A Theory of Dynamic Contracting with Financial Constraints*

Ilia Krasikov Rohit Lamba

December 2017

First draft: December 2016

Abstract

We study a dynamic principal-agent model where the agent has access to a persistent private technology but is strapped for cash. Financial constraints are generated by the periodic interaction between incentives (private information) and a strong notion of feasibility (being strapped for cash). This interaction produces dynamic distortions that are a sum of two effects: backloading of incentives and illiquidity. Bad technology shocks increase distortions and monotonically push the optimal contract further away from efficiency. An endogenous number of good shocks is required for the contract to become liquid, and then eventually efficient. Efficiency is an absorbing state that is reached almost surely. Persistence of private information increases the variance of total economic surplus generated by the model, and decreases the rate at which surplus converges to its efficient value. The key predictions continue to hold in the continuous time setting. A simple economic implementation of the optimal contract is also provided.

1 Introduction

There is a large empirical literature explaining inability of start-ups or small businesses to raise capital. For example, Campello et al. [2010] conducted a survey of 1050 CFOs across the US, Europe and Asia, and found considerable impact of financial constraints on firm behavior in the aftermath of the Great Recession. Banerjee and Duflo [2014] exploit a change in policy by the Indian government to show that most firms in their study were credit constrained, and a relaxation of the same led to a spurt in growth. In an elegant note Nobuhiro Kiyotaki advocates "a mechanism design approach to illustrate how different environments of private information and limited commitment generate different financial frictions," with an aim to provide theoretical constructs to the wide prevalence of financial constraints.[1] In the spirit of the said agenda, this paper posits financial constraints as a product of the interaction between (i) persistent private information, and (ii) limitations on the ability of agents to generate timed cash flows.

*Krasikov: Pennsylvania State University, izk113@psu.edu; Lamba: Pennsylvania State University, rlamba@psu.edu. We are indebted to Nageeb Ali for his advice and encouragement. Thanks also to Felix Bierbrauer, Aislinn Bohren, Jinwen Wang, Nima Haghpanah, Vijay Krishna, George Mailath, Bruno Strulovici, and to the seminar participants at Northwestern University, University of Pittsburgh, Pennsylvania economic theory conference, Vanderbilt mechanism design conference, Midwest economic theory conference at University of Michigan and SED conference in Edinburgh for their comments.

[1] Kiyotaki [2012].

Most mechanism design models that feature repeated interactions have invoked intertemporal notions of individual rationality as the feasibility constraint, wherein every period the sum of agent's current and future expected payoffs is restricted to be positive (see surveys by Vohra [2012], Krähmer and Strausz [2015a] and Bergemann and Välimäki [2017]). The agent can therefore be required to pay large upfront capital with the promise of being repaid later. This allows the principal to almost unbridledly backload payments and relax incentive constraints. However, in many real situations, eg. in supply contracts, managerial compensation, provision of public goods and regulation, the agent may not be able to borrow money or forgo payments. We model this stronger feasibility constraint by restricting the agent's stage (or per-period) utility to be positive.

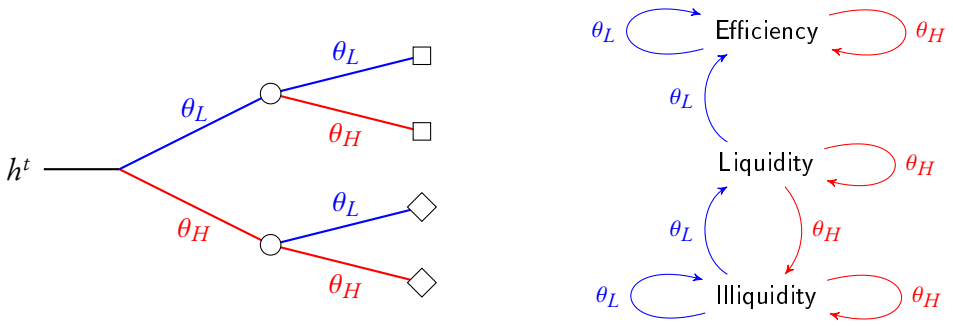
The aforementioned restriction on payoffs has been studied in the dynamic financial contracting literature (see surveys by Biais, Mariotti, and Rochet [2013], and Sannikov [2013]), motivated therein primarily as a limited liability constraint. Most of this literature focusses on moral hazard or cash flow diversion with iid technologies as the driving agency friction. In contrast, this paper looks at a dynamic adverse selection (or screening) model with persistence in technology, which is a natural assumption in long-term contracting from an empirical perspective.[2] We will discuss in detail how the predictions of our model change as a function of persistence.

Following Kiyotaki [2012]'s cue, we term the economic force generated by the interaction of this stronger feasibility restriction and private information as the *financial constraint*. The big picture question then is- when do these financial constraints bind and when they do, what dynamic inefficiencies do they generate? In asking and then trying to answer this question, we add to the dynamic mechanism design literature a deeper understanding of the role of financial constraints, and to the financial contracting literature a deeper understanding of the role of persistence in agency frictions. On the desire to unify the key economic ideas across these two bodies of work Sannikov [2013] writes "While several common themes emerge, in general there is no unified way to analyze settings of dynamic adverse selection and moral hazard, and this area is ripe for future research." This paper seeks to take a step in that direction by providing a tractable framework that delivers clear insights on the short-run and long-run outcomes of contracting in a dynamic screening model with endogenously binding financial constraints.

The results of the paper can be divided into four parts. First is the structure of the optimal contract, the degree of inefficiency in the evolution of allocation and economic surplus. Second is a plausible mechanism that implements the optimal contract through working capital and an eventual takeover. Third is unpacking the economic content of the novel elements- (i) stronger feasibility requirement versus no restriction on transfers, and (ii) the role of persistence in agency frictions in generating financial constraints. Fourth is the relationship between our discrete time model and its continuous time counterpart. We expound upon each after briefly describing the model.

The formal model is as follows. A big firm (principal) repeatedly producing a final good contracts with a smaller firm (agent) that supplies an important input. Each period, the small firm privately observes either a low ("good shock") or high ("bad shock") marginal cost; after being drawn from a prior, costs evolve according to an exogenous two state Markov process. The small firm lacks capital- it cannot borrow money or collateralize its assets. In simple terms it is *strapped for*

[2]For example, İmrohoroğlu and Tüzel [2014] find the average persistence in total factor productivity of firms in Compustat data from 1962 to 2009 to be 0.7.



(a) $q(\theta_H|h^t,\theta_L) > q(\theta_H|h^t) > q(\theta_H|h^t,\theta_H)$ (b) Evolution of the optimal contract

Figure 1: The optimal contract

cash. Mathematically, this restriction demands positivity of stage utility of the small firm. The big firm is tasked with designing a contract which sets supply of inputs by the small firm, and payments for its production.

Structure of the optimal contract. A Pareto-optimal contract chooses allocations and transfers that satisfy incentive compatibility and cash-strapped constraints to maximize the profit of the big firm while ensuring a minimum ex ante payoff for the small firm. Figure 1a depicts a typical sequence of technology shocks. For a history of costs realizations h^t and current cost θ_i, let $q(\theta_i|h^t)$ and $U(\theta_i|h^t)$ be the allocation and expected utility of the small firm. At this point, if the marginal cost of incentive provision is zero, then $q(\theta_i|h^t) = q^e(\theta_i)$, that is the (statically) efficient quantity is supplied. If it is positive, then $q(\theta_i|h^t) = q^e(\theta_i) - d(\theta_i|h^t)$ where d measures the history dependent optimal distortion. As is standard, the low cost type always supplies the efficient quantity: $q(\theta_L|h^t) = q^e(\theta_L)$.[3] On the other hand, each "bad shock" increases optimal distortions: $q(\theta_H|h^t,\theta_H) < q(\theta_H|h^t) < q^e(\theta_H)$. This is in striking contrast to dynamic mechanisms without financial constraints that emphasize progressively decreasing distortions along all histories (see Besanko [1985] and Battaglini [2005]) or on average (see Garrett, Pavan, and Toikka [2017]). In addition, the realization of a "good shock" decreases the optimal distortion: $q(\theta_H|h^t) < q(\theta_H|h^t,\theta_L)$. An endogenous number of consecutive "good shocks", say $n(h^t)$, is required for the optimal distortion to reach zero. For every additional "bad shock", as distortions increase, the number increases: $n(h^t,\theta_H) \geq n(h^t)$. Once the optimal distortion reaches zero it stays at zero, that is, efficiency is an absorbing state. In the long run, the efficient contract is supplied almost surely.[4]

With reference to Figure 1a, the expected utilities of both the low and high cost types go up after a "good shock" and go down after a "bad shock". That is, as long as the contract is inefficient, $(U(\theta_L|h^t,\theta_H),U(\theta_H|h^t,\theta_H)) \leq (U(\theta_L|h^t),U(\theta_H|h^t)) \leq (U(\theta_L|h^t,\theta_L),U(\theta_H|h^t,\theta_L))$. Two thresholds on the vector of expected utilities divide the evolution of the optimal contract into three regions- illiquidity, liquidity and efficiency; see Figure 1b. The contract typically starts in the illiquid region- both incentive and cash-strapped constraints bind to produce financial constraints

[3] A low cost realization is better for economic surplus than a high cost realization.

