannot arise in our model. The potential reason for quitting in our model is to cover up the fraudulent collection of unemployment benefits before the verification period. Our optimal mechanism provides incentives for the agent not to delay reporting his transition to employment and not to conceal his earnings.

Overpayment due to quits is small relative to the overpayment due to concealed earnings (see Table 1). Our mechanism deters fraud due to both concealed earnings and quits.

8 Stochastic Verification

Our monitoring mechanism in the previous sections was restricted to deterministic verification. Here we consider a more general mechanism where the principal verifies randomly after receiving the unemployment report. Conditional on the unemployment report at t, the principal chooses the monitoring Poisson rate $p(t) \geq 0$. That is, over a period of length dt, the principal monitors with probability $p(t)dt$ and she does not monitor with probability $1 - p(t)dt$. (Since our model is in continuous time, $p(t)$ is not the monitoring probability.)

We assume that if a worker is monitored and caught cheating, he has to pay a finite penalty forever. With infinite penalty, an arbitrarily small monitoring probability would deliver the full-information constant consumption. In our model, if the principal can choose any finite penalty between 0 and $\phi > 0$, he would always choose ϕ. Henceforth, we assume that the finite penalty is ϕ units of the consumption good, forever.

Similar to (10) and (11), the promise-keeping constraint and incentive constraint are

$$U' = r(U-u) - \pi(E-U) - p(\tilde{U}-U), \tag{24}$$

$$E' \leq rE - re^{-\rho w}u - p(e^{\rho\phi}-1)E, \tag{25}$$

where $\tilde{U}$ is the unemployed agent's continuation utility after monitoring. Because the probability that monitoring does not occur in $[0,t)$ is $e^{-\int_0^t p(s)ds}$, the principal's objective is

$$\int_0^\infty e^{-(r+\pi)t-\int_0^t p(s)ds}\left(\pi c(E(t))+rc(u(t))+p(t)(\gamma+C(\tilde{U}(t)))\right)dt. \tag{26}$$

The principal chooses the utilities $\{U(t),E(t),u(t),\tilde{U}(t);t\geq 0\}$ and the arrival rates of monitoring $\{p(t);t\geq 0\}$ to minimize (26) subject to (24), (25), and the constraint $E(t)\geq U(t)$, $\forall t\geq 0$.

Since the penalty for a worker with high promised utility is the same as that for a worker with low promised utility, we obtain a scaling property similar to the one in Section 4.1. Thus, the incentives to conceal earnings are the same for workers with different promised utilities. Similar to our model with deterministic verification, we show in Proposition 4 that the optimal stochastic verification mechanism consists of cycles. See Appendix D for the proof.

PROPOSITION 4 *There exists an $N>0$ such that the principal monitors the unemployed with a constant arrival rate $p>0$ if and only if $t\geq N$. Before N, the time path $(U(\cdot),E(\cdot))$ converges to the 45-degree line; after N, it moves along the 45-degree line toward $(-\infty,-\infty)$ until the agent is randomly drawn to be verified. After the verification, (U,E) jumps to a new state $(\tilde{U},\tilde{E})$ and a new cycle starts.*

The unemployed worker is in one of two states: (i) not monitored (i.e., $p(t)=0$) or (ii) randomly drawn to be monitored (i.e., $p(t)\equiv p>0$). Within each cycle, an unemployed worker is initially in the not-monitored state. He is moved to the random monitoring state if the duration of his unemployment report exceeds the threshold N. If he is randomly drawn to be monitored, then he is moved to the not-monitored state after being monitored, and a new cycle begins. While the date of monitoring is stochastic, the threshold duration is not. That is, within each cycle, the principal guarantees that the worker will not be monitored until the threshold duration is reached, similar to the deterministic verification case.

The intuition for why the worker is not monitored before the threshold duration is as follows. The Unemployment Insurance agency has access to two instruments: tax/subsidy and monitoring. Recall that at verification the true employment status is revealed, and E

is reset to a level such that its shadow price is zero, which means that, immediately after monitoring, the employment tax can be varied at no cost. The cost of the tax/subsidy instrument is lower than the cost of monitoring, $\gamma > 0$, immediately after monitoring, and remains so until some threshold unemployment duration is reached. Hence, it is optimal to use only the tax/subsidy instrument for the provision of incentives before the threshold.

REMARK 3 *The absence of verification until a threshold duration is unlikely to be robust to other types of penalties. For instance, in Popov (2009) there is an exogenous lower bound on the worker's continuation utility and a worker who is caught cheating is pushed to this lower bound. So the penalty for a worker with high continuation utility is larger than that for a worker with low continuation utility. With hidden i.i.d. income, he shows that the verification probability is always positive.*

The stochastic monitoring mechanism clearly dominates the deterministic mechanism characterized in Section 6. To see this, consider a stochastic monitoring scheme in which the arrival rate of monitoring is higher than p for workers in the random monitoring state. Denote this higher arrival rate as $\tilde{p}$. Proposition 4 implies that $\tilde{p}$ is suboptimal. By continuity, the limiting scheme as $\tilde{p} \to \infty$ should also be suboptimal. This limiting scheme is exactly the deterministic monitoring mechanism.

We argue below that the key insights on the use of tax/subsidy and monitoring instruments in the suboptimal deterministic mechanism are nearly identical to the insights from the optimal stochastic mechanism. We describe in detail the similarities and differences between the implications of the two mechanisms.

8.1 Comparison of Monitoring with the Deterministic Case

First, both the stochastic and deterministic mechanisms have the feature that monitoring does not occur before a threshold unemployment duration; m_1 in the deterministic case and N in the stochastic case. These thresholds, however, could be different; i.e., in general $m_1 \neq N$.

Second, both mechanisms feature cycles. In the deterministic case, after m_1 a new cycle begins, with exactly the same length as the previous cycle. Similarly, in the stochastic case,

after monitoring occurs a new cycle begins and verification does not occur again before the threshold N is reached. The exact date when the monitoring occurs in the stochastic case is random. This is because, after N monitoring arrives according to a Poisson process and, hence, the exact length of each cycle depends on when the worker is actually verified. As in the deterministic case, however, the value of N is the same in each cycle.

8.2 Comparison of Tax/Subsidy with the Deterministic Case

Consumptions in the stochastic monitoring case are similar to those in the deterministic case. Within each cycle, before the threshold N, the patterns of consumption are identical to (c^E, c^U) in Figures 4 and 5. After N, if a worker is monitored and verified to be truly unemployed, then the unemployment benefits jump down, as in the deterministic case.

The only difference is that in the deterministic case, continuation utilities and consumptions are reset when the threshold m_1 is reached. In the stochastic case, after the threshold N and before the monitoring actually arrives, continuation utilities and consumptions smoothly decline with the duration of unemployment. The decreasing continuation utilities and the monitoring (and finite punishment) jointly provide incentives for truth telling; the worker is indifferent between reporting a job offer and committing fraud.

8.3 Quantitative Analysis

To illustrate our optimal contract, we follow Hopenhayn and Nicolini (1997) closely and perform a quantitative exercise similar to theirs. We let the agents in our model face a stylized version of the U.S. unemployment insurance system. We calibrate the model to match the observed rate of concealed earnings fraud. We then compute the gain to switching to the optimal mechanism in our model.

To perform this exercise, we have to add some heterogeneity to our model; otherwise everyone would cheat or no one would cheat, and we would not be able to match the observed rate of concealed earnings fraud. We assume that the workers are heterogeneous in the wages they earn and, hence, the replacement rate for unemployment benefits. Concretely, we assume that the wage distribution is lognormal with parameters μ_w and σ_w^2.

