Abstract

This paper studies a dynamic agency model in which the agent privately observes the firm’s cash flows that are subject to persistent shocks. We characterize the optimal contract by continuation utilities contingent on the agent’s report today and tomorrow. The optimal contract can be implemented by a contingent credit line, stock options, and equity. In contrast to the iid case, we find: (i) investment is possibly efficient in the constrained firm, and is varying with cash flow in the unconstrained firm; (ii) the firm possibly experiences longer time of being financially constrained; (iii) the agent receives cash payment less than what he can divert from cash flow and investors hold more equity stake; (iv) compensation to the agent is via stock options and equity, the combination of which depends on persistence level; (v) firm credit line limits are contingent on compliance with a cash flow covenant and are history dependent.

Keywords: Dynamic agency, Markovian information, Investment, Payout, Capital structure.

JEL Classification: D82, D86, D92, G32, G35.

(∗) We thank Marina Halac, Adriano Rampini, Philipp Sadowski, Lukas Schmid, S. Viswanathan, Neil Wallace, and seminar participants at Duke University, New York University, University of Rochester, Johns Hopkins University, Georgia State University, Indiana University Bloomington, and 2015 Columbia/Duke/MIT/Northwestern IO Theory Conference. Krishna is grateful to the National Science Foundation for financial support through grant SES-1132193. The most recent version of the paper is available at http://ssrn.com/abstract=2721428.

(†) Simon School of Business, University of Rochester <shiming.fu@simon.rochester.edu>

(‡) Department of Economics, Florida State University <vkrishna@fsu.edu>
1. Introduction

There is considerable evidence that funding for firms, especially young firms, is far from efficient, and that firms must grow over time into their optimal size. In particular, financing constraints greatly affect firm size, growth, and a young firm’s prospects. Starting from Jensen and Meckling (1976), a large literature has focused on the conflicts of interests between investors and agents as a key friction that constrains firm financing and investment. Agency problems arise because agents have more information about their firm behavior and their own actions than the outside investors. Financial contracting therefore characterizes the securities and firm policies designed to mitigate agency conflicts.

An influential recent literature analyzes agency problems in dynamic contexts. For instance, Clementi and Hopenhayn (2006), and DeMarzo and Fishman (2007b) characterize the optimal long-term contract and firm dynamics when agency frictions are involved. These models have advantage in analyzing how firm investment, payout, capital structure and other policies evolve, and how firms grow over time. However, they typically assume by convenience that agents’ private information about firm behavior is iid over time, because considering persistence has proved challenging in this class of models. In practice, firms’ profitability or other economic behavior, which is agents’ private information in these models, exhibits high autocorrelation. For instance, Gomes (2001) calibrated that autocorrelation of firm productivity shocks is 0.62 at the annual frequency. Many other researches show even higher numbers. So having private information about firm behavior today will definitely help agents to better predict firms’ future prospects. Moreover, if we want to take the dynamic agency models seriously or quantify these models, then adopting the persistent assumption becomes crucial.

In this paper, we incorporate persistent private information in a dynamic agency model of firm financing. The firm consists of a risk-neutral agent and risk-neutral investors. The agent has the expertise to operate the firm but does not have funds. The investors provide funds to launch the firm and finance firm’s risky investment in each

(1) Gertler and Gilchrist (1994) find that manufacturing significantly declines in small firms when monetary policy tightens. Beck, Demirgüç-Kunt and Maksimovic (2005) documents that financial constraints create obstacle to the growth of small firms, and that small firms benefit most from financial development.
(3) Nikolov and Schmid (2012) estimate a q-theoretic dynamic agency model with iid cash flow shocks. They find that agency conflict between investors and agent needs to be substantial in order to ration-alize observed firm financing and investment policies.
The agency problem is that the cash flows from investment projects are privately observed by the agent. So the agent has incentive to divert firm cash for his own consumption. The key element is that firm cash flows are subject to positively correlated shocks which follow a two-state Markov process.

We develop a method to solve the optimal contract in this environment. When the cash flow shocks are iid, it is well known that promising the agent continuation values contingent on his report today is sufficient for truth telling. However, this approach is no longer sufficient in the persistent environment. This is because agent’s report today also contains private belief about the likelihood of future shocks. To induce truth telling, the contract has to take into account the agent’s private benefit from i) diverting today’s cash flow and ii) distorting investors’ belief about future shocks. We show a sufficient method is to promise the agent different contingent continuation values from tomorrow onward according to his report today. In other words, the continuation values will be contingent on both today and tomorrow’s report. By characterizing and implementing the optimal contract, we find that the persistent environment implies several distinct features regarding firm investment, capital structure, compensation, and growth dynamics.

In the iid case, investment is only determined by the credit availability of the firm. When the firm is young and constrained, its investment is very sensitive to cash flow shock. But investment becomes constant (at the efficient level) when the firm is unconstrained. Such dynamics cannot be reconciled with the evidence shown by Kaplan and Zingales (1997): less financially constrained firms exhibit greater investment to cash flow sensitivity. In our model, besides the financial slack of the firm, investment is also determined by investors’ belief about the likelihood of good shock in the future, which in turn depends on the current cash flow shock. So investment is varying (at efficient levels) with firm cash flow even when it is unconstrained.

A key feature of the implementation in our model is to finance the firm by credit line contingent on maintaining cash flow covenant. On the contrary, the iid models are always implemented by an uncontingent credit line with fixed limit. In our model, because the agent has private belief about future cash flow shocks, the variations in continuation values (or the unused credit lines in the implementation) has to be no less than the current cash flow plus a dynamic information rent. Due to this new element of dynamic information rent (which disappears in the iid case), credit limit has to be adjusted according to cash flow shocks in order to implement our optimal contract.

Credit line is an important way of firm financing and liquidity management. According to Demiroglu and James (2011), drawdowns of credit line account for 75% of bank lending to firms and 63% of corporate debt. In our model, when the firm violates cash flow covenant at any time, its current credit limit will immediately drop, and its
future contingent credit limits will also be reduced. These results very well rationalize how the credit line is provided in practice. Empirically, cash flow based covenants are typically written into the credit line contract. Also most credit lines have material adverse change (MAC) clauses which permit lenders to withhold funds if a borrower’s credit quality deteriorates significantly. According to Sufi (2009), a covenant violation is associated with a 15% to 25% drop in the availability of total line of credit.

The long-run incentive provision and implied compensation scheme in our model are also in stark contrast to the iid models. In dynamic contexts, the variation of future investments and the associated information rents provide incentives for the agent to report cash flow truthfully today. In the iid case, investment and associated information rent become constant when the firm is unconstrained. The only way to provide incentive in the long-run is to issue the agent the same amount of cash payment as what he can divert. So agent’s payment is exogenously determined and linear in firm performance. In the persistent case, since it is efficient to reduce investment level after a bad shock, investors have an added threat to punish the agent if he were to divert cash flow. Hence the agent only receives an endogenous payment less than what he can divert.

The agent’s payment in our model is shown to be convex in firm performance or stock price. In particular, the agent gets larger fraction of firm payouts as firm performance improves. If the persistence level is high, the agent only gets positive payment when good performances are observed in a row or when the highest stock price is achieved. Otherwise, investors obtain all the firm payouts. We show that this convex payment scheme can be implemented by a combination of stock options and equity, the composition of which depends on the persistence level. And as persistence level increases, stock option payoff accounts for a larger portion of the agent’s payment. On the contrary, because compensation is always linear in firm performance in the iid models, agent only holds equity and options never play a role.

Empirically, stock options are a popular way of compensating executives and em-

---

(4) According to Demiroglu and James (2011), coverage and debt-to-cash flow covenants are the two most common financial covenants. Coverage covenants require that a borrower’s coverage ratio (typically the ratio of EBITDA to fixed charges or interest expenses) remain above a minimum and debt-to-cash flow covenants restrict borrowing if the ratio of debt-to-cash flows exceeds a preset maximum.

(5) Both DeMarzo and Sannikov (2006) and DeMarzo and Fishman (2007b) assume that agent can divert $\lambda \leq 1$ fraction of firm cash flow and hence is compensated this amount by cash when the firm pays off its debt. Clementi and Hopenhayn (2006) and our model correspond to the case of $\lambda = 1$, and can also be extended to the case of $\lambda < 1$.

(6) For instance, the agent in Clementi and Hopenhayn (2006) is the firm’s residual claimant and holds all the equity. Although we also assume the agent can divert all the cash flow, he is not the residual claimant and investors in our model also hold an equity stake.
ployees. 71% of the 250 largest U.S. companies in the Standard & Poor’s 500 Index use stock options as incentive grant.\(^7\) According to Larcker 2008, payment from stock options accounts for 27\% (the largest component) of CEO compensation in the top 4000 U.S. companies. Bergman and Jenter (2007) also shows that stock options plan is the most common method for employee compensation below the executive rank.

Regarding the relation between firm size and age in the growth dynamics, our model better matches the empirical finding of Hurst and Pugsley (2011): most small firms stay in small size for long time and therefore are old firms. Because of persistence, the firm in our model is more likely to receive many bad shocks in a row and becomes more financially constrained. Moreover, investments after bad shocks will be kept low, slowing down firm growth and extending the stage of the firm being constrained. Therefore, the firm in our model possibly experiences longer time of being constrained than in the iid case, and stays in small or inefficient size for longer time.

In our model the constrained firm can temporarily invest at an efficient (first best) level before it becomes unconstrained; while in the iid case, investment is always distorted downward before the firm is unconstrained. This happens when the persistence level is high and a bad shock previously hit the firm. If that is the case, by misreporting the agent can divert little cash today because today’s investment level is low after a bad shock. Moreover, the agent will be punished hard by low level of future investments after reporting a bad shock today. So the agent is possibly better off by reporting truthfully. Or in other words, the incentive compatibility constraint is possibly slack. If this situation happens, the investment will be efficient even before the firm becomes unconstrained.

Consistent with the implications from the iid models, our model also predicts that payouts is delayed until the firm is unconstrained and investments are positively correlated over time. These predictions are robust for all persistence levels. Delaying payments is optimal since it reserves more liquidity in the firm to improve future investment efficiency. And since in the dynamic context firm policies always exhibit history dependence, investments will be positively correlated regardless of the persistence level. But of course, high persistence implies high autocorrelation of investment.

We review the relevant literature in Section 2. Section 3 introduces the model. Section 4 discusses the sequential contracts and our approach of recursively formulating the problem. Section 5 describes the optimal contract for the constrained firm as it pertains to compensation, investment, and the evolution of continuation utility, while

\(^7\) See the Frederic W. Cook survey of long-term incentive grant at: \url{http://www.fwcook.com/alert_letters/The_2014_Top_250_Report_Long-Term_Incentive_Grant_Practices_for_Executives.pdf}
Section 7 looks at the optimal contract for a mature firm. Section 6 shows that every firm will eventually mature (we ignore the possibility that the firm will be liquidated), while Section 8 discusses empirical implications and stylized facts. Section 9 discusses the initialization of the contract, the nature of path dependence, and the possibility of liquidating the firm. Section 10 concludes, and all proofs are in the appendices.

2. Related Literature

Our work builds on a literature that studies the financing of firms under asymmetric information, typically assuming that the agent can divert cash flows without the principal’s knowledge. An early and seminal paper in this literature is Bolton and Scharfstein (1990), who study a two-period model, where the threat of early termination provides incentives in the first period. Fully dynamic versions of CFD models are Clementi and Hopenhayn (2006), Biais et al. (2007), and DeMarzo and Fishman (2007a), where the latter two emphasize the implementation of the optimal contract via standard securities. All these papers, regardless of time horizon, consider iid shocks to the output process. We consider the same discrete time economic environment as Clementi and Hopenhayn (2006), except that we allow for persistence in the shocks to the output process.

Infinite horizon (iid) screening models were first studied by Thomas and Worrall (1990), who introduce recursive methods to such problems, and show that by using the utility promised to the agent as a state variable, the optimal contract can be reduced to a Markov decision process for the principal. Although the literature on firm financing has focused on the iid case, there is nonetheless a literature on dynamic screening with Markovian types. The recursive approach is emphasized by Fernandes and Phelan (2000), who note that promised utility alone is inadequate in the Markovian case. To recursively formulate the problem, they use two \textit{ex ante} promised utilities, one from truth-telling and the other from lying. Although we also use a vector of promised utilities, they are \textit{interim}, contingent on the production shock in the period. Our state variables are easier to specify the domain of the dynamic programming, allowing for analytical characterization.

Doepke and Townsend (2006) extends the environment to incorporate both hidden states and hidden actions. They focus on how to reduce constraints by imposing off-path utility bounds and how to numerically solve their model. Instead, we focus on analytical characterization and implementation of the optimal contract. Tchistyi (2013) studies an environment similar as ours except that his model has finite periods and no investment.

(8) Such models are therefore referred to as \textit{cash flow diversion} (CFD) models.
Although Tchistyj (2013) uses only one ex ante promised utility, a time varying functional has to be defined that transforms the agent’s on-path utility to the utility from lying. Kapička (2013) uses a first-order approach to study an environment with continuum of states that are persistent. If the information rent is monotone, Kapička (2013) shows that the state variables can be reduced to two numbers: continuation utility and marginal continuation utility. However, the validity of the first-order approach is hard to verify.

Battaglini (2005) considers the problem of a principal sells some quantity of a good to a consumer, whose valuation for the good follows a two-point Markov process. Pavan, Segal and Toikka (2014) and Eső and Szentes (2013) study mechanism design in dynamic quasilinear environment where the agent’s persistent private information are described by "impulse response functions". The key difference from our model is that our agent has no cash and is subject to limited liability, while the agent in these models has deep pocket, and only has a participation constraint that must be respected at each point in time. With no limited liability constraint, these models have the same feature: the principal’s expected payoff from implementing an allocation is the same as if he could observe the agent’s orthogonalized private information after the initial period. It is precisely the inability of the principal to extract future information rents that makes our model economically interesting.

While not directly related to the principal-agent literature, Halac and Yared (2014) consider the problem of a government that has time-inconsistent preferences. The government privately observes shocks that follow a two-state Markov process. Using the same techniques as in this paper, Bloedel and Krishna (2014) study the question of immiseration in a problem of risk-sharing where the agent’s taste shock follows a Markov process. Independently, Guo and Hörner (2014) use the same techniques to study mechanism design without monetary transfers.

3. Model

A principal with deep pockets has access to an investment opportunity. In order to avail herself of this opportunity, she needs the managerial skills of an agent. The agent has no funds to operate the project and is therefore dependent on the principal’s funds for operational costs. Time is discrete, the horizon is infinite, both the principal and agent are risk neutral, and both discount the future at a common rate of $\delta \in (0, 1)$.

The project, which we shall also refer to as the firm, requires an investment $k_t \geq 0$ in every period. Capital depreciates completely, and so cannot be carried over to subsequent periods. The return on capital is random, and is $\theta(s) R(k)$, where $s \in S := \{b, g\}$
and $\theta$ is a $\{0,1\}$-valued random variable. The function $R : \mathbb{R}^+ \to \mathbb{R}^+$ is increasing, strictly concave, and continuously differentiable, with $R(0) = 0$ and the usual Inada conditions, i.e., $\lim_{k \downarrow 0} R'(k) = \infty$, and $\lim_{k \uparrow \infty} R'(k) = 0$.

The states $s \in S$ evolves according to a Markov process, and $\theta : S \to \{0,1\}$ is defined as $\theta(b) = 0$, while $\theta(g) = 1$. Thus, $\theta$ represents a random production shock; a return of $R(k)$ occurs if the shock is good, while a return of 0 occurs if the shock is bad. Conditional on today’s shock being $s \in \{b,g\} =: S$, the probability of having a good shock tomorrow is $p_s$. The transition probabilities for the production shocks is given by the matrix

$$
\begin{bmatrix}
  & b & g \\
  b & 1 - p_b & p_b \\
  g & 1 - p_g & p_g \\
\end{bmatrix}
$$

We shall assume that $\Delta := p_g - p_b \geq 0$, i.e., the Markov process is positively persistent. The case where $\Delta = 0$ corresponds to the iid case.\(^9\) We also assume that $p_b, p_g \in (0,1)$, which ensures that the Markov process has a unique ergodic measure and has neither absorbing nor transient sets.

The agency problem arises because (i) the principal cannot observe the output while the agent can, and (ii) the agent is cash constrained. The agent’s lack of funds implies that he needs, as mentioned above, working capital from the principal. We assume, therefore, that the principal cannot extract payments in excess of revenues from the agent, so the agent is protected by limited liability. These are the twin frictions of the model. If the agent were not cash constrained, the principal could simply sell him the firm. If the output were observable, the principal would pay the agent the ‘minimum wage’ of 0, i.e., offer him just enough to stay with the firm, while she retains all the revenue. Thus (and this is true for any degree of persistence), it is the combination of limited liability constraints and privately observed cash flow that gives rise to a non-trivial contracting problem.

The cumulative information available to the principal at time $t$ consists of the investments the principal has made and the amount of cash that the agent has transferred back to her in all prior periods. A contract (formally defined below) conditions investment and cash transfers (conditional on positive output) in any period on all previous cash transfers by the agent, and all previous investments by the principal. We assume

\(^9\) However, while $\Delta > 0$ implies the process is persistent, the levels of $p_b$ and $p_g$, and not just their difference, are also relevant for the evolution of investment etc in the optimal contract. A further discussion of degrees of persistence, especially as it pertains to contract structure, relies on more notation and terminology, and so is postponed to Section 4.5 provides a categorisation of degrees of persistence.
throughout that the agent cannot save cash made available to him in any period. In other words, all savings are done on behalf of the agent by the principal as part of the contract.

The timing runs as follows: At the beginning of time, at $t = 0$, the principal offers the agent in infinite horizon contract that he may accept or reject. If he rejects the offer, the principal and agent go their separate ways, and their interaction ends. If the contract is accepted, it is executed. The agent can leave at any time to an outside option worth 0 without further penalty. The principal fully commits to the contract. As mentioned before, the only significant difference between our model and that studied in Clementi and Hopenhayn (2006) is that we allow for persistence in the production shocks, while they restrict attention to the case where production shocks are iid. There is one other, minor, difference. Clementi and Hopenhayn (2006) allow for the project to be scrapped at any time for a value that is divided between the principal and the agent according to some formula that is history dependent and optimally chosen. For simplicity, we set the scrap value to zero. Our principal results go through in the case of a positive scrap value, albeit with some straightforward modifications. In particular, the properties of the mature firm are independent of the existence or level of a scrap value (as long as the scrap value of the firm does not exceed the value of the firm when it is run efficiently).

4. Contracts

A dynamic contract conditions investments and cash transfers on the history of all previous cash transfers and investments. By the Revelation Principle, we may equivalently think of the agent as reporting the current shock as being good or bad (which corresponds to positive and zero output respectively), so that a sequence of reports now constitutes a history. Of course, contracts can be written so as to condition on entire histories of reports. But, as it turns out, we may restrict attention to a class of recursive contracts, and this restriction is without loss of generality (in that there is no loss of social surplus or utility to either principal or agent). These recursive contracts are described next.
4.1. Recursive Contracts

Before we describe the recursive formulation of the principal’s problem, it is useful to reconsider the recursive formulation in the special case of iid states.\textsuperscript{10} As is well known, in the iid case, there is no loss in generality in indexing contracts according to promised utility \( v \), which is the agent’s expected lifetime utility from the contract. Upon entering a period with a specific level of promised utility \( v \), the agent’s report will result in a new level of promised utility \( w \) beginning the next period.

However, such an approach is inadequate in the Markovian case because the agent has private information about his preferences over future streams of utility. (Recall that today’s state dictates the probability distribution over tomorrow’s states, and today’s state is only known to the agent.) If the agent has lied in a period, and if the principal has promised utility \( v \) (in the form of future cash flows), the agent will assess a different expected utility from the stream of future cash flows than \( v \). This results in the principal and agent having different beliefs about how much utility is promised to the agent and, more importantly, the principal does not have enough instruments to successfully screen the agent and elicit this information.

We consider contracts where the principal indexes contracts by a pair of interim or contingent utilities and the previous period’s shock, which determines the distribution over the current period’s shock. The contingent utilities are represented by a vector \( v = (v_b, v_g) \) with the interpretation that if the agent reports a bad shock in the current period, he will get \( v_b \) utiles, but receives \( v_g \) utiles if he reports a good shock. This means promised utility can no longer serve as the state variable for the principal’s recursive problem (at least not without significant loss of generality and loss of utility to the principal).

We now describe how our approach differs from the iid case with promised utility. Suppose the agent enters the period with promised contingent utilities \( v = (v_b, v_g) \). As we shall see below (in the promise keeping constraints \([PK_b]\) and \([PK_g]\)), the agent receives his promised utility in terms of state (report) contingent payments and continuation contingent utilities. Thus, and this is crucial, even if the agent lied in the last period, contingent on today’s shock being \( g \) (say), his lifetime interim utility is still \( v_g \). This is because past information is now rendered payoff irrelevant (which follows directly from the assumption that states follow a Markov process). Thus, our formulation ensures that the agent cannot benefit from multiple deviations and, moreover, truth telling is optimal after every history. Also, observe that after every history, the contingent promised utiles are common knowledge. This ensures that contingent promised

\(\textsuperscript{10}\) Because this approach is now well understood, our description will be informal.
utilities can serve as a state variable for a recursive formulation.

It goes without saying that the equivalence of the proposed recursive formulation, with what we refer to as *contingent* promised utilities, and the sequential contract needs proof. However, the proof is very similar to the proof offered in Theorem 2.1 of Fernandes and Phelan (2000) and so is omitted.\(^{11}\)

### 4.2. Constraints

Given a pair of contingent utilities \(v = (v_b, v_g) \in \mathbb{R}^2\) with the last period being in state \(s\) (at least as far as the principal believes), the principal chooses an investment policy \(k(v, s) \in \mathbb{R}\), transfers \(m_b(v, s), m_g(v, s) \in \mathbb{R}\), and continuation contingent utilities \(w_b = (w_{bb}, w_{bg}) \in \mathbb{R}^2\) and \(w_g = (w_{gb}, w_{gg}) \in \mathbb{R}^2\) subject to promise keeping, incentive, and limited liability constraints. The promise keeping constraints are

\[
\begin{align*}
\text{[PK}_b\text{]} & : v_b = -m_b + \delta \mathbb{E}^b[w] \\
\text{[PK}_g\text{]} & : v_g = R(k) - m_g + \delta \mathbb{E}^g[w]
\end{align*}
\]

where \(\mathbb{E}^s[w] = (1 - p_s)w_b + p_s w_g\) for \(s = b, g\), represents the agent’s expected utility from the vector \(w \in \mathbb{R}^2\) when the current shock is \(s\). Persistence implies that, in general, \(\mathbb{E}^b[w] \neq \mathbb{E}^g[w]\).