[4] Beyond these qualitative properties, we pin down the optimal limit contract in closed form, as the Markov process governing agent's evolution of types converges to the identity matrix. The analysis provide intuition for the structure of the optimal dynamic contract with highly persistent agency frictions.

that bite. A low cost type either keeps the contract in illiquidity or can transition it to liquidity. A high cost type decreases the expected utility of the small firm which keeps it illiquid. After an endogenous number of low cost realizations, the expected utility of the small firm reaches a critical threshold at which the big firm agrees to lax the cash-strapped constraint and provide the small firm with some credit. This is called the liquid region. Liquidity is not an absorbing state, a high cost realization can push the small firm back into illiquidity. The liquid region forms a penultimate zone towards efficiency. Once liquid, the realization of one more low cost pushes expected utility of the small firm beyond the second threshold, which propels the optimal contract into the absorbing state of efficiency.[5]

At a technical level, we use a mixture of sequential and recursive approaches to characterize the optimal contract. A novelty we bring to the table is the existence of a "shell", a subset of the recursive domain which houses the optimal constrained contract. The recursive domain is too large to make crisp predictions about the evolution of the contract. We show that as long as the optimal contract is inefficient, the expected utility of the agent must always lie in this shell. To the best of our knowledge, this is a new feature of dynamic contracts. It allows us to show all the aforementioned monotonicity properties of the evolution of the optimal contract. We also provide a simple price-theoretic explanation of the construction of the shell.

An economic implementation. At the end of every period, the two dimensional vector of the continuation utility of the big firm (its profit) and that of the small firm (its promised rent) constitutes the *capital structure* of the "meta firm" borne out of the economic partnership. The sum of the continuation utilities constitutes the economic surplus, termed as the net present value of the meta firm. The value of the meta firm and its capital structure evolve endogenously with the realization of technology shocks.[6]

In the illiquid region, the cash-strapped constraint binds and the big firm only provides *working capital* to the small firm. Through a sequence of consecutive low cost realizations, the small firm has to earn its way into liquidity. In the liquid region, the big firm promises to *take-over* the small firm on the realization of one more low cost type for a determinable strike price. Thereafter, the small firm operates in-house, producing the efficient quantity.

The value of the meta firm is increasing in the share of the agent. A "good shock" decreases the optimal distortion which in turn increases the information rent and surplus. A "bad shock" on the other hand, increases optimal distortions which reduces the information rent of the agent and *downsizes* the meta firm. Therefore, as the agent assumes a greater stake of the total surplus, incentives align with bargaining power to reduce agency frictions and increase the size of pie. The efficient contract marks a mature "meta firm" that has been able to overcome financial constraints-both incentive and cash-strapped constraints are slack.

The paper provides a framework for the design of real contracts. In the face of uncertain technology and credit constraints, contracts should provide incentives through delayed payments. The

[5]Complimentarily, in a cash-flow diversion model of dynamic contracting, Fu and Krishna [2017] also establish a penultimate zone towards efficiency. The monotonicity of the allocation, expected utility and number of shocks required to get to efficiency, though, are all unique to our paper.

[6]In the corporate finance view of our model, the Modigilliani-Miller Theorem does not hold since capital structure matters for the value of the firm. However, given efficiency is attained almost surely in the long-run, the theorem does hold asymptotically,

terms should be flexible- they should sequentially deteriorate in the event of bad outcomes, and improve with good outcomes. After a certain number of "successes", the "agent" should be deemed credit worthy, and if he continues to show results, the "principal" should take-over the technology and promote it in-house.

The role of financial constraints and persistence of private information. Allowing for a long-term contract helps mitigate the problem of agency frictions- this mitigation is achieved by backloading incentives. Financial constraints, though, restrict the extent of backloading. Dynamic distortions in our framework are an additive sum of two effects: backloading of incentives and illiquidity due to financial constraints; the latter increases with each "bad shock", overturning the standard result of decreasing distortions. Efficiency, perhaps surprisingly, is still a certainty, though the path towards it is much more constrained in comparison to the standard dynamic mechanism design model sans financial constraints.

We also provide a conceptual foundation for a productive role of limited liability for small businesses. We ask- when should the positivity of stage utility of the small firm be interpreted as a limited liability constraint (as opposed to a credit constraint)? The answer depends on whether the agent does better in the benchmark model without financial constraints or in our model with financial constraints. Consider the principal profit maximizing contract on the Pareto frontier in which the big firm has all the bargaining power. The ex ante expected utility of the small firm from the contract is determined endogenously as part of the optimum. We show that in the iid limit the ex ante expected utility of the agent is higher in our model than in the benchmark, and in the perfectly persistent limit the ranking depends on other parameters. Broadly, the benchmark model may produce a higher utility for the agent for highly persistent private information and low levels of discounting. When the agent is reasonably patient, being "protected by limited liability" helps him in a utilitarian sense and is a meaningful interpretation.

Finally, we take this model as representative of firm dynamics in an economy with financial constraints and show how persistence makes a difference to the predictions in comparison to an iid model. We make three broad points. The fraction of financially constrained firms in the short-run is monotonically increasing in the persistence of technology shocks. The average rate at which firms converge to the state of being unconstrained is decreasing in persistence- the iid model predicts a quick dissolution of financial constraints. And, variance in the total value of both constrained and unconstrained firms is larger with persistence. The standard dynamic financial contracting literature that operates in the iid world would miss all these, empirically important, comparative statics.

The model in continuous time. As a bridge between the literatures on dynamic mechanism design and dynamic financial contracting, we write down our model in continuous time as well- most models in the former are formulated in discrete time settings (see Bergemann and Strack [2015] for a exception) while those in the latter typically adopt continuous time techniques. We do not operate in the usual Brownian setting (see for example Williams [2011]), rather we employ the continuous time analog of the two state discrete time Markov process. The solution to the problem is described by two Hamiltonian-Jacobi-Bellman equations.

All key features from the discrete time setting continue to hold except one- liquidity is now

synonymous with efficiency. There is no penultimate zone where the cash-strapped constraint slacks before the contract becomes efficient. This is quite intuitive since there is no notion of a "period" in the continuous time setting. The two thresholds in the discrete time model, separated by the realization of an additional "good shock", merge as time between interactions goes to zero. The optimal contract still lives in a shell, and is characterized by monotonicity of allocation and expected utility. Moreover, the continuos time counterpart of the cash-flow diversion model (for example Clementi and Hopenhayn [2006]) is analogous to our model, which helps unify results across the two approaches.

2 Model

The key economic forces in dynamic contracting with persistent private information can be formulated through various related models. We choose the repeated version of the marginal cost screening model, based on Laffont and Martimort [2002]. A big firm (principal) specializing in a final good requires a non-durable input that is produced by a smaller firm (agent) every period at a cost θq, where θ is the small firm's private information.[7] The principal values the final good at $V(q)$, where $V : \mathbb{R}_+ \to \mathbb{R}_+$ satisfies the Inada conditions.[8] She pays a price p to the agent for supplying her the intermediate good (or input), and the utility of both is linear in the price.[9] Therefore, the per-period (or stage) utility for the principal and agent is given by $V(q) - p$ and $p - \theta q$, respectively. The contract lasts for T periods, where $T \leq \infty$. There is a common discount factor δ.

The marginal cost θ, oft referred to as the agent's type, can take on two values: $\Theta = \{\theta_L, \theta_H\}$, where $0 < \theta_L < \theta_H$. It is drawn from a prior $\mu = (\mu_L, \mu_H)$, and then evolves according to a Markov process: $f(\theta_L|\theta_i) = \alpha_i$, and $f(\theta_H|\theta_i) = 1-\alpha_i$ for $i = H, L$. Distributions have full support: $\mu >> 0$ and $f >> 0$. The Markov process is assumed to be "persistent": $\alpha_L \geq \alpha_H$, and for simplicity of exposition, from hereon we will assume a symmetric Markov process: $\alpha_L = 1 - \alpha_H = \alpha \geq \frac{1}{2}$. In the appendix, we consider the general asymmetric case.

At the start of every period, the agent privately learns his marginal cost, θ. Given the Markovian nature of shocks, θ is also informative about future types. Both the principal and the agent can commit to a dynamic contract. Invoking the revelation principle, therefore, we can focus on the direct mechanism. Every period the agent reports his marginal type to the principal. The principal offers a menu of history dependent price-quantity pairs to the agent. Her objective is to maximize her expected profit subject to incentive and feasibility constraints. We solve for the entire Pareto frontier by introducing as a parameter the agent's minimum ex ante share of the total economic surplus, v_0. The set of parameters is thus given by $\Gamma = \{V(.), \Theta, \mu, \alpha, \delta, v_0\}$.

Formally the mechanism is: $m = \langle \mathbf{p}, \mathbf{q} \rangle = \left(p\left(\hat{\theta}_t \left| h^{t-1} \right.\right), q\left(\hat{\theta}_t \left| h^{t-1} \right.\right) \right)_{t=1}^{T}$, where h^{t-1} and $\hat{\theta}_t$ are, respectively, the history of reports up to $t-1$ and current report at time t.[10] The reported history h^t is recursively defined as $h^t = \left\{ h^{t-1}, \hat{\theta}_t \right\}$ starting with $h^0 = \emptyset$. The set of possible histories at

[7] We can introduce a fixed cost of production: $c(\theta, q) = \theta q + F$ without changing any of our results. For simplicity it is normalized to zero: $F = 0$.

[8] Technically: (i) $V'(q) > 0, V''(q) < 0$ for all $q \geq 0$, (ii) $V(0) = 0$, (iii) $\lim_{q \to 0} V'(q) = \infty$, $\lim_{q \to \infty} V'(q) = 0$.

[9] Throughout, the principal will be referred to as a 'she' and the agent as a 'he'.