The BAM data provides earnings information for an individual's previous employment (the earnings that determine the unemployment benefits for the individual). In the 2007 sample of BAM data, the mean weekly wage is \$692 and the coefficient of variation is 0.79. Using these data moments, we calibrate $\mu_w = 6.296$ and $\sigma_w^2 = 0.488$. By construction, the earnings in the BAM data are only for those who collect unemployment benefits. Instead of using the BAM data we could use the CPS data on earnings for the entire employed population to calibrate the wage distribution in the model. However, individuals collecting unemployment benefits generally earn less (while employed) than the individuals in the entire employed population.[9]

We calculate the unemployment benefits as a function of wages, again using the BAM 2007 data: $\ln(\text{unemployment benefits}) = 1.31 + 0.65 \ln(\text{wages})$.

We assume that the model period is 1 week and that the interest rate $r = 0.001$. Since the average duration of unemployment in 2007 is 16.85 weeks, we calibrate the job arrival rate to be $\pi = 1/16.85$. The monitoring cost γ is calibrated as follows. On average, the BAM investigators spend 12.6 hours per case and the average wage of the investigators is \$43 in 2012 (the only year when such data is available). So, adjusting the average wage to 2007 dollars, we calibrate γ to be \$501. We calibrate the value of absolute risk aversion ρ such that the *relative risk aversion* for the average wage earner is 2. Since the average wage is \$692 in our sample, $\rho = 2/692$.

We then calibrate the probability of monitoring and the penalty in the U.S. system if caught cheating to match two targets: fraction of people committing concealed earnings fraud and fraction of people caught cheating among those committing the fraud.

With CARA preferences, wage heterogeneity is not relevant for matching the two targets, but it is relevant for computing the distribution of initial promised utility in the baseline. In the counterfactual, we take these initial promised utilities as given, calculate the optimal monitoring and benefits, and then compute the cost of delivering the initial promised utilities. The job arrival rate, wage distribution, and penalty are held fixed at the same values as the baseline calibration.

[9]The mean weekly wage among employed workers, in the March 2007 CPS, is \$861 and the coefficient of variation is 1.27.

The results imply that, measured in present value, the cost of optimal monitoring is 60 percent of the cost in the current U.S. system. In the optimal contract (averaging across the initial promised utilities), $N = 11.64$ weeks. That is, the planner guarantees that monitoring does not occur for roughly the first 12 weeks of the unemployment spell and, thus, reduces the monitoring cost with an efficient use of the monitoring technology.

To determine the magnitude of the gain from switching to the optimal mechanism, suppose that the planner is restricted to use the same amount of resources as the current U.S. system. How much additional utility can the planner deliver to the average worker? The answer is a utility gain equivalent to a 1.55% more consumption at every date, relative to the U.S. system. This gain arises from two sources: (i) improved consumption smoothing between employed and unemployed states and (ii) reduced monitoring costs or higher consumption on average. The U.S. system spends only 0.24 percent its resources on monitoring the average worker and spends the rest on unemployment benefits (net of wages), but the same resources are allocated differently in the optimal contract: 0.17 percent is spent on monitoring the average worker and the rest is spent on unemployment benefits. Thus, almost all of the gain in our model comes from improved consumption smoothing.

There are some obvious limitations to this analysis. Most notably, our exercise is a partial equilibrium analysis, as in Hopenhayn and Nicolini (1997). To fully quantify the welfare gains from adopting the optimal contract, we have to conduct a general equilibrium analysis incorporating transition from employment to unemployment and disciplining the model with aggregate worker flows.

9 Conclusion

The most prevalent incentive problem in the U.S. unemployment insurance system is that individuals collect unemployment benefits while being gainfully employed. We examine a model of optimal unemployment insurance where a worker can conceal his employment status and the Unemployment Insurance authority has a technology to verify his employment status. We find that the optimal interval between consecutive monitoring periods is a constant, independent of history. The optimal employment tax is nonmonotonic, in-

creasing between verifications and decreasing immediately after a verification. The optimal unemployment benefits decline with unemployment duration with sharp declines after each verification. Our optimal contract also prevents fraud from quits.

Unemployment insurance in our model is a form of social insurance protecting workers against the risk of job loss. Acemoglu and Shimer (1999, 2000), Shimer and Werning (2008), and Alvarez-Parra and Sanchez (2009) explore another role of unemployment insurance. They examine environments with heterogeneous jobs, and unemployment insurance helps the worker wait for the appropriate job. Some jobs have higher productivity than others, but such job opportunities arrive less frequently. Unemployment benefits help workers wait for more productive matches and endure longer unemployment durations. The benefits in these environments affect the aggregate composition of jobs. An interesting direction for future research is to extend our environment to multiple jobs and examine optimal monitoring in the presence of the alternative role of unemployment insurance.

Finally, our model does not include any job retention effort. Incorporating the job retention effort into our model requires employment to be stochastic. If workers can conceal earnings, their hidden income could affect their job retention effort. Analyzing interaction between effort and fraud is another interesting direction for future research.

References

Acemoglu, D., and R. Shimer (1999): "Efficient Unemployment Insurance," *Journal of Political Economy*, 107(5), 893–928.

——— (2000): "Productivity Gains from Unemployment Insurance," *European Economic Review*, 44(7), 1195–1224.

Aliprantis, C., and O. Burkinshaw (1990): *Principles of Real Analysis, Second Edition*. Academic Press, Inc., San Diego, CA, United States.

Alvarez-Parra, F., and J. M. Sanchez (2009): "Unemployment Insurance with a Hidden Labor Market," *Journal of Monetary Economics*, 56(7), 954–967.

Ashenfelter, O., D. Ashmore, and O. Deschenes (2005): "Do Unemployment Insurance Recipients Actively Seek Work? Evidence from Randomized Trials in Four U.S. States," *Journal of Econometrics*, 125(1-2), 53–75.

Atkeson, A., and R. E. Lucas (1995): "Efficiency and Equality in a Simple Model of Efficient Unemployment Insurance," *Journal of Economic Theory*, 66(1), 64–88.

Baily, M. (1978): "Some Aspects of Optimal Unemployment Insurance," *Journal of Public Economics*, 10(3), 379–402.

Fuller, D. L., B. Ravikumar, and Y. Zhang (2013): "Unemployment Insurance Fraud and Optimal Monitoring," *Working Paper 2012-024C, Federal Reserve Bank of St. Louis*.

Gauthier-Loiselle, M. (2011): "Find a Job Now, Start Working Later Does Unemployment Insurance Subsidize Leisure?," *Working Paper, Princeton University*.

Golosov, M., and A. Tsyvinski (2006): "Designing Optimal Disability Insurance: A Case for Asset Testing," *Journal of Political Economy*, 114(2), 257–279.

Hansen, G., and A. Imrohoroglu (1992): "The Role of Unemployment Insurance in an Economy with Liquidity Constraints and Moral Hazard," *Journal of Political Economy*, 100(1), 118–142.

Hopenhayn, H., and J. P. Nicolini (1997): "Optimal Unemployment Insurance," *Journal of Political Economy*, 105(2), 412–438.

——— (2009): "Optimal Unemployment Insurance and Employment History," *Review of Economic Studies*, 76(3), 1049–1070.

Pavoni, N. (2007): "On Optimal Unemployment Compensation," *Journal of Monetary Economics*, 54(6), 1612–1630.

Popov, L. (2009): "Stochastic Costly State Verification and Dynamic Contracts," *Working Paper, University of Virginia*.

Ravikumar, B., and Y. Zhang (2012): "Optimal Auditing and Insurance in a Dynamic Model of Tax Compliance," *Theoretical Economics*, 7(2), 241–282.