Clearly, the only incentive constraint that need be considered is when the agent incorrectly reports the state as being bad rather than good, which is written as

\[
\text{[IC]} : v_g \geq R(k) - m_b + \delta \mathbb{E}^b[w]
\]

where \(\mathbb{E}^g[w]\) represents the agent’s expected continuation utility from reporting the current shock as bad when it is actually good. The limited liability constraints are

\[
\text{[LL]} : m_g \leq R(k) \quad \text{and} \quad m_b \leq 0
\]

Throughout we impose the feasibility constraint that \(k \geq 0\) without comment. Using the promise keeping constraints \([\text{PK}_b]\) and \([\text{PK}_g]\), the incentive constraint \([\text{IC}]\) can be written somewhat more simply as

\[
\text{[IC}\ast\text{]} : v_g - v_b \geq R(k) + \delta \Delta(w_{bg} - w_{bb})
\]

\(^{11}\) Fernandes and Phelan (2000) have a slightly different formulation, where the state variables are promised utility and a *threat-point* utility, where the latter evaluates the agent’s expected utility from cash streams if he has lied in the last period. Notice that both the promised and threat-point utilities are ex ante utilities, while our contingent utilities are interim in nature. Apart from this difference, the two approaches are essentially identical. Nevertheless, we shall see below that contingent utilities are somewhat easier to interpret, and so render themselves more suitable for the application considered in this paper;
The sum $R(k) + \delta \Delta (w_{bg} - w_{bb})$ represents the agent’s information rent conditional on being promised $v$ contingent utilities. This sum permits a natural decomposition into an iid and a Markovian component.

Clearly, $R(k)$ is the iid (or static) information rent while $\Delta (w_{bg} - w_{bb})$ is the Markovian (or dynamic) information rent which is 0 in the iid case, i.e., if $\Delta = 0$. Thus, the constraint [IC*] crystallizes the effect of Markovian states. If production shocks are iid, $\Delta = 0$ and [IC*] reduces to $v_g - v_b \geq R(k)$. As we shall see below, we must necessarily have $w_{bg} \geq w_{bb}$, which implies $\Delta (w_{bg} - w_{bb}) \geq 0$, so that with persistence, the incentive constraint is tighter. To see this more starkly, suppose $w_b$ is feasible for a low value of $\Delta$. Then, keeping $k$ fixed, an increase in $\Delta$ renders $w_b$ infeasible insofar as it leads to a violation of [IC*].

It is easy to see that given the promise keeping constraints [PK_b] and [PK_g], the constraints [IC] and [IC*] are equivalent. In what follows we shall work with both constraints, while being explicit about which version of the incentive constraint is under consideration. Having described the state variables and constraints for our recursive formulation, we now describe more carefully the domain for the principal’s problem.

4.3. A Recursive Domain

Given that cash flows for the agent are always non-negative, it is clear that any vector of contingent utilities that can be realized must also be non-negative. But our other constraints impose even more restrictions on the feasible $(v_b, v_g)$. Formally, we say that the tuple $(k, m_i, w_i)_{i=b,g}$ implements $(v_b, v_g)$ if $(k, m_i, w_i)$ satisfies the incentive compatibility, promise keeping, and limited liability constraints.

As noted above, because cash flows are non-negative, the only feasible choices of $w_i$ must lie in $\mathbb{R}^+_2$. However, even with the restriction that $w_i \in \mathbb{R}^+_2$, not every $v \in \mathbb{R}^+_2$ is implementable. Indeed, we show in Lemma A.1 that no $v = (v_b, 0)$ with $v_b > 0$ is implementable.

To serve as the domain for a recursive problem, the set of feasible utilities that can be implemented must have the property that the contingent continuation utilities must also lie in this feasible set. In other words, what is required is a set $V \subset \mathbb{R}^+_2$ such that for any $v \in V$, there exists a collection $(k, m_i, w_i)$ that implements $v$ and has $w_i \in V$ for $i = b, g$. Such a set $V$ always exists — take, for instance, $V = \{0\}$. However, there exists

---

(12) Strictly speaking, our notion of implementability should also include a transversality condition to ensure that promised utilities are actually delivered. For instance, to show formally that contingent utilities can never be negative because of limited liability constraints, one must use a transversality argument. However, because all the contractual variables considered in this paper will actually lie in a compact set, we eschew references to transversality conditions.
a much larger (indeed, a largest), non-trivial set, as described next, with this property, and this set will serve as the domain for our recursive formulation of the principal’s problem.\(^\text{(13)}\)

![Figure 2: Recursive Domain V](image)

**Proposition 4.1.** There exists a largest set \( V \subset \mathbb{R}^2_+ \) such that every \( v \in V \) is implemented by some \((k, m_i, w_i)\) with \( w_i \in V \) for \( i = b, g \). In particular, \( V := \{ (v_b, v_g) \in \mathbb{R}^2_+ : v_g \geq v_b \} \) (see figure 2).

A recursive contract \((k, m_i, w_i)\) therefore conditions current investment and contingent transfers on the current state \((v, s)\) and also determines the continuation state \( w_i \) for \( i = b, g \) beginning in the next period. Of course, one could consider more general sequential contracts. But every such sequential contract is outcome equivalent to a recursive contract indexed by \((v, s)\). Therefore, there is no loss of generality in restricting attention to recursive contracts with state space \( V \times S \).\(^\text{(14)}\)

In what follows, for each \((v, s) \in V \times S\), let

\[
\Gamma(v, s) := \{ (k, m_i, w_i) : (k, m_i, w_i) \text{ implements } v \text{ and } w_i \in V \}
\]

that is, \( \Gamma(v, s) \) denotes the set of feasible contractual variables \((k, m_i, w_i)\) that satisfy \([\text{PK}_b], [\text{PK}_g], [\text{IC}], \text{ and } [\text{LL}]\) and have \( w_i \in V \). Because \( \Gamma(v, s) \) is independent of \( s \), we shall, when there is no cause for confusion, denote this set by \( \Gamma(v) \).

### 4.4. Optimal Contracts

We consider the recursive problem of maximizing firm surplus (which, in this setting, is precisely the social surplus), and use the agent’s contingent utilities (and the previous

\(^{(13)}\) Proposition 4.1 is an analogue of Lemma 2.2 in Fernandes and Phelan (2000). In the terminology of Abreu, Pearce and Stacchetti (1990), \( V \) is self-generating. Indeed, the proof of Proposition 5.2 consists of showing that \( V \) is the (largest) fixed point of an appropriate mapping.

\(^{(14)}\) The proof is similar to the proof of Theorem 2.1 in Fernandes and Phelan (2000) and so is omitted.
period’s state) as state variables for our dynamic program. (Note that because both principal and agent are risk neutral, social surplus and hence firm surplus is invariant to monetary transfers.) Towards this end, we let $Q(v,s)$ denote the surplus of the firm when the previous period’s shock was $s$, and when the agent enters the period with contingent utility $v = (v_b,v_g)$. The principal’s expected utility at the beginning of each period is then $Q(v,s) - E^t[v]$, ie, firm surplus net of expected utility promised to the agent.

In what follows, we shall denote $\partial Q/\partial v_b$ by $Q_b$ and $\partial Q/\partial v_g$ by $Q_g$. An optimal contract is a solution to the firm’s recursive maximization problem. We shall denote an optimal contract by the optimal policy $(k,m_i,w_i)$, $(i = b,g)$, of the firm’s problem. Our first result establishes the existence of the firm’s value function, as well as some of its properties. It also shows that an optimal contract exists by virtue of being the policy function for the firm’s problem.

**Theorem 1.** The firm’s discounted surplus under an optimal contract $(k,m_i,w_i)$ is given by a unique, concave, and continuously differentiable function $Q : V \times S \rightarrow \mathbb{R}$ that satisfies

$$[VF] \quad Q(v,s) = \max_{(k,m_i,w_i)} \left[ -k + p_s(R(k) + \delta Q(w_g,g)) + (1 - p_s)\delta Q(w_b,b) \right]$$

subject to $(k,m_i,w_i) \in \Gamma(v,s)$. The contract $(k,m_i,w_i)$ is continuous in $(v,s)$. Moreover, the following hold.

(a) $Q(0,s) = 0$, $Q_g((v,v),s) = \infty$ and $Q_b((0,v),s) = \infty$ for all $v \geq 0$.

(b) $\sup_{v \in V} Q(v,s) < \infty$ for $s = b,g$.

(c) $Q_g(v,s) \geq 0$ and $Q_b(v,s) \geq 0$ for all $(v,s) \in V \times S$, though $Q_b(v,s)$ is sometimes negative.

(d) $Q(v,g) \geq Q(v,b)$ for all $v \in V$.

(e) $Q(w_g(v,s), g) \geq Q(w_b(v,s), b)$ for all $(v,s) \in V \times S$.

(f) $Q(\cdot,s)$ is supermodular in $v$ for all $s \in S$.

The existence, uniqueness, concavity and differentiability properties of the surplus function $Q$ are standard, as is the continuity of the policy function. By never investing and immediately paying the agent his expected promised utilities, which effectively shuts down the firm, the principal can always reduce the value of the firm to 0, which is then a lower bound on the value of the firm. If the principal could operate the firm herself or equivalently, if she could observe cash flows, she would invest the efficient amount $\tilde{k}_s$ (which solves $p_sR(\tilde{k}_s) = 1$) in each period, and retain all cash flows. The resulting expected (discounted) cash flows clearly provide an upper bound to the value of the firm.
It is easy to see that \( \Gamma(\mathbf{0}, s) = \{(k = 0, m_i = 0, w_i = 0) : i = b, g\} \). In other words, with promised contingent utilities of \((0, 0)\), the principal cannot invest in the current period nor promise utilities in the future. Therefore, \( Q(\mathbf{0}, s) = 0 \).

The values of the partial derivatives in part (a) depend crucially on the assumption that \( R'(0) = \infty \) while part (d) reflects the simple fact that shocks are serially positively correlated and that good shocks are good for the firm.

The main observations with regards to contingent (promised) utilities are in parts (c), (e), and (f). Part (c) says that increasing \( v_g \) is always beneficial to firm surplus, while an increase in the contingent utility \( v_b \) does not always increase firm surplus. If \( v_b \) is very low relative to \( v_g \), there is little effect on \([IC^*]\), but we can raise \( w_b \) as dictated by \([PK_b]\). If \( v_b \) is sufficiently high, then the primary impact of raising \( v_b \) is on tightening \([IC^*]\). Thus, increasing \( v_b \) may be beneficial, but it can also reduce the value of the firm because it constrains feasible levels of output. Part (e) says that success in the present period increases firm surplus in the next period. Finally, part (f) notes that for each \( s \), \( Q(v, s) \) is supermodular in \( v \), i.e., \( v_b \) and \( v_g \) are complementary instruments for the firm. For a fixed \( v_g \), increasing \( v_b \) reduces the downside risk to the firm, because the smaller \( v_b \) is, the lower the size of the firm in the next period. On the other hand, increasing \( v_b \) tightens the incentive constraint \([IC^*]\). As is clear from \([IC^*]\), this second effect is less pronounced when \( v_g \) is higher, so \( Q(v, s) \) is supermodular in \( v \).

As noted above, an optimal contract is a solution to the firm’s recursive maximization problem \([VF]\). Theorem 1 says that an optimal contract \((k, m_i, w_i)\) exists and that it is continuous in \((v, s)\), but says little more. In the next section, we shall study in greater detail the structure of the optimal contract.

We now consider the first best (efficient) firm surplus. This is precisely the case where there are no agency problems and the principal operates the firm. Then, the efficient investment level \( \bar{k}_s \) in each period is the unique \( k \) that solves \( p_k R'(k) = 1 \). This implies that the efficient level of firm surplus in state \( s \) is \( \bar{Q}(s), s = b, g \), where

\[
\bar{Q}(b) = -\bar{k}_b + p_b \left[R(\bar{k}_b) + \delta \bar{Q}(g)\right] + (1 - p_b) \delta \bar{Q}(b)
\]

\[
\bar{Q}(g) = -\bar{k}_g + p_g \left[R(\bar{k}_g) + \delta \bar{Q}(g)\right] + (1 - p_g) \delta \bar{Q}(b)
\]

The efficient investment level \( \bar{k}_s \) solves \( p_k R'(\bar{k}_s) = 1 \). These two equations allow us to explicitly calculate \( \bar{Q}(b) \) and \( \bar{Q}(g) \). What is relevant for us is that \( \bar{Q}(s) \) represents an upper bound for the value of the firm in state \( s \), and that it entails perpetual efficient investment and production.

Let \( E_s := \{v \in V : Q(v, s) = \bar{Q}(s)\} \) denote the efficient or unconstrained sets of contingent utilities. This is the set of contingent utilities that ensure perpetual efficient investment and production. The following proposition describes sets of efficient contin-
Proposition 4.2. For each \(s = b, g\), the sets \(E_s\) defined above are non-empty, closed, convex, and satisfy the following.

(a) There exists \(\bar{v}_s \in E_s\) such that \(v \in E_s\) implies \(v \geq \bar{v}_s\).

(b) For each \(v \in E_s\), \(k(v, s) = \bar{k}_s\) and \(w_i(v, s) \in E_i, i \in \{b, g\}\).

(c) For each \(v \in E_s\), \(v_g - v_b \geq R(\bar{k}_s) + \delta \Delta \max \left[ \frac{\delta v_{bg} - v_b}{\delta(1-p_b)}, \frac{R(k_b)}{1-\delta\Delta} \right]\).

Intuitively, \(\bar{v}_s\) represents the lowest levels of contingent utility that the agent must have in order to obtain perpetual efficient investment in state \(s\). In other words, \(\bar{v}_s\) is the smallest level of contingent utility needed so that financing constraints no longer bind. Indeed, if \(v < \bar{v}_s\), then \(v \notin E_s\). That output must be perpetually efficient follows from the fact that \(\hat{Q}|_{E_s} = \tilde{Q}(s)\); if investment were ever to be inefficient at some date and after some history, then the present discounted value of the firm would be strictly less than the value of the efficient firm, namely \(\tilde{Q}(s)\).

Lemma 7.1 in Section 7 below explicitly describes \(\bar{v}_s\), which in turn gives us the unconstrained sets \(E_s\) explicitly. For now, what is important is the observation that we get efficient investment and firm surplus is thereby maximized if (i) \(v_b\) is sufficiently high, and (ii) \(v_g - v_b\) is sufficiently high. Requirement (ii) is peculiar to our formulation in terms of contingent utilities. This says that the difference \(v_g - v_b\) must also be sufficiently great, because this relaxes the incentive constraint \([IC^*]\), thereby permitting efficient investment. Part (c) of Proposition 4.2 indicates precisely the minimal size that the difference \(v_g - v_b\) needs to achieve, in order to be in \(E_s\). It is clear that (i) is a necessary property, because if contingent utility \(v_b\) is very low, then by \([PK_b]\), the vector...
w_b must be small and, perforce, \( w_{bg} - w_{bb} \) must also be very small. But this violates property (ii) in the subsequent period in the event of a bad shock.

Having described threshold levels of contingent utility, we now turn our attention to the properties of the optimal contract in the early stages of the contract, when the firm is constrained, ie, prior to reaching the threshold levels of contingent utility. But further results, especially properties of the financially unconstrained firm, are sensitive to the degree of persistence. Unfortunately, in our setting, persistence is not completely captured by \( \Delta := p_g - p_b \). The following section provides a categorisation of different degrees of persistence.

### 4.5. Degrees of Persistence

Recall that \( \Delta := p_g - p_b > 0 \) amounts to saying shocks being positively persistent. However, the same value of \( \Delta \) may correspond to different combinations of \((p_b, p_g)\). Indeed, lowering both \( p_g \) and \( p_b \) by the same amount leaves \( \Delta \) unchanged, but can change the structure of the optimal contract, because the probability of success now, and in the future, is uniformly lower. It is easy to see from equations [4.1] and [4.2] that such a change would reduce firm surplus while leaving \( \Delta \) intact.

Therefore, when we speak of levels of persistence, it is important to keep both \( \Delta \) as well as \( p_g \) (or \( p_b \)) in mind. Towards that end, recall that \( \hat{k}_g \) is the unique \( k \) that solves \( p_g R'(k) = 1 \), and is the efficient amount of investment following a good shock. Now define the functions \( \hat{k}(p_g) \) and \( \psi(p_g) \) as

\[
\hat{k}(p_g) := R^{-1} \left( \frac{\delta p_g R(\bar{k}_g)}{1 + \delta p_g} \right) \quad \text{and} \quad \psi(p_g) := p_g - \frac{1}{R'(\hat{k}(p_g))}
\]

We show in Appendix F that \( \psi(p_g) \in [0, p_g] \) for all \( p_g \in [0, 1] \) and that it is well defined. With these definitions in hand, we can partition the space \( \Pi := \{(p_b, p_g) \in [0, 1]^2 : p_b \leq p_g\} \) by defining the sets

- \( B_- := \{(p_b, p_g) \in \Pi : \Delta < \psi(p_g)\} \),
- \( B_0 := \{(p_b, p_g) \in \Pi : \Delta = \psi(p_g)\} \), and
- \( B_+ := \{(p_b, p_g) \in \Pi : \Delta > \psi(p_g)\} \).

The set \( B_+ \) represents the case where persistent level is high, \( B_- \) represents the low persistence case, and \( B_0 \) is the boundary of the two. We will see below that the optimal contract exhibits different properties in these different cases.

### 5. The Constrained Firm

An important consequence of the characterization in Theorem 1 is that the firm surplus \( Q(v, s) \) is the value for a concave programming problem (after an appropriate change
of variables). Thus, an optimal contract is a solution to the relevant first order conditions (which are necessary and sufficient). In the Appendix, we derive these optimal conditions along with the envelope conditions. The agency problem which arises due to private information causes the firm’s financing constraint. The optimal contract determines how the financing constraint evolves over time, as well as its compensation, financing, and investment policies. We shall characterize them in turn in the constrained stage of the firm.

5.1. Initialization

When the firm is initiated, the choice of the contingent utility vector determines the expected payoffs of the principal and the agent throughout the lifetime of the firm. So where the firm initiates depends on the bargaining powers of the two parties. We assume that the principal has the full bargaining power, or alternatively entrepreneurs are competitive. Hence the principal chooses the initial contingent utilities (i.e. initial state of the firm) to maximize his expected profit. Having belief about the probability of success is \( p_s \) (i.e. \( s_0 = s \)), the principal chooses the initial contingent utility vector \( v \) to maximize his expected profit \( Q(v, s) - \mathbb{E}[v] \). We denote this optimal choice as \( v^o_s \).

Since the policies and the evolution of the optimal contract depend on where it starts, it is important to characterize where \( (v^o_s, s) \) locates. Obviously, the principal can choose the initial state so that the firm surplus is maximized. However, that clearly gives too much rent to the agent. The following result shows that the firm is initiated at a state where marginally raising the agent’s promised utilities (either \( v_g \) or \( v_b \) or both) is beneficial to the firm surplus. And this property holds as the optimal contract evolves before reaching the unconstrained stage. It is convenient to denote the set of such states as \( H := \{(v, s) : Q_b(v, s) > 0, Q_g(v, s) > 0 \} \). This set is plotted as the interior of dotted and dark gray shaded areas in figure 3.

**Proposition 5.1.** Given \( s_0 = s \in S \), the optimal contract is initialized at \( (v^o_s, s) \in H \) and stays in this set before reaching the unconstrained stage after a good shock. Moreover, the principal gets higher expected profits if ex-ante information indicates a greater chance of success.

The persistent shock in this model implies that the expected firm surplus depends on the principal’s initial belief \( s_0 \). Accordingly, the principal will optimally choose different starting points of the contract and obtain different expected profits. In the venture capital investments, The dependence of initial expected profits on the surrounding business climate is a subject that has received much attention because of the volatility
in venture capital investments. In this model, because the expected firm surplus is higher when the initial information indicates a greater chance of success, i.e. $s_0 = g$, the principal also gets higher expected profit by choosing the initial state optimally.

5.2. Optimal Compensation

This part provides us some insight into the compensation structure when firm financing is constrained. According to the contract, the principal gets the transfer $m_i$ each period when shock $i$ currently occurs. The rest of the firm cash flow $\theta(i)R(k) - m_i$ is the agent’s compensation. We now characterize the optimal transfer or compensation policy at different states of the set $H$.

**Proposition 5.2.** Let $(k, m_i, w_i)$ be the optimal policy at any state $(v, s) \in H$:

(a) $m_b = 0$;
(b) $v_g < \delta \mathbb{E}[\bar{v}_g]$ implies $R(k) - m_g = 0$;
(c) $v_g \geq \delta \mathbb{E}[\bar{v}_g]$ if and only if $w_g \in E_g$;
(d) There exists an optimal contract with maximal rent such that contingent utilities are never greater than $\bar{v}_s$, and $R(k) - m_g > 0$ if and only if $v_g > \delta \mathbb{E}[\bar{v}_g]$.

When a bad shock occurs, the firm has zero cash flow ($\theta(b) = 0$). But it’s still possible that the agent gets paid, since the limited liability constraint only requires the transfer $m_b$ to be non-positive. Part (a) of Proposition 5.2 says that it is actually optimal to not compensate the agent in this case. When a good shock occurs, the firm has cash flow $R(k)$. Part (b) of the proposition says that the agent does not get paid if the contingent utility $v_g$ is smaller than the value $\delta \mathbb{E}[\bar{v}_g]$. Part (c) of the proposition indicates that $\delta \mathbb{E}[\bar{v}_g]$ can be interpreted as the one-step boundary in the sense that a good shock in the present period will send the firm to the unconstrained stage if $v_g$ is larger than this boundary.

Since both parties are risk-neutral and they share a common discount factor, the optimal transfers $m_i$ and contingent utilities $w_i$ may be indeterminate. Part (d) shows that there exists a useful class of contracts, namely, the maximal rent contracts which involve the earliest possible payment to the agent. In these contracts, contingent utilities never rise above $\bar{v}_s$ and the agent gets cash compensation when $v_g$ is larger than the one-step boundary.

Put differently, the proposition states that all information rents are back loaded — the principal initially keeps all the revenue from production, and only eventually does the agent get a share of the proceeds. By withholding cash payments, the principal

---

(15) See Gompers et al. 2008
(16) Actually $k$ is always unique and $m_i, w_i$ are also unique when $v_g$ is below the one-step boundary.
ensures that whatever utility accrues to the agent stays within the firm, and therefore relaxes the limited liability constraint. This leads to more efficient investments in future periods. If however, the principal makes a cash payment to the agent, then that cash is lost forever. More importantly, Proposition 5.2 shows that the back loading of information rents is a robust property, because it holds regardless of the degree of persistence. In particular, it holds in the iid case; see, for instance, Proposition 3 of Clementi and Hopenhayn (2006).

Another feature of the compensation policy is that the trade-off between incentivizing the agent through cash versus contingent utilities is unrelated to the value of \( v_b \) or the previous shock \( s \). To see why this is the case, note that by the promise keeping constraint \([\text{PK}_g]\) cash compensation amounts to \( v_g - \delta E^g [w_g] \). Moreover, we can obtain the policy \( w_g \) by maximizing \( Q(w_g, g) \) subject to \( \delta E^g [w_g] \leq v_g \), since \( w_g \) only appears in the constraint \([\text{PK}_g]\). In other words, \( w_g \) is only a function of \( v_g \) and so is the cash compensation.

5.3. Evolution of the Optimal Contract

We have established that the states induced by the optimal contract always stay in set \( H \) where increasing either contingent utility \( v_b \) or \( v_g \) is beneficial for the firm surplus. We now characterize how the contingent utility vectors \( w_g \) and \( w_b \) evolve.

**Proposition 5.3.** Let \( w_b, w_g \) be the optimal contingent utilities at any state \((v, s) \in H\).

(a) \( w_b < \tilde{v}_b \), and \( w_g \leq \tilde{v}_g \);

(b) \( w_{gg} > v_g \), and \( w_{gg}, w_{gb} \) both increases in \( v_g \);

(c) \( w_{bb} < v_b / \delta \), and if \( v_g \geq av_b \) at \( s = b \), then \( w_{bg} \geq aw_{bb}, w_{bb} \leq v_b \) where \( a = \tilde{v}_{bg} / \tilde{v}_{bb} \).