[10] At the cost of minimal confusion, subscripts will be used interchangeably for time and L/H. Moreover, as is standard, the contract is restricted to lie in l^∞.

time t is denoted by H^t. Define the private history of the agent to be $h_A^t = \left\{h_A^{t-1}, \theta_t, \hat{\theta}_{t-1}\right\}$, starting from $h_A^0 = \{\theta_1\}$, where $\hat{\theta}_t$ and θ_t are the reported and actual types, respectively. Fixing the set of parameters Γ, for a given direct mechanism m, we have a dynamic decision problem described by $\langle m, \Gamma\rangle$ in which the strategy for the agent, $(\sigma_t)_{t=1}^T$, is simply a function that maps his private history into an announcement every period: $h_A^t \mapsto \sigma_t(h_A^t) \in \Theta$.

Define the stage and expected utility of the agent (under truthful reporting) after any history of the contract tree to be

$$u(\theta_t|h^{t-1}) = p(\theta_t|h^{t-1}) - \theta_t q(\theta_t|h^{t-1})$$

$$U(\theta_t|h^{t-1}) = u(\theta_t|h^{t-1}) + \delta\mathbb{E}\left[U(\tilde{\theta}_{t+1}|h^{t-1}, \theta_t) \mid \theta_t\right]$$

It is straightforward to show that the contract space can equivalently be expressed as $\langle \mathbf{u}, \mathbf{q}\rangle$ or $\langle \mathbf{U}, \mathbf{q}\rangle$. We shall use the three formulations interchangeably.

The constraints on the space of contracts can be divided into two categories- incentives and feasibility. The contract $\langle \mathbf{U}, \mathbf{q}\rangle$ is said to be *incentive compatible* if truthful reporting is profitable for the agent. Using the one shot deviation principle, formally, $\forall\ h^{t-1} \in H^{t-1}\ \forall\ t$:

$$U(\theta_t|h^{t-1}) \geq p(\hat{\theta}_t|h^{t-1}) - \theta_t q(\hat{\theta}_t|h^{t-1}) + \delta\mathbb{E}\left[U(\tilde{\theta}_{t+1}|h^{t-1}, \hat{\theta}_t) \mid \theta_t\right]$$

for all $\theta_t, \hat{\theta}_t \in \Theta$. The Markovian assumption on stochastic evolution of types ensures that the agent wants to report truthfully even if he has lied in the past.

Two types of feasibility constraints are explored in the paper- individual rationality and strapped for cash. A contract is said to be *individually rational* if it offers each type of the agent a non-negative expected utility after every history. Formally:

$$U(\theta_t|h^{t-1}) \geq 0 \quad \forall\ \theta_t \in \Theta,\ h^{t-1} \in H^{t-1}, \forall\ t$$

And, the contract is said to be *cash-strapped* if it must provide each type of the agent a non-negative stage utility at every history. Formally:

$$u(\theta_t|h^{t-1}) \geq 0 \quad \forall\ \theta_t \in \Theta,\ h^{t-1} \in H^{t-1}, \forall\ t$$

Individual rationality keeps the expected "lifetime" utility of the agent in every period above an exogenous level normalized to zero. However, the agent can be asked to forgo payments or deposit upfront capital with a promise of being compensated for it later. The cash-strapped constraint precludes contracts with such delayed promises; it is a restriction on the magnitude of per period transfers.[11] We want the reader to view it primarily as a credit constraint. The small firm needs regular payments to produce the intermediate good and cannot pledge any collateral to borrow money.[12,13]

[11] Note that if a contract is strapped for cash it necessarily satisfies individual rationality while the opposite is not necessarily true.

[12] An alternate way to express the cash-strapped constraint would be $p_t \geq C$ for all t, that is payment to the small firm or the agent has to be above a minimum constant amount every period. We allow the boundary to move with the supply contract, that is $p_t \geq \theta_t q_t$ for tractability. Think of a third party that disburses short term credit to the smaller firm that is able to verify its type and amount of production. Our results can be generalized to the constraint $p_t \geq C$.

[13] Just as the incentive constraint, the cash-strapped constraint holds both "on" and "off" path. Even if the agent may

This constraint has been invoked in various guises in the financial contracting literature, especially in models of moral hazard and cash flow diversion (see for example Clementi and Hopenhayn [2006], Biais et al. [2007], and Myerson [2012]). It is motivated therein primarily as a limited liability restriction. We provide a comparison between the two interpretations- limited liability and being credit constrained- in Section 5.2.

3 Optimal contract

In order to appreciate the role of each of moving part in the model, and simplify the technical exposition that follows, we divide this section into six parts. First, we briefly exposit the optimal contract in the benchmark model with individual rationality. Second, we solve the two-period model with financial constraints to communicate the key forces in a simple way. Third, extending the sequential characterization, we point to an informative Lagrangian approach. Fourth, we solve for the optimal contract using a recursive formulation of the problem. Fifth, we put all the insights together and present the main theorem that provides a precise characterization of the optimal contract. And, sixth, we pin down the optimal limit contract as types become perfectly persistent.

Define $s(\theta, q) = V(q) - \theta q$ to be the static surplus, succinctly expressed as $s(\theta) = V(q(\theta)) - \theta q(\theta)$ for the direct mechanism. It is straightforward to note that the *efficient quantity* that maximizes the surplus is given by $V'(q^e(\theta)) = \theta$. Moreover, let $\overline{S} = \sum_{t=1}^{T} \delta^{t-1} \mathbb{E}\left[s(\tilde{\theta}_t)\right]$ be the (ex ante) expected surplus. The principal's problem, $(\mathcal{P}^*)$, can be stated as:

$$(\mathcal{P}^*) \qquad \max_{\langle \mathbf{U},\mathbf{q} \rangle} \quad \overline{S} - [\mu_L U(\theta_L) + \mu_H U(\theta_H)]$$

subject to $\mathbf{q} \geq 0$,

$$\text{(PK):} \quad \mu_L U(\theta_L) + \mu_H U(\theta_H) \geq v_0, \text{ and}$$

$$IC_L(h^{t-1}), IC_H(h^{t-1}), C_L(h^{t-1}), C_H(h^{t-1}) \; \forall \; h^{t-1} \in H^{t-1} \; \forall \; t$$

where (PK) is the ex ante promise keeping constraint, and $IC_i(h^{t-1})$ and $C_i(h^{t-1})$ are the incentive and cash-strapped constraints, respectively, for type θ_i in period t after history h^{t-1}. Note that v_0 parameterizes the bargaining power of the agent, and maps the Pareto frontier. Since quantity is always non-negative at the optimum, we shall drop that constraint.

We consider a relaxed problem where we ignore $IC_H(h^{t-1})$ for all histories. A justification of this is provided in Section 9.10, including sufficient conditions for global optimality. The principal's relaxed problem, $(\mathcal{RP}^*)$, reads as follows:

$$(\mathcal{RP}^*) \qquad R^*(v_0) = \max_{\langle \mathbf{U},\mathbf{q} \rangle} \quad \overline{S} - [\mu_L U(\theta_L) + \mu_H U(\theta_H)]$$

subject to (PK), and

$$IC_L(h^{t-1}), C_L(h^{t-1}), C_H(h^{t-1}) \; \forall \; h^{t-1} \in H^{t-1} \; \forall \; t$$

where $R^*(v_0)$ is the value of the objective- the principal's profit at the constrained optimum. The ex ante economic surplus generated by the optimal contract shall be denoted by $S^*(v_0) = R^*(v_0) +$

have misreported in the past, the principal delivers a non-negative stage utility to him if he is truthful today.

$\max\{\underline{v}, v_0\}$.

The Myersonian quest herein is to write down an optimization problem equivalent to $(\mathcal{RP}^*)$ where a subset of binding incentive constraints is used to eliminate $\mathbf{U}$, and the objective and all remaining constraints are expressed only in terms of $\mathbf{q}$. Pointwise optimization of allocations along all histories then yields the efficient quantity for the low cost type: $q(\theta_L|h^{t-1}) = q^e(\theta_L)$, and for the high cost type:

$$\underbrace{\mathbb{P}(h^{t-1}, \theta_H)\left(V'(q(\theta_H|h^{t-1})) - \theta_H\right)}_{\text{marginal benefit}} = \underbrace{r(h^{t-1})}_{\text{marginal cost}} \qquad (*)$$

where the left hand side of equation $(*)$ represents the expected marginal benefit of allocating quantity $q(\theta_H|h^{t-1})$, and right hand side represents the marginal cost of incentive provision (or information rent) at history h^{t-1}. The optimal allocation is implicity defined by the function $Q : \mathbb{R}_+ \to \mathbb{R}_+$:

$$V'(Q(x)) = \theta_H + x\Delta\theta \qquad (1)$$

where $x(h^{t-1})\Delta\theta = \frac{r(\theta_H^{t-1})}{\mathbb{P}(h^{t-1},\theta_H)}$ is the optimal distortion, which we represent succinctly as $q(\theta_H|h^{t-1}) = q^e(\theta_H) - d(\theta_H|h^{t-1})$. Note that $d \uparrow x$, that is d increases with an increase in x. We shall precisely characterize r and hence x, as a function of the parameters of the model, which in turn pins down the evolution of optimal quantities and expected utilities.

3.1 Benchmark: dynamic model without financial constraints

Before presenting novel results with the cash-strapped constraint, we consider the model where the agent has access to deep pockets. The problem looks exactly the same as $(\mathcal{P}^*)$, except that $C_i(h^{t-1})$ is replaced by $IR_i(h^{t-1})$ for $i = L, H$ and for all h^{t-1}. Dynamic contracting models of this form have been studied amongst others by Courty and Li [2000], Battaglini [2005] and Pavan, Segal, and Toikka [2014]. In our framework, the optimal allocation, $\mathbf{q}^\#$, is characterized by two facts.