Setty, O. (2011): "Optimal Unemployment Insurance with Monitoring," *Working Paper, MPRA*.

Shavell, S., and L. Weiss (1979): "The Optimal Payment of Unemployment Insurance Benefits over Time," *Journal of Political Economy*, 87(6), 1347–1362.

Shimer, R., and I. Werning (2008): "Liquidity and Insurance for the Unemployed," *American Economic Review*, 98(5), 1922–42.

Wang, C., and S. Williamson (2002): "Moral Hazard, Optimal Unemployment Insurance, and Experience Rating," *Journal of Monetary Economics*, 49(7), 1337–1371.

Zhang, Y. (2009): "Dynamic Contracting with Persistent Shocks," *Journal of Economic Theory*, 144(2), 635–675.

Appendix A Data

Fraud and Overpayments Table A.1 details the various types of fraud overpayments from 2005 – 2009, averaged over all U.S. states. Concealed earnings fraud is the dominant source of overpayments in every year.

Table A.1: Fraud Overpayments

	Percent of Total Fraud Overpayments				
Cause	2005	2006	2007	2008	2009
Concealed Earnings	62.64	54.40	60.06	67.32	65.89
Insufficient Job Search	4.55	4.15	4.95	3.02	2.75
Refused Suitable Offer	0.63	0.36	0.80	0.36	0.77
Quits	12.78	16.41	7.06	5.04	5.14
Fired	4.27	4.60	13.29	12.69	9.61
Unavailable for Work	4.94	6.95	4.17	4.60	7.38
Other	10.20	13.14	9.67	6.97	8.46
Total	100.00	100.00	100.00	100.00	100.00

Source: Benefit Accuracy Measurement Program, U.S. Department of Labor

The unemployment insurance system might incur another form of overpayment if workers strategically delay the start date of employment. That is, workers might accept a job offer but agree to start the job after their unemployment benefits have expired. Gauthier-Loiselle (2011) documents that unemployment insurance expenditures are higher in Canada because of such cases. In the U.S., this is not considered fraud. Thus, the BAM data include no information on such cases, so they are not included in the fraud overpayments statistics.

Overpayments due to Insufficient Search In Table 1 in Section 2, the overpayments due to concealed earnings fraud were almost twelve times the overpayments due to insufficient search fraud. Do the data understate the incidence of insufficient search? Recall that the BAM program measures only the extensive margin — whether the individual submits the required number of applications. It is possible that the unmeasured intensive margin — effort that turns an application into a job offer — is large enough to make the overpayments due to insufficient search comparable in magnitude to the overpayments due to concealed earnings. The following facts, however, suggest that the unmeasured component is unlikely to be large.

1. Measured overpayments due to insufficient search have been declining: In 1988 they accounted for 34 percent of the total overpayments due to all fraud, whereas in 2007 they accounted for less than 5 percent. (The corresponding numbers for concealed earnings fraud were 41 percent and more than 60 percent.)

2. The job search requirements that make an unemployed person eligible for benefits have *increased* over time, so the decline in the measured component is not due to changes in eligibility criteria. Hence, for the insufficient search overpayments to be the same in 2007

as those measured in 1988, the unmeasured component has to be almost six times that of the measured component in 2007.

3. If unmeasured efforts to translate a job application into a job offer were substantially higher in 2007, then the increase in efforts should imply a substantially higher transition rate from unemployment to employment. However, the transition rate is roughly constant: The quarterly rate was 0.31 for the period 1988-1997 and 0.33 for 1998-2007.

From a normative point of view, as noted in Section 1, the prevailing quantitative theory prescribes an intensive margin search effort that is less than the effort exerted under the current unemployment insurance program in the U.S. In other words, insufficient search is not a critical incentive problem in the U.S. (Using evidence from randomized trials in four U.S. sites, Ashenfelter, Ashmore, and Deschenes (2005) find that insufficient job search is not a significant source of unemployment insurance overpayments.)

Appendix B Microfoundations for $E(t) \geq U(t)$

Suppose that the worker can privately refuse a job offer. The timing in each period is as follows. The stochastic job opportunity arrives and the worker either receives an offer or does not. He then chooses to report the offer (if any) to the principal. Conditional on the report of an offer, the principal recommends the worker to either accept or reject the offer. The worker then chooses whether to follow the principal's recommendation. (In contrast, job acceptance is implicitly imposed in our model in Section 3.) Conditional on the report, the principal assigns current and future consumptions.

In such a job-refusal model, it is optimal for the principal to always recommend to the worker who reports an offer to accept the offer. Recommending "accept" minimizes the cost of delivering the promised utility since the worker's consumption is constant upon job acceptance and the principal gets the perpetual wage. Recommending "reject" means that the continuation contract involves additional uncertainty of job offers, reports, and incentive constraints. So the consumption cost of delivering the same promised utility is higher under "reject." Recall that, unlike Atkeson and Lucas (1995), we do not have disutility to working so it is optimal to always recommend "accept."

The incentive compatibility for an agent with a job offer is as follows. If he reports his offer and receives a recommendation to accept, he strictly prefers "accept" to "reject." This is because rejecting the offer would not make him eligible for any unemployment insurance benefits, but would make him lose his wage income. If the agent does not report his offer, then either he rejects the offer and obtains $U(t)$, or he accepts the offer and commits fraud (i.e., he works and collects unemployment benefits at the same time). For the agent to truthfully report his offer, the utility of reporting and accepting the offer, $E(t)$, must be higher than both $U(t)$ and the utility he obtains by committing Concealed Earnings fraud. These incentive compatibility constraints are exactly conditions (2) and (3) in our model in Section 3.

Appendix C Proofs

PROOF OF LEMMA 1: Suppose that a contract $\sigma \equiv \{(U(t), E(t), u(t), c^U(t), c^E(t), m_i); t \geq 0, i \geq 1\}$ delivers the continuation utility U. Then, a contract

$$\sigma_\alpha \equiv \{(\alpha U(t), \alpha E(t), \alpha u(t), c^U(t) - \log(\alpha)/\rho, c^E(t) - \log(\alpha)/\rho, m_i); t \geq 0, i \geq 1\}$$

delivers αU. The reverse is also true. Further, σ is incentive compatible if and only if σ_α is incentive compatible. Therefore, $\{(U^*(t), E^*(t), u^*(t), c^{U*}(t), c^{E*}(t), m_i^*); t \geq 0, i \geq 1\}$ is the optimal contract to deliver U if and only if

$$\{(\alpha U^*(t), \alpha E^*(t), \alpha u^*(t), c^{U*}(t) - \log(\alpha)/\rho, c^{E*}(t) - \log(\alpha)/\rho, m_i^*); t \geq 0, i \geq 1\}$$

is the optimal contract to deliver αU. □

LEMMA C.1 *The promise-keeping constraint* (1) *and the incentive constraint* (6) *hold for all* $0 \leq t < s \leq m_1$ *if and only if*

$$U(s) - U(t) = \int_t^s ((r+\pi)U(x) - \pi E(x) - ru(x))\, dx, \tag{27}$$

$$E(s) - E(t) \leq \int_t^s \left(rE(x) - re^{-\rho w}u(x)\right) dx, \tag{28}$$

hold for all $0 \leq t < s \leq m_1$. *Taking the limit as* s *goes to* t *yields the differential equations* (10) *and* (11).

PROOF. We only show the equivalence between (6) and (28), since the equivalence between (1) and (27) can be obtained similarly by replacing the inequalities below with equalities.