The contingent utilities are always below their thresholds when firm financing is constrained. When good shock occurs, the new contingent utility vector \( w_g \) will be \( \tilde{v}_g \) when \( v_g \) is above the one-step boundary. Otherwise, the agent does not enjoy any cash compensation and by the promise keeping constraint \( \delta E^g [w_g] = v_g \). Because \( w_{gg} \) is greater than \( w_{gb} \), the utility contingent on good shocks always becomes larger, i.e. \( w_{gg} > v_g \). As pointed out earlier, the policy \( w_g \) is only a function of \( v_g \), which implies that \( w_g \) always locates on the same curve in \( V \) as illustrated in the right panel of figure 4. Moreover, the supermodularity of \( Q \) implies that \( w_{gb} \) and \( w_{gg} \) are complements in determining firm surplus. Thus, \( w_{gg} \) and \( w_{gb} \) both become larger as \( v_g \) increases. So after good shocks, the new contingent utility vector will keep increasing until reaches the unconstrained threshold \( \tilde{v}_g \).

When bad shock occurs, because \( w_{bb} \) is greater than \( w_{gb} \), the promise keeping constraint \([\text{PK}_b]\) implies \( w_{bb} < v_b / \delta \). Moreover, part (c) of proposition 5.3 shows that if
the firm enters the region where \( v_g \geq a v_b \) after a bad shock (the area above the dashed line in the left panel of figure 4), the policy \( w_b \) always satisfies \( w_{bb} \leq v_b \) and the contract will stay in that region until another good shock occurs. In other words, though \( w_{bb} \) might be larger than \( v_b \) after some history (when \((v, b)\) locates below the dashed line), it cannot always be the case. Otherwise, \( w_{bb} \) would go above the threshold level \( \bar{v}_{bb} \) and violate part (a) of the proposition. So if the firm keeps experiencing bad shocks, it will enter the region where \( v_g \geq a v_b \) and evolve towards the origin.

As in the iid case, the optimal contract exhibits strong path dependence in that its evolution depends on the sequence of shocks. However, there is a stronger form of path dependence in the present setting, because the evolution of the contract now depends not only on the continuation utilities but also on the previous shock \( s \). It’s very clear from figure 4 that the contract evolves in either the left or the right panel with different features contingent on bad or good shock. Also, the agent prefers the good shocks come sooner. In that case, the firm becomes financially unconstrained in relatively shorter time.

We have established in Proposition 5.1 that it’s only possible for the constrained firm to reach the efficient sets after a good shock. This is also true in the iid case, as noted by Clementi and Hopenhayn (2006). However, if \( \Delta \) is sufficiently large, and this is unique to the persistent case, it necessarily takes two consecutive good production shocks for the firm to become unconstrained.

(16) In this simulation \( p_b = 0.2, p_g = 0.6 \). The left panel is the growth path of contingent utilities. In the right panel, the green and red dots are investment levels, and the green bars are compensations to the agent.
Proposition 5.4. In the high persistence case ($\Delta > \psi(p_g)$), whenever a bad shock occurs the firm needs to experience at least two consecutive good shocks before becoming unconstrained.

Proposition 5.4 shows another unique feature of the contract regarding path dependence in the persistent environment. The result makes intuitive sense because if the shocks are sufficiently persistent, then after a bad shock, the agent’s contingent utility is sufficiently low that it cannot reach $E_g$ in one step. In this sense, it is potentially takes longer time for the firm in our model to become unconstrained than in iid models.

5.4. Optimal Financing and Investment

The presence of private information implies that investment may well be inefficient, since investment size determines the static information rent which adds to the cost of financing. In general, the degree of underinvestment depends on the agent’s contingent utilities $v$ and the previous shock $s$. In figure 3, the dark gray shaded region indicates underinvestment and the dotted region indicates efficient investment. The boundary of these regions are two upward sloping curves in the left and the right panels.

Lemma 5.5. Given any state $(v, s) \in V \times S$, the optimal investment $k(v, g) \geq k(v, b)$. Moreover, $k(v, s)$ increases in $v_g$ and decreases in $v_b$.

Lemma 5.5 shows that the investment policy exhibits monotonicity over the state variables. Since the feasible policies under state $(v, b)$ and state $(v, g)$ are the same but the latter state implies higher probability of getting positive cash flow, investment is higher when last period had a good shock. Moreover, larger $v_g$ and smaller $v_b$ implies higher investment, because the change of contingent utilities in these directions relaxes the incentive constraint.

It’s easy to see that for any given $v_b$ and $s$, if we set $v_g = v_b$, then [IC$^*$] implies the only feasible investment is $k = 0$, since $w_{bg} - w_{bb}$ on the right hand side of [IC$^*$] is nonnegative. But if we set $v_g$ to be sufficiently large, then [IC$^*$] will be slack even at the efficient investment level $\bar{k}_s$. This is because the largest possible value of $w_{bg} - w_{bb}$ (which appears on the right hand side of [IC$^*$]) is constrained by $v_b$ through the promise keeping constraint [PK$^*_b$]. Decrease in $v_b$ will enlarge the left hand side of [IC$^*$] and shrink the largest possible value of $w_{bg} - w_{bb}$. Both of these channels lead to higher cash flow and investment.

Next we examine properties of the investment policy along the optimal states of the contract. A key feature of the iid models is that firm investment is always inefficient in the constrained stage. For instance, in Clementi and Hopenhayn 2006, the investment
size is tied to the variation of the agent’s continuation utilities. Because these variations are costly (surplus function is concave), optimal investment will be kept below the first best level. The following result shows that when shocks are close to iid, this investment pattern still holds. However, when shocks are far from iid, investment can be temporarily efficient though the firm is not fully unconstrained in the sense that some future investment will still be inefficient.

**Proposition 5.6.** Given any state \((v, s) \in H\), suppose shock \(i \in S\) occurs and the optimal contract specifies the new state as \(w_i(v, s)\) and the investment at the new state as \(k(w_i, i)\).

(a) When \(i = g\) and \(w_g \notin E_g\), investment is always inefficient, i.e. \(k(w_g, g) < \bar{k}_g\).

(b) When \(i = b\), investment can be efficient or not depending on persistence level and history. If persistence is low (\(\Delta \leq \psi(p_g)\)), then \(k(w_b, b) < \bar{k}_b\). If persistence is high (\(\Delta > \psi(p_g)\)), then there exists a region around and below \(\bar{v}_b\) where \(w_b\) possibly locates and where \(k(w_b, b) = \bar{k}_b\).

Proposition 5.6 (b) shows that in the high persistent case, the optimal investment can be temporarily efficient. In the left panel of figure 3, we can see that \(\bar{v}_b\) is immediately surrounded by efficient investment states. When the contingent utility vector evolves to that region, investment will be \(\bar{k}_b\). This is possible because policy continuity implies that if \((v, g)\) is sufficiently close to \((\bar{v}_g, g)\) after some history, then a bad shock will result in a new state \((w_b, b)\) very close to \((\bar{v}_b, b)\). In that case, we show in the Appendix that firm investment will stay at \(\bar{k}_b\) until a good shock occurs. After that, investment will become inefficient (less than \(\bar{k}_g\)) again according to part (a) of the proposition. This feature clearly distinguishes our implication from the iid case where efficient investment at any date implies efficiency thereafter.

To understand this result, let us consider the agent’s incentive when the contract evolves to the vicinity of \((\bar{v}_b, b)\) after some history. If cash flow \(R(k)\) realizes today and the agent lies, he can divert the cash flow and obtain expected utility (from tomorrow onward) \(\mathbb{E}^x[w_b] \simeq \mathbb{E}^x[\bar{v}_b]\). If the agent reports truthfully, he gets zero today and expected utility \(\mathbb{E}^x[w_g] \simeq \mathbb{E}^x[\bar{v}_g]\). When persistence is high, the agent strictly prefers truth-telling even at the efficient investment level. First, the static information rent \(R(k)\) is very small because \(\bar{k}_b\) is small even though it is efficient. Second, since efficient investment and firm surplus are very volatile with high persistence, \(\mathbb{E}^x[\bar{v}_g]\) is much larger than \(\mathbb{E}^x[\bar{v}_b]\). Hence investing \(\bar{k}_b\) satisfies incentive compatibility and is optimal.

Of course, with persistence, one would suspect, and our numerical calculation show this to be true for some parameter values, that increased persistence leads to increased volatility in investment. This is natural because the random shocks is an additional channel of inducing volatility. Moreover, with high persistence, because the
states are highly correlated, investment is also highly correlated across periods. While these observations are intuitive, these properties are nevertheless difficult to establish analytically in our discrete-time framework.

6. Long-Run Convergence

In the iid model of Clementi and Hopenhayn (2006) where ex-ante promised utility is the state variable, the long-run properties of the contract are uncovered using the observation that the derivative of the surplus function is a martingale. This observation was first made (also in an iid setting) by Thomas and Worrall (1990). As we are working with contingent (interim) utilities, the martingale relevant to us is the directional derivative of $Q$ at $v$ in the direction $(1,1)$. Theorem 1 ensures that $Q$ is differentiable everywhere, so the directional derivative is $D_{(1,1)}Q(v,s) = Q_b(v,s) + Q_g(v,s)$.

**Theorem 2.** Starting at $(v,s) \in V \times S$, an optimal contract induces a process $D_{(1,1)}Q$ that is a non-negative martingale. The martingale $D_{(1,1)}Q$ converges to 0 in finite time almost surely. In a maximal rent contract, contingent utility converges to $\bar{v}_g$ in finite time almost surely, and then cycles between $\bar{v}_b$ and $\bar{v}_g$, with transitions according to the Markov process on $S$.

To understand why $D_{(1,1)}Q$ must be a martingale, let us consider an increase in $v$ by $(\epsilon, \epsilon)$ for a sufficiently small $\epsilon > 0$. One way to accomplish this is by increasing each of $w_b$ and $w_g$ by $(\epsilon/\delta, \epsilon/\delta)$. It is easy to see that such a change to continuation contingent utilities is incentive compatible and also satisfies promise keeping. The resulting increase in firm surplus is approximately $(1 - p_s)D_{(1,1)}Q(w_b,b) + p_sD_{(1,1)}Q(w_g,g)$. An envelope argument shows that these changes in $w_b, w_g$ are locally optimal, and so the above increase in firm surplus must be $D_{(1,1)}Q(v,s)$, ie, $D_{(1,1)}Q$ is a martingale.

It follows from Theorem 1 that the martingale $D_{(1,1)}Q$ is nonnegative. Doob’s Martingale Convergence Theorem — see, for instance, Theorem 2 on p517 of Shiryaev (1995) — ensures that $D_{(1,1)}Q$ converges to a non-negative and integrable random variable. The proof of Theorem 2 shows that along almost every path, if the value of $D_{(1,1)}Q$ is always positive, then it must vary after ‘good-good’ and ‘good-bad’ shock pairs. (In other words, the martingale ‘splits’ after good shocks. Contrast this with the iid case, where the martingale splits whenever it is non-zero.) Because these shock pairs occur infinitely often almost surely, the martingale must converge to 0 almost along every path. Proposition 4.2 now implies that the induced contingent utilities will eventually locate in the unconstrained set $E_s$.

While the conclusion of (martingale) convergence is also drawn in the iid case, there is an important difference. In the iid case, if the value function is strictly concave,
there is a one-to-one relationship between the derivative and ex ante promised utility. In the case of persistence, knowing \( D_{(1,1)} Q(v, s) = c \) for some \( c > 0 \) does not pin down \( v \). Instead, it only gives us a set of points (typically a curve in \( V \)) where the directional derivative is \( c \). This makes the convergence argument more subtle.

Theorem 2 also states that the firm becomes unconstrained in finite time almost surely. In other words, the value of \( D_{(1,1)} Q \) cannot stay positive but very close to zero for infinite amounts of time. If that happens, contingent utility vector \( w_g \) must necessarily be in a small neighborhood below \( \bar{v}_g \) after a good shock. An additional good shock will then send continuation contingent utility vector to \( \bar{v}_g \) where the martingale is zero. Thus, all convergence occurs in finite time almost surely.

In the maximal rent contract, the contingent utility transitions initially to \( \bar{v}_g \) in finite time, and then cycles between \( \bar{v}_g \) and \( \bar{v}_b \). This property is another major difference between the iid case and the persistence case. In the former, \( \bar{v}_b = \bar{v}_g \), and so no firm dynamics exist in the long run. However, in the persistent case, the existence of non-trivial dynamics even after the firm is unconstrained leads to interesting long-run policy implications that we now examine.

### 7. The Unconstrained Firm

A unconstrained firm is one that has efficient financing and investment and pays dividends to its equity holders. The structure of the unconstrained firm is stark in the iid case: the agent is the residual claimant of the firm and firm investment and capital structure is constant over time. These properties are seldom seen in practice. As we shall see below, when private information displays persistence, none of these conclusions hold.

In the unconstrained firm, it is easy to see that the investment is \( \bar{k}_s \) and the contingent utility vector is \( \bar{v}_s \) depends on the Markov state \( s \). The optimal transfers and the explicit values of \( \bar{v}_s \) are pinned down by the binding constraints. Obviously, the promise keeping conditions \([PK_s]\) and \([PK_g]\) must hold. But, as we shall see below, whether or not the incentive compatibility \([IC]\) and limited liability \([LL]\) conditions bind depends on the degree of persistence.

As we show in the Appendix, the incentive compatibility \([IC]\) constraint always binds when persistence level is low, ie, \((p_b, p_g) \in B_{-}\). (In particular, as in Clementi and Hopenhayn 2006, \([IC]\) holds as an equality in the iid case.) However, when persistence is high, so \((p_b, p_g) \in B_{+}\), \([IC]\) is slack and the limited liability constraint holds as an equality. Given the binding constraints at various values of \((p_b, p_g)\), we are now in a position to explicitly calculate the threshold contingent utility vector.

**Lemma 7.1.** The threshold levels of contingent utility are as follows.
(a) For \((p_b, p_g) \in B_+\),

\[
\tilde{v}_b = (\tilde{v}_{bb}, \tilde{v}_{bg}) = \left( \frac{\delta p_b \tilde{v}_{bg}}{1 - \delta (1 - p_b)}, \frac{\delta p_g [1 - \delta(1 - p_b)] R(\tilde{k}_g)}{(1 + \delta p_g)(1 - \delta)(1 - \delta \Delta)} \right),
\]

\[
\tilde{v}_g = (\tilde{v}_{gb}, \tilde{v}_{gg}) = \left( \frac{\delta p_b \tilde{v}_{bg}}{1 - \delta (1 - p_b)}, \tilde{v}_{bg} + \frac{R(\tilde{k}_g)}{1 + \delta p_g} \right).
\]

(b) For \((p_b, p_g) \in B_-\),

\[
\tilde{v}_b = (\tilde{v}_{bb}, \tilde{v}_{bg}) = \left( \frac{\delta p_b \tilde{v}_{bg}}{1 - \delta (1 - p_b)}, \frac{[1 - \delta(1 - p_b)] R(\tilde{k}_b)}{(1 - \delta)(1 - \delta \Delta)} \right),
\]

\[
\tilde{v}_g = (\tilde{v}_{gb}, \tilde{v}_{gg}) = \left( \frac{\delta p_b \tilde{v}_{bg}}{1 - \delta (1 - p_b)}, \frac{[p_g - \delta \Delta] \tilde{v}_{bg} + R(\tilde{k}_g)}{1 - \delta (1 - p_b)} \right).
\]

It is clear these expressions rely on the degree of persistence. But the common feature shown in Lemma 7.1 is that \(\tilde{v}_b \neq \tilde{v}_g\), as long as persistence is positive. In particular, the result shows that \(\tilde{v}_{bb} = \tilde{v}_{gb}\) and \(\tilde{v}_{bg} < \tilde{v}_{gg}\). If, however, shocks are iid \((\Delta = 0)\), the following contingent utilities are

\[
\tilde{v}^{\text{id}} = \left( \frac{\delta p R(\tilde{k})}{1 - \delta}, \frac{(1 - \delta + \delta p) R(\tilde{k})}{1 - \delta} \right)
\]

which are constant over time and independent of the previous shock. Notice that the ex ante expected utility from the interim (contingent) utility vector is precisely \(\tilde{v}^{\text{id}} = \mathbb{E}^p[\tilde{v}^{\text{id}}] = pR(\tilde{k})/(1 - \delta)\), just as in Clementi and Hopenhayn (2006).

Applying the binding constraints at various values of \((p_b, p_g)\), we can also obtain explicit expressions of the optimal policies in the unconstrained firm.

**Theorem 3.** Suppose the firm is in the unconstrained stage, the maximal rent contract \((k, m_b, w)\) takes the form:

- \(k(\tilde{v}_s, s) = \tilde{k}_s\)
- \(w_g(\tilde{v}_s, s) = \tilde{v}_g, w_b(\tilde{v}_s, s) = \tilde{v}_b\)
- \(m_{sb} = 0\)
- \(m_{bg} = R(\tilde{k}_b), m_{gg} = \frac{\delta p_g R(\tilde{k}_g)}{1 + \delta p_g} < R(\tilde{k}_g), \quad \text{if } (p_b, p_g) \in B_+\)
- \(m_{bg} = m_{gg} = \delta p_g [R(\tilde{k}_g) - R(\tilde{k}_b)] < R(\tilde{k}_b), \quad \text{if } (p_b, p_g) \in B_-\)

where \(m_{sb} = m_b(\tilde{v}_s, s)\) and \(m_{sg} = m_g(\tilde{v}_s, s)\).

Theorem 3 shows the striking features of the transfer or compensation policy in the unconstrained firm. If the shocks are iid, the transfer to the principal is always zero and the agent gets the entire cash flow, which makes him the residual claimant to the firm. However, for any degree of positive persistence \((\Delta > 0)\), the principal always gets some positive part of the cash flow even though the agent can divert everything, and
the agent is no longer the residual claimant. Thus, the theorem highlights the sensitivity of the qualitative features of the contract to the degree of persistence. Figure 5 clearly illustrates these features by showing cash flow fractions obtained by the principal and by the agent respectively for different persistent levels.

\[ \text{Cash Flow} \]

\[ \text{Payment to Financier} \]

\[ \text{Payment to Agent} \]

**Figure 5: Cash Flow, Transfer and Payment in Unconstrained Firm**

The fact that the agent is no longer the residual claimant is reflected in the observation that whenever \( \Delta > 0 \), \( v_g > v_b \). The intuition behind this is when shocks are persistent, if the agent misreports a good shock to be a bad one, investment in the subsequent period will be reduced to \( \bar{k}_b \) instead of \( \bar{k}_g \). But payments to the agent entirely reflect his information rent, and a lower investment reduces the agent’s future information rent, making him less willing to lie today. Hence, the principal can carve out a transfer even after a good shock by (rationally) threatening, and committing, to reduce investment, and thereby information rent, if a bad shock is ever reported.

Indeed, if persistence is sufficiently large, so \( (p_b, p_g) \in B_+ \), then contingent on the previous shock being bad, the agent transfers all the cash flow to the principal, while retaining a positive (but strictly less than one) fraction only if the previous shock was good. In other words, when persistence is high, the agent is compensated only when two good shocks occur in a row. This is reflected in the right panel of Figure 5 where the red line becomes zero for \( \Delta > 0.4 \). The high persistence case highlights the agent’s compensation is convex in firm performance, because he obtains larger fraction of cash flow when it is \( R(\bar{k}_g) \) (after ‘good-good’ shock pair) than when it is \( R(\bar{k}_b) \) (after ‘bad-good’ shock pair).

8. **Implementation**

In this section, we examine how to implement the optimal contract using a set of standard financial securities. Since the financial securities can be held by widely dispersed investors or intermediaries, it is not necessary to rely on a single principal to execute the contract. At time 0, the firm issues equity and obtains a credit line with cash flow
covenant from the investors. The firm raises its initial capital from issuing equity to the investors and from drawing on the credit line. The agent is compensated by holding equity and stock options.

**Equity**: equity holders receive dividend payments made by the firm. Dividends are paid by the available cash or credit and at the discretion of the agent.

**Stock options**: the agent has the option of buying the firm’s stock at the specified strike price and then selling it back to the firm at the market value. The market value of the firm’s stock is defined as the present value of expected cash flows.

**Credit line**: revolving credits provided to the firm with cash flow covenant. The firm complies with the credit line covenant whenever its cash flow is positive, and fails to do so otherwise. The credit line limits are contingent on compliance with the covenant. Any balance on the credit account is charged an interest rate $r = 1/\delta - 1$ each period. The agent controls how to use the available credits at any time.

### 8.1. Equity and Stock Options

Recall in the optimal contract, the agent is compensated cash when the firm is unconstrained. More importantly, the compensation policy has two distinct features when deviating from the iid case. First, the agent’s compensation accounts for a larger fraction of firm cash flow after good-good shock (when cash flow is $R(\bar{k}_g)$) than after bad-good shock (when cash flow is $R(\bar{k}_b)$). This means the compensation is convex in firm performance. But when shocks are iid ($\Delta = 0$), the agent always gets a constant fraction of cash flow (100%), meaning compensation is linear in firm performance. Second, the principal gets positive amount of cash flows in the unconstrained stage when shocks are persistent ($\Delta > 0$), but gets nothing when shocks are iid. So compared with the iid case, the agent holds a smaller fraction of firm value and the principal holds a larger fraction.

We now show that the agent’s compensation can be implemented by equity and stock options.

In this part, we assume that the firm has paid off its credit balance and started payout. In the unconstrained stage, since the compensating balance pays interest $\bar{k}_s$ each period, the total cash flow is $R(\bar{k}_s)$ if a good shock occurs today and is 0 if a bad shock occurs today. Accordingly, we define the firm value as the present value of its cash flows. Hence the firm values are contingent on the current and last periods’ shocks and are denoted as $Z_{sb}, Z_{sg}$. Their values are jointly determined by:

\[
\begin{align*}
Z_{bg} & = R(\bar{k}_b) + \delta[p_g Z_{gg} + (1-p_g) Z_{gb}] \\
Z_{gg} & = R(\bar{k}_g) + \delta[p_g Z_{gg} + (1-p_g) Z_{gb}] \\
Z_{bb} & = Z_{gb} = \delta[p_b Z_{bg} + (1-p_b) Z_{bb}]
\end{align*}
\]
The cum-dividend stock price is exactly $Z_{sb}, Z_{sg}$ if we normalize the total equity share to be one. We provide their explicit values in the Appendix. The agent’s compensation consists of two parts. First, he is granted an equity share of $\lambda \in [0, 1]$. The remaining $1 - \lambda$ share is held by investors. Second, at the beginning of each period the agent is granted an option to purchase the firm stock at price $K$ which expires at the end of the period. If the option is exercised, then the firm will issue one additional share of stock to the agent and buys it back at the current stock price. So the option pays out $\max\{Z - K, 0\}$. The rest of firm cash flow will be issued as dividend to both the agent and investors according to their equity holdings. The following result shows that by appropriately designing the strike price $K$ and the equity share $\lambda$, we can compensate the agent the exact amount as in the optimal contract.