Facts. Let $\theta_H^t \in H^t$ represent the history where each report until period t has been θ_H. The optimal allocation in the benchmark model:

1. becomes efficient forever as soon as the agent becomes a low cost type: $q^\#(\theta_L|h^{t-1}) = q^e(\theta_L)$ $\forall\ h^{t-1}$, and $q^\#(\theta_H|h^{t-1}) = q^e(\theta_H)$ $\forall\ h^{t-1} \neq \theta_H^{t-1}$
2. has decreasing distortions along the constant high cost history: $q^\#(\theta_H|\theta_H^{t-1}) = q^e(\theta_H) - d(\theta_H|\theta_H^{t-1})$, where $d(\theta_H|\theta_H^{t-1})$ is decreasing in t.

Battaglini [2005] terms these *generalized no distortion at the top* and *vanishing distortions at the bottom*, respectively. Drawing from equation $(*)$, $r(h^{t-1}) = 0$ for all $h^{t-1} \neq \theta_H^{t-1}$. Once a "good shock" arrives, the marginal cost of incentive provision becomes zero. On the other hand, $d(\theta_H|\theta_H^{t-1})$ decreases over time since the marginal cost of incentive provision decreases along the history of constant high costs which we'll refer to often as the "lowest history".[14]

The key economic force to take note here is that the feasibility constraint binds only once- IR_H is the only individual rationality constraint that binds at the optimum. Therefore, multiple prices

[14]Formal details of the facts are provided in the appendix in Section 9.1.

implement the optimal allocation with a restriction that first period expected utilities- $U(\theta_L)$ and $U(\theta_H)$- are uniquely pinned down. How much is this instrument of free movement of trnasfers across time exploited by the principal?

3.2 Two period model

Consider problem ($\mathcal{RP}^*$) for $T = 2$. In addition to (PK), we have to consider the following set of constraints:

$$IC_L, C_L, C_H, \text{ and } IC_L(\theta_i), C_L(\theta_i), C_H(\theta_i) \text{ for } i = L, H$$

It can be shown that $C_L(\theta_L)$ and $C_L(\theta_H)$ are implied by the other constraints. IC_L, $IC_L(\theta_H)$, C_H and $C_H(\theta_H)$ all bind at the optimum, and C_L can bind sometimes. Moreover, $IC_L(\theta_L)$ and $C_H(\theta_L)$ can be assumed to hold as an equality, they bind if C_L does. Using the set of binding constraints, we want to express $\mu_L U(\theta_L) + \mu_H U(\theta_H)$ as a function of quantities. IC_L is the key first period constraint which binds at the optimum:

$$\begin{aligned} U(\theta_L) &= \Delta\theta q(\theta_H) + u(\theta_H) + \delta\left[\alpha u(\theta_L|\theta_H) + (1-\alpha)u(\theta_H|\theta_H)\right] \\ &= U(\theta_H) + \Delta\theta q(\theta_H) + \delta(2\alpha - 1)\left[u(\theta_L|\theta_H) - u(\theta_H|\theta_H)\right] \end{aligned}$$

The term $(2\alpha - 1)$ is essentially the impact of misreport by agent on his expected utility in period 2. Using the second period binding incentive constraint $IC_L(\theta_H)$, we can rewrite this equation in the form of an "envelope formula":

$$\frac{U(\theta_L) - U(\theta_H)}{\Delta\theta} = q(\theta_H) + \delta\left(\frac{2\alpha - 1}{\alpha}\right)\alpha q(\theta_H|\theta_H) \tag{2}$$

Equation (2) is a mini version of a much more general formula derived for continuous type spaces in Pavan, Segal, and Toikka [2014].[15] The term $\left(\frac{2\alpha-1}{\alpha}\right)$ has been referred to variedly in the literature as the informativeness measure, impulse response and dynamic distortion.

In the benchmark model, the binding IR_H constraint would deliver zero expected utility for the high cost type. However, in the presence of the cash-strapped constraint we have[16]

$$U(\theta_H) = \delta(1-\alpha)\Delta\theta q(\theta_H|\theta_H) \tag{3}$$

Through equations (2) and (3), it is clear that being strapped for cash interacts with the incentive constraint to ensure information rents need to be paid to both the low and high cost types. Additively:

$$\begin{aligned} \mu_L U(\theta_L) + \mu_H U(\theta_H) &= \Delta\theta\mu_L q(\theta_H) + \delta\Delta\theta\left(\mu_L\alpha + \mu_H(1-\alpha)\right) q(\theta_H|\theta_H) \\ &= \Delta\theta\mathbb{P}(\theta_1 = \theta_L) q(\theta_H) + \delta\Delta\theta\mathbb{P}(\theta_2 = \theta_L) q(\theta_H|\theta_H) \end{aligned} \tag{4}$$

where $\mathbb{P}(\theta_t = \theta_L)$ is the ex ante probability of being the low cost type in period t. Define the

[15] Battaglini and Lamba [2017] derive the same formula for discrete types. Esö and Szentes [2017] also have a nice derivation of the result for continuous types.

[16] Equation (3) is generated by the binding C_H, $C_H(\theta_H)$ and $IC_L(\theta_H)$ constraints: $U(\theta_H) = u(\theta_H) + \delta\left[(1-\alpha)u(\theta_L|\theta_H) + \alpha u(\theta_H|\theta_H)\right] = \delta(1-\alpha)\Delta\theta q(\theta_H|\theta_H)$.

threshold generated by equation (4) for the efficient quantity to be $\overline{v}$:

$$\overline{v} = \Delta\theta \sum_{t=1}^{2} \delta^{t-1}\mathbb{P}(\theta_t = \theta_L) q^e(\theta_H|\theta_H^{t-1}) = \Delta\theta \sum_{t=1}^{2} \delta^{t-1}\mathbb{P}(\theta_t = \theta_L) q^e(\theta_H) \tag{5}$$

and that generated by the optimal contract when we ignore (PK) to be $\underline{v}$.[17]

Finally, C_L can be expressed as follows[18]

$$C_L: \quad q(\theta_H) + \delta\alpha q(\theta_H|\theta_H) \geq \delta\alpha q(\theta_H|\theta_L) \tag{6}$$

If ignored, parameters leading to large distortions along the high cost history would ensure that C_L is violated. So we introduce this as a constraint in terms of quantities. We can now pin down the optimal allocation rule. The principal chooses $\mathbf{q}$ to maximize $\overline{S} - [\mu_L U(\theta_L) + \mu_H U(\theta_H)]$ subject to (PK) and C_L, where $\mu_L U(\theta_L) + \mu_H U(\theta_H)$ is given by equation (4). The precise closed form solution is provided in Section 9.2 in the appendix. Here we deliver the basic economic message.

Proposition 1. *The optimal contract, $\mathbf{q}^*$, with promised utility $v_0 \in [0, \overline{v})$, is characterized by the following allocation rule:*

1. $q^*(\theta_L|h) = q^e(\theta_L)$ *for* $h = \emptyset, \theta_L, \theta_H$.
2. $q^*(\theta_H|\theta_L) = q^e(\theta_H) - d(\theta_H|\theta_L)$, *where* $d(\theta_H|\theta_L) \geq 0$, *and* $d(\theta_H|\theta_L) > 0 \Leftrightarrow C_L$ *binds.*
3. $q^*(\theta_H|h) = q^e(\theta_H) - d(\theta_H|h)$ *for* $h = \emptyset, H$, *where* $d(\theta_H|\theta_H) > d(\theta_H) > 0$.

It always profitable to supply the efficient quantity to the low cost type for the marginal cost of this provision is zero. Using the framework of equation $(*)$, for the high cost type, the marginal cost of incentive provision (and hence dynamic distortion) is an additive sum of two economic forces:

$$r(h) = \underbrace{\text{backloading of incentives}(h)}_{r_1(h):\text{benchmark marginal cost}} + \underbrace{\text{financial constraints}(h)}_{r_2(h):\text{added marginal cost}}$$

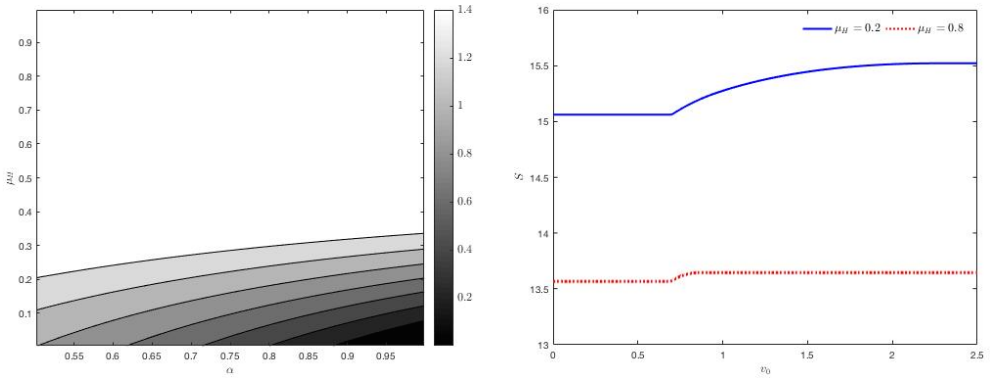
where $h = \emptyset, \theta_L, \theta_H$. In the benchmark model, $r_2(h) = 0 \; \forall \; h$. Since backloading of incentives is costless after a "good shock", $r_1(\theta_L) = 0$. With financial constraints, however, $r_2(\theta_L) > 0$ when C_L binds. Hence, $q^*(\theta_H|\theta_L) < q^e(\theta_H)$ and generalized no distortion at the top does not hold. Next, $r_2(\emptyset) < r_2(\theta_H)$, that is the impact of financial constraints strengthens for consecutive "bad shocks", culminating into $q^*(\theta_H|\theta_H) < q^*(\theta_H) < q^e(\theta_H)$. This overturns the vanishing distortions at the bottom result to increasing distortions at the bottom.[19]

When does C_L bind? Equation (6) clearly establishes that low values of $q(\theta_H)$ and $q(\theta_H|\theta_H)$ would violate C_L. To compensate, we must simultaneously distort $q(\theta_H|\theta_L)$ downwards, and $q(\theta_H)$ and $q(\theta_H|\theta_H)$ upwards, in proportion to the shadow price imposed by the constraint. Figure 2a shows the parametric range for which C_L binds. In the $\mu_H \times \alpha$ rectangle, it plots $q^*(\theta_H|\theta_L)$- shades

[17] $\underline{v}$ refers to the agent's expected utility on the principal profit maximizing point of the Pareto frontier. If $v_0 < \underline{v}$, then (PK) does not bind.