Necessity: If (6) holds for all $t < s$, then

$$\begin{aligned}
& E(t) + \int_t^s \left(rE(x) - re^{-\rho w}u(x)\right) dx \\
\geq\ & \int_t^s e^{-r(x-t)} re^{-\rho w} u(x) dx + e^{-r(s-t)} E(s) \\
& + \int_t^s \left(r \left(\int_x^s e^{-r(\eta - x)} re^{-\rho w} u(\eta) d\eta + e^{-r(s-x)} E(s) \right) - re^{-\rho w} u(x) \right) dx \\
=\ & \left(e^{-r(s-t)} + \int_t^s re^{-r(s-x)} dx \right) E(s) + \int_t^s \left(e^{-r(x-t)} - 1 \right) re^{-\rho w} u(x) dx \\
& + \int_t^s r \left(\int_x^s e^{-r(\eta-x)} re^{-\rho w} u(\eta) d\eta \right) dx \\
=\ & E(s) + \int_t^s \left(e^{-r(x-t)} - 1 \right) re^{-\rho w} u(x) dx + \int_t^s \left(\int_t^\eta re^{-r(\eta - x)} dx \right) re^{-\rho w} u(\eta) d\eta \\
=\ & E(s) + \int_t^s \left(e^{-r(x-t)} - 1 \right) re^{-\rho w} u(x) dx + \int_t^s \left(1 - e^{-r(\eta - t)} \right) re^{-\rho w} u(\eta) d\eta \\
=\ & E(s).
\end{aligned}$$

Hence, inequality (28) is verified.

Sufficiency: Define an absolutely continuous function $f(\cdot)$ as

$$f(s) \equiv \int_t^s e^{-r(x-t)} r e^{-\rho w} u(x) dx + e^{-r(s-t)} \left(E(t) + \int_t^s \left(rE(x) - re^{-\rho w} u(x) \right) dx \right).$$

Because f is absolutely continuous, it is differentiable almost everywhere (a.e.), and

$$\begin{aligned} f'(s) &= e^{-r(s-t)} r e^{-\rho w} u(s) - r e^{-r(s-t)} \left(E(t) + \int_t^s \left(rE(x) - re^{-\rho w} u(x) \right) dx \right) \\ &\quad + e^{-r(s-t)} \left(rE(s) - re^{-\rho w} u(s) \right) \\ &= r e^{-r(s-t)} \left(E(s) - E(t) - \int_t^s \left(rE(x) - re^{-\rho w} u(x) \right) dx \right), a.e. \end{aligned}$$

If (28) holds, then $f'(s) \leq 0$ a.e. Then, it follows from Theorem 29.15 in Aliprantis and Burkinshaw (1990) that

$$f(s) = f(t) + \int_t^s f'(x) dx \leq f(t) = E(t).$$

Therefore,

$$\int_t^s e^{-r(x-t)} r e^{-\rho w} u(x) dx + e^{-r(s-t)} E(s) \leq f(s) \leq E(t),$$

which verifies inequality (6). □

Proof of Lemma 2: If (19), (20) and (21) all hold, we can substitute them into $(\Phi U + \lambda E)'$ and obtain

$$\begin{aligned} (\Phi U + \lambda E)' &= \Phi U' + \lambda' E + \lambda E' \\ &= \Phi \left((r+\pi) U - \pi E - ru \right) + \pi \left(\Phi - c'(E) + \lambda \right) E + \lambda (rE - re^{-\rho w} u) \\ &= (r+\pi) (\Phi U + \lambda E) - \pi c'(E) E - r(\Phi + e^{-\rho w} \lambda) u. \end{aligned}$$

Because $-c'(E)E = \rho^{-1}$ and $-(\rho u)^{-1} = c'(u) = \Phi + e^{-\rho w}\lambda$, we have

$$(\Phi U + \lambda E)' = (r+\pi) \left(\Phi U + \lambda E + \rho^{-1} \right). \tag{29}$$

Because $\Phi U(0) + \lambda(0) E(0) + \rho^{-1} = 0$, it follows from (29) that $\Phi U(t) + \lambda(t) E(t) + \rho^{-1} = 0$ for all $t \in [0, m_1]$.

On the other hand, if (20) and (21) hold and

$$\Phi U(t) + \lambda(t) E(t) + \rho^{-1} = 0, \quad \forall t \in [0, m_1],$$

then $(\Phi U + \lambda E)' = 0$ for all $t \in [0, m_1]$. Then (19) can be derived by reversing the above steps. □

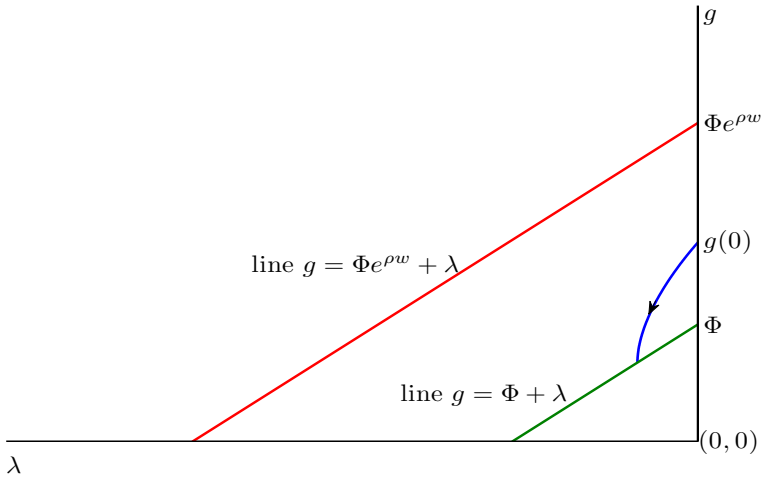

Figure 7: Phase Diagram for (λ, g).

Proof of Lemma 3: First, it is convenient to transform the state variable E, which may approach $-\infty$, into a bounded one. To do so, we replace E with

$$g \equiv c'(E) = -(\rho E)^{-1}.$$

Now, the ODE system consists of (21) and

$$g' = \frac{E'}{\rho E^2} = \frac{rg^2}{\Phi e^{\rho w} + \lambda} - rg, \tag{30}$$

with boundary condition $g(m_1) = \Phi + \lambda(m_1)$ (Figure 7 shows the phase diagram). Let $m(g(0))$ be the time to hit the straight line $g = \Phi + \lambda$ starting with $(\lambda(0) = 0, g(0))$.

Second, we show that $\lim_{g(0)\downarrow\Phi} m(g(0)) = 0$. If $\lambda = 0$ and $g = \Phi$, then

$$\begin{aligned}(g-\lambda)'(t) &= \left(\frac{rg^2}{\Phi e^{\rho w} + \lambda} - rg + \pi(g - \lambda - \Phi)\right)\Big|_{(\lambda,g)=(0,\Phi)} \\ &= \frac{r\Phi^2}{\Phi e^{\rho w}} - r\Phi < 0.\end{aligned}$$

Continuity of the ODE system (21), (30) implies that $(g-\lambda)'(t) < 0$ in a small neighborhood of $(0, \Phi)$. If $\lambda(0) = 0$ and $g(0)$ approaches Φ from above, then $g(0) - \lambda(0) - \Phi$ approaches zero. Since the solution curve starting with $(0, g(0))$ will remain in the small neighborhood of $(0, \Phi)$ for a while, it will decrease and hit the line $g = \Phi + \lambda$ quickly if $g(0) - \lambda(0) - \Phi$ is sufficiently small.