**Proposition 8.1.** The agent gets the same compensation as in the optimal contract if he is granted: i) an equity share of $\lambda = \frac{R(k_b) - \bar{m}_{bg}}{R(k_b)}$; ii) stock options with strike price $K = Z_{bg} + \bar{m}_{gg} - \bar{m}_{bg}$.

Figure 6 shows how the stock prices and option payoffs change on a simulated path when the firm is unconstrained. We can see that the option is in the money only after “good-good” shocks. This is because the defined strike price in Proposition 8.1 satisfies $K > Z_{bg} > Z_{sb}$. Moreover, when shocks are iid, the transfers and stock prices satisfy $\bar{m}_{gg} = \bar{m}_{bg}$ and $Z_{bg} = Z_{gg}$. So in the iid case, the option is always out of money and is never exercised, implying stock options have no role in compensation.

**Corollary 8.2.** The agent’s equity share $\lambda$ is decreasing in persistence level $\Delta$, and strictly decreasing when $p \in B_-$.

(17) In the left panel, $p_b = 0.2, p_g = 0.48$. In the right panel, $p_b = 0.2, p_g = 0.8$. 

29
Corollary 8.2 highlights that the agent does not hold all the firm’s equity though he can divert the whole cash flow. Figure 7 compares option and equity compensations as persistent level varies. It plots these two sources of compensation when a good shock occurs today contingent on yesterday’s shock being good or bad. As persistence increases, the option compensation accounts for a larger fraction of the agent’s total pay. When the persistence level is sufficiently high (i.e., \( p \in B_+ \)), the agent is only paid by stock options and holds zero equity.

The above results also highlight the sensitivity of the optimal compensation structure implied by the iid models. As long as \( \Delta > 0 \), the option compensation always plays a role, and the agent is not the residual claimant of the firm.

### 8.2. Contingent Credit Line

To implement the optimal contract, we first find an alternative way to represent the state variables \((v, s)\). The binary variable \( s \) simply indicates whether the firm complies with its covenant or not. If shocks are iid, then the only state variable in the contract will be the ex-ante continuation utility. With a fixed credit line limit, the credit account balance corresponds one to one with the continuation utility. However, since the persistent model has a pair of contingent utilities \( v_b, v_g \) as its state variables, it’s impossible to represent them using only the credit balance.

One way to deal with this issue is to let the credit limits potentially vary with the firm’s performance history. In this implementation, investors immediately adjust the credit limit to a low level \( C^d > 0 \) when the firm fails to comply with its covenant, and to a high level \( C^u > C^d \) otherwise. Using the credit account balance and these time-varying limits we can implement the optimal contract.

Suppose staring certain period the firm has credit balance \( M \). The agent will draw

---

(18) In both panels, \( p_b \) is fixed at 0.2 and \( p_g \) increases from 0.2 to 0.9.
down $k$ from the credit line to invest. If the firm fails to comply with the covenant, it will have available credit $C^d - (M + k)$. Accordingly, if the firm complies with the covenant this period, then it can repay the credit balance using its cash flow $R(k)$ and ends up with available credit $C^u - [M + k - R(k)]$. Since the agent can always draw down and divert all the available credits, his continuation payoffs are

\[ v_b = C^d - [M + k(v, s)] \]
\[ v_g = C^u - [M + k(v, s) - R(k(v, s))] \]

Obviously we can not uniquely determine $C^d, C^u$ and $M$ for a pair of payoffs $v_b, v_g$. So in this implementation, we fix the low credit limit $C^d = \bar{v}_b$ after any performance history. Then \[8.4\] \[8.5\] implies that there is a unique pair of $C^u, M$ for a given pair of $v_b, v_g$. In other words, it is equivalent to use either $(C^u, M, s)$ or $(v_b, v_g, s)$ as state variables. And contingent on whether the firm complies with the covenant or not today, the next period’s state variables are either $(C^u, M, g)$ or $(C^u, M, b)$.

We can simplify \[8.4\] \[8.5\] to get the relation between these two credit limits as $C^u = C^d + v_g - v_b - R(k)$. The wedge between $C^u$ and $C^d$ arises from the dynamic information rent which is endogenously varying over time. Since we keep $C^d$ constant, $C^u$ has to be time-varying. If we eliminate the persistence, then the dynamic information rent will always be zero and these two limits will collapse to one constant number. In the model, if we set $\Delta = 0$ then $[1C^*]$ always holds as equality in the optimal contract. Hence $v_g - v_b - R(k) = 0$ and $C^u = C^d$. In that case, this implementation becomes a similar mechanism as in DeMarzo and Fishman (2007b).

The credit balance evolves according to withdraws and repayments. The agent draws down credit to make investment and repays credit balance when positive return occurs. The compensating balance also pays interest that deducts the credit balance. If the firm fails to comply with covenant this period, then starting next period, its credit balance will be

\[ M_b = (1 + r)(M + k) - a_b \]

If the firm complies with the covenant this period, then starting next period its credit balance will be

\[ M_g = (1 + r)[M + k - R(k)] - a_g \]

In order to implement the optimal contract, the credit limit and balance have to evolve in a way that available credits are matched with contingent utilities from next period on. In particular, given the current state $(C^u, M, s)$ and shock $i \in \{b, g\}$, the next
period’s credit balance $M_i$ satisfies:

$$w_{ib}(C^u, M, s) = \bar{v}_b - (M_i + k_i)$$

The equations [8.6] to [8.8] that characterize the credit balance evolution also uniquely pins down the compensating balance payments $a_b, a_g$.

**Lemma 8.3.** Suppose the agent truthfully reports cash flow and does not pay out until credit balance is paid off. Then continuation utilities in the optimal contract evolve the same way as the available credits under the following mechanism:

$$a_i = r\bar{v}_b - x_i(C^u, M, s)$$

$$C^u_i = M_i - R(k_i) + w_{ig}(C^u, M, s)$$

where $x_b = p_b(w_{bg} - w_{bb}) - k_b$, and $x_g = p_g(w_{gg} - w_{gb}) - \Delta(w_{bg} - w_{bb}) - k_g$.

The left panel in Figure 8 shows how the credit limit evolves on a simulated path. If the covenant is maintained, the credit limit will be the level specified in the upper scatter plot. If the covenant is violated, the credit limit will immediately drop to the lower line (1.462). Moreover, we can see that after bad shock or covenant violation expected credit limit next period also drops.

**Theorem 4.** The optimal contract can be implemented by a combination of equity share, stock options, and contingent credit line:

(i) The agent holds the fraction $\lambda$ of the outstanding equity, and one period stock options with stick price $K$.

(ii) The credit line limit is set as $\bar{v}_b$ if cash flow covenant is violated or $C^L$ if the covenant is maintained. Depending today’s shock $i$, the compensating balance pays interest $D_i$ tomorrow and upper limit is set as $C^L_i$ tomorrow.

(19) $p_b = .4$ and $p_g = .6$. 

---

**Figure 8:** Contingent Credit Line and Balance

The left panel in Figure 8 shows how the credit limit evolves on a simulated path. If the covenant is maintained, the credit limit will be the level specified in the upper scatter plot. If the covenant is violated, the credit limit will immediately drop to the lower line (1.462). Moreover, we can see that after bad shock or covenant violation expected credit limit next period also drops.
It is incentive compatible for the agent to truthfully report cash flow and use it to pay credit balance before initiating payout. Once the credit balance is fully repaid, cash flows are issued as option payoff and dividends.

9. Conclusion

In this paper, we explore the question of how a firm is financed when its cash flows are privately observed by an agent who operates the firm. The new ingredient of our model is that firm cash flows are subject to persistent shocks. Many studies have already shown that adopting this assumption is crucial if we want to take the dynamic agency models seriously and quantify them. However, it is a theoretical challenge to consider persistent shock in this class of models, because private information of current cash flow means the agent is also better informed of the firm’s future. We formulate the mechanism design as a recursive problem in which the agent’s continuation utility contingent on today’s and tomorrow’s reports (which is a vector) is a state variable, along with the beliefs about high revenue in the current period. With this recursive approach, we can analytically characterize the firm’s financing, investment and compensation policies and show that they depend crucially on the degree of persistence.

We show that when the firm is initiated in the optimal contract, it faces financing constraints and hence its investment is inefficient. The incentive scheme involves the agent being compensated exclusively through adjustments to his continuation utility until reaching some thresholds. After that the firm no longer faces financing constraints, investments are forever optimal, and the agent may get cash payments. Apart from the iid case, investment may be temporarily efficient (after a bad shock) before the firm becomes fully unconstrained. Moreover, the firm may need to receive a sequence of good shocks in a row in order to reach the unconstrained stage, which means the firm may be stuck in the constrained stage for much longer time than in the iid case.

By identifying the right martingale, we show that the firm converges to the unconstrained stage in finite periods with probability one. And when it becomes unconstrained, its investments cycle between the efficient levels according to its shocks. The agent gets cash payments but less than what he can divert. In the high persistent case, the agent may not even get any pay after good shocks. This also implies the principal holds more stake in the firm than the iid case.

To better understand the distinct features of the persistent model, we show an implementation using credit line with contingent limits, stock options and equity. After bad shocks, the credit line limit will drop to a lower lever immediately and the future expected credit line limits will also drop. Moreover, the agent gets paid by holding
equity and stock options in the implementation. As the persistence increases, option payments accounts for larger fraction of the agent’s compensation.

Using the new recursive approach, we can better characterize firm policies in the persistent environment. It is easy to see that persistent private information has rich implications and potentially helps the literature better understand issues in firm financing, investment, compensation, and growth.
Appendices

A. Proofs from Section 4

A.1. The Recursive Domain

We begin with a demonstration that not all points \( v \in \mathbb{R}^2_+ \) are implementable.

Lemma A.1. Let \( v = (v_b, 0) \) where \( v_b > 0 \). Then, such a \( v \) is not implementable with \( w \in \mathbb{R}^2_+ \).

Proof. Notice \([PK_g]\) requires that

\[
0 = R(k) - m_g + \delta \left[ (1 - p_g)w_{gb} + p_gw_{gg} \right]
\]

By \([LL]\), we know that \( R(k) - m_g \geq 0 \), and by assumption, \( w_g \in \mathbb{R}^2_+ \), which implies \( (1 - p_g)w_{gb} + p_gw_{gg} \geq 0 \). Therefore, it must be that \( R(k) = m_g \), and \( w_g = (0, 0) \). Now notice that by \([IC]\), we obtain

\[
0 \geq R(k) - m_b + \delta \left[ (1 - p_g)w_{bb} + p_gw_{bg} \right]
\]

As noted above, \( w_b \in \mathbb{R}^2_+ \), and \( R(k) \geq 0 \). By \([LL]\), we also have \( m_b \leq 0 \), which implies \( 0 \geq R(k) - m_b + \delta \left[ (1 - p_g)w_{bb} + p_gw_{bg} \right] \geq 0 \), ie, \( R(k) = m_b = k = 0 \) and \( w_b = (0, 0) \). Therefore, by \([PK_g]\), we must have \( v_b = -m_b + \delta \left[ (1 - p_b)w_{bb} + p_bw_{bg} \right] = 0 \). But this contradicts our assumption that \( v_b > 0 \). Thus, \((v_b, 0)\) with \( v_b > 0 \) is not implementable, or equivalently, is infeasible.

We now present the proof of Proposition 4.1. It is easy to see that the set of contingent utilities \( v \in \mathbb{R}^2_+ \) that can be implemented by \((k, m_i, w_i)\) with \( w_i \in \mathbb{R}^2_+ \) is a closed and convex cone. Therefore, in our search for a suitable domain, it suffices to restrict attention to closed and convex cones.

Recall (from section 4.3) that \((k, m_i, w_i)_{i=b,g}\) implements \((v_b, v_g)\) if \((k, m_i, w_i)\) satisfies the incentive compatibility, promise keeping, and limited liability constraints. Let \( \mathcal{K} \) denote the space of closed and convex cones that are subsets of \( \mathbb{R}^2_+ \). Following Abreu, Pearce and Stacchetti (1990), we define the operator \( \Phi : \mathcal{K} \to \mathcal{K} \) as follows: for \( C \in \mathcal{K} \), let

\[
\Phi(C) := \{ v \in \mathbb{R}^2_+ : \exists (k, m_i, w_i) \text{ that implements } v \text{ and has } w_i \in C, i = b, g \}
\]

In other words, \( \Phi(C) \) consists of all implementable contingent utilities \( v \) wherein the continuation contingent utilities \( w_i \) lie in the set \( C \). Clearly, any recursive program must
only consider contingent utilities \( v \) that lie in a set \( C \) such that \( C \) is a fixed point of \( \Phi \), so that all present contingent utilities as well as future continuation contingent utilities lie in the same set. Essentially, Proposition 4.1 delineates such a set.

**Proof of Proposition 4.1.** It is easy to see that \( \Phi \) is well defined, that is, \( \Phi \) maps closed and convex cones to closed and convex cones. Let \( \alpha \in [0, 1] \), and define \( C_{\alpha} := \{(v_b, v_g) \in \mathbb{R}_+^2 : v_g \geq \alpha v_b\} \). Let \( v \in \mathbb{R}_+^2 \) be such that \((k,m_i,w_i)\) implements \( v \) with the restriction that \( w_i \in C_{\alpha} \). The set of all such \( v \) is precisely the set \( \Phi(C_{\alpha}) \).

By \[ PK_b \], we obtain

\[
v_b = -m_b + \delta \left[(1 - p_b)w_{bb} + p_bw_{bg}\right] \\
\geq -(1 - p_b + p_b \alpha)m_b + \delta [(1 - p_b)w_{bb} + p_b \alpha w_{bb}] \\
= (1 - p_b + p_b \alpha)(\delta w_{bb} - m_b)
\]

where the inequality follows from the assumption that \( w_{bg} \geq \alpha w_{bb} \), from \[ LL \] which requires that \( m_b \leq 0 \), and from the fact that \( 1 - p_b(1 - \alpha) \leq 1 \). This implies

\[
m_b - \delta w_{bb} \geq -v_b/(1 - p_b + p_b \alpha)
\]

Notice that \[ PK_b \] can be written as \( \delta p_b(w_{bg} - w_{bb}) = v_b + (m_b - \delta w_{bb}) \), which implies

\[
\delta (w_{bg} - w_{bb}) \geq \frac{v_b}{p_b} \left[ 1 - \frac{1}{\frac{1 - p_g + p_g \alpha}{1 - p_b + p_b \alpha}} \right]
\]

\[ A.1 \]

Plugging this into \[ IC^* \], we obtain

\[
v_g \geq v_b + R(k) + \delta \Delta (w_{bg} - w_{bb}) \\
\geq v_b \left[ 1 - \frac{(1 - \alpha)\Delta}{1 - p_b + p_b \alpha} \right] \\
= v_b \left[ \frac{1 - p_g + p_g \alpha}{1 - p_b + p_b \alpha} \right] \\
=: \alpha' v_b
\]

where the first inequality is merely \[ IC^* \] and the second inequality follows from \[ A.1 \] and the fact that \( R(k) \geq 0 \).

Thus, if continuation contingent utilities \( w_i \) are constrained to lie in the set \( C_{\alpha} \), then the set of implementable \( v \) must lie in the set \( C_{\alpha'} \), where \( \alpha' = (1 - p_g(1 - \alpha))/(1 - p_b(1 - \alpha)) \). In particular, any \( v \in C_{\alpha'} \) can be implemented by \((k,m_i,w_i)\) with \( w_i \in C_{\alpha} \) for \( i = b, g \).
We claim that if \( \alpha \in [0, 1) \), then \( \alpha' > \alpha \). To see this, notice that

\[
\alpha' = \frac{1 - p_g(1 - \alpha)}{1 - p_b(1 - \alpha)} > \alpha
\]

iff

\[
1 - p_g(1 - \alpha) > \alpha - \alpha p_b(1 - \alpha)
\]

iff

\[
(1 - \alpha)(1 - p_g) > -\alpha p_b(1 - \alpha)
\]

iff

\[
(1 - p_g) > -\alpha p_b
\]

which always holds because \( p_b, p_g \in (0, 1) \) and \( \alpha \in [0, 1) \). Therefore, for any such \( \alpha \in [0, 1) \), \( \Phi(C_\alpha) = C_{\alpha'} \subseteq C_\alpha \). Notice that \( \Phi^n(C_0) = \bigcap_{k \leq n} \Phi^k(C_0) = C_{\alpha_n} \), where \( \Phi^n(C_0) := \Phi(\Phi^{n-1}(C_0)) \), \( \alpha_n = \frac{1 - p_g(1 - \alpha_{n-1})}{1 - p_b(1 - \alpha_{n-1})} \), and \( \alpha_0 = 0 \). This means iterating the operator \( \Phi \) from \( C_0 = \mathbb{R}_+^2 \) induces a strictly increasing sequence \( \{\alpha_n\}_{n=0}^\infty \in [0, 1) \), and a corresponding sequence of strictly nested sets \( C_{\alpha_n} \). It is easy to see that \( \lim_{n \to \infty} \alpha_n = 1 \), and therefore, \( \lim_{n \to \infty} \Phi^n(C_0) = C_1 = V \).

To see that \( V := \{(v_b, v_g) \in \mathbb{R}_+^2 : v_g \geq v_b\} \) is a fixed point of \( \Phi \), we apply the operator \( \Phi \) to \( V \). Take any continuation utility \( v \in V \), and consider the policy \( (k, m, w) \) that satisfies \( R(k) = m_g = v_g - v_b, m_b = 0, w_i g = w_i b = v_i / \delta \). Since \( (k, m, w) \) implements \( v \) and \( w \in V \), we must have \( v \in \Phi(V) \), which means \( V = \Phi(V) \). By construction, \( V \) is the largest fixed point of \( \Phi \), which completes the proof.

\[\Box\]

A.2. Proof of Theorem 1

Proof of Theorem 1. (a) Because the only feasible policy at state \((0, 0)\) is

\[(k, m_b, m_g, w_b, w_g) = (0, 0, 0, 0, 0)\]

we must have \( Q(0, s) = 0 \).

We first show that \( Q_g(0, s) = \infty \), and \( D_{(1, 1)} Q(0, s) = \infty \). Then we use these two facts to show \( Q_b((0, v), s) = \infty \). Note that

\[
Q_g(0, s) = \lim_{\varepsilon \to 0} \frac{Q((0, \varepsilon), g)}{\varepsilon}
\]

\[
\geq \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left[ -R^{-1}(\varepsilon) + p_s \varepsilon + \delta p_s Q \left( \left( 0, \frac{\varepsilon}{\delta p_g}, g \right) \right) \right]
\]

\[\text{[A.2]}\]

\[
= p_s + \frac{p_s}{p_g} Q_g(0, g)
\]

The inequality is because

\[
(k, m_b, m_g, w_b, w_g) = \left( R^{-1}(\varepsilon), 0, \varepsilon, 0, \left( 0, \frac{\varepsilon}{\delta p_g} \right) \right) \in \Gamma((0, \varepsilon), s)
\]

37
Because \( p_g > 0 \), we must have \( Q_g(0, g) = \infty \) for [A.2] to hold when \( s = g \). Then \( Q_g(0, b) = \infty \) is implied by [A.2] when \( s = b \).

Next,

\[
D_{(1,1)} Q(0, s) = \lim_{\varepsilon \to 0} \frac{D_{(1,1)} Q((\varepsilon, \varepsilon), s)}{\varepsilon} \\
\geq \lim_{\varepsilon \to 0} \frac{\delta}{\varepsilon} \left[ p_s Q \left( \left(0, \frac{\varepsilon}{\delta p_g}\right), g \right) + (1 - p_s) Q \left( \left(\frac{\varepsilon}{\delta}, \frac{\varepsilon}{\delta}\right), b \right) \right] \\
= \frac{p_s}{p_g} Q_g(0, g) + (1 - p_s) D_{(1,1)} Q(0, b)
\]

[A.3]

The inequality is because

\[
(k, m_b, m_g, w_b, w_g) = \left(0, 0, 0, \left(\frac{\varepsilon}{\delta}, \frac{\varepsilon}{\delta}\right)\right) \in \Gamma((\varepsilon, \varepsilon), b)
\]

Because \( Q_g(0, g) = \infty \), evaluating \( D_{(1,1)} Q(0, b) \) using [A.3] (at \( s = b \)) implies \( D_{(1,1)} Q(0, b) = \infty \). Then evaluating \( D_{(1,1)} Q(0, g) \) using [A.3] (at \( s = g \)) implies \( D_{(1,1)} Q(0, g) = \infty \).

Next, let \( (k, 0, m_g, 0, w_g) \) be the optimal policy at \((0, v), s\) for any \( v > 0 \). The optimal \( m_b = w_{bh} = 0 \) is implied by [PK3] at \( v_b = 0 \). Then we have

\[
\left(k, 0, m_g, \left(\frac{\varepsilon}{\delta}, \frac{\varepsilon}{\delta}\right), w_g\right) \in \Gamma((\varepsilon, v), s)
\]

which implies

\[
Q_b((0, v), s) = \lim_{\varepsilon \to 0} \frac{Q((\varepsilon, v), s) - Q((0, v), s)}{\varepsilon} \\
\geq \lim_{\varepsilon \to 0} \frac{(1 - p_s)\delta}{\varepsilon} \left[ Q \left( \left(\frac{\varepsilon}{\delta}, \frac{\varepsilon}{\delta}\right), b \right) - Q(0, b) \right] \\
= (1 - p_s) D_{(1,1)} Q(0, b) = \infty
\]

Next, let’s define \( Q_b((v, v), s) \) as the left derivative since only \( v_g \geq v_b \) is feasible.

For arbitrary small \( \varepsilon > 0 \) and \( v' = v - \frac{1 - p_b}{1 - p_g} R(\varepsilon) \), we get

\[
Q((v', v), b) = Q((v, v), b) \\
\geq - \varepsilon + p_b R(\varepsilon) + (1 - p_b)\delta \left[ Q \left( \left(w_{bg}, w_w\right), b \right) - Q((w, w), b) \right] \geq 0
\]

[A.4]

where \( w = \frac{v + m_b}{\delta} \), and \( w_{bg} = w - \frac{R(\varepsilon)}{\delta(1 - p_g)} \). Divide both sides of [A.4] by \( \frac{1 - p_b}{1 - p_g} R(\varepsilon) \), and take the limit as \( \varepsilon \) converges to zero to obtain:

\[
-A_b((v, v), b) \geq \frac{p_b(1 - p_g)}{1 - p_b} - Q_b((w, w), b)
\]

[A.5]

Because \( Q_b((w, w), s) \leq 0 \) by Theorem 1 (c) of the Supplement Appendix, we know

\[
-A_b((v, v), b) \geq \frac{p_b(1 - p_g)}{1 - p_b}. \text{The same argument shows that } -Q_b((w, w), b) \geq \frac{p_b(1 - p_g)}{1 - p_b}. \text{ So}
\]

38
it must be that $-Q_b((v,v),b) \geq \frac{2p_b(1-p_b)}{1-p_b}$. Repeating this procedure, we get the result 
$-Q_b((v,v),b) \geq \frac{np_b(1-p_b)}{1-p_b}$ for any $n \in \mathbb{N}$. Hence, we must have $Q_b((v,v),b) = -\infty$.