[18] Because $u(\theta_L) = U(\theta_L) - \delta\left[\alpha u(\theta_L|\theta_L) + (1-\alpha)u(\theta_H|\theta_L)\right] = \Delta\theta q(\theta_H) - \delta\Delta\theta\alpha\left[q(\theta_H|\theta_H) - q(\theta_H|\theta_L)\right]$.

[19] We have $r_1(\emptyset) > r_1(\theta_H)$ and $r_2(\emptyset) < r_2(\theta_H)$. But, in the aggregate $\frac{r(\emptyset)}{\mu_H} < \frac{r(\theta_H)}{\mu_H\alpha}$, which using equation $(*)$ ensures that distortions increase at the "lowest" history.



(a) $q^*(\theta_H|\theta_L)$ as a function of μ_H, α (b) S^* as a function of v_0

Figure 2: $q^*(\theta_H|\theta_L)$ and $S^*(v_0)$ for the two period model

represent numerical values as shown on the vertical key on the right. Darker the shade, the greater is the optimal distortion, and the lightest region corresponds to the efficient quantity. It is clear that low values of μ_H and high values of α lead to the largest liquidity crunch.[20]

What about the utility of the agent? It is obvious from equation (2) that the first period expected utility of the low cost type is higher than that of the high cost type. However, once a low or high cost is realized, we can say more about how the vector of utilities evolves. For a "good shock" the next period's utility is larger for both types than that for a "bad shock". Summarized in Corollary 1, this observation furnishes the underpinning of the recursive approach we shall employ to solve the general model.

Corollary 1. *In the optimal contract,* $(u^*(\theta_L|\theta_L), u^*(\theta_H|\theta_L)) \geqslant (u^*(\theta_L|\theta_H), u^*(\theta_H|\theta_H))$.

Finally, a higher v_0 shifts bargaining power towards the possessor of private information, increasing the total size of the pie. This reduces optimal distortions and increases total surplus. Formally, let $R^*(v_0)$ and $S^*(v_0)$ respectively be the principal's ex ante expected utility and the total economic surplus generated by the optimal allocation rule in Proposition 1. The economic relationship between the principal and the agent creates a "meta firm". Each point $\left(R^*(v_0), \max\left\{\underline{v}, v_0\right\}\right)$ on the Pareto frontier corresponds to a capital structure composed of their respective shares with total value $S^*(v_0)$. Figure 2b plots the value of the "meta firm" as a function of v_0. As we increase v_0, total economic surplus at first remains flat till $\underline{v}$, then rises to its efficient value S^e at $\overline{v}$, and for $v_0 \geq \overline{v}$ it stays constant.

Corollary 2. $S^*(v_0)$ *is increasing in* v_0*, and strictly so for* $v_0 \in [\underline{v}, \overline{v}]$.

The two period model illustrates the key economic forces at play. It allows us to make educated guesses about what general results may look like. Are optimal distortions increasing in the number of consecutive "bad shocks"? Would these distortions be reduced by a "good shock"? Is efficiency is a certainty in the long run? What is the path to liquidity and efficiency? How does the size of the meta firm evolve? What role does persistence play in short and long run?

[20] Unless specified otherwise, throughout the paper: $V(q) = 10\sqrt{q}$, $\alpha = 0.75$, $\theta_L = 3, \theta_H = 4$, $v_0 = 0$. For the two period model we also assume $\delta = 1$.

3.3 Sequential approach

A natural next step is to the extend the two period analysis to the T period model, that is solve $(\mathcal{RP}^*)$ for a general time horizon. Using a subset of binding constraints, we can express $\mu_L U(\theta_L) + \mu_H U(\theta_H)$ and all remaining cash-strapped constraints in terms of quantities. To this end, the first step is to inductively apply the binding constraints $IC_L(h^{t-1})$ and $C_L(h^{t-1})$ and generalize equation (4). Total expected utility for the agent is given by:

$$\mu_L U(\theta_L) + \mu_H U(\theta_H) = \Delta\theta \sum_{t=1}^{T} \delta^{t-1} \mathbb{P}(\theta_t = \theta_L) q(\theta_H | \theta_H^{t-1}) \tag{7}$$

Define the threshold generated by equation (7) for the efficient quantity to be $\overline{v}$ and that generated by the optimal contract when we ignore (PK) to be $\underline{v}$. It is easy to see that optimal contract is always efficient for $v_0 \geq \overline{v}$ and it selects the principal optimal contract for all $v_0 < \underline{v}$. Moreover, $C_L(h^{t-1})$ can be stated in terms of the quantity vector, the closed form expression is provided in the appendix. Then, as in the two period model, the principal's problem can be reduced to choosing $\mathbf{q}$ to maximize $\overline{S} - [\mu_L U(\theta_L) + \mu_H U(\theta_H)]$ subject to (PK) and $C_L(h^{t-1})$, where $\mu_L U(\theta_L) + \mu_H U(\theta_H)$ is given by equation (7). Introducing Lagrange multipliers for (PK) and all the $C_L(h^{t-1})$ constraints, we can then write down the optimal allocation rule, see Section 9.3 in the appendix.

The sequential approach allows us to exposit the additive structure of distortions, check for the validity of the relaxed problem approach, and as we will see later, pin down the optimal limit contract as $\alpha \to 1$. Beyond this, it is hard to derive general arguments about the nature of dynamic distortions because the Lagrange multipliers are endogenous and jointly determined at the optimum. We turn to the recursive approach to reduce the curse of dimensionality and provide a precise characterization of the optimal contract.

3.4 Recursive formulation: a full characterization

The recursive approach to dynamic contracting, understood at least since Green [1987], and Spear and Srivastava [1989], allows us to characterize the optimal contract using "promised utility" of the agent as a state variable. As noted by Fernandes and Phelan [2000], with *Markovian* agency frictions, the recursive domain of "promised utility" need not be one-dimensional, its dimensionality depends on the cardinality of the type space.

Assume throughout that $T = \infty$. Let $S(h^{t-1})$ be the expected total surplus generated by the sequential contract from period t onwards:

$$S(h^{t-1}) = \sum_{s=0}^{\infty} \delta^s \mathbb{E}\left[s(\tilde{\theta}_{t+s}, q(\tilde{\theta}_{t+s} | \tilde{h}^{t+s-1})) \mid \tilde{h}^{t+s} \in H^{t+s}|_{h^{t-1}} \right]$$

where $H^t|_{h^\tau}$ for $\tau \leq t$ is set of all histories of length t whose first τ elements are h^τ.

Suppose that the agent reported (h^{t-1}, θ_j) truthfully and the principal is committed to deliver exactly w_i to the agent of type θ_i at this date. Then, for $\mathbf{w} = (w_L, w_H) \in \mathbb{R}^2$, define $Q_j^*(\mathbf{w})$ for $j = L, H$ to be

$$\boxed{\star} \qquad Q_j^*(\mathbf{w}) = \max_{\langle \mathbf{U}, \mathbf{q} \rangle} \quad S(h^{t-1}, \theta_j)$$

subject to $w_i = U(\theta_i|h^{t-1}, \theta_j)$ for $i = L, H$, and

$$IC_L(h^{t+s}), C_L(h^{t+s}), C_H(h^{t+s}) \; \forall \; h^{t+s} \in H^{t+s}|_{(h^{t-1},\theta_j)} \quad \forall \; s$$

Here $Q_j^*(\mathbf{w})$ is the maximal surplus (and hence maximal expected profit for the principal) generated by the optimal contract given that the previous period type was θ_j, and the agent has to be provided an expected utility vector $\mathbf{w}$. It is standard practice to show that conditional on $\mathbf{w}$, the optimal value is independent of h^{t-1}, thus the sparse expression $Q_j^*(\mathbf{w})$. That is, all the history dependence is encoded in the two-dimensional expected utility and last period's type, j. Let W be the largest set of $\mathbf{w}$ such that the constraints set above is non-empty. Again, this set does not depend on h^{t-1}.[21]

The problem from $t = 2$ is *recursive* and it reads as follows:

$$(\mathcal{RF}) \qquad Q_j^*(\mathbf{w}) = \max_{\langle \mathbf{z_L}, \mathbf{z_H}, \mathbf{q} \rangle} \quad \alpha_j[s(\theta_L, q_L) + \delta Q_L^*(\mathbf{z_L})] + (1 - \alpha_j)[s(\theta_H, q_H) + \delta Q_H^*(\mathbf{z_H})]$$

subject to $\langle \mathbf{z_L}, \mathbf{z_H}, \mathbf{q} \rangle \in W^2 \times \mathbb{R}_+^2$, and

$$\begin{aligned} w_L - w_H &\geq \Delta\theta q_H + \delta(2\alpha - 1)(z_{HL} - z_{HH}) \\ w_L &\geq \delta[\alpha z_{LL} + (1 - \alpha) z_{LH}] \\ w_H &\geq \delta[(1 - \alpha) z_{HL} + \alpha z_{HH}] \end{aligned}$$

where $\alpha_L = 1 - \alpha_H = \alpha$, and by $(\mathcal{RF})$ we mean recursive formulation. Note that $\mathbf{q} = (q_L, q_H)$ is the optimal allocation rule, $\mathbf{z_L} = (z_{LL}, z_{LH})$ is the expected utility vector of the agent for the next period if his type today is θ_L, and $\mathbf{z_H} = (z_{HL}, z_{HH})$ is the expected utility vector of the agent for the next period if his type today is θ_H. Given these choice variables, the first constraint is the incentive constraint for the low cost type, and next two are cash-strapped constraints for the low and high cost types, respectively.