Third, we show that $m(g(0))$ is strictly increasing in $g(0)$. Consider two paths that start with initial conditions $(0, g_1(0))$ and $(0, g_2(0))$, where $\Phi < g_1(0) < g_2(0)$. We will show that $g_1(t) - \lambda_1(t) < g_2(t) - \lambda_2(t)$ for all t. By contradiction, suppose $(g_1 - \lambda_1)(t) = (g_2 - \lambda_2)(t)$

for the first time at $t = t^*$. Because the two paths cannot cross, we cannot have that $g_1(t^*) \leq g_2(t^*)$. Then $g_1(t^*) > g_2(t^*)$ and $\lambda_1(t^*) > \lambda_2(t^*)$. Hence

$$\begin{aligned}(g_1 - \lambda_1)'(t^*) &= -\frac{rg_1}{\Phi e^{\rho w} + \lambda_1}(\Phi e^{\rho w} + \lambda_1 - g_1) - \pi(\Phi + \lambda_1 - g_1) \\ &< -\frac{rg_2}{\Phi e^{\rho w} + \lambda_2}(\Phi e^{\rho w} + \lambda_2 - g_2) - \pi(\Phi + \lambda_2 - g_2) \\ &= (g_2 - \lambda_2)'(t^*),\end{aligned}$$

where the inequality follows from $\frac{g_1}{\Phi e^{\rho w}+\lambda_1} > \frac{g_2}{\Phi e^{\rho w}+\lambda_2}$. That $(g_1 - \lambda_1)'(t^*) < (g_2 - \lambda_2)'(t^*)$ contradicts the facts that $(g_1 - \lambda_1)(t^*) = (g_2 - \lambda_2)(t^*)$ and $(g_1 - \lambda_1)(t) < (g_2 - \lambda_2)(t)$ for all $t < t^*$. Thus $g_1(t) - \lambda_1(t) < g_2(t) - \lambda_2(t)$ for all t, and the path $(\lambda_1(t), g_1(t))$ reaches $g = \Phi + \lambda$ sooner.

Finally, we show there exists a unique $g(0)$ to satisfy $m(g(0)) = m_1$ for any $m_1 > 0$. The second step in this proof shows that $\lim_{g(0)\downarrow\Phi} m(g(0)) = 0$. Part (ii) in Lemma C.2 (page 43) shows that $m(g(0))$ can be arbitrarily large with high values of $g(0)$. Hence, the existence of a unique solution to $m(g(0)) = m_1$ follows from the intermediate value theorem and the monotonicity of $m(g(0))$ in $g(0)$. □

Proof of Proposition 2: First, we show that E, c^U, U, and $\frac{U}{E}$ all fall on $[0, m_1]$. It follows from $g'(t) < 0$ that $E'(t) = \rho E^2(t) g'(t) < 0$. Equation (13) implies that $u'(t) = \frac{e^{-\rho w}\lambda'(t)}{c''(u)} < 0$, or $(c^U)'(t) < 0$. Equation (22) implies that $U'(t) = -\Phi^{-1}(\lambda(t)E(t))' < 0$. Equation (22) also implies that $\frac{U}{E} = \Phi^{-1}(g - \lambda)$. Hence part (i) in Lemma C.2 implies that $\left(\frac{U}{E}\right)'(t) < 0$.

Second, to see the downward jump in $c^U(\cdot)$ at m_1, we show that

$$\lim_{t\uparrow m_1} c'(u(t)) > \lim_{t\downarrow m_1} c'(u(t)).$$

The left side is $\Phi + e^{-\rho w}\lambda(m_1)$ according to (13). To obtain the right side, we apply (13) to the interval $[m_1, 2m_1)$, and obtain

$$c'(u(t)) = C'(U(m_1)) + e^{-\rho w}\tilde{\lambda}(t), \quad t \geq m_1,$$

where $\tilde{\lambda}$ denotes the multiplier λ for the problem on the interval $[m_1, 2m_1)$. Because $\tilde{\lambda}(m_1) = 0$, we have $\lim_{t\downarrow m_1} c'(u(t)) = c'(u(m_1)) = C'(U(m_1)) + 0 = \Phi + \lambda(m_1)$. Therefore,

$$\lim_{t\uparrow m_1} c'(u(t)) = \Phi + e^{-\rho w}\lambda(m_1) > \Phi + \lambda(m_1) = \lim_{t\downarrow m_1} c'(u(t)).$$

□

Proof of Proposition 3: First, because $\frac{\Phi+e^{-\rho w}\lambda}{\Phi+\lambda}$ decreases in λ, and $\lambda(m_1)$ decreases in $g(0)$ and m_1, there is a unique value for $g(0)$ (as well as m_1) for a given ψ.

Second, to show that (23) is sufficient, we prove that

$$\mathscr{C}'(m_1) \begin{cases} < 0, & m_1 < m_1^*; \\ > 0, & m_1 > m_1^*. \end{cases}$$

This is because $\frac{\Phi+e^{-\rho w}\lambda(m_1)}{\Phi+\lambda(m_1)}$ strictly increases in m_1: $\frac{\Phi+e^{-\rho w}\lambda(m_1)}{\Phi+\lambda(m_1)}$ decreases in $\lambda(m_1)$ and the proof of Lemma 3 shows that $\lambda(m_1)$ decreases in $g(0)$ and m_1. □

Details in the computation of $\mathscr{C}'(m_1)$

Rewrite $\mathscr{C}(m_1)$ as

$$\begin{aligned}&\int_0^{m_1} e^{-(r+\pi)t}\big(\pi c(E^{m_1}) + rc(u^{m_1}) + \Phi((r+\pi)U^{m_1} - \pi E^{m_1} - ru^{m_1} - (U^{m_1})')\\&+\lambda^{m_1}(rE^{m_1} - re^{-\rho w}u^{m_1} - (E^{m_1})')\big)dt + e^{-(r+\pi)m_1}\left(\gamma + \psi + c(U^{m_1}(m_1))\right)\\&+e^{-(r+\pi)m_1}\lambda^{m_1}(m_1)(E^{m_1}(m_1) - U^{m_1}(m_1)),\end{aligned}$$

where we put a superscript m_1 on $U(\cdot)$, $E(\cdot)$, $u(\cdot)$, and $\lambda(\cdot)$ because these optimal paths rely on m_1. We use the Envelope theorem to simplify the computation of $\mathscr{C}'(m_1)$. Since $U^{m_1}(t)$, $E^{m_1}(t)$, $u^{m_1}(t)$ are already optimally chosen at each t, we may view them as fixed when we vary m_1. Further, $U^{m_1}(m_1)$ and $E^{m_1}(m_1)$ can be viewed as varying only with the terminal date in the parenthesis.[10] Viewed in this light, a small increment of m_1 is just an extrapolation of all time paths over a longer duration of unemployment, while the paths themselves are fixed. That is, we view all superscripts as being fixed and omit them when we calculate derivatives. Because $E(m_1) - U(m_1) = 0$, we have

$$\begin{aligned}\mathscr{C}'(m_1) &= e^{-(r+\pi)m_1}\Big(\pi c(E(m_1)) + rc(u(m_1)) - (r+\pi)(\gamma + \psi + c(U(m_1)))\\&\quad +c'(U(m_1))U'(m_1) + \lambda(m_1)\left(E'(m_1) - U'(m_1)\right)\Big).\end{aligned}$$

It follows from $c'(U(m_1)) = \Phi + \lambda(m_1)$, $\lambda'(m_1) = 0$ and Lemma 2 that

$$\begin{aligned}&c'(U(m_1))U'(m_1) + \lambda(m_1)\left(E'(m_1) - U'(m_1)\right)\\=\ &\Phi U'(m_1) + \lambda(m_1)E'(m_1) = \left(\Phi U(m_1) + \lambda(m_1)E(m_1)\right)' = 0.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathscr{C}'(m_1) &= e^{-(r+\pi)m_1}\Big(\pi c(E(m_1)) + rc(u(m_1)) - (r+\pi)(\gamma + \psi + c(U(m_1)))\Big)\\&= e^{-(r+\pi)m_1}\left(r\rho^{-1}\log\left(\frac{\Phi + e^{-\rho w}\lambda(m_1)}{\Phi + \lambda(m_1)}\right) - (r+\pi)(\gamma+\psi)\right).\end{aligned}$$

Fixed-point condition for ψ

The condition for ψ is that ψ is the fixed point of operator T, i.e.,

$$\psi + c(U(0)) = T(\psi) + c(U(0)) \equiv \min_{\sigma} C(\sigma).$$

We obtain ψ from the first-order condition (23) for m_1,

$$\psi = \frac{r\rho^{-1}}{r+\pi}\log\left(\frac{\Phi + e^{-\rho w}\lambda(m_1)}{\Phi + \lambda(m_1)}\right) - \gamma.$$

[10] This is because $U^{\tilde{m}_1}(m_1)$ and $E^{\tilde{m}_1}(m_1)$ can be viewed as being fixed when we vary $\tilde{m}_1$.