Now let $(k,m_b,m_g,(w,w'),w_g)$ be the optimal policy at $((v,v),s)$ for some $v > 0$, where $w = \frac{v + m_g}{\delta}$. Also let $w_{bb} = w - \varepsilon, m'_b = m_b - (1 - p_b)\delta \varepsilon$, and $m'_g = m_g - \varepsilon$. Then we have 

$$(k,m'_b,m'_g,(w - \varepsilon,w),w_g) \in \Gamma((v,v + \varepsilon),s)$$

So optimality at $((v,v + \varepsilon),s)$ implies 

$$Q_g((v,v),s) \geq \lim_{\varepsilon \to 0} \frac{\delta(1 - p_s)}{\varepsilon} [Q((w - \varepsilon,w),b) - Q((w,w),b)]$$

$$= -\delta(1 - p_s)Q_b((w,w),b) = \infty$$

(b) Because it is feasible to invest zero in each period, make transfers $m_g = -v_g, m_b = -v_b$ in the first period, and make no transfer in all subsequent periods, we know $Q(v,s) \geq 0$. Moreover, the surplus $Q(v,s)$ is bounded above by the efficient surplus $\bar{Q}(s)$. So let $M = \max\{\bar{Q}(g), \bar{Q}(b)\}$, then $Q(v,s) \leq M$ for any $(v,s) \in V \times S$.

(c) Take any $(v,s) \in V \times S$. Let $(k,m_b,m_g,w_b,w_g)$ be the optimal policy at $(v,s)$, $v' = v + (0,\varepsilon)$ for any $\varepsilon > 0$, and $m'_g = m_g - \varepsilon$. Then the policy 

$$(k,m_b,m'_g,w_b,w_g) \in \Gamma(v',s)$$

because the specified change in states and policy increase both sides of [PK$\varepsilon$] by $\varepsilon$ and only increase the left hand side of [IC$^*$]. Moreover, because the repayment $m_g$ does not appear in the objective of [VF], we must have $Q(v',s) \geq Q(v,s)$, implying $Q_g(v,s) \geq 0$. Moreover 

$$Q_b(v,s) + Q_g(v,s) = \lim_{\varepsilon \to 0} \frac{Q(v + \varepsilon(1,1),s) - Q(v,s)}{\varepsilon} \geq 0$$

where the inequality is by Theorem 1 (e) of the Supplementary Appendix. Part (c) of the that Theorem also implies that $Q_b(v,s) < 0$ if the value of $v_g$ is very close to $v_b$.

Part (d) (e) and (f) are shown in Theorem 1 (d) and (g) of the Supplementary Appendix. \hfill \Box

### A.3. Proof of Proposition 4.2

**Lemma A.2.** The efficient surpluses of the firm are 

$$\bar{Q}(b) = \frac{1 - p_g\delta}{(1 - \delta)(1 - \Delta \delta)} [p_b R(\bar{k}_b) - \bar{k}_b] + \frac{p_b\delta}{(1 - \delta)(1 - \Delta \delta)} [p_g R(\bar{k}_g) - \bar{k}_g]$$

$$\bar{Q}(g) = \frac{(1 - p_g)\delta}{(1 - \delta)(1 - \Delta \delta)} [p_b R(\bar{k}_b) - \bar{k}_b] + \frac{1 - \delta + p_g\delta}{(1 - \delta)(1 - \Delta \delta)} [p_g R(\bar{k}_g) - \bar{k}_g]$$

39
**Proof.** We simply solve equations [4.1] and [4.2] that jointly determine $\hat{Q}(s)$.

To properly define the threshold contingent utilities, we use the following procedure. First, we fix the level of $v_b$ and find the smallest value of $v_g$ at which $Q_g$ becomes zero. This defines a cutoff curve as functions of $v_b$ along which $Q_g$ is zero. Second, we find the smallest value of $v_b$ at which $Q_b$ becomes zero along the defined cutoff curve.

**Lemma A.3.** For each $v_b \geq 0$, there exists $f_s(v_b) > v_b$ such that $Q_g(v_b, s) = 0$ if $v_g \geq f_s(v_b)$, and $Q_g(v_b, s) > 0$ if $v_b \leq v_g < f_s(v_b)$. Moreover, $f_s(v_b)$ is increasing in $v_b$.

**Proof.** First, we show that for any $v_b \geq 0$, there exists some $v_g > v_b$ such that $Q_g(v, s) = 0$. Theorem 1 (b) of the Supplementary Appendix shows that there exists some $\hat{v}_s$ that satisfies $\hat{v}_g > \hat{v}_{sb}$ and for any $v'$ in a small neighborhood around $\hat{v}_s$, $Q(v', s) = \hat{Q}$. This implies $Q_g(\hat{v}_s, s) = 0$. For any $v_b < \hat{v}_{sb}$, the supermodularity of $Q$ implies that $0 = Q_g(\hat{v}_s, s) \geq Q_g((v_b, \hat{v}_g), s) \geq 0$. The last inequality is by Theorem 1 (c). So $Q_g((v_b, \hat{v}_g), s) = 0$. For any $v_b > \hat{v}_{sb}$, we consider $v_g = \hat{v}_g + (v_b - \hat{v}_{sb})$. Theorem 1 (b) of the Supplementary Appendix shows that for any $v'$ in a small neighborhood around $v$, $Q(v', s) \geq \hat{Q}(\hat{v}_s, s) = Q(s)$, which implies $Q_g(v, s) = 0$. Next, we fix any $v_b \geq 0$ and define

[A.6] \[ f_s(v_b) := \min\{x \geq v_b : Q_g((x, v_b), s) = 0\} \]

Now we show that $f_s(\cdot)$ is increasing. Take any $v_b$ and $v'_b$ such that $0 \leq v_b < v'_b$. Supermodularity of $Q$ implies $0 \leq Q_g((v_b, f_s(v'_b)), s) \leq Q_g((v'_b, f_s(v'_b)), s) = 0$. So we have $Q_g((v_b, f_s(v'_b)), s) = 0$, which further implies that $f_s(v_b) \leq f_s(v'_b)$ by the definition of $f_s(\cdot)$.

**Lemma A.4.** For any $v_b \geq 0$, we have $Q_b((v_b, f_s(v_b)), s) \geq 0$. For some $v_b \geq 0$, we have $Q_b((\hat{v}_b, f_s(\hat{v}_b)), s) = 0$.

**Proof.** Note that the definition of $f_s(\cdot)$ means $Q_g((v_b, f_s(v_b)), s) = 0$. So we have $Q_b((v_b, f_s(v_b)), s) = Q_b((v_b, f_s(v_b)), s) + Q_g((v_b, f_s(v_b)), s) \geq 0$ for any $v_b \geq 0$. The inequality is implied by Theorem 1 (c). Moreover, Theorem 1 (b) of the Supplementary Appendix implies that there exists some $\hat{v}_s \in V$ such that $Q_b(\hat{v}_s, s) = Q_g(\hat{v}_s, s) = 0$. By the definition of $f_s(\cdot)$, we know $f_s(\hat{v}_{sb}) \leq \hat{v}_g$. Supermodularity of $Q$ then implies $Q_b((\hat{v}_{sb}, f_s(\hat{v}_{sb})), s) \leq Q_b(\hat{v}_s, s) = 0$. So it has to be that $Q_b((\hat{v}_{sb}, f_s(\hat{v}_{sb})), s) = 0$.

We are now ready to define the threshold contingent utilities and the unconstrained sets. By Lemma A.4, we can define

[A.7] \[ \bar{v}_{sb} = \min\{v_b \geq 0 : Q_b((v_b, f_s(v_b)), s) = 0\} \quad \text{and} \quad \bar{v}_g = f_s(\bar{v}_{sb}) \]
where \( \bar{v}_{sb} \) is the threshold of continuation utility contingent on bad shock this period (subscript \( s \) denotes last period’s shock), while \( \bar{v}_{sg} \) is the threshold continuation utility contingent on good shock this period. From Theorem 1 (a), \( Q_b((0, f_s(0)), s) = \infty \). So by definition \( \bar{v}_{sb} > 0 \). Moreover, we define the unconstrained sets of contingent utilities as

\[
E_s^\ast := \{ v \in V : v_b \geq \bar{v}_{sb}, v_s \geq f_s(v_b) \}
\]

**Lemma A.5.** Let \( E_s^\ast \) be defined as in [A.8]. Then, \( v \in E_s^\ast \) if, and only if, \( Q_b(v, s) = 0 = Q_g(v, s) \). Moreover, \( E_s^\ast = E_s := \{ v \in V : Q(v, s) = \tilde{Q}(s) \} \), and \( E_s \) is closed and convex.

**Proof.** We shall first prove the ‘if’ part. To see this, let \( v \in E_s^\ast \). Hence, by the definition of \( \bar{v}_{sb} \) in [A.7], and the second inequality is because \( Q_b \) is concave and continuously differentiable, \( Q \) achieves its maximum value at any such point.

If \( v_g < f_s(v_b) \), then the definition of \( f_s \) in [A.6] implies \( Q_g(v, s) > 0 \), a contradiction. If \( v_b < \bar{v}_{sb} \) and \( v_s \geq f_s(v_b) \), then \( 0 < Q_b(v_b, f_s(v_b)), s \leq Q_b(v, s) \). The first inequality is by the definition of \( \bar{v}_{sb} \) in [A.7], and the second inequality is because \( Q_b \) is increasing in \( v_g \) from part (f) of Theorem 1. Hence, \( v \in E_s^\ast \).

To see the ‘only if’ part, consider any \( v \in E_s^\ast \). By the definition of \( E_s^\ast \), \( v_g \geq f_s(v_b) \). Hence, by the definition of \( f_s \), \( Q_g(v, s) = 0 \). Then we know \( Q_b(v, s) \geq 0 \), because \( Q_b(v, s) + Q_g(v, s) \geq 0 \) by Theorem 1 (c). Moreover, concavity of \( Q \) implies \( Q_b(v, s) \leq Q_b((\bar{v}_{sb}, v_g), s) \). So if we can show \( Q_b((\bar{v}_{sb}, v_g), s) = 0 \), then we can also conclude \( Q_b(v, s) = 0 \). Now take any \( \hat{v}_g \in V \) and any sufficiently small \( \varepsilon > 0 \). Monotonicity of \( f_s \) implies

\[
f_s(\bar{v}_{sb} - \varepsilon) \leq f_s(\bar{v}_{sb}) = \bar{v}_{sg} \leq \hat{v}_g
\]

So by the definition of \( f_s \), \( Q_g((\bar{v}_{sb}, \hat{v}_g), s) = Q_g((\bar{v}_{sb} - \varepsilon, \hat{v}_g), s) = 0 \). Since these relations hold for arbitrary \( \hat{v}_g \), we have

\[
Q((\bar{v}_{sb}, v_g), s) = Q(\bar{v}, s), \quad Q((\bar{v}_{sb} - \varepsilon, v_g), s) = Q((\bar{v}_{sb} - \varepsilon, \bar{v}_{sg}), s)
\]

which further implies

\[
Q_b((\bar{v}_{sb}, v_g), s) = \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \left[ Q((\bar{v}_{sb}, v_g), s) - Q((\bar{v}_{sb} - \varepsilon, v_g), s) \right]
\]

\[
= \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \left[ Q((\bar{v}_{sb}, \bar{v}_{sg}), s) - Q((\bar{v}_{sb} - \varepsilon, \bar{v}_{sg}), s) \right]
\]

\[
= Q_b(\bar{v}, s) = 0
\]

where the last equality is by the definition of \( \bar{v}_{sb} \). Therefore, we have \( Q_b(v, s) = 0 \), as claimed.

All that remains is to show that \( E_s = E_s^\ast \). Equivalently, this amounts to show that \( \max_{v \in V} Q(v, s) = \tilde{Q}(s) \). This is established in Lemma 1.2 of the Supplementary
Appendix. The concavity and continuity of $Q$ now imply that $E_s$ is closed and convex, which completes the proof.

We now proceed to the proof of Proposition 4.2.

**Proof of Proposition 4.2.** It is immediate from Lemma A.5 that $E_s$ is non-empty, closed, and convex. Then, part (a) follows immediately from the definition of $E_s^*$ in [A.8] and from Lemma A.5 which proves that $E_s^* = E_s$.

To see (b), take any $v \in E_s$ and observe that because $Q$ is concave, $Q_b(v, s) = Q_b(v, s) = 0$ implies $Q(v, s)$ achieves its maximum $\bar{Q}(s)$. By the definition of $Q(v, s)$ and $\bar{Q}(s)$ we have

\[
\bar{Q}(s) = -\bar{k}_s + p_s[R(\bar{k}_s) + \delta \bar{Q}(g)] + (1 - p_s)\delta \bar{Q}(b) = -k(v, s) + p_s[R(k(v, s)) + \delta Q(w_g(v, s), g)] + (1 - p_s)\delta Q(w_b(v, s), b) = Q(v, s)
\]

which implies that we must have $k(v, s) = \bar{k}_s$, $Q(w_g(v, s), s) = \bar{Q}(g)$, and $Q(w_b(v, s), s) = \bar{Q}(b)$. Lemma A.5 then implies $w_g(v, s) \in E_g$ and $w_b(v, s) \in E_b$.

For part (c), we shall establish some intermediate claims.

(1) Let $v \in E_b$. We show that $v_g - v_b \geq R(\bar{k}_b) + \delta \Delta \frac{R(\bar{k}_b)}{1 - \delta \Delta}$ is necessary to obtain efficient firm surplus at $(v, b)$. Let $(k, m_i, w_i)$ be the optimal policy at $(v, b)$. We know $k = \bar{k}_b$ from part (c). The constraint [IC*] at $(v, b)$ implies $v_g - v_b \geq R(\bar{k}_b) + \delta \Delta (w_{bg} - w_{bb}) \geq R(\bar{k}_b)$. Moreover, we must also have $w_b \in E_b$. Otherwise, $Q(v, b)$ will be smaller than the first best surplus $\bar{Q}(b)$, a contradiction. The constraint [IC*] at $(w_b, b)$ implies that $w_{bg} - w_{bb} \geq \bar{R}(\bar{k}_b)$. So we have $v_g - v_b \geq R(\bar{k}_b) + \delta \Delta R(\bar{k}_b)$. Repeating this procedure we obtain $v_g - v_b \geq (1 + \delta \Delta + \delta^2 \Delta^2 + \ldots) R(\bar{k}_b) = R(\bar{k}_b) + \delta \Delta R(\bar{k}_b)$ from the first step.

(2) Let $v \in E_g$ and let $(k, m_i, w_i)$ be the optimal policy at $(v, g)$. Similar argument shows that $k = \bar{k}_g$, and $w_b \in E_b$. So [IC*] at $(v, g)$ implies $v_g - v_b \geq R(\bar{k}_g) + \delta \Delta (w_{bg} - w_{bb}) \geq R(\bar{k}_g) + \delta \Delta R(\bar{k}_b)$, because $w_{bg} - w_{bb} \geq \frac{R(\bar{k}_b)}{1 - \delta \Delta}$ from the first step.

(3) Take any $v \in E_s$ for $s = b, g$. Let $(k, m_i, w_i)$ be the optimal policy at $(v, s)$. We show that $w_{bg} - w_{bb} \geq \frac{\delta \bar{v}_{bg} - \bar{v}_b}{\delta (1 - p_b)}$. Suppose not. Then we can derive

\[
\bar{w}_{bg} - \bar{w}_{bb} < \frac{\delta \bar{v}_{bg} - \bar{v}_b}{\delta (1 - p_b)} = \frac{\delta \bar{v}_{bg} + m_b - \delta [p_b w_{bg} + (1 - p_b)w_{bb}]}{\delta (1 - p_b)}
\]

where the equality is from [PK*]. Rearranging [A.9] we get $w_{bg} < \bar{v}_{bg} + \frac{m_b}{\Delta} \leq \bar{v}_{bg}$. This means $w_b \in V \setminus E_b$ by part (a). Hence, $Q(w, b) < \bar{Q}(b)$, implying $Q(v, s) < \bar{Q}(s)$, a contradiction with $v \in E_s$. Since $k = \bar{k}_s$, [IC*] at $(v, s)$ implies that $v_g - v_b \geq R(\bar{k}_s) + \delta \Delta \frac{\delta \bar{v}_{bg} - \bar{v}_b}{\delta (1 - p_b)}$.

Combining (1) to (3), we conclude that it is necessary to satisfy $v_g - v_b \geq R(\bar{k}_s) + \delta \Delta \max \left[ \frac{\delta \bar{v}_{bg} - \bar{v}_b}{\delta (1 - p_b)}, \frac{R(\bar{k}_b)}{1 - \delta \Delta} \right]$ for any $v \in E_s$. 

42
B. Auxiliary Problem

To proceed the proof in Section 5 and beyond, it is convenient to define an auxiliary problem as follows:

\[ [P3] \quad \Psi(y, s) = \max_{x_g \geq x_b \geq 0} \delta Q(x, s) \]

subject to \( y \geq \delta(p_s x_g + (1 - p_s)x_b) \)

where \( y \geq 0 \) and \( s = b, g \). To ease notation, let \( x(y, s) \) be a solution to problem \([P3]\).

Lemma B.1. Function \( \Psi(y, s) \) defined in \([P3]\) has the following properties:

(a) \( \Psi(y, s) \) is continuously differentiable, increasing, and concave in \( y \).

(b) \( \Psi_y(y, s) = D_{(1,1)} Q(x(y, s)) \).

(c) When \( y \geq \delta E^y[\bar{v}_s] \), \( \Psi(y, s) = \delta \bar{Q}(s) \), \( \Psi_y(y, s) = 0 \), and \( x(y, s) \in E_s \).

(d) When \( y < \delta E^y[\bar{v}_s] \), \( \Psi_y(y, s) \) is strictly increasing in \( y \), and the constraint of \([P3]\) is active, ie, the Lagrange multiplier for the constraint is strictly positive (which implies that the constraint holds as an equality).

Proof. (a) By the Theorem of the Maximum, \( \Psi(\cdot, s) \) is continuous. Continuous differentiability follows from an argument similar to that used to prove that \( Q \) is continuously differentiable, ie, the Beneveniste-Scheinkman Theorem — see Theorem 4.11 of Stokey, Lucas and Prescott (1989). Notice that raising \( y \) always relaxes the constraint in problem \([P3]\), so we have \( \Psi_y(y, s) \geq 0 \). Moreover, because \([P3]\) has a concave objective (shown in part (a) of Theorem 1) and a convex constraint, \( \Psi(y, s) \) is concave in \( y \).

(b) Theorem 1 implies that any \( x(v, s) \) must lie in the interior of \( V \). Let \( \gamma \) be the Lagrange multiplier of \([P3]\). The first order conditions and the envelope condition for problem \([P3]\) are

\[ [B.1] \quad Q_b(x(y, s), s) = (1 - p_s)\gamma \quad \text{and} \quad Q_g(x(y, s), s) = p_s \gamma, \]

\[ [B.2] \quad \Psi_y(y, s) = \gamma \]

It is easy to see from \([B.1]\) and \([B.2]\) that \( \Psi_y(y, s) = D_{(1,1)} Q(x(y, s), s) \).

(c) When \( y \geq \delta E^y[\bar{v}_s] \), \( x(y, s) = \bar{v}_s \) is feasible in \([P3]\). And because \( Q(\cdot, s) \) reaches its upper bound \( \bar{Q}(s) \) at \( \bar{v}_s \), we must have \( \Psi(y, s) = \delta \bar{Q}(s) \). Obviously, \( \Psi_y(y, s) = 0 \), because \( \Psi(y, s) \) is a constant. Proposition 4.2 implies that \( x(y, s) \in E_s \).

(d) Suppose there exists some \( \tilde{y}_s < \delta E^y[\bar{v}_s] \) such that \( \Psi_y(\tilde{y}_s, s) = 0 \). Concavity of \( \Psi \) implies that \( \Psi(\tilde{y}_s, s) = \Psi(\bar{y}_s, s) = \delta \bar{Q}(s) \), where \( \bar{y}_s = \delta E^y[\bar{v}_s] \). From Proposition 4.2, we know \( x(\bar{y}_s, s) \geq \bar{v}_s \). However, this implies \( \bar{y}_s \geq \delta E^y[\bar{v}_s] \) by the constraint of
Lemma B.2. For any \((v, s) \in V \times S\), the optimal policy \(w^*(v, s)\) satisfies
(a) \(w^*_g(v, s)\) is a solution to problem \([P3]\) at \((v_g, g)\), and is independent of \(v_b, s\).
(b) \(w^*_g(v, s) \in E_g\) when \(v_g \geq \delta \mathbb{E}^s[\bar{v}_s]\).
(c) If \(\alpha(v, s) = 0\), then \(w^*_b(v, s)\) is a solution to problem \([P3]\) at \((v_b, b)\).

Proof. Part (a) is a direct implication of Lemma 1.8 in the Supplement Appendix. Part (b) simply follows from the result (c) in Lemma B.1, because \(w^*_g(v, s)\) is a solution to problem \([P3]\) at \((v_g, g)\). Part (c) holds simply because when \([IC]\) does not bind at \((v, s)\), \(w^*_b(v, s)\) has to solve the problem without \([IC]\).

Lemma B.3. The function \(\Psi(\cdot, s)\) is strictly concave on the set \([0, \delta \mathbb{E}^s[\bar{v}_s]]\).

Proof. Let \(\bar{y}_s = \delta \mathbb{E}^s[\bar{v}_s]\). Take any distinct \(\bar{y}, \tilde{y} \in [0, \bar{y}_s]\). From Lemma B.1, the constraint for Problem \([P3]\) binds at both \(y = \bar{y}\) and \(y = \tilde{y}\). So \(\bar{y} \neq \tilde{y}\) implies \(x(\bar{y}, s) \neq x(\tilde{y}, s)\). Moreover, from equations \([B.1]\) and \([B.2]\), we can see that

\[
\min [Q_b(x(y, s), s), Q_b(x(y, s), s)] > 0
\]

when \(y < \bar{y}_s\), because \(\Psi(y, s) > 0\). This means \((x(\bar{y}, s), s), (x(\tilde{y}, s), s) \in H\). Hence,

\[
\theta \Psi(\bar{y}, s) + (1 - \theta) \Psi(\tilde{y}, s) = \delta [\theta Q(x^*(\bar{y}, s), s) + (1 - \theta) Q(x^*(\tilde{y}, s), s)] \\
\leq \delta Q(\theta x^*(\bar{y}, s), s) + (1 - \theta) x^*(\tilde{y}, s), s) \\
\leq \Psi(\theta \bar{y} + (1 - \theta) \tilde{y}, s)
\]

where the second line is implied by Proposition 2.5 in the Supplement Appendix.

C. Optimal Conditions and Directional Derivative

In this section, we first derive the first order conditions an envelope conditions that are necessary and sufficient for the firm’s maximization problem \([VF]\). We then show the directional derivative of the surplus function is a nonnegative martingale and examine how it evolves in the optimal contract.

In what follows, \(\eta^*_g(v, s)\) and \(\eta^*_b(v, s)\) are the Lagrange multipliers for the promise keeping constraints \([PK_g]\) and \([PK_b]\), \(\alpha(v, s)\) is the Lagrange multiplier for the incentive
Proof. The process $D_{(1,1)}$ of a contract is a nonnegative martingale.

**Lemma C.1.** The process $D_{(1,1)} Q(v,s) = Q_b(v,s) + Q_g(v,s)$ induced by the optimal contract is a nonnegative martingale.