At date $t = 1$ the problem is different for two reasons. First, the belief is equal to the prior, and second the contract has not been yet initialized.

$$(\mathcal{RF}_0) \qquad R^*(v_0) = \max_{\langle \mathbf{w}, \mathbf{z_L}, \mathbf{z_H}, \mathbf{q} \rangle} \quad \mu_L[s(\theta_L, q_L) + \delta Q_L^*(\mathbf{z_L})] + \mu_H[s(\theta_H, q_H) + \delta Q_H^*(\mathbf{z_H})] - \bar{U}$$

subject to $\langle \mathbf{w}, \mathbf{z_L}, \mathbf{z_H}, \mathbf{q} \rangle \in W^3 \times \mathbb{R}_+^2$, and

$$\begin{aligned} \bar{U} &= \mu_L w_L + \mu_H w_H \geq v_0 \\ w_L - w_H &\geq \Delta\theta q_H + \delta(2\alpha - 1)(z_{HL} - z_{HH}) \\ w_L &\geq \delta[\alpha z_{LL} + (1 - \alpha) z_{LH}] \\ w_H &\geq \delta[(1 - \alpha) z_{HL} + \alpha z_{HH}] \end{aligned}$$

where Q_L^* and Q_H^* are calculated in $(\mathcal{RF})$.

The recursive domain W, that is the set of all possible expected utilities that generate themselves in an incentive compatible and feasible manner, is precisely the positive orthant above the 45 degree

[21] Given the time structure of the problem it can be shown that W is also independent of j, see Claim 1 in the appendix.

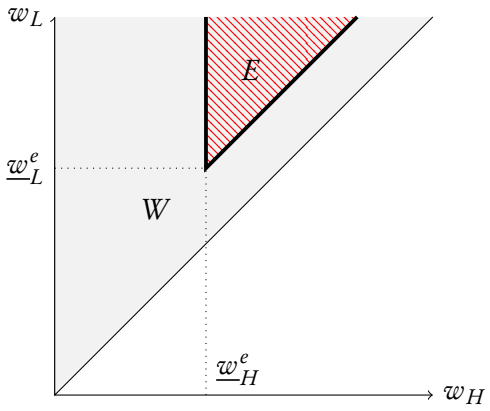


Figure 3: Recursive domain W and the efficient set E.

line.[22] Figure 3 plots the recursive domain. Next, let E be the largest subset of W such that the constraints set in $\boxed{\star}$ is non-empty when $\mathbf{z_i} \in E$ and $q_i = q^e(\theta_i)$ for $i = L, H$. We term this the efficient set. E too is self-generating, hence an absorbing subset. Figure 3 plots the efficient set E - it is characterized by its lowest point $\underline{\mathbf{w}}^{\mathbf{e}}$ and two rays.[23] Note that $\overline{v} = \mu_L \underline{w}^e_L + \mu_H \underline{w}^e_H$.

With some work, we can also show that the Bellman operator has a unique, continuous bounded fixed point Q^* which is concave, supermodular and continuously differentiable. Importantly the value functions in the sequential and recursive problems coincide.

Shape of the optimal contract

The optimal contract is characterized by the first-order and envelope conditions. These are provided in the appendix. Here we geometrically explain the structure of expected utilities that arise as part of the optimal contract. First, there exists a threshold $\underline{w}^{liq}_L$ on the expected utility of the low cost type, above which the contract becomes liquid: the Lagrange multiplier on the second constraint in $(\mathcal{RF})$, say ρ_L, is zero for $w_L \geq \underline{w}^{liq}_L$. Critically, this threshold lies below the efficient level: $\underline{w}^{liq}_L < \underline{w}^e_L$, see Figure 4a. Also, we show that the for any w such that $w_L \geq \underline{w}^{liq}_L$, $\mathbf{z_L} \in E$: once in the liquid region, an additional realization of the low cost type pushes the optimal contract into the efficient region. Therefore, $\rho_L = 0$ forms a penultimate zone towards efficiency.[24]

Next, we draw the two level curves that enclose the optimal contract in the inefficient region. To understand their geometry, think of simple price theory. Cull the following sub-problem from $(\mathcal{RF})$:

$$\max_{\mathbf{z_L}} \quad Q^*_L(\mathbf{z_L}) \quad s.t. \quad w_L = \delta\left[\alpha z_{LL} + (1-\alpha) z_{LH}\right]$$

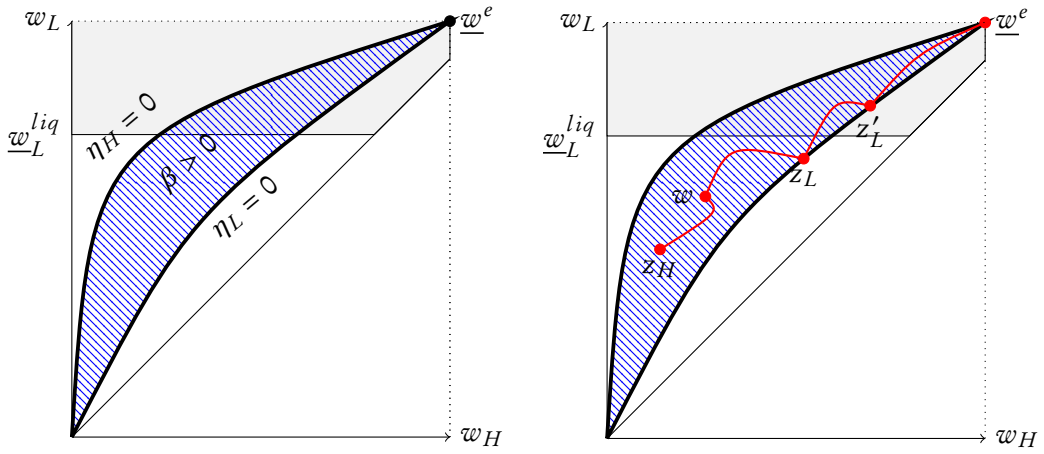
η_L is locus of points where "marginal rate of substitution = relative prices" for this optimization problem. That is,

$$\eta_L(\mathbf{w}) = 0 \Leftrightarrow \frac{D_L Q^*_L(\mathbf{w})}{\alpha} = \frac{D_H Q^*_L(\mathbf{w})}{1-\alpha}$$

[22] Self generation refers to the idea of identifying the largest possible set such that given an expected utility vector $\mathbf{w}$ in that set, there exists some feasible policy choice $q, \mathbf{z_L}, \mathbf{z_H}$ such that $\mathbf{z_L}, \mathbf{z_H}$ also lie in the set.

[23] Just like W, E too is independent of j.

[24] In the benchmark model $\underline{w}^{liq}_L = 0$, therefore, one "good shock" propels the contract into efficiency.



(a) Liquidity threshold and the Shell, B (b) Evolution of the optimal contract

Figure 4: Constructing the space wherein the optimal contract lies

where D_i is the directional derivative of Q_j^* for $i = L, H$. Figure 4a plots $\eta_L = 0$. We show that it is an increasing curve that joins the origin and $\underline{\mathbf{w}}^{\mathbf{e}}$. Similarly, cull the following problem from $(\mathcal{RF})$:

$$\max_{\mathbf{z_H}} \quad Q_H^*(\mathbf{z_H}) \quad s.t. \quad w_H = \delta\left[(1-\alpha)z_{HL} + \alpha z_{HH}\right]$$

to generate a locus η_H that satisfies:

$$\eta_H(\mathbf{w}) = 0 \Leftrightarrow \frac{D_H Q_H^*(\mathbf{w})}{1-\alpha} = \frac{D_L Q_H^*(\mathbf{w})}{\alpha}$$

Figure 4a plots $\eta_H = 0$. We show that it is an increasing curve that joins the origin and $\underline{\mathbf{w}}^{\mathbf{e}}$. Moreover, it lies above $\eta_L = 0$. The optimal constrained contract resides on or in the interior of the curves, a space we call the shell (B). The shell is characterized by binding incentive constraints, $\beta > 0$. Figure 7d numerically constructs the shell for different values of α.