We obtain $T(\psi)$ from the HJB equation for the cost function at time zero

$$\begin{aligned} T(\psi) + c(U(0)) &= \frac{\pi c(E(0)) + rc(u(0)) + \Phi\left((r+\pi)U(0) - \pi E(0) - ru(0)\right)}{r+\pi} \\ &= \frac{\pi}{r+\pi}\left(\frac{\Phi}{g(0)} - \log\left(\frac{\Phi}{g(0)}\right) - 1\right) + c(U(0)). \end{aligned}$$

The fixed-point condition $\psi = T(\psi)$ is rewritten as

$$(r+\pi)\gamma = r\rho^{-1}\log\left(\frac{\Phi + e^{-\rho w}\lambda(m_1)}{\Phi + \lambda(m_1)}\right) - \pi\left(\frac{\Phi}{g(0)} - \log\left(\frac{\Phi}{g(0)}\right) - 1\right). \tag{31}$$

PROPOSITION 5 *The path that satisfies (31) exists and is unique.*

PROOF. The existence of a path that satisfies (31) follows from the intermediate value theorem and the fact that right side of (31) is either extremely large or extremely small if we vary $g(0)$. To see this, note that the proof of Lemma 3 shows that $\lim_{g(0)\downarrow\Phi} m_1 = 0 = \lim_{g(0)\downarrow\Phi}\lambda(m_1)$. Therefore,

$$\lim_{g(0)\downarrow\Phi} r\rho^{-1}\log\left(\frac{\Phi + e^{-\rho w}\lambda(m_1)}{\Phi + \lambda(m_1)}\right) - \pi\left(\frac{\Phi}{g(0)} - \log\left(\frac{\Phi}{g(0)}\right) - 1\right) = 0.$$

On the other hand, the proof of part (ii) of Lemma C.2 shows the existence of paths with $\lambda(m_1)$ approaching $-\Phi$ and $g(0) \in (\Phi, \Phi e^{\rho w})$. For these paths, $\log\left(\frac{\Phi+e^{-\rho w}\lambda(m_1)}{\Phi+\lambda(m_1)}\right)$ can be arbitrarily large, while $\frac{\Phi}{g(0)}$ remains bounded.

The uniqueness can be shown by contradiction. Suppose there are two paths satisfying (31). Associated with the two paths are two fixed points, $\psi < \tilde{\psi}$. Because the principal facing $\tilde{\psi}$ may monitor at $m_1(\psi) > 0$ and adopt the optimal consumption paths under ψ,

$$T(\tilde{\psi}) \le \psi + e^{-(r+\pi)m_1(\psi)}(\tilde{\psi} - \psi) < \tilde{\psi},$$

which contradicts the fact that $\tilde{\psi}$ is a fixed point. □

LEMMA C.2 *Consider the ODE system (21), (30) with time running backwards, that is,*

$$\lambda' = \pi(g - \Phi - \lambda), \tag{32}$$

$$g' = rg - \frac{rg^2}{\Phi e^{\rho w} + \lambda}. \tag{33}$$

Suppose the initial condition is $(\lambda(0), g(0) = \Phi + \lambda(0))$, $-\Phi < \lambda(0) < 0$, *and* $m^-(\lambda(0))$ *denotes the first time to hit the g-axis, i.e.,* $m^-(\lambda(0)) = \min_t\{t > 0 : \lambda(t) = 0\}$.

(i) $(g-\lambda)'(t) > 0$ *for all* $t \in [0, m^-(\lambda(0))]$.

(ii) $m^-(\lambda(0))$ *is finite, and* $\lim_{\lambda(0)\downarrow-\Phi} m^-(\lambda(0)) = \infty$.

PROOF.

(i) The path starting with $(\lambda(0), g(0) = \Phi + \lambda(0))$ has

$$\begin{aligned} \lambda'(0) &= \pi(g(0) - \Phi - \lambda(0)) = 0, \\ g'(0) &= rg(0) - \frac{rg(0)^2}{\Phi e^{\rho w} + \lambda(0)} > 0. \end{aligned}$$

Hence it moves beyond $g = \Phi + \lambda$ at time zero and satisfies $\Phi + \lambda < g < \Phi e^{\rho w} + \lambda$ before reaching the g-axis. If $\Phi + \lambda < g < \Phi e^{\rho w} + \lambda$, then $g' > 0$ and $\lambda' > 0$.

To show that $(g - \lambda)'(t) > 0$ for all $t \in [0, m^-(\lambda(0))]$, suppose to the contrary that $(g - \lambda)'(s) \leq 0$ for some s. Let $t^* = \min_s\{s > 0 : (g - \lambda)'(s) \leq 0\}$. It is easily seen that $(g-\lambda)'(t^*) = 0$ and $(g-\lambda)''(t^*) \leq 0$. Since $(g-\lambda)' = rg - \frac{rg^2}{\Phi e^{\rho w}+\lambda} - \pi(g-\Phi-\lambda)$,

$$\begin{aligned} (g - \lambda)''(t^*) &= \left(r - \frac{2rg(\Phi e^{\rho w} + \lambda)}{(\Phi e^{\rho w} + \lambda)^2} - \pi\right) g'(t^*) + \left(\frac{rg^2}{(\Phi e^{\rho w} + \lambda)^2} + \pi\right) \lambda'(t^*) \\ &= \left(r + \frac{rg^2 - 2rg(\Phi e^{\rho w} + \lambda)}{(\Phi e^{\rho w} + \lambda)^2}\right) g'(t^*) \\ &= r\frac{(\Phi e^{\rho w} + \lambda - g)^2}{(\Phi e^{\rho w} + \lambda)^2} g'(t^*) > 0, \end{aligned}$$

where the second equality follows from $g'(t^*) = \lambda'(t^*)$. This contradicts that $(g - \lambda)''(t^*) \leq 0$.

(ii) First, we show that $m^-(\lambda(0))$ is finite. We know from part (i) that $\lambda' > 0$. It follows from (32) and $(g - \lambda)' > 0$ in part (i) that

$$\lambda'' = \pi(g - \lambda)' > 0.$$

Hence starting from $\lambda(0) < 0$, $\lambda(t)$ accelerates and will reach zero in finite time.