**Proof.** Take any $(v,s) \in V \times S$. Adding the first order conditions [FOCwb] to [FOCwg], and using envelope conditions [Envb] and [Envg] to substitute $\eta_i(v,s)$, we get

\[
(1 - p_b) D_{(1,1)} Q(w_b(v,s),b) + p_g Q_{(1,1)} g(w_g(v,s),g) = D_{(1,1)} Q(v,s)
\]

Moreover, by Theorem 1 (c) $D_{(1,1)} Q(v,s) \geq 0$. Thus, the process $D_{(1,1)} Q$ is a nonnegative martingale.

Using the martingale relation [C.1] and the first order conditions, we can characterize the evolution of the directional derivative martingale $D_{(1,1)} Q$ on any path induced by the optimal contract, as in the following Lemma.

**Lemma C.2.** For the consecutive shocks ‘good-good’, ‘good-bad’, ‘bad-good’, and ‘bad-bad’ and for an initial state $(v,s)$, let the states induced by the optimal contract at $(v,s)$ be $w_i = w_i(v,s)$, and let the states induced by the optimal contract at $(w_g,g)$ and $(w_b,b)$ be $w_i^g = w_i(w_g,g)$ and $w_i^b = w_i(w_b,b)$ respectively. Then, the martingale $D_{(1,1)} Q$ evolves according to the following relations

\[
\begin{align*}
D_{(1,1)} Q(w_i^g,g) &= D_{(1,1)} Q(w_g,g) - \frac{1}{p_g} \alpha(w_g,g) \\
D_{(1,1)} Q(w_i^b,b) &= D_{(1,1)} Q(w_g,g) + \frac{1}{1 - p_g} \alpha(w_g,g) \\
D_{(1,1)} Q(w_i^b,g) &= D_{(1,1)} Q(w_b,b) - \frac{(1 - p_b) \alpha(w_b,b) - \Delta \alpha(v,s)}{p_b(1 - p_b)} \\
D_{(1,1)} Q(w_i^b,b) &= D_{(1,1)} Q(w_b,b) + \frac{(1 - p_b) \alpha(w_b,b) - \Delta \alpha(v,s)}{(1 - p_b)(1 - p_b)}
\end{align*}
\]
Proof. First, from $[\text{Env}_g]$ and $[\text{FOC}_{wg}]$, we get:

\[ C.6 \quad p_s[\eta_g(w_g, g) - \alpha(w_g, g)] = p_sQ_g(w_g, g) - p_s\alpha(w_g, g) \]
\[ = p_g[\eta_g(v, s) - \alpha(v, s)] - p_s\alpha(w_g, g) \]

Add $[\text{FOC}_{wb}]$ and $[\text{FOC}_{wg}]$ at state $(v, s)$ to get:

\[ C.7 \quad p_sD_{(1,1)}Q(w_g, g) = \eta_g(v, s) - \alpha(v, s) \]

Add $[\text{FOC}_{wb}]$ and $[\text{FOC}_{wg}]$ at state $(w_g, g)$ to get:

\[ C.8 \quad p_gD_{(1,1)}Q(w_g^b, g) = \eta_g(w_g, g) - \alpha(w_g, g) \]

Then combine $[C.6]$, $[C.7]$, $[C.8]$ and rearrange to get $[C.2]$. Next, from $[\text{Env}_g]$ and $[\text{FOC}_{wb}]$, we get

\[ C.9 \quad (1 - p_s)[\eta_b(w_b, b) + \alpha(w_b, b)] \]
\[ = (1 - p_s)Q_b(w_b, b) + (1 - p_s)\alpha(w_b, b) \]
\[ = (1 - p_b)[\eta_b(v, s) + \alpha(v, s)] - \Delta\alpha(v, s) + (1 - p_s)\alpha(w_b, b) \]

Add $[\text{FOC}_{wb}]$ and $[\text{FOC}_{wb}]$ at state $(v, s)$ to get:

\[ C.10 \quad (1 - p_s)D_{(1,1)}Q(w_b, b) = \eta_b(v, s) + \alpha(v, s) \]

Add $[\text{FOC}_{wb}]$ and $[\text{FOC}_{wb}]$ at state $(w_b, b)$ to get:

\[ C.11 \quad (1 - p_b)D_{(1,1)}Q(w_b^b, b) = \eta_b(w_b, b) + \alpha(w_b, b) \]

Then combine $[C.9],[C.10]$, $[C.11]$ and rearrange to get $[C.5]$. Finally, $[C.1]$ at state $(w_g, g)$ implies:

\[ C.12 \quad (1 - p_g)D_{(1,1)}Q(w_b^g, b) + p_gD_{(1,1)}Q(w_g^g, g) = D_{(1,1)}Q(w_g, g) \]

$[C.1]$ at state $(w_b, b)$ implies:

\[ C.13 \quad (1 - p_b)D_{(1,1)}Q(w_b^b, b) + p_bD_{(1,1)}Q(w_b^b, g) = D_{(1,1)}Q(w_b, b) \]

Then we can combine $[C.2]$ $[C.12]$ and rearrange to obtain $[C.3]$. Similarly, we can combine $[C.5]$ $[C.13]$ and rearrange to obtain $[C.4]$. □

Lemma C.3. For any $(v, s) \in V \times S$, we must have either $w_g(v, s) \in H$ or $w_g(v, s) \in E_g$. For any $(v, s) \in H$, we must have $w_b(v, s) \in H$. 

46
\textbf{Proof.} Take any \((v, s) \in V \times S\). From a little manipulation of the first order conditions \([\text{FOC}_w^g]\) and \([\text{FOC}_s^g]\), we see that \(p_g Q_b(w_g(v, s), g) = (1 - p_g) Q_g(w_g(v, s), g)\). So we must either have
\[
\min\left[Q_b(w_g(v, s), g), Q_g(w_g(v, s), g)\right] > 0
\]
or have \(Q_b(w_g(v, s), g) = 0 = Q_g(w_g(v, s), g)\). The former case means \(w_g(v, s) \in H\) by definition of \(H\), while the latter case means that \(w_g(v, s) \in E_g\) by part (b) of Proposition 4.2.

Now consider any \((v, s) \in H\). From \([\text{FOC}_w^b]\) and \([\text{FOC}_s^b]\) we know
\[
(1 - p_s) Q_b(w_b(v, s), b) \geq (1 - p_b) Q_b(v, s) > 0
\]
\[
(1 - p_s) Q_g(w_b(v, s), b) \geq p_b Q_b(v, s) > 0
\]
The first inequality in both lines is because \(\alpha(v, s) \geq 0\). Hence \((w_b(v, s), b) \in H\). \qed

We show in the following that the directional derivative must split, ie, it goes down after a good shock and goes up after a bad shock, if the previous period had a good shock.

\textbf{Lemma C.4.} Suppose the optimal contract starts at \((v, s)\) and evolves to the state \((w_g, g)\) satisfying \(D_{(1,1)}Q(w_g, g) > 0\) after a good shock. Then the directional derivative goes down after another good shock and goes up after another bad shock, ie, \(D_{(1,1)}Q(w_g^g, g) < D_{(1,1)}Q(w_g, g)\) and \(D_{(1,1)}Q(w_g^b, b) > D_{(1,1)}Q(w_g, g)\).

\textbf{Proof.} We shall show in two cases that \(D_{(1,1)}Q(w_g^g, g) < D_{(1,1)}Q(w_g, g)\). The conclusion that \(D_{(1,1)}Q(w_g^b, b) > D_{(1,1)}Q(w_g, g)\) then simply follows from the martingale equation \([\text{C.1}]\) at the state \((w_g, g)\).

First, we consider the case of \(w_{gg} < \delta \mathbb{E}^g[v_g]\). From \([\text{PK}_g]\), we know that \(v_g \leq \delta \mathbb{E}^g[w_g] \leq \delta w_{gg} < w_{gg}\), implying \(v_g < w_{gg} \leq \delta \mathbb{E}^g[v_g]\). By Lemma B.3, \(\Psi_y(w_{gg}, g) < \Psi_y(v_g, g)\), because \(\Psi(\cdot, g)\) is strictly concave on \((0, \delta \mathbb{E}^g[v_g])\). By Lemmas B.1 and B.2, we see that \(D_{(1,1)}Q(w_g, g) = \Psi_y(v_g, g)\) and \(D_{(1,1)}Q(w_g^g, g) = \Psi_y(w_{gg}, g)\). So the conclusion follows.

Second, we consider the case of \(w_{gg} \geq \delta \mathbb{E}^g[v_g]\). From part (b) of Lemma B.2, we know that \(w_g^g \in E_g\). So the left hand side of \([\text{FOC}_w^g]\) at \((w_g, g)\) is zero implying \(\eta_g(w_g, g) = \alpha(w_g, g)\). The condition \(D_{(1,1)}Q(w_g, g) > 0\) implies that \((w_g, g) \in H\) by Lemma C.3. Hence, \(\alpha(w_g, g) = Q_g(w_g, g) > 0\). And equation \([\text{C.2}]\) simply means \(D_{(1,1)}Q(w_g^g, g) < D_{(1,1)}Q(w_g, g)\). \qed
D. Proofs from Section 5

Proof of Proposition 5.1. Define the expected profit of the principal at initialization of the contract as

\[ D(s) = \max_{v \in V} Q(v, s) - [p_s v_g + (1 - p_s) v_{sb}] \]

where \( s = s_0 \). From the first order condition of [IN], we know the optimal initial state \((v^o_s, s)\) satisfies

\[ Q_g(v^o_s, s) = p_s, \quad Q_b(v^o_s, s) = 1 - p_s \]

which implies \((v^o_s, s) \in H\) by the definition of the set \( H \). Then Lemma C.3 implies that the optimal contract can only leave \( H \) after a good shock and before that it stays in \( H \).

Let \((k, m_i, w_i)\) be the optimal policy at \((v^o_b, b)\). Then

\[
D(b) = -k + p_b R(k) + \delta p_b [Q(w, g) - Q(w, b)] + \delta Q(w, b) - p_b v^o_{bb} - (1 - p_b) v^o_{bb}
\]

\[
\leq -k + p_b \{ R(k) + \delta [Q(w, g) - Q(w, b)] - (v^o_{bg} - v^o_{bb}) \} + \delta [Q(w, b) - p_b w_{bg} - (1 - p_b) w_{bb}]
\]

The inequality is because [PKb] at \((v^o_b, b)\) implies \(-v^o_{bb} \leq -[p_b w_{bg} + (1 - p_b) w_{bb}]\). So we can rewrite the above inequality as

\[
D(b) \geq D(b) - \delta [Q(w, b) - p_b w_{bg} - (1 - p_b) w_{bb}] + k > 0
\]

The second inequality is because \( \delta < 1, w_b \in V, \) and \( D(b) \) is the maximum of problem [IN] at \( s = b \). Moreover, because \( v^o_b \in V \), optimality of [IN] at \( s = g \) implies

\[
D(g) - D(b) \geq Q(v^o_b, g) - p_g v^o_{bg} - (1 - p_g) v^o_{bb} - [Q(v^o_b, b) - p_b v^o_{bg} - (1 - p_b) v^o_{bb}]
\]

\[
= Q(v^o_b, g) - Q(v^o_b, b) - \Delta(v^o_{bg} - v^o_{bb})
\]

\[
\geq \Delta \{ R(k) + \delta [Q(w, g) - Q(w, b)] - (v^o_{bg} - v^o_{bb}) \} > 0
\]

The first inequality in the last line is implied by \((k, m_i, w_i) \in \Gamma(v^o_b, g)\). And the second inequality in the last line is by [D.2].

The condition derived in Lemma 1.6 of Supplement Appendix regarding Lagrange multipliers also holds for Lagrange multipliers in problem [VF]. This is because the function \( P \) in [P1] (defined in Supplement Appendix) satisfies all the properties of the function \( Q(v, s) \) in [VF]. Therefore we know

[D.3] \( \eta_b(v, s) + \alpha(v, s) - \mu_b(v, s) \geq 0 \), \( m_b(v, s) [\eta_b(v, s) + \alpha(v, s) - \mu_b(v, s)] = 0 \)

[D.4] \( \eta_g(v, s) - \alpha(v, s) - \mu_g(v, s) = 0 \)
Proof of Proposition 5.2. Take any \((v,s) \in H\).

(a) If \(\mu_b(v,s) > 0\), then complementary slackness implies \(m_b(v,s) = 0\). If \(\mu_b(v,s) = 0\), then \(\eta_b(v,s) + \alpha(v,s) - \mu_b(v,s) > 0\), because \(\eta_b(v,s) > 0\) when \(v \in H\) and \(\alpha(v,s) \geq 0\). Then [D.3] implies \(m_b(v,s) = 0\).

(b) From Lemma B.2 (a), \(w_g(v,s)\) is a solution of \([P3]\) at \((v_g,g)\). Suppose \(v_g < \delta E^g[\bar{v}_g]\).

Then from Lemma B.1 (d), the constraint of \([P3]\) at \((v_g,g)\) must bind, which implies \(m_g(v,s) = R(k(v,s))\) by \([PK_g]\) at \((v,s)\).

(c) Suppose \(v_g \geq \delta E^g[\bar{v}_g]\). From Lemma B.1 (c), \(w_g(v,s) \in E_g\). That is the contingent utility reaches the unconstrained set after a good shock. Now suppose \(w_g(v,s) \in E_g\). From Proposition 4.2 (a), \(w_{gg}(v,s) \geq \bar{v}_{gg}\) and \(w_{gb}(v,s) \geq \bar{v}_{gb}\). Then by \([PK_g]\) at \((v,s)\), we know

\[v_g \geq \delta [p_g w_{gg}(v,s) + (1 - p_g) w_{gb}(v,s)] \geq \delta E^g[\bar{v}_g]\]

(d) Maximum rent contract implies \(w_g(v,s) = \bar{v}_g\) whenever \(w_g(v,s) \in E_g\). Then from part (c), we conclude that \(v_g \geq \delta E^g[\bar{v}_g]\) if and only if \(w_g(v,s) = \bar{v}_g\). On one hand, if \(v_g > \delta E^g[\bar{v}_g]\) then \([PK_g]\) implies

\[D.5\]

\[R(k(v,s)) - m_g(v,s) = v_g - \delta E^g[\bar{v}_g]\]

because \(w_g(v,s) = \bar{v}_g\). Hence, \(R(k(v,s)) - m_g(v,s) > 0\). On the other hand, if \(R(k(v,s)) - m_g(v,s) > 0\), then from the result of part (b) we must have \(v_g \geq \delta E^g[\bar{v}_g]\), which implies \(w_g(v,s) = \bar{v}_g\) in the maximum rent contract. And \([PK_g]\) implies \([D.5]\) holds. Hence, \(v_g > \delta E^g[\bar{v}_g]\).

Before proceeding to characterize other policies of the optimal contract, let us show where the set \(H\) locates and some useful properties this set.

Lemma D.1. For any \(v \in V\) with \(v_b \geq \bar{v}_{sb}\), \(Q_b(v,s) \leq 0\).

Proof. Note that by concavity of \(Q\) it suffices to show that \(Q_b((\bar{v}_{sb},v_g),s) \leq 0\). When \(v_g \geq \bar{v}_{sg}\), \((\bar{v}_{sb},v_g) \in E_s\) simply implies \(Q_b((\bar{v}_{sb},v_g),s) = 0\) by part (b) of Proposition 4.2. When \(v_g < \bar{v}_{sg}\), because \(Q\) is supermodular by part (f) of Theorem 1, we have \(Q_b((\bar{v}_{sb},v_g),s) \leq Q_b((\bar{v}_{sb},\bar{v}_{sg}),s) = 0\).

Lemma D.2. Let \(G := \{(v,s) \in V \times S : v < \bar{v}_s\}\), and let \(cl(H), cl(G)\) be the closure of set \(H\) and set \(G\) respectively. Then,

(a) \(H \subset G\);
(b) For any \(s \in S\), \((\bar{v}_s,s) \in cl(H)\);
(c) For any \((v,s) \in cl(H), Q_g(v,s) + Q_b(v,s) = 0\) implies \(v = \bar{v}_s\).
Proof. (a) Take any \((v,s) \in H\). By definition \(Q_b(v,s), Q_g(v,s) > 0\). So Lemma D.1 implies \(v_b < \bar{v}_b\). Suppose \(v_g \geq \bar{v}_g\). Then
\[
0 \leq Q_g(v,s) \leq Q_g(\bar{v}_b, v_g), s) = 0
\]
The first inequalities is from Theorem 1 (c) that \(Q_g\) is positive. The second inequality is Theorem 1 (f) that \(Q_g\) is increasing in \(v_b\). The last equality is because \((\bar{v}_b, v_g) \in E_s\) and by Proposition 4.2 we know \((\bar{v}_b, v_g) \in E_s\) which implies \(Q_g(\bar{v}_b, v_g), s) = 0\). Hence \(Q_g(v,s) = 0\), a contradiction.

(b) Part (a) implies \(c(l(H) \subseteq c(l(G))\). Suppose \((\bar{v}_s, s) \notin c(l(H))\) for some \(s \in \{b, g\}\). Then there must exist some \(\varepsilon > 0\) such that \(N_\varepsilon \cap c(l(H)) = \emptyset\) where \(N_\varepsilon : = \{(v,s) : \bar{v}_s - (\varepsilon, \varepsilon) \leq v < \bar{v}_s\}\). However, there must exist \(\eta > 0\) such that \((x, f_s(x)) \in N_\varepsilon\) for any \(x \in [\bar{v}_b - \eta, \bar{v}_b]\). This is because \(f_s(\cdot)\) (as defined in [A.6]) is increasing and satisfies \(f_s(\bar{v}_b) = \bar{v}_g\) by Lemma A.3. And by the definition in [A.7], \(Q_b((x, f_s(x)), s) > 0\) for all such \(x\). Then the definition of \(f_s\) and continuity of \(Q_b(\cdot, s)\) imply that there exists some region slightly below the curve \(\{v : v_g = f_s(v_b), v_b \in [\bar{v}_b - \eta, \bar{v}_b]\} and locates in \(N_\varepsilon\) that satisfies \(Q_g(v_s) > 0, Q_b(v,s) > 0\). This forms a contradiction.

Hence, \((\bar{v}_s, s) \in c(l(H))\).

(c) Take any \((\hat{v}, s) \in c(l(H))\) such that \(D_{(1,1)} Q(\hat{v}, s) = 0\). Then \(\hat{v} \leq \bar{v}_s\) and \(Q_b(\hat{v}, s) = Q_g(\bar{v}_s, s) = 0\). By the definition of \(E_s\), we know \(\hat{v} \in E_s\), and hence, \(\hat{v} \geq \bar{v}_s\). So we must have \(\hat{v} = \bar{v}_s\).

Proof of Proposition 5.3. Let \((k, m_i, w_i)\) be the optimal policy at \((v,s) \in H\).

(a) Combing the results from Lemma C.3 and Lemma D.2 (a), we know: \(w_b \in H\) and hence \(w_b < \bar{v}_b\); moreover, \(w_g \in H\) or \(w_g \in E_g\) and hence \(w_g \leq \bar{v}_g\).

(b) (i) If \(v_g < \delta \mathbb{E}[\bar{v}_g]\), then Proposition 5.2 (a) implies \(m_g = R(k)\). From the constraint \([P_k]\), \(v_g = \delta \mathbb{E}[w_g] \leq \delta w_{gg}\). Hence, \(w_{gg} \geq v_g\). If \(v_g \geq \mathbb{E}[\bar{v}_g]\), then Proposition 5.2 (c) implies \(w_{gg} \geq \bar{v}_{gg}\). Because \(v \in H\), we know \(v_g < \bar{v}_{gg}\) by part Lemma D.2 (a). Hence, \(w_{gg} > v_g\).

(ii) Let \(\hat{v} = v + (0, \varepsilon)\) for some sufficiently small \(\varepsilon > 0\). And let \((\hat{k}, \hat{m}_i, \hat{w}_i)\) be the optimal policy at \((\hat{v}, s)\). If \(v_g \geq \mathbb{E}[\bar{v}_g]\), then obviously \(\hat{v}_g > \mathbb{E}[\bar{v}_g]\). Hence, \(w_g = \hat{w}_g = \bar{v}_g\) which means the optimal contingent utility vector stay constant (weakly increasing) after a good shock. If \(v_g < \mathbb{E}[\bar{v}_g]\), then \(\hat{v}_g < \mathbb{E}[\bar{v}_g]\) since \(\varepsilon\) is small. Then \(m_g = R(k)\) and \(\hat{m}_g = R(\hat{k})\). From \([P_k]\) at \((v,s)\) and \((\hat{v}, s)\), either \(\hat{w}_{gg} > w_{gg}\) or \(\hat{w}_{gb} > w_{gb}\) or both hold. In addition, from \([B.1]\), \([B.2]\) and Lemma B.3, we must have \(Q_g(\hat{w}_g, g) < Q_g(w_g, g)\) and \(Q_b(\hat{w}_g, g) < Q_b(w_g, g)\). Now suppose \(\hat{w}_{gg} \leq w_{gg}\) and \(\hat{w}_{gb} > w_{gb}\). By the supermodularity and concavity of \(Q\), we know \(Q_g(\hat{w}_g, g) \geq Q_g(w_g, g)\), a contradiction. Similarly, \(\hat{w}_{gb} \leq w_{gb}\) and \(\hat{w}_{gg} > w_{gg}\) will also lead to a contradiction. So we must have \(\hat{w}_g \gg w_g\).
(c) By \([PK_b]\) at \((v,s)\) and the fact \(w_{bg} \geq w_{bb}\), we know \(v_b \geq \delta E^b[w_b] \geq \delta w_{bb}\). Now assume \(v_g > av_b\) and \(s = b\). We will show \(w_{bb} \leq v_b\). By the definition of \(a\), \(v_{bg} = a v_{bb}\).

We will verify in Theorem 3 that \(v_{bg} = \delta E^b[v_b]\), which then implies \(\delta E^b[(1, a)] = 1\) or \(a = 1 - \delta + p_b \delta\). Thus, \(v_g \geq av_b\) is equivalent to \(\delta E^b[v] \geq v_b\). And by \([PK_b]\), we know \(E^b[v] \geq E^b[w_b]\). Suppose \(w_{bb} > v_b\). Then it must be \(w_{bg} < v_g\). The strict concavity and supermodularity of \(Q\) in \(H\) then imply \(Q_b(w_b, b) < Q_b(v, b)\). However, from \([FOC_{wb}]\), \(Q_b(w_b, b) \geq Q_b(v, b)\) because \(\alpha(v, b) \geq 0\), a contradiction. Thus, we must have \(w_{bb} \leq v_b\). And by \([PK_b]\), \(w_{bb} \leq \delta E^b[w_b]\), which can be rewritten as \(w_{bg} \geq aw_{bb}\).

Proof of Proposition 5.4. When \(\Delta > \psi(p_g)\), Theorem 3 will verify that the contingent utility thresholds satisfy \(v_{bg} = \delta E^b[v_b]\) and \(v_b < v_g\). Now consider starting at state \((v, s) \in H\) and a bad shock occurs. We know from Lemma C.1 that \(w_b(v, s) \in H\). And by Lemma D.2, \(w_{bg} < v_{bg}\). Hence, \(w_{bg} \leq \delta E^b[v_b] < \delta E^b[v_g]\). Then by Proposition 5.2 (c), we know \(w_g[w_b(v, s), b] \notin E_g\). Or the contingent utility vector will not reach the constraint set \(E_g\) after another good shock. In other words, to reach \(E_g\), the firm needs to experience at least two good shocks in a row.