Figure 4b gives an example of the evolution of expected utility. Starting at $\mathbf{w}$, it moves to $\mathbf{z_L}$ on the realization of a low cost type, and to $\mathbf{z_H}$ on high cost type. In fact the example is chosen so that on the realization of two consecutive low cost types, the contract becomes liquid (at $\mathbf{z'_L}$), and therefore a third "good shock" carries it over to the efficient region, E.[25]

3.5 The main result

Collecting all the key insights from the sequential and recursive approaches, we enlist a complete characterization of the optimal contract. For consistency, we state all the results in the lexicon of the sequential model. We assume that the prior is first-order stochastically ranked around its Markov evolution: $1-\alpha \leq \mu_L \leq \alpha$. This is strictly more general than assuming a "seed type", that is $\mu_L = \alpha$

[25] As Figure 4b depicts, the realization of a low cost realizations always chooses the expected utility vector in the northeast direction *on* the locus $\eta_L = 0$, that is $\mathbf{z_L}$ and $\mathbf{z'_L}$ lie *on* the curve. Whereas, realization of a high cost type chooses a point in the southwest direction in the *interior* of the shell.

or $\mu_L = 1 - \alpha$ which is a standard in the recursive contracting literature.[26]

Theorem 1. *Let $T = \infty$. The optimal contract, $\langle \mathbf{U}^*, \mathbf{q}^* \rangle$ (solution to ($\mathcal{RP}^*$)), is characterized by the following properties.*

A **Optimal distortions***:*

1. *Optimal contract is downward distorted: , $q^*(\theta_L|h^{t-1}) = q^e(\theta_L)$, and $q^*(\theta_H|h^{t-1}) = q^e(\theta_H) - d(\theta_H|h^{t-1})$, where $d(\theta_H|h^{t-1}) \geq 0$.*
2. *Distortions are strictly increasing: for $q^*(\theta_H|h^{t-1}) < q^e(\theta_H)$, $d(\theta_H|h^{t-1}, \theta_H^s)$ is strictly increasing in s.*
3. *Distortions are muted after a "good shock": for $q^*(\theta_H|h^{t-1}) < q^e(\theta_H)$, $d(\theta_H|h^{t-1}, \theta_L) < d(\theta_H|h^{t-1})$.*

B **Expected utility***:*

4. *Expected utility increases (decreases) with a low (high) cost type: for $q^*(\theta_H|h^{t-1}) < q^e(\theta_H)$, $(U^*(\theta_L|h^{t-1}, \theta_H), U^*(\theta_H|h^{t-1}, \theta_H)) \leq (U^*(\theta_L|h^{t-1}), U^*(\theta_H|h^{t-1})) \ll (U^*(\theta_L|h^{t-1}, \theta_L), U^*(\theta_H|h^{t-1}, \theta_L))$.*

C **Liquidity***:*

5. *The contract becomes liquid above a fixed threshold which is below the efficient level: $\exists\ \underline{w}_L^{liq} < \underline{w}_L^e$ such that for $U^*(\theta_L|h^{t-1}) \geq w_L^{liq}$, $C_L(h^{t-1})$ is slack.*

D **Efficiency***:*

6. *Efficiency is an absorbing state: $d(\theta_H|h^{t-1}) = 0 \Rightarrow d(\theta_H|h^{t+s-1}) = 0$, and $(U^*(\theta_L|h^{t-1}), U^*(\theta_H|h^{t-1})) \in E \Rightarrow (U^*(\theta_L|h^{t+s-1}), U^*(\theta_H|h^{t+s-1})) \in E\ \forall\ h^{t+s-1} \in H^{t+s-1}|_{h^{t-1}}$.*
7. *An endogenous and monotonic number of "good shocks" are required for efficiency: $\exists\ n^*(h^{t-1}) \in \mathcal{N}$ such that $d(\theta_H|h^{t-1}, \theta_L^{n^*}) = 0$, and $n^*(h^{t-1}, \theta_H) \geq n^*(h^{t-1})$.*
8. *Efficiency is achieved through a penultimate slacking of the cash-strapped constraint: $C_L(h^{t-1})$ binds $\Leftrightarrow n^*(h^{t-1}) \geq 2$, and $C_L(h^{t-1})$ is slack $\Leftrightarrow d(\theta_H|h^{t-1}, \theta_L) = 0$.*

E **Long run***:*

9. *Efficiency is a certainty: $d(\theta_H|h^{t+s-1}) \xrightarrow[s\to\infty]{} 0$, and $(U^*(\theta_L|h^{t+s-1}), U^*(\theta_H|h^{t+s-1})) \xrightarrow[s\to\infty]{} \mathbf{w} \in E$ almost surely.*

Part A of the theorem characterizes the optimal allocation rule through dynamic distortions produced by the periodic interaction between incentives and cash-strapped constraints. The low cost type supplies the efficient quantity, and the high cost's supply is distorted downwards. Using the framework of equation $(*)$, the evolution of $r(h^{t-1})$ determines the optimal distortions d, which drives the theorem.

While the contract is inefficient, every further high cost realization strictly increases the dynamic distortions, thereby strictly decreasing the optimal quantity. In Figure 1a, quantity along the history (h^{t-1}, θ_H^2) is less than quantity along (h^{t-1}, θ_H). This is in stark contrast to the standard results in dynamic mechanism design (without financial constraints) that emphasize decreasing distortions over time. Moreover, a "good shock" reduces the optimal distortions- quantity along $(h^{t-1}, \theta_L, \theta_H)$

[26]The assumption $1 - \alpha \leq \mu_L \leq \alpha$ is made to ensure that the optimal contract starts in shell defined in Section 3.4. If these conditions are not satisfied then the optimal contract enters the shell the moment it gets a low cost realization. All our points still hold expect for the "lowest history" of continued high cost realizations.

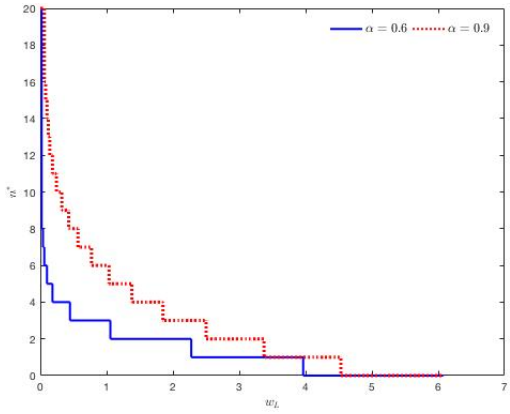


Figure 5: Number of consecutive low cost types required to reach E

is greater than that along (h^{t-1}, θ_H). These rankings of optimal distortions form the bedrock of our analysis.[27,28]

Part B tracks the optimal path of expected utility. For any history of types, for the inefficient contract, expected utility strictly increases along both dimensions after a "good shock" and reduces after a "bad shock". In terms of Figure 1a, vector of squares is larger than circles which is larger diamonds. And, in the two dimensional domain of expected utility a "good shock" takes it in the northeast direction and "bad shock" in the southwest; always within the shell that we constructed in Section 3.4.

Part C characterizes liquidity. The interval $[\underline{w}_L^{liq}, \underline{w}_L^e)$ for $U^*(\theta_L|h^{t-1})$ witnesses slacking of $C_L(h^{t-1})$. The region can only be attained through a "good shock" (given the contract does not start in this region). It is important to note that liquidity is not an absorbing state, and it is not synonymous with efficiency. Even in the liquid region, a "bad shock" can revert the contract back into the illiquid region, $[0, \underline{w}_L^{liq})$.

The path to efficiency is completed in part D. Efficiency is an absorbing state (point 6). Once $q(\theta_H|h^{t-1}) = q^e(\theta_H)$, the contract is efficient thereafter: $q(\theta_H|h^{t-1}, h^s) = q^e(\theta_H)$. That is, once the distortion for the high cost type reduces to zero, it stays zero. Next the efficient region, E, can be attained only through an endogenous number of "good shocks". This number depends on the path of reported types. In particular, on a high cost realization, the number of types required to reach the efficient region increases (point 7). Figure 5 plots $n^*(w_L)$- the number of consecutive low cost realizations required to reach the efficient region as a function of w_L which encodes all the history dependence required for n^*. As w_L decreases, n^* can become quite large, in fact $n^* \to \infty$ as $w_L \to 0$. Monotonicity in the endogenous number of shocks is an intuitive but formally novel addition to dynamic contracting.

In addition, for most parametric settings (where expected utility starts in the region $[0, \underline{w}_L^{liq})$), the efficient region is achieved through a penultimate liquid region- $\left[\underline{w}_L^{liq}, \underline{w}_L^e\right)$ (point 8). Once the

[27] In the benchmark model, even when the first-order approach fails, distortions decrease to zero on average, see Garrett and Pavan [2015] and Garrett, Pavan, and Toikka [2017]. It is possible for distortions to increase over time when the private information is regarding the parametrization of the Markov process itself, see for example Boleslavsky and Said [2013]. Our paper invokes the standard model but changes the feasibility constraints to produce opposite result.

[28] In the appendix we prove Theorem 1 for general asymmetric Markov evolution. There the quantities are strictly decreasing and distortions strictly increasing along consecutive "bad shocks", after scaling for the asymmetry through appropriate weights.

cash-strapped constraint is slack, efficiency is attained through one more low cost type.[29] Complimentarily, Fu and Krishna [2017] show that this result also holds in the Markov extension of the cash flow diversion model of Clementi and Hopenhayn [2006]. We show in Section 6 that this particular aspect of the optimal contract, however, does not hold in the continuous time version of the model for there is no notion of a "time period", and we get $\underline{w}_L^{liq} = \underline{w}_L^e$.

Finally in part E, we close the theorem with the certainty of efficiency in the long-run. At any point on the contract tree (and hence any level of expected utility), the contract will converge to the efficient region almost surely. The expected utility of the agent turns out to be a martingale, and we use the martingale convergence theorem to establish the certainty of efficiency. A look at Figure 4b throws up two plausible scenarios- expected utility of the agent gets sucked into zero or that it converges to efficiency, the former is suboptimal for the principal. In fact, at any point on the recursive domain, it requires infinitely many bad shocks to get to zero, and finitely many good shocks to get to efficiency.