Second, we show that $\lim_{\lambda(0)\downarrow -\Phi} m^-(\lambda(0)) = \infty$. If $\lambda(0) = -\Phi$ and $g(0) = 0$, then

$$\begin{aligned} \lambda'(0) &= \pi(g(0) - \Phi - \lambda(0)) = 0, \\ g'(0) &= rg(0) - \frac{rg(0)^2}{\Phi e^{\rho w} + \lambda(0)} = 0. \end{aligned}$$

Continuity of the ODE system (32), (33) implies that (λ, g) will stay in a small neighborhood of $(-\Phi, 0)$ for a long duration if $\lambda(0)$ is sufficiently close to $-\Phi$ and $g(0) = \Phi + \lambda(0)$. Therefore, $\lim_{\lambda(0)\downarrow -\Phi} m^-(\lambda(0)) = \infty$.

□

Appendix D Stochastic Verification

D.1 Construction of a Contract

To prove Proposition 4, we first construct a contract σ^* in which $E(t) > U(t)$ implies $p(t) = 0$, and $E(t) = U(t)$ implies $p(t) > 0$. This contract has the features described in Proposition 4, and in the next section we verify it is indeed optimal.

First, since the principal does not monitor in this contract when $E > U$, we still use the ODE system (20), (21) to find a solution path in the interval $[0, N]$, where N satisfies

$$-\int_0^N \lambda(t)\left(rE - re^{-\rho w}u\right)dt - \lambda(N)(e^{\rho\phi} - 1)E(N) + \gamma = 0. \tag{34}$$

The two boundary conditions for the ODE system (20), (21) are still $\lambda(0) = 0$ and $E(N) = -\rho^{-1}(\Phi + \lambda(N))^{-1}$.

Lemma 4 *The N that satisfies (34) exists and is unique.*

Proof. For uniqueness, we show that $f(N) \equiv -\int_0^N \lambda(t)\left(rE - re^{-\rho w}u\right)dt - \lambda(N)(e^{\phi} - 1)E(N)$ decreases with N. Since both $\lambda(N)$ and $E(N)$ are negative and decreasing with N, $-\lambda(N)(e^{\phi} - 1)E(N)$ decreases with N. Moreover,

$$-\lambda\left(rE - re^{-\rho w}u\right) = \frac{r|\lambda|}{g(\Phi e^{\rho w} + \lambda)}(g - \lambda - \Phi e^{\rho w}).$$

For fixed t, $\frac{r|\lambda|}{g(\Phi e^{\rho w}+\lambda)}$ increases with N, while $(g - \lambda - \Phi e^{\rho w})$ is more negative with higher N. Therefore, $-\int_0^N \lambda\left(rE - re^{-\rho w}u\right)dt$ decreases with N too. For existence, note that $\lim_{N\to 0} f(N) = 0$. Because $\lim_{N\to\infty} \lambda(N) = -\Phi$ and $\lim_{N\to\infty} E(N) = -\infty$, we have $\lim_{N\to\infty} f(N) = -\infty$. □

Second, choose $p > 0$ after N so that the state vector stays on the 45-degree line before the monitoring arrives, i.e., $U(t) = E(t)$ for all $t \geq N$. Choosing $\tilde{U}(N) = U(0) = -\frac{1}{\rho\Phi}$ and solving the equation $U'(N) = E'(N)$, we have

$$p = \frac{r(1 - e^{-\rho w})(\Phi + e^{-\rho w}\lambda(N))^{-1}}{e^{\rho\phi}(\Phi + \lambda(N))^{-1} - \Phi^{-1}} > 0. \tag{35}$$

Note that p is independent of Φ. This also implies that $p > 0$ is time invariant after N because $U(t) = E(t)$ for $t \geq N$.

Third, the constructed solution path defines a contract σ^* as follows. For each $t \in [0, N]$, the policy $u(t)$ is obtained by the first-order condition (13)

$$u(t) = -\frac{1}{\rho(\Phi + e^{-\rho w}\lambda(t))}. \tag{36}$$

If $t \geq N$, then the state vector moves along the 45-degree line, and $u(t)$ is always proportional to $(U(t), E(t))$. That is, for all $t \geq N$,

$$\frac{u'(t)}{u(t)} = \frac{E'(t)}{E(t)} = \frac{U'(t)}{U(t)} = r - \frac{r(\Phi + \lambda(N))}{\Phi + e^{-\rho w}\lambda(N)} + p\left(1 - \frac{\Phi + \lambda(N)}{\Phi}\right) > 0. \tag{37}$$

The contract σ^* is defined by (34–37), and the property that the continuation contract after a monitoring at $t \geq N$ starts a new cycle, in which the continuation utility is $\tilde{U}(t) = \frac{\Phi+\lambda(N)}{\Phi} U(t)$ instead of $U(0)$. In this construction, σ^* has the features mentioned in Proposition 4.

D.2 Optimality of the Contract

First, using the path obtained in Lemma 4, we construct a cost function C as

$$\begin{aligned}(r+\pi)C(U(t),E(t)) &= \pi c(E(t)) + rc(u(t)) + \Phi((r+\pi)U(t) - \pi E(t) - ru(t)) \\ &\quad +\lambda(t)(rE(t) - re^{-\rho w}u(t)).\end{aligned} \tag{38}$$

LEMMA 5 *$C_U(U(t),E(t)) = \Phi$, and $C_E(U(t),E(t)) = \lambda(t)$.*

PROOF. Differentiate (38) with respect to t, we have

$$(r+\pi)(C_U U'(t) + C_E E'(t)) = \pi c'(E)E'(t) + \Phi((r+\pi)U'(t) - \pi E'(t)) + \lambda(t)rE'(t) + \lambda'(t)E'(t),$$

which, after substituting $\lambda'(t) = \pi(\Phi - c'(E) + \lambda)$, becomes

$$C_U U'(t) + C_E E'(t) = \Phi U'(t) + \lambda(t)E'(t).$$

Homogeneity of $C(\cdot,\cdot)$ implies that $C_U U(t) + C_E E(t) + \rho^{-1} = 0 = \Phi U(t) + \lambda(t)E(t) + \rho^{-1}$. Because the vectors $(U'(t), E'(t))$ and $(U(t), E(t))$ are linearly independent (we have shown that $\left(\frac{U}{E}\right)'(t) < 0$ in the proof of Proposition 2, which is $\frac{E'(t)}{E(t)} > \frac{U'(t)}{U(t)}$), we have $C_U = \Phi$ and $C_E = \lambda(t)$. □

Second, we verify that the cost function C satisfies the HJB equation:

$$\begin{aligned}(r+\pi)C(U,E) = \min_{u,p,\tilde{U},\tilde{E}} \quad &\Big\{rc(u) + \pi c(E) + p\left(C(\tilde{U},\tilde{E}) + \gamma - C(U,E)\right) \\ &+C_U\left(r(U-u) - \pi(E-U) - p(\tilde{U}-U)\right) \\ &+C_E\left(rE - re^{-\rho w}u - p(e^{\rho\phi}-1)E\right)\Big\},\end{aligned} \tag{39}$$

where $(\tilde{U},\tilde{E})$ is the new state vector the principal chooses after the next monitoring.