Proof of Lemma 5.5. (a) Take any \(v \in V\). To ease notation, let \(k = k(v, g), w_i = w_i(v, g)\), and denote \(\hat{k} = k(v, b), \hat{w}_i = w_i(v, b)\). Suppose \(\hat{k} > k\). Since \(\Gamma(v, g) = \Gamma(v, b)\), optimality at \((v, g)\) implies

\[
-k + p_g[R(k) + \delta Q(w_g, g)] + (1 - p_g)\delta Q(w_b, b) \\
\geq -\hat{k} + p_g[R(\hat{k}) + \delta Q(\hat{w}_g, g)] + (1 - p_g)\delta Q(\hat{w}_b, b)
\]

[D.6]

From part (a) of Lemma [B.2], \(Q(w_g, g) = Q(\hat{w}_g, g)\). Then [D.6] implies \(Q(w_b, b) > Q(\hat{w}_b, b)\). Moreover, optimality at \((v, b)\) implies

\[
-k + p_b[R(k) + \delta Q(w_g, g)] + (1 - p_b)\delta Q(w_b, b) \\
\geq -\hat{k} + p_b[R(\hat{k}) + \delta Q(\hat{w}_g, g)] + (1 - p_b)\delta Q(\hat{w}_b, b)
\]

[D.7]

Add [D.6] [D.7] and rearrange to get:

\[
\delta[Q(\hat{w}_b, b) - Q(w_b, b)] \geq R(\hat{k}) - R(k) > 0
\]

which forms a contradiction.

(b) This result is shown in Theorem 1 (f) of the Supplement Appendix.

To characterize the investment dynamics, we show in the following two lemmas that for a certain value of \(v_g\), there exists a cutoff value \(h_s(v_g)\) such that optimal investment is efficient if \(v_g \geq h_s(v_g)\) and inefficient if \(v_g < h_s(v_g)\).
Lemma D.3. For any \( v \in V \) with \( v_g \geq R(\bar{k}_s) + \frac{p_p v_b}{p_b}, k(v, s) = \bar{k}_s \).

Proof. Take any \( v \in V \) with \( v_g \geq R(\bar{k}_s) + \frac{p_p v_b}{p_b} \). Then \( v_g - v_b \geq R(\bar{k}_s) + \frac{\Delta v_b}{p_b} \). By \([PK_3]\) at \((v, s)\),

\[
\delta[p_b(w_{bg}(v, s) - w_{bb}(v, s)) + w_{bb}(v, s)] \leq v_b
\]

\[
\implies w_{bg}(v, s) - w_{bb}(v, s) \leq \frac{v_b}{\delta p_b}
\]

\[
\implies R(\bar{k}_s) + \delta \Delta(w_{bg} - w_{bb}) \leq R(\bar{k}_s) + \frac{\Delta v_b}{p_b} \leq v_g - v_b
\]

The second line is because \( w_{bb}(v, s) \geq 0 \). The last line implies it’s always feasible to invest efficiently, i.e. \((\bar{k}_s, m_t(v, s), w_i(v, s)) \in \Gamma(v, s)\). Hence, \( k(v, s) = \bar{k}_s \). 

Lemma D.4. For any \( v_b \geq 0 \), there exists \( h_s(v_b) > v_b \) such that \( v_g \geq h_s(v_b) \) implies \( k(v, s) = \bar{k}_s \), and \( v_b \leq v_g < h_s(v_b) \) implies \( k(v, s) < \bar{k}_s \). Moreover, \( h_s(v_b) \) is increasing and satisfies \( h_s(0) = R(\bar{k}_s), h_s(\bar{v}_s) \leq \bar{v}_g \).

Proof. Take any \( v_b > 0 \). If \( v_g \) is sufficiently close to \( v_b \), then investment \( k(v, s) \) will be sufficiently close to zero by \([IC^*]\). If \( v_g \) is sufficiently large, by Lemma D.3, we know \( k(v, s) = \bar{k}_s \). Hence, for a certain value of \( v_b \) we can define \( h_s(v_b) = \inf \{ v \in V : v_g \geq v_b, k(v, s) = \bar{k}_s \} \). Since \( k(v, s) \) is increasing in \( v_g \) by Lemma 5.5, we have \( k(v, s) = \bar{k}_s \) if \( v_g \geq h_s(v_b) \).

Consider \((v, s) = ((0, R(\bar{k}_s)), s)\). Since the only feasible \( w_b \) at \((v, s)\) is \( 0 \) and \( v_g = R(\bar{k}_s) \), we know \( k(v, s) = \bar{k}_s \). Since \((\bar{v}_s, s) = \bar{k}_s \), we know \( h_s(\bar{v}_s) \leq \bar{v}_g \). Now we show \( h_s(v_b) \) is increasing. Take any \( v_b, v_b' \) with \( v_b > v_b' \geq 0 \). Suppose \( h_s(v_b') < h_s(v_b) \). Then there exists \( \varepsilon > 0 \) such that \( h_s(v_b') < h_s(v_b) - \varepsilon \). By definition of \( h_s \),

\[
k[(v_b', h_s(v_b'))] = \bar{k}_s > k[(v_b, h_s(v_b)) - \varepsilon] \geq k[(v_b', h_s(v_b'))]
\]

The last line is because \( k(v, s) \) is decreasing in \( v_b \) by Lemma 5.5. Then we must have \( h_s(v_b') > h_s(v_b) - \varepsilon \). Otherwise, \( k[(v_b', h_s(v_b'))] \leq k[(v_b', h_s(v_b) - \varepsilon)] \) because \( k(v, s) \) is increasing in \( v_g \). This is a contradiction. Hence, \( h_s(v_b') \geq h_s(v_b) \)

To characterize the investment dynamics along optimal states, we show in the following three lemmas that whenever the contract evolves to the efficient investment region after a bad shock it will stay in that region until a good shock occurs. Moreover, note that we can use \([D.4]\) to rewrite \([FOC^k]\) as:

\[
[FOC^k] \quad R'(k(v, s)) = 1/[p_s - \alpha(v, s)]
\]

Lemma D.5. For any \((v, s) \in H\) such that \( v_g \geq h_s(v_b) \), \( Q_b(v, s) = Q_b[(v_b, h_s(v_b), s)] \) and \( Q_g(v, s) = Q_g[(0, v_g), s] \).
Proof. Since \( v_g \geq h_s(v_b) \geq h_s(0) \) by the monotonicity of \( h_s(\cdot) \), we know \( \alpha(v,s) = \alpha[(v_b,h_s(v_b)),s] = \alpha[(0,v_g),s] = 0 \). From Lemma B.2 (c), \( w_b(v,s) = w_b[(v_b,h_s(v_b)),s] \), which implies the left hand side of [C.10] is the same as state \((v,s)\) or \([(v_b,h_s(v_b)),s]\). Hence [C.10] implies \( Q_b(v,s) = \alpha[(v_b,h_s(v_b)),s] \). Similarly, Lemma B.2 (a) implies \( w_g(v,s) = w_g[(0,v_g),s] \). So the left hand side of [C.7] is the same at state \((v,s)\) or \([(0,v_g),s]\). Hence [C.7] implies \( Q_g(v,s) = Q_g[(0,v_g),s] \).

Lemma D.6. Fix \( s \in S, Q_b[(v_b,h_s(v_b)),s] \) strictly decreases in \( v_b \) on the set \([0,\delta\mathbb{E}^b[\hat{v}_b]]\).

Proof. Take any \( v_b < \hat{v}_b \leq \delta\mathbb{E}^b[\hat{v}_b] \). And let \( v = (v_b,h_s(v_b)) \) and \( \hat{v} = (\hat{v},h_s(\hat{v}_b)) \). By the definition of \( h_s, \alpha(v,s) = \alpha(\hat{v},s) = 0 \). Then from Lemma B.2 (c), we know that \( w_b(v,s), w_b(\hat{v},s) \) are solutions to problem \([P3]\) at \((v_b,b), (\hat{v}_b,b)\) respectively. In addition, from Lemma B.1 (b) and [C.10] we know

\[
(1 - p_s)\Psi_1(v_b,b) = (1 - p_s)D_{\beta(1,1)}Q[w_b(v,s),b] = Q_b(v,s)
\]

\[
(1 - p_s)\Psi_2(\hat{v}_b,b) = (1 - p_s)D_{\beta(1,1)}Q[w_b(\hat{v},s),b] = Q_b(\hat{v},s)
\]

From Lemma B.3, \( \Psi(\cdot,b) \) is strictly concave on the set \([0,\delta\mathbb{E}^b[\hat{v}_b]]\). Then the above equations imply that \( Q_b(\hat{v},s) < Q_b(v,s) \).

Lemma D.7. For any \((v,s) \in H, w_b = w_b(v,s)\) and \( w_b = w_b[w_b(v,s),b] \). If \( w_{bg} \geq h_b(w_{bb}) \), then \( w_{bg} \geq h_b(w_{bb}) \).

Proof. From Lemma C.3, \( w_b, w_b \in H \) and hence \( w_{bb}, w_{bb} < \hat{v}_{bb} \). We will verify in Lemma 7.1 that \( \hat{v}_{bb} = \delta\mathbb{E}^b[\hat{v}_b] \). Now take \( w_{bg} \geq h_b(w_{bb}) \) as the given fact and suppose \( w_{bg} < h_b(w_{bb}) \). In the following two cases, we show this assumption always leads to contradictions. First, consider the case of \( w_{bb} > w_{bb} \). We can obtain:

\[
Q_b(w_{bb},b) \leq Q_b[(w_{bb},h_b(w_{bb})),b] < Q_b[(w_{bb},h_b(w_{bb})),b] \leq Q_b(w_b,b)
\]

The first and the last inequality is by the supermodularity of \( Q \) and the assumptions. The second inequality is by Lemma D.6 and \( w_{bb} \geq w_{bb} \). However, \([FOCw_{bb}]\) at state \((w_b,b)\) implies \( Q_b(w_{bb},b) = Q_b(w_b,b) \) because \( \alpha(w_b,b) = 0 \), which forms a contradiction with above inequality.

Second, consider the case of \( w_{bb} \leq w_{bb} \). By the assumptions and by the monotonicity of \( h_b(\cdot) \), we know \( w_{bb} < h_b(w_{bb}) \leq h_b(w_{bb}) \leq w_{bg} \). Then we can obtain:

\[
Q_b[(0,w_{bg}),b] < Q_b[(0,w_{bb}),b] \leq Q_b(w_{bb},b)
\]

The first inequality is because \( Q(v,b) \) is strictly concave in \( v_g \) on the set \([(0,w_{bg}),(0,w_{bg})]\) by Lemma 2.3 of the Supplement Appendix. The second inequality is because of the
supermodularity of $Q$. Moreover, Lemma D.5 implies $Q_s([0,w_{bg}],b) = Q_b(b,w_b)$. Combining with [D.8] we know $Q_s(w_b,b) < Q_s(w_b^p,b)$. However, from the first order conditions we can derive

[D.9] \[ Q_s(w_b^p,b) = \frac{p_b}{1-p_b}Q_b(w_b,b) = \frac{p_b}{1-p_b}Q_b(v,s) + \frac{p_b(1-p_b)}{(1-p_b)(1-p_s)}\alpha(v,s) \leq \frac{p_b}{1-p_b}Q_b(v,s) + \frac{p_g}{1-p_s}\alpha(v,s) = Q_b(v,s), \]

The first equality is from $[\text{FOC}_w]$ at state $(w_b,b)$. The second and the last equalities are from $[\text{FOC}_b]$ and $[\text{FOC}_w]$ at state $(v,s)$ respectively. This forms a contradiction. □

**Corollary D.8.** Suppose $w_{bg} \geq h_b(w_{bb})$ in Lemma D.7. Then $w_{bb}^b = w_{bb}$ and $w_{bg}^b \geq w_{bg}$.

**Proof.** Suppose $w_{bb}^b \neq w_{bb}$. To ease notation, we denote $\hat{w}_b = (w_{bb}, h_b(w_{bb}))$ and $\hat{w}_b^b = (w_{bb}^b, h_b(w_{bb}^b))$. Since $w_{bg} \geq h_b(w_{bb})$ by assumption and $w_{bb} \geq h_b(w_{bb})$ by Lemma D.7, we know from Lemma D.5 that $Q_b(w_b,b) = Q_b(\hat{w}_b,b)$ and $Q_b(w_b^b,b) = Q_b(\hat{w}_b^b,b)$.

Because $w_{bb}^b \neq w_{bb}$, Lemma D.6 then implies $Q_b(w_b^b,b) \neq Q_b(\hat{w}_b^b,b)$, which further means $Q_b(w_b,b) \neq Q_b(w_b^b,b)$. This forms a contradiction with $[\text{FOC}_w]$ at state $(w_b,b)$ which implies $Q_b(w_b^b,b) = Q_b(\hat{w}_b^b,b)$ because $\alpha(w_b,b) = 0$. Thus, $w_{bb}^b = w_{bb}$.

Now suppose $w_{bg}^b < w_{bg}$. Then the strict concavity of $Q(v,s)$ in $v_g$ on the set $[w_{bb}^b, w_{bb}]$ (see Lemma 2.3 of the Supplement Appendix) implies $Q_s(w_{bb}^b,b) > Q_s(w_{bb},b)$ which violates [D.9]. Thus, we must have $w_{bg}^b \geq w_{bg}$. □

**Proof of Proposition 5.6.** (a) Let $w_g = w_g(v,s)$ and $w_g^g = w_g(v,s), g)$. From Lemma C.3, we know $w_g \notin E_g$ implies $w_g \in H$. So $Q_g(w_g,g) > 0$ and $D_{1(1)}Q(w_g,g) > 0$. First, let us consider the case $w_g^g \in E_g$. Then $Q_g(w_g^g,g)$ which is the left hand of $[\text{FOC}_w^g]$ at state $(w_g, g)$ is zero. So from the right hand side of $[\text{FOC}_w^g]$ we get $\alpha(w_g^g,g) = Q_g(w_g^g,g) > 0$, which implies $k(w_g,g) < \tilde{g}$ by $[\text{FOC}_g]$. Second, consider $w_g^g \notin E_g$. Then Lemma C.4 shows that $D_{1(1)}Q(w_g,g) > D_{1(1)}Q(w_g^g,g)$. Hence by [C.2] we know $\alpha(w_g,g) > 0$ which implies $k(w_g,g) < \tilde{g}$.

(b) We will verify in Lemma 7.1 that $\Delta \leq \psi(p_g)$ implies $\bar{v}_{bb} > \frac{\delta p_k R(\bar{k}_b)}{1-\delta |1-\delta \Delta|}$, and $\Delta > \psi(p_g)$ implies $\bar{v}_{bb} > \frac{\delta p_k R(\bar{k}_b)}{1-\delta |1-\delta \Delta|}$. (i) Consider the former case where $\Delta \leq \psi(p_g)$. Take $(v,s) \in H$ and let $w_b = w_b(v,s), w_b^b = w_b[w_b(v,s), b]$. Suppose $w_{bg} \geq h_s(w_{bb})$. Then from $[\text{PK}_b]$ and $[\text{IC}^*]$ at state $(w_b,b)$ and from Corollary D.8, we get

\[
w_{bb} \geq \delta E^b[w_b^b] \geq \delta E^b[w_b] \\quad w_{bg} - w_{bb} \geq R(\bar{k}_b) + \delta \Delta[w_{bg} - w_{bb}] \geq R(\bar{k}_b) + \delta \Delta[w_{bg} - w_{bb}] \]

54
which together imply \( w_{bb}(v, s) \geq \frac{\delta p_B R(\bar{k}_b)}{(1 - \delta)(1 - \delta \Delta)} = \bar{v}_{bb} \). However, this is a contradiction with \((w_b, b) \in H\) which requires \( w_{bb} < \bar{v}_{bb} \) by Lemma D.2 (a).

(ii) Now consider the latter case of \( \Delta > \psi(p_s) \). Theorem 3 in Section 7 will show that at \((\bar{v}_b, b)\) the constraint \([IC]\) holds as strict inequality. From Lemma D.2 (b) we know \((\bar{v}_s, s) \in cl(H)\), which means we can find a sequence \(\{(v^{(n)}, g)\}_{n=1}^\infty \in H\) that converges to \((\bar{v}_g, g)\). Since \(w_b(\bar{v}_g, g) = \bar{v}_b\) and \(w_b(\cdot, g)\) is continuous, there exists \(N\) such that \([IC]\) holds as strict inequality at state \([w_b(v^{(n)}, g), b]\) if \(n \geq N\). This means \(k[w_b(v^{(n)}, g)] = \bar{k}_b\) if \(n \geq N\). In other words, there must exit a neighborhood \(F \subset \{(v, b): v \in V, \bar{v}_b - (\varepsilon, \varepsilon) < v < \bar{v}_b\} \cap H\) for some sufficiently small \(\varepsilon > 0\) such that \(k(v, b) = \bar{k}_b\) if \((v, b) \in F\). Moreover, contingent utility vector can possibly evolve to this neighborhood \(F\) after a bad shock.

\[ \square \]

### E. Proofs from Section 6

Define the process \((v^{(t)}, s_{t-1})_{t=0}^\infty\) to be the states induced by the optimal contract starting at some \((v^{(0)}, s_{-1}) \in V \times S\). In the high persistence case the optimal contract may need enough good shocks (at least two) to reach efficient sets. So to establish the result that efficient sets are achieved in finite time, we need to consider sequences with two good shocks in a row happen infinitely often. We begin with some preliminary lemma.

**Lemma E.1.** The non-negative martingale \(\left( D_{(1,1)} Q(v^{(t)}, s_{t-1}) \right)_{t=0}^\infty\) converges to 0 almost surely.

**Proof.** By Lemma C.1, the process \(\left( D_{(1,1)} Q(v^{(t)}, s_{t-1}) \right)_{t=0}^\infty\) is a nonnegative martingale. Then, the Martingale Convergence Theorem — see, for instance, Theorem 1 and Corollary 3 on pp 508–509 of Shiryaev (1995) — ensures \(D_{(1,1)} Q(v^{(t)}, s_{t-1})\) converges almost surely to a non-negative and integrable random variable.

Because \(D_{(1,1)} Q(v^{(t)}, s_{t-1})\) is a non-negative martingale, it follows immediately that if \(D_{(1,1)} Q(v^{(t)}, s_{t-1}) = 0\) for some (possibly random) time \(\tau\), then \(D_{(1,1)} Q(v^{(t)}, s_{t-1}) = 0\) at all subsequent times \(t \geq \tau\). Therefore, in order to show that \(D_{(1,1)} Q(v^{(t)}, s_{t-1})\) converges to 0 almost surely, it suffices to consider the dates and states where \(D_{(1,1)} Q(v^{(t)}, s_{t-1}) > 0\).

Consider a path where good shocks occur infinitely many times and let \(a = \lim_{t \to \infty} D_{(1,1)} Q(v^{(t)}, s_{t-1})\). Suppose \(a > 0\). Then following the observation in the previous paragraph, we find that \(D_{(1,1)} Q(v^{(t)}, s_{t-1}) > 0\) for all \(t \geq 0\). Lemma C.3 then implies \(w_b(v^{(t)}, s_{t-1}) \in H\) for all \(t \geq 0\).

Now consider the subsequence on this path that has only good shocks. This subsequence stays in \(H\) forever. Moreover, by Lemma D.2 (a), this subsequence is bounded and must have a convergent subsequence \((v^{(t)}, g)_{t=0}^\infty\) with limit \((\bar{v}, g)\). By definition,
$s_{t_{\tau-1}} = g$, and it is either the case that $s_{t_{\tau}} = b$ infinitely often (ie, the ‘good-bad’ shock pair occurs infinitely often), or the case that $s_{t_{\tau}} = g$ infinitely often (ie, the ‘good-good’ shock pair occurs infinitely often) along this subsequence.

In the former case, let $(v(\tau'), g)_{\tau'=0}^{\tau}$ be a subsequence of $(v(\tau'), g)_{\tau'=0}^{\tau}$ with $s_{\tau'-1} = g$ and $s_{\tau'} = b$. This means $v(\tau'+1) = w_b(v(\tau'), g)$. The continuity of $w_b(\cdot, g)$ implies $\lim_{\tau' \to \infty} v(\tau'+1) = w_b(\hat{v}, g)$. Moreover, because $D_{(1,1)}Q$ is continuous,

$$\lim_{\tau' \to \infty} D_{(1,1)}Q(v(\tau'), g) = D_{(1,1)}Q(\hat{v}, g) = a > 0 \quad \text{and} \quad \lim_{\tau' \to \infty} D_{(1,1)}Q(v(\tau'+1), b) = D_{(1,1)}Q(w_b(\hat{v}, g), b) = a > 0$$

But now recall Lemma C.4, which implies that if $D_{(1,1)}Q(\hat{v}, g) > 0$, then we know $D_{(1,1)}Q(\hat{v}, g) < D_{(1,1)}Q(w_b(\hat{v}, g), b)$, which contradicts the displayed equations above.

In the latter case, we can use the same argument to show that $D_{(1,1)}Q(\hat{v}, g) = D_{(1,1)}Q(w_g(\hat{v}, g), g) = a$. But this case also contradicts Lemma C.4. Therefore, we must have $a = 0$.

Finally, by Proposition 3.2 of the Supplement Appendix, paths with only finitely many good shocks have measure zero, so we have $\lim_{\tau \to \infty} D_{(1,1)}Q(v(t), s_{t-1}) = 0$ almost surely.

**Proof of Theorem 2.** Lemma E.1 shows that the martingale $(D_{(1,1)}Q(v(t), s_{t-1}))_{t=0}^{\infty}$ converges to 0 almost surely. We now show that this convergence occurs in finite time almost surely. Towards this end, consider a path with the property that $D_{(1,1)}Q(v(t), s_{t-1}) > 0$ for all $t < \infty$ and where the ‘good-good’ shock pair occurs infinitely often. Take a subsequence $(v(\gamma), s_{\gamma-1})_{\gamma=0}^{\infty}$ such that $s_{\gamma-1} = s_{\gamma} = g$. From Lemma C.3, $(v(\gamma), s_{\gamma-1}) \in H$ for all $t < \infty$ along this path.

Let $\varepsilon > 0$ be some sufficiently small number. Note the set $C := \{v \in V : (v, g) \in \text{cl}(H), v_g \leq \bar{v}_{gg} - \varepsilon\}$ is a compact set that. Also, we know from Lemma D.2 (b) that $(\bar{v}, s)$ are the only points in $\text{cl}(H)$ that have zero directional derivative. The continuity of $D_{(1,1)}Q(\cdot, g)$ implies that there exists some $\eta > 0$ such that $D_{(1,1)}Q(v, g) \geq \eta$ for all $v \in C$. Moreover, because $\lim_{\tau \to \infty} D_{(1,1)}Q(v(\gamma), g) = 0$ by Lemma E.1, there must exist some $T$ such that $v(\gamma) \not\in C$ for any $t \geq T$. In other words, $v_{gg} - v_{g}^{(\gamma)} < \varepsilon$ for all $t \geq T$.

But this means $(v^{(\gamma)}, s_{\gamma-1}) \not\in A_{1,g}$, because $\varepsilon$ is sufficiently close to zero. And since $s_T = g$ by construction, we must have $(v^{(\gamma)}, s_{\gamma-1}) \in E_g$ and $D_{(1,1)}Q(v^{(\gamma)}, g) = 0$, a contradiction.