LEMMA 6 *The $C(\cdot,\cdot)$ defined in (38) satisfies (39).*

PROOF. The only differences between (38) and (39) are the terms associated with arrival rate p, which will be shown to be zero in this proof. Fix a $t \in [0,N]$ and consider the HJB equation at $(U(t),E(t))$. The first-order condition for $\tilde{U}$ implies that $\tilde{U} = U(0)$. Then we have

$$\begin{aligned}&C(\tilde{U},\tilde{E}) + \gamma - C(U,E) - \Phi(\tilde{U}-U) - C_E(e^{\rho\phi}-1)E \\ =\ &-\int_0^t \lambda(s)\left(rE(s) - re^{-\rho w}u(s)\right)ds - \lambda(t)(e^{\phi}-1)E(t) + \gamma.\end{aligned}$$

The above is decreasing in t because $\lambda(t) < 0$, and $E(t) < 0$ both decrease in t. Moreover, the integral $-\int_0^t \lambda(s)\left(rE(s) - re^{-\rho w}u(s)\right) ds$ decreases in t because

$$rE(t) - re^{-\rho w}u(t) = E'(t) = \rho E^2(t)g'(t) < 0.$$

Therefore, the definition of N in (34) implies that

$$C(\tilde{U},\tilde{E}) + \gamma - C(U,E) - \Phi(\tilde{U}-U) - C_E(e^{\rho\phi}-1)E \begin{cases} > 0, & \text{if } t < N, \\ = 0, & \text{if } t = N. \end{cases}$$

This implies that

$$\min_{p\geq 0} p\left(C(\tilde{U},\tilde{E}) + \gamma - C(U,E) - \Phi(\tilde{U}-U) - C_E(e^{\rho\phi}-1)E\right) = 0,$$

which finishes the proof. □

Finally, to complete the proof of Proposition 4, we show that the contract σ^* is optimal.

Proof of Proposition 4: Because the technique of using the HJB equation to verify optimality is standard, we spare the reader of detailed steps. Given the initial promised utilities (U, E), we need to verify that

(i) The cost of the contract σ^* is $C(U, E)$.

(ii) The costs of other I.C. contracts are weakly higher than $C(U, E)$.

We only verify (ii) here, since the proof for (i) can be obtained simply by replacing the following inequalities with equalities.

To see that the cost of an I.C. contract $\left\{(\tilde{c}^E(t), \tilde{c}^U(t), \tilde{p}(t)); t \geq 0\right\}$ is higher than $C(U, E)$, define

$$\begin{aligned} h(T) &= \int_0^T e^{-(r+\pi)t - \int_0^t \tilde{p}(x)dx}\left(\pi c(\tilde{E}(t)) + r\tilde{c}^U(t) + \tilde{p}(t)\left(C(\tilde{U}(t),\tilde{E}(t)) + \gamma\right)\right)dt \\ &\quad + e^{-(r+\pi)T - \int_0^T \tilde{p}(x)dx} C(U(T), E(T)). \end{aligned}$$

The HJB equation implies that $f'(T) \geq 0$. Therefore, $h(T)$ increases in T, and

$$C(U, E) = h(0) \leq h(T).$$

Taking limit $T \to \infty$, we have

$$C(U,E) \leq \int_0^\infty e^{-(r+\pi)t - \int_0^t \tilde{p}(x)dx}\left(\pi c(\tilde{E}(t)) + r\tilde{c}^U(t) + \tilde{p}(t)\left(C(\tilde{U}(t),\tilde{E}(t)) + \gamma\right)\right)dt,$$

which can be rewritten as

$$\begin{aligned} C(U,E) &\leq E\left[\int_0^{\tau_1} e^{-rt}\left(\pi c(\tilde{E}(t)) + r\tilde{c}^U(t)\right)dt\right] + E\left[e^{-r\tau_1}\gamma\right] \\ &\quad + E\left[e^{-r\tau_1} C(\tilde{U}(\tau_1), \tilde{E}(\tau_1))\right], \end{aligned}$$

where τ_1 is the first monitoring time and $(\tilde{U}(\tau_1), \tilde{E}(\tau_1))$ is the state vector immediately after monitoring. Inductively, we obtain

$$\begin{aligned} C(U,E) &\leq E\left[\int_0^{\tau_n} e^{-rt}\left(\pi c(\tilde{E}(t)) + r\tilde{c}^U(t)\right)dt\right] + E\left[\sum_{i=1}^{n} e^{-r\tau_i}\gamma\right] \\ &\quad + E\left[e^{-r\tau_n} C(\tilde{U}(\tau_n), \tilde{E}(\tau_n))\right], \end{aligned}$$

where τ_n is the nth monitoring time. Without loss of generality, we may assume that $\lim_{n\to\infty}\tau_n = \infty$ almost surely (otherwise the principal monitors infinitely many times in finite time and the monitoring cost is infinity). Taking limit $n \to \infty$ yields

$$C(U,E) \leq E\left[\int_0^{\infty} e^{-rt}\left(\pi c(\tilde{E}(t)) + r\tilde{c}^U(t)\right)dt\right] + E\left[\sum_{i=1}^{\infty} e^{-r\tau_i}\gamma\right].$$

□

Appendix E Imperfect Detection

This section presents a version of the stochastic verification model where detection is imperfect. Specifically, there is a positive probability $\varpi > 0$ of monitoring error. In the event of monitoring error, an unemployed worker is labeled as employed. If an unemployed worker is monitored after reporting unemployment, the principal observes either an unemployed signal $\mathcal{U}$ with probability $1-\varpi$ or an employed signal $\mathcal{E}$ with probability ϖ. On the other hand, there is no monitoring error that labels an employed worker as being unemployed, i.e., if an employed worker is monitored after reporting unemployment, the principal observes $\mathcal{E}$ with probability one.

The timing of the problem is similar to the stochastic verification case in Section 8. The planner still chooses the arrival rate of monitoring, $p(t)$, conditional on the report of unemployment in period t. There are, however, two differences in the case of imperfect detection. First, the planner assigns continuation utilities based not only on whether or not monitoring occurs (as above) *but also* on the signal from monitoring. Let $U_{\mathcal{U}}(t)$ and $U_{\mathcal{E}}(t)$ be the continuation utilities of a monitored unemployed worker with signals $\mathcal{U}$ and $\mathcal{E}$ at t, respectively. Let $E_{\mathcal{E}}(t)$ be the continuation utility of a monitored employed worker (whose signal can only be $\mathcal{E}$) at t. Finally, $E_{\mathcal{U}}(t)$ is the continuation utility of a monitored unemployed worker with signal $\mathcal{U}$ who transited to employment immediately after being monitored. Second, the penalty is exogenous in the case of perfect detection above, but is endogenous with imperfect detection.

Similar to (24) and (25), the promise-keeping constraint and incentive constraint are

$$U' = r(U-u) - \pi(E-U) - p\left[(1-\varpi)U_{\mathcal{U}} + \varpi U_{\mathcal{E}} - U\right], \tag{40}$$

$$E' \leq rE - re^{-\rho w}u - p(E_{\mathcal{E}} - E). \tag{41}$$

There are two differences between these two equations and (24) and (25). First, the promise-keeping constraint (40) incorporates the possibility that an unemployed worker may be

labeled as employed after monitoring. Second, in (25) the last term on the right-hand side results from the exogenous and finite penalty, ϕ, whereas in (41) the last term allows the penalty $E_{\mathcal{E}}$ to be endogenous.

The main results from the perfection detection case and stochastic monitoring still hold here. That is, the optimal monitoring mechanism consists of cycles. Within each cycle, there exists some N such that the planner sets $p = 0$ before N, and then monitors at rate p thereafter. Formally we state the following proposition.

PROPOSITION 6 *There exists an $N > 0$ such that the principal monitors the unemployed with a constant arrival rate $p > 0$ if and only if $t \geq N$. Before N, the time path $(U(\cdot), E(\cdot))$ converges to the 45-degree line; after N, the utility pair $(U(t), E(t))$ remains stationary (i.e., $U(t) = E(t) = U(N) = E(N)$ for all $t \geq N$) until the worker is randomly drawn to be monitored. If the observed signal from monitoring is $\mathcal{E}$, the worker is punished, $U_{\mathcal{E}} = E_{\mathcal{E}} < U(N)$. If the signal is $\mathcal{U}$, the worker is rewarded, $U_{\mathcal{U}} > U(N)$, and the contract enters a new cycle.*