In Proposition 3.2 of the Supplement Appendix, we show that the path with infinitely many ‘good-good’ shocks occurs almost surely. Hence, the argument above shows $D_{(1,1)}Q(v(t), s_{t-1}) = 0$ for some $t < \infty$ almost surely. Let $\tau = \min \{t : D_{(1,1)}Q(v(t), s_{t-1}) = 0\}$. From Proposition 5.1, we know $(v^{\tau}, s_{\tau-1}) \in E_{g_{s_{\tau-1}}}$ for any $t \geq \tau$. So in a maximum rent contract, the contingent utilities $v^{\tau} = \bar{v}_{s_{\tau-1}}$ for all $t \geq \tau$. 

56
F. Proofs from Section 7

In this section, we show the level of threshold contingent utilities and repayments of the unconstrained firm.

For a given $p_g$, we define a pair $(\varphi(p_g), \hat{k}(p_g))$ on $(0, 1]$ implicitly by

\[ \varphi(p_g)R'[\hat{k}(p_g)] = 1 \]

\[ R[\hat{k}(p_g)] = \frac{\delta p_gR(\bar{k}_g)}{1 + \delta p_g} \]

where $\bar{k}_g$ is defined as the unique $k$ that satisfies $p_g R'(k) = 1$. It is clear that for a given $p_g$, the pair $(\varphi(p_g), \hat{k}(p_g))$ is unique.

**Lemma F.1.** The functions $\hat{k}$ and $\varphi$ are continuous. For any $p_g > 0$, we have $0 < \varphi(p_g) < p_g$, and $\varphi(p_g)$ increases in $p_g$. Moreover, $\lim_{p_g \to 0} \varphi(p_g) = 0$.

*Proof.* It is clear from [F.2] that $\hat{k}$ is continuous and $\hat{k}(0) = 0$. Continuity of $\varphi$ for all $p_g > 0$ follows immediately from [F.1]. Because $R'(0) = \infty$, we see that $\lim_{p_g \to 0} \varphi(p_g) = 0$.

From [F.2], $\hat{k}(p_g) < \bar{k}_g$, since $\frac{\delta p_g}{1 + \delta p_g} < 1$. Concavity of $R$ then implies that $R'[\hat{k}(p_g)] > R'[\bar{k}_g]$. From [F.1], $\varphi(p_g)R'[\hat{k}(p_g)] = p_g R'(\bar{k}_g) = 1$, implying that $0 < \varphi(p_g) < p_g$. Moreover, $\frac{\delta p_g R[\bar{k}_g]}{1 + \delta p_g}$ increases in $p_g$. Hence, $\hat{k}(p_g)$ and $\varphi(p_g) = 1/R'(\hat{k}(p_g))$ both increase in $p_g$. \[\square\]

Using the cutoff function $\varphi(p_g)$, we can partition the space of transition probability $\{ (p_b, p_g) : 0 < p_b \leq p_g < 1 \}$ as

- $B_- = \{ p : \varphi(p_g) < p_b \leq p_g \}$
- $B_+ = \{ p : p_b = \varphi(p_g) < p_g \}$
- $B_+ = \{ p : p_b < \varphi(p_g) < p_g \}$

The set $B_-$ characterizes the case of low persistence, while the set $B_+$ characterizes the case of high persistence.

**Lemma F.2.** The partition of transition probabilities can be characterized as follows:

(a) $p \in B_-$ if, and only if, $\delta p_g[R(\bar{k}_g) - R(\bar{k}_b)] < R(\bar{k}_b)$.

(b) $p \in B_-$ if, and only if, $\delta p_g[R(\bar{k}_g) - R(\bar{k}_b)] = R(\bar{k}_b)$.

(c) $p \in B_+$ if, and only if, $\delta p_g[R(\bar{k}_g) - R(\bar{k}_b)] > R(\bar{k}_b)$.

*Proof.* To see (a), notice that from [F.1], we find $\varphi(p_g)R'[\hat{k}(p_g)] = p_b R'(\bar{k}_b) = 1$. Then $p \in B_-$ implies $R'[\hat{k}(p_g)] > R'(\bar{k}_b)$. Concavity of $R$ then implies $\hat{k}(p_g) < \bar{k}_b$. From [F.2], $R[\hat{k}(p_g)] = \frac{\delta p_g R(\bar{k}_g)}{1 + \delta p_g} < R(\bar{k}_b)$. We can rearrange the last inequality to obtain $\delta p_g[R(\bar{k}_g) - R(\bar{k}_b)] < R(\bar{k}_b)$.

57
Conversely, if we know \( \delta p_g[R(\hat{k}_g) - R(\hat{k}_b)] < R(\hat{k}_b) \), then \( R(\hat{k}_b) > \frac{\delta p_g R(\hat{k}_g)}{1 + \delta p_g} = R[\hat{k}(p_g)] \) by [F2]. Hence, \( \hat{k}_b > \hat{k}(p_g) \) which implies \( R'[\hat{k}(p_g)] > R'(\hat{k}_b) \). But \( \hat{k}_b \) (uniquely) satisfies \( p_g R'(\hat{k}_b) = 1 \), which implies \( \varphi(p_g) < p_b \), ie, \( p \in B_- \). The other parts are established by similar arguments.

Now, we can define \( \psi(p_g) := p_g - \varphi(p_g) \) for a given \( p_g \) and rewrite the above partitions as \( B_- = \{ p : \Delta < \psi(p_g) \} \), \( B_0 = \{ p : \Delta = \psi(p_g) \} \), and \( B_+ = \{ p : \Delta > \psi(p_g) \} \).

**Lemma F.3.** At state \((\hat{v}_s, s)\), the constraints [IC] and [LL] for \( s = g \) cannot both hold as strict inequalities.

**Proof.** Suppose not. Then at state \((\hat{v}_s, s)\), we have

\[
\hat{v}_{sg} - \hat{v}_{sb} > R(\hat{k}_s) + \delta \Delta(\hat{v}_{bg} - \hat{v}_{bb}), \quad R(\hat{k}_s) > \hat{m}_{sg}
\]

So there exist some \( \varepsilon_{IC}, \varepsilon_{LL} > 0 \) such that

\[
(\hat{v}_{sg} - \varepsilon_{IC}) - \hat{v}_{sb} = R(\hat{k}_s) + \delta \Delta(\hat{v}_{bg} - \hat{v}_{bb}), \quad R(\hat{k}_s) = \hat{m}_{sg} + \varepsilon_{LL}
\]

Let \( \varepsilon := \min[\varepsilon_{IC}, \varepsilon_{LL}] \), and using [PKb] at state \((\hat{v}_s, s)\), notice that

\[
(\hat{v}_{sg} - \varepsilon) - \hat{v}_{sb} \geq R(\hat{k}_s) + \delta \Delta(\hat{v}_{bg} - \hat{v}_{bb}), \quad R(\hat{k}_s) \geq \hat{m}_{sg} + \varepsilon
\]

with at least one of them holding as equality.

Let \( \hat{v}_s = (\hat{v}_{sg} - \varepsilon, \hat{v}_{sb}) \), and \( \hat{m}_{sg} = \hat{m}_{sg} + \varepsilon \). Then the policy \((\hat{k}_s, \hat{m}_{sb}, \hat{m}_{sg}, w_b = \hat{v}_b, w_g = \hat{v}_g) \in \Gamma(\hat{v}_s, s)\). The contracts at \((\hat{v}_s, s)\) and at \((\hat{v}_s, s)\) only differ in their transfers, which implies that \( Q(\hat{v}_s, s) = \hat{Q}(s) \). However, \( \hat{v}_s < \hat{v}_s \) by constructions, which implies that \( \hat{v}_s \notin E_s \), by part (a) of Proposition 4.2. This forms a contradiction because part (d) of Proposition 4.2 implies that for such \((\hat{v}_s, s)\), we must necessarily have \( Q(\hat{v}_s, s) < \hat{Q}(s) \). \( \square \)

**Lemma F.4.** In the mature firm at state \((\hat{v}_s, s)\), the agent gets zero payment when a bad shock occurs in the current period, ie, \( \hat{m}_{sb} = 0 \).

**Proof.** Suppose \( \hat{m}_{sb} < 0 \). Then there exists some \( \varepsilon > 0 \) such that \( \hat{m}_{sb} + \varepsilon < 0 \). From [PKb] at state \((\hat{v}_s, s)\) we obtain

\[
\hat{v}_{sb} - \varepsilon = - (\hat{m}_{sb} + \varepsilon) + \delta \mathbb{E}[\hat{v}_b]
\]

From [IC*] at state \((\hat{v}_s, s)\) we find

\[
\hat{v}_{sg} - (\hat{v}_{sb} - \varepsilon) > \hat{v}_{sg} - \hat{v}_{sb} \geq R(\hat{k}_s) + \delta \Delta(\hat{v}_{bg} - \hat{v}_{bb})
\]

58
Let \( \hat{v}_s = (\bar{v}_{sg}, \bar{v}_{sb} - \varepsilon) \) and \( \hat{m}_{sb} = \bar{m}_{sb} + \varepsilon \). Then the policy \((\bar{k}_s, \bar{m}_{sb}, \bar{m}_{sg}, w_b = \bar{v}_b, w_g = \bar{v}_g) \in \Gamma(\hat{v}_s, s)\). The contracts at \((\hat{v}_s, s)\) and at \((\bar{v}_s, s)\) only differ in their transfers, which implies that \( Q(\hat{v}_s, s) = \bar{Q}(s) \). However, \( \bar{v}_s < \hat{v}_s \) by constructions, which implies that \( \hat{v}_s \notin \mathcal{E}_s \), by part (a) of Proposition 4.2. This forms a contradiction because part (d) of Proposition 4.2 implies that for such \((\hat{v}_s, s)\), we must necessarily have \( Q(\hat{v}_s, s) < \bar{Q}(s) \).

Lemma F.5. The threshold contingent utilities satisfy: \( \bar{v}_{bb} = \bar{v}_{bg}, \bar{v}_{gb} < \bar{v}_{gg} \).

Proof. The right hand side of \([PK_s]\) at \((\hat{v}_s, s)\) is not contingent on \( s \) because \( \bar{m}_{sb} = 0 \). So we must have \( \bar{v}_{sb} = \delta \bar{v}_{gb}^{bb}[\bar{v}_b] \), which implies that \( \bar{v}_{gb} = \bar{v}_{bb} \). Using this relation and \( \bar{m}_{sb} = 0 \), we can rewrite \([IC]\) at \((\hat{v}_s, s)\) as

\[
[F.3] \quad \bar{m}_{sg} \leq \delta p_g(\bar{v}_{gg} - \bar{v}_{bg})
\]

Moreover, subtracting \([PK_s]\) at state \((\hat{v}_b, b)\) from \([PK_s]\) at state \((\bar{v}_g, g)\), we can get:

\[
[F.4] \quad (\bar{m}_{gg} - \bar{m}_{bg}) = R(\bar{k}_g) - R(\bar{k}_b) - (\bar{v}_{gg} - \bar{v}_{bg})
\]

Suppose \( \bar{v}_{gg} \leq \bar{v}_{bg} \). Then \([F.3]\) implies that \( \bar{m}_{gg}, \bar{m}_{bg} \leq 0 \). In addition, \([F.4]\) implies that \( 0 < R(\bar{k}_g) - R(\bar{k}_b) \leq \bar{m}_{gg} - \bar{m}_{bg} \). So we must have \( \bar{m}_{bg} < \bar{m}_{gg} \leq \delta p_g(\bar{v}_{gg} - \bar{v}_{bg}) \leq 0 \), by \([F.3]\). However, this means \( \bar{m}_{bg} < R(\bar{k}_b) \) and \( \bar{m}_{bg} < \delta p_g(\bar{v}_{gg} - \bar{v}_{bg}) \). In other words, \([IC]\) and \([LL]\) for \( g \) both hold as strict inequality at \((\bar{v}_b, b)\), a contradiction of Lemma F.3.

Proof of Theorem 3. We have already shown \( \bar{m}_{sb} = 0 \) in Lemma F.4. So we are only left to show the conditions satisfied by \( \bar{m}_{sg} \). We proceed the proof by the following four parts.

(i) We first establish that \( \bar{m}_{gg} = \delta p_g(\bar{v}_{gg} - \bar{v}_{bg}) < R(\bar{k}_g) \). Suppose \( \bar{m}_{gg} = R(\bar{k}_g) \). From \([F.4]\), \( \bar{m}_{bg} - R(\bar{k}_b) = \bar{v}_{gg} - \bar{v}_{bg} > 0 \), which means \( \bar{m}_{bg} > R(\bar{k}_b) \). But this violates \([LL]\) for \( b \). So we must have \( \bar{m}_{gg} < R(\bar{k}_g) \). From Lemma F.3, \([IC]\) must hold as equality at \((\bar{v}_g, g)\). Therefore, from display \([F.3]\), we must have \( \bar{m}_{gg} = \delta p_g(\bar{v}_{gg} - \bar{v}_{bg}) \).

(ii) We now show that \( \bar{m}_{bg} = \delta p_g(\bar{v}_{gg} - \bar{v}_{bg}) < R(\bar{k}_b) \) when \( p \in B_- \). Suppose, instead, that \( \bar{m}_{bg} = R(\bar{k}_b) \). By \([F.4]\),

\[
\bar{v}_{gg} - \bar{v}_{bg} = R(\bar{k}_g) - \bar{m}_{gg} = R(\bar{k}_g) - \delta p_g(\bar{v}_{gg} - \bar{v}_{bg})
\]

The last equality is from part (i) above. We then obtain

\[
\delta p_g(\bar{v}_{gg} - \bar{v}_{bg}) = \frac{\delta p_g R(\bar{k}_g)}{1 + \delta p_g} < R(\bar{k}_b) = \bar{m}_{bg}
\]

The inequality is from Lemma F.2 when \( p \in B_- \). However, comparing this display with \([F.3]\), we find \( \bar{m}_{bg} > \delta p_g(\bar{v}_{gg} - \bar{v}_{bg}) \), ie, \([IC]\) at \((\bar{v}_b, b)\) is violated. Therefore, we must have \( \bar{m}_{bg} < R(\bar{k}_b) \). By Lemma F.3, \([IC]\) must hold as equality at \((\bar{v}_b, b)\). Hence, from \([F.3]\), we obtain \( \bar{m}_{bg} = \delta p_g(\bar{v}_{gg} - \bar{v}_{bg}) \).
(iii) Next, we show that $\bar{m}_{bg} = R(\bar{k}_b) < \delta p_g(\bar{v}_{gg} - \bar{v}_{bg})$ when $p \in B_+$. Suppose, to the contrary, that $\bar{m}_{bg} = \delta p_g(\bar{v}_{gg} - \bar{v}_{bg})$. Then, from part (i) above, we have $\bar{m}_{bg} = \bar{m}_{gg}$. And display [F.4] implies $\bar{v}_{gg} - \bar{v}_{bg} = R(\bar{k}_g) - R(\bar{k}_b)$. So $\bar{m}_{bg} = \delta p_g[R(\bar{k}_g) - R(\bar{k}_b)] > R(\bar{k}_b)$, where the inequality is from Lemma F.2 when $p \in B_+$. However, this means [LL] for $g$ is violated at $(\bar{v}_b, b)$, a contradiction. Therefore, we must have $\bar{m}_{bg} < \delta p_g(\bar{v}_{gg} - \bar{v}_{bg})$, by [F.3]. This means the [IC] holds as strict inequality at $(\bar{v}_b, b)$. By Lemma F.3, [LL] for $g$ must hold as equality at $(\bar{v}_b, b)$, which means $\bar{m}_{bg} = R(\bar{k}_b)$.

(iv) Finally, we show that $\bar{m}_{bg} = R(\bar{k}_b) = \delta p_g(\bar{v}_{gg} - \bar{v}_{bg})$ when $p \in B_-$. In this case, the same procedure as in part (ii) and (iii) above shows that $\bar{m}_{bg} = R(\bar{k}_b)$ and $\bar{m}_{bg} = \delta p_g(\bar{v}_{gg} - \bar{v}_{bg})$ imply one another. So the only way that Lemma F.3 is satisfied is to have both [IC] and [LL] for $g$ both hold as equality at $(\bar{v}_b, b)$. Hence, the conclusion holds. 

Proof of Lemma 7.1. From [PKₐ] and [PKₜ] at $(\bar{v}_b, b)$ and $(\bar{v}_g, g)$ respectively, we can get

[F.5] $\bar{v}_b = \delta E^b[\bar{v}_b]$

[F.6] $\bar{v}_{bg} = R(\bar{k}_b) - \bar{m}_{bg} + \delta E^g[\bar{v}_g]$

[F.7] $\bar{v}_{gg} = R(\bar{k}_g) - \bar{m}_{gg} + \delta E^g[\bar{v}_g]$

where $\bar{v}_{sb} = \bar{v}_b$, because $\bar{m}_{sb} = 0$.

(a) When $\Delta \geq \psi(p_g)$, or equivalently $p \in B_+$ or $B_-$, Theorem 3 implies:

[F.8] $\bar{m}_{bg} = R(\bar{k}_b), \quad \bar{m}_{gg} = \delta p_g(\bar{v}_{gg} - \bar{v}_{bg})$

Combining [F.5] to [F.8] and eliminating $\bar{m}_{sg}$, we can obtain the specified solution of $\bar{v}_b, \bar{v}_{bg}, \bar{v}_{gg}$ in the Lemma.

(b) When $\Delta \leq \psi(p_g)$, or equivalently $p \in B_-$ or $B_-$, Theorem 3 implies:

[F.9] $\bar{m}_{bg} = \bar{m}_{gg} = \delta p_g(\bar{v}_{gg} - \bar{v}_{bg})$

Combining [F.5] [F.6] [F.7] [F.9] and eliminating $\bar{m}_{sg}$, we can obtain the specified solution in the Lemma.

\[\Box\]
G. Proofs from Section 8

Lemma G.1. When the firm is unconstrained, the present values of expected cash flows are:

\[ Z_{bg} = \frac{R(\tilde{k}_b) + p_g \delta [R(\tilde{k}_g) - R(\tilde{k}_b)]}{1 - p_g \delta} \]

\[ Z_{gg} = Z_{bg} + R(\tilde{k}_g) - R(\tilde{k}_b) \]

\[ Z_{bb} = Z_{gb} = \frac{\bar{p}_b \delta}{1 - \delta + \delta \bar{p}_b} Z_{bg} \]

Proof. We simply solve \( Z_{ab}, Z_{sg} \) from [8.1] to [8.3] and obtain \([Z_{bg}]\) to \([Z_{ab}]\). □

Proof of Proposition 8.1. Note that \( K \geq Z_{bg} > Z_b \). The first inequality is because \( \bar{m}_{gg} \geq \bar{m}_{bg} \). The second inequality is from \([Z_{ab}]\). And from \([Z_{gg}]\), we get

\[ Z_{gg} - K = [R(\tilde{k}_g) - \bar{m}_{gg}] - [R(\tilde{k}_b) - \bar{m}_{bg}] > 0 \]

because \( R(\tilde{k}_g) > \bar{m}_{gg} \) and \( R(\tilde{k}_b) \geq \bar{m}_{bg} \) by Theorem 3. So the stock option is in the money only when good-good shock occurs.

Now let us consider the total compensation to the agent. When a bad shock occurs today, the firm has zero cash flow and obviously the agent gets zero payment. When a good shock occurs today, depending on yesterday’s shock \( s \), the agent’s total compensation is

\[ \lambda [R(\tilde{k}_s) - \max(Z_{sg} - K, 0)] + \max(Z_{sg} - K, 0) \]

\[ = \lambda R(\tilde{k}_s) + (1 - \lambda) \max(Z_{sg} - K, 0) \]

We show in the following cases that the total compensation in \([G.2]\) is equal to what the agent obtains in the optimal contract, ie \( R(\tilde{k}_s) - \bar{m}_{sg} \).

First, consider \( s = b \). Since the option is out of money, \([G.2]\) reduces to \( \lambda R(\tilde{k}_b) \) which equals \( R(\tilde{k}_b) - \bar{m}_{bg} \) by the construction of \( \lambda \).

Second, consider \( s = g \). We can rewrite \([G.2]\) as

\[ \lambda R(\tilde{k}_g) + (1 - \lambda)(Z_{gg} - K) \]

\[ = \lambda R(\tilde{k}_g) + (1 - \lambda)[Z_{gg} - Z_{bg} - (\bar{m}_{gg} - \bar{m}_{bg})] \]

\[ = \lambda R(\tilde{k}_g) + (1 - \lambda)[R(\tilde{k}_g) - \bar{m}_{gg} - (R(\tilde{k}_b) - \bar{m}_{bg})] \]

\[ = R(\tilde{k}_g) + (1 - \lambda)[\bar{m}_{bg} - \bar{m}_{gg} - R(\tilde{k}_b)] \]

The first line is by \([G.1]\). The second line is simply plugging in the value of \( K \). And the third line is by \([Z_{gg}]\). If \( \Delta \leq \psi(p_g) \), then \( \bar{m}_{bg} = \bar{m}_{gg} \) by Theorem 3. So \([G.3]\) reduces to
\[ R(\bar{k}_g) - (1 - \lambda)R(\bar{k}_b) = R(\bar{k}_g) - \bar{m}_{gg} \] by plugging in \( \lambda \). If \( \Delta > \psi(p_g) \), then \( R(\bar{k}_b) = \bar{m}_{bg} \) by Theorem 3. So \( \lambda = 0 \) and [G.3] reduces to \( R(\bar{k}_g) - \bar{m}_{gg} \).

In all cases, the agent’s total compensation is equivalent to what he obtains in the optimal contract. 

**Proof of Corollary 8.2.** From Proposition 8.1, the agent’s equity share \( \lambda = 1 - \frac{\bar{m}_{bg}}{R(\bar{k}_b)} \).

Then by Theorem 3, \( \lambda = 0 \) when \( p \in B_+ \); and \( \lambda = 1 - \delta p_g \frac{R(\bar{k}_g)}{R(\bar{k}_b)} - 1 \) when \( p \in B_- \).

The conclusion holds trivially in the former case. Consider the latter case. As \( \Delta \) increases, either \( p_g \) increases or \( p_b \) decreases or both. This means either \( \bar{k}_g \) increases or \( \bar{k}_b \) decreases or both, which means \( \frac{R(\bar{k}_g)}{R(\bar{k}_b)} \) strictly increases. Hence, \( \lambda \) strictly decreases in \( \Delta \). 

**Proof of Theorem 4.** We only need to show that the agent has no incentive to draw down available credit to payout or misreport cash flow. First, at any time \( t \), the agent will not draw down available credit to either pay out or consume. He at most gets \( C_L - M + R(k) \) or \( \bar{v}_b - M \) depending on which shock occurs. But these contingent values are equal to the payoffs that the agent can obtain from waiting until the credit balance is fully repaid before issuing payout: \( v_b \), or \( v_g \).

Second, the agent has no incentive to misreport cash flow as well. By misreporting at time \( t \), the agent can divert \( R(k) \). Since the account balance and credit limit will evolve according to [8.6], [8.7] and [8.10], the available credit in period \( t + 1 \) will be \( (w_{bb}, w_{bg}) \). So the expected continuation value of misreporting is \( R(k) + \delta [p_g w_{bg} + (1 - p_g) w_{bb}] \). And by the incentive compatible condition in the optimal contract, the agent must get no less by truthfully reporting cash flow.
References


Larcker, David F. 2008. ‘Executive Compensation and Incentives’. *Stanford Graduate School of Business*. (Cit. on p. 5).