3.6 Optimal limit contract

While Theorem 1 provides a fairly precise characterization, the optimal contract can be completely pinned down in the limit as the persistence in types converges to unity. This analysis, which is presented in the next proposition, sheds further light on the structure of dynamic contracts with highly persistent agency frictions. Recall the function Q defined in equation (1).

Proposition 2. *Denote the Lagrange multiplier on* (PK) *by* λ*. There exists* $N \in \mathbb{N}$ *and a sequence* $\{d_n\}$*, functions of* $\Gamma\backslash\{\alpha\}$*, such that the optimal limit contract can be described as follows:*

1. $\lim_{\alpha\to 1} q^*(\theta_H|h^{t-1}) = Q(d_n)$*, such that* h^{t-1} *contains* n *draws of* θ_L*, where*

$$\begin{aligned} Q(d_0) &= Q\left((1-\lambda)\frac{\mu_L}{\mu_H}\right) \\ Q(d_n) &= \frac{1}{\delta}Q\left(d_{n-1}\right) \leq q^e(\theta_H) \quad \textit{for} \quad 1 \leq n \leq N \\ Q(d_n) &= q^e(\theta_H) \quad \textit{for} \quad n > N \end{aligned}$$

2. *It takes* N *low cost draws (in any order) to become liquid, and* $N+1$ *to reach efficiency where* N *is the largest positive integer that satisfies:* $Q\left((1-\lambda)\frac{\mu_L}{\mu_H}\right) \leq \delta^N q^e(\theta_H)$.

Therefore, the optimal distortion is highest at the start, and then decreases multiplicatively for every "good shock" as a function of the discount factor. The higher the discount factor, the larger is the number of "good shocks" required to attain efficiency- the principal finds it profitable to spread distortions over time. Intuitively, for high levels of persistence in private information, each "good shock" has a larger positive effect than the corresponding negative effect of a "bad shock".

[29] Summarizing, in terms of equation $(*)$, $r(h^t)$ in positive if and only if (i) $h^t = \theta_H^t$, or (ii) $h^t = (h^{s-1}, \theta_L, \theta_H^{t-s})$ for $1 \leq s \leq t-1$ and $C_L(h^{s-1})$ binds, or (iii) $\theta_t = \theta_L$, $r(h^{t-1}) > 0$ and $C_L(h^{t-1})$ is slack.

4 Dynamics of payments

There are three salient features of an incentive compatible payment schedule that implements the optimal allocation. First, as long as the optimal contract is illiquid, delayed payments are optimal. Second, at any given history, promised utility is marked up after a "good shock" and marked down after a "bad shock", both in proportion to the history dependent information rent. And, third, the total economic surplus is increasing in both the share of the agent and his utility spread.

The dynamics of payments are as follows. As long we are in the illiquid region, $u^*(\theta_L|h^{t-1}) = u^*(\theta_H|h^{t-1}) = 0$, and these are uniquely determined by the binding cash-strapped constraints. If we are in the liquid region, $u^*(\theta_H|h^{t-1}) = 0$ and $u^*(\theta_L|h^{t-1})$ is chosen to provide the low cost type with positive utility according to inductively binding incentive compatibility constraints.[30] We define the mechanism formally in the next proposition. For $j = L, H$, let $v_j^e = \alpha_j \underline{w}_L^e + (1-\alpha_j)\underline{w}_H^e$ be the promised utility offered to the agent at the lowest point of the efficiency set.

Proposition 3. *Suppose $v_0 \leq \overline{v}$. Given optimal allocation rule $\mathbf{q}^*$, the following transfer rule implements it:*

$$u^*(\theta_H|h^{t-1}) = 0 \quad \textit{and} \quad u^*(\theta_L|h^{t-1}) = \max\left\{U^*(\theta_L|h^{t-1}) - \delta v_L^e, 0\right\} \; \forall \; h^{t-1} \; \forall \; t$$

Suppose $v_0 > \overline{v}$. Then $\mathbf{q}^* = \mathbf{q}^{\mathbf{e}}$ and the following transfers rule implements it: the principal makes an initial transfer of $\eta = v_0 - \overline{v}$ to the agent, and then follows transfers as described above.

Next, at any given promised utility, defined by $v^*(\theta_j|h^{t-1}) = \alpha_j U^*(\theta_L|h^{t-1}, \theta_j) + (1-\alpha_j)$ $U^*(\theta_H|h^{t-1}, \theta_j)$, the expected utility increases after a "good shock" and decreases after a bad one, both in proportion to the utility spread: $U^*(\theta_L|h^{t-1}, \theta_j) = U^*(\theta_H|h^{t-1}, \theta_j) + I(h^{t-1}, \theta_j)$, where

$$I(h^{t-1}) = \Delta\theta \sum_{s=1}^{\infty} \delta^{s-1}(2\alpha - 1)^{s-1} q^*(\theta_H|h^{t-1}, \theta_H^{s-1}) \tag{8}$$

is the dynamic information operator which measures the difference between expected utility offered to the low and high types respectively through the binding incentive constraints.

An intuitive way to think about our mechanism is the following. In the illiquid region, the principal only compensates the agent with working capital. Each "good shock" marks up the expected utility and each bad one marks it down. In liquid region the principal loosens her purse for the first time, by providing $u^*(\theta_L|h^{t-1}) > 0$, with a push towards efficiency in the event of another "good shock". Once the contract becomes efficient, and hence the information rent of the agent is maximal and stationary, the big firm (principal) can simply the take-over the small firm (agent) by allowing it to operate "in-house". The price of the take-over is precisely the expected utility of the agent at the time it becomes efficient, viz. $\underline{w}_L^e$. After the take-over the big firm simply provides working capital (sans the information rent) every period. When $v_0 > \overline{v}$, the contract is efficient: $\mathbf{q}^* = \mathbf{q}^{\mathbf{e}}$. Therefore, the take-over can happen at the inception.

Finally, we show that the value of the "meta firm" is increasing in the share of the agent and his utility spread. The former is simply a statement of aligning incentives with bargaining power.

[30] $\mathbf{U}^*$ is uniquely defined in the illiquid region, and in the liquid and efficient regions we choose expected utility to satisfy $IC_L(h^{t-1})$ as an equality. Once $\mathbf{U}^*$ is fixed, $\mathbf{u}^*$ is determined through definitional identities.

As the agent assumes a greater stake of the total surplus the effect of agency frictions is reduced which increases the size of pie towards its efficient value. The latter though may not be immediately obvious. The optimal utility spread increases on the realization of a "good shock". This makes the provision of incentives easier, thereby decreasing the optimal distortion and increasing the surplus. Upon reaching efficiency, both the utility spread and value of the "meta firm" become constant for each type. It should be noted that while the monotonic relationship of economic surplus with promised utility is a global one, that with utility spread in only valid along the optimum.

Proposition 4. *For the optimal contract, total economic surplus is positively correlated with promised utility and utility spread:*

$$S^*(h^{t-1}) \uparrow v^*(h^{t-1}) \quad \textit{and} \quad S^*(h^{t-1}) \uparrow I(h^{t-1})$$

5 Role of financial constraints and persistence in private information

There are three conceptual points that emanate from studying this dynamic screening model with persistent private information and cash-strapped constraint- (i) the interaction of incentive constraint with stronger feasibility restrictions generates novel dynamic distortions, (ii) a foundation for when positivity of stage utility can be interpreted as a limited liability restriction that is beneficial for the agent, and (iii) the impact of persistence in agency frictions on the evolution of the optimal contract and economic surplus.

5.1 Cash-strapped versus individual rationality

The elegance of capturing real economic frictions in much of mechanism design is embedded in the interaction of various incentive and feasibility constraints.[31] Figure 6 exhibits the interaction of incentives and feasibility in our model. Each time cash-strapped constraint for the high cost type binds, its interaction with the incentive constraint generates distortions that propagate infinitely along the sequence of high types from then on. In the benchmark model described in Section 3.1, this interaction happens only in the first period for only the first period individual rationality constraint binds. This is because there is no restriction on the extent to which the agent's payoffs can be backloaded. Hence, the propagation of distortions happens once, along the lowest history, whose effect mitigates over time leading to a decreasing sequence of distortions. However, in our model the cash-strapped constraint potentially interacts with the incentive constraint on a sustained basis which keeps adding to the distortions that would propagate along a sequence of consecutive "bad shocks". Dynamic distortions are now a sum of two components: backloading of incentives to the extent possible and illiquidity due to financial constraints; the latter increases with each "bad shock", overturning the standard result of decreasing distortions.

Moreover, since the cash strapped constraint of the low cost type binds in the illiquid region, it too interacts with the incentive constraints to sustain distortions, which propagate along a consecutive sequence of "bad shocks" even when "good shocks" have been realized before it. An endogenous

[31] For example think of how the interaction of three constraints- incentive compatibility, individual rationality and budget balance- produces impossibility of efficiency in a bilateral trade setting in Myerson and Satterthwaite [1983].

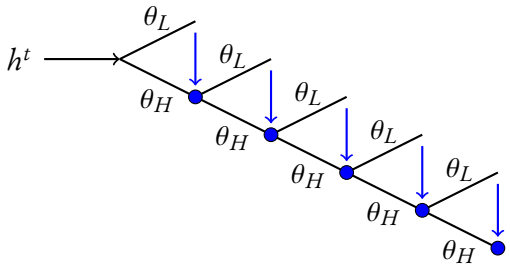


Figure 6: Propagation of distortions

number of "good shocks" are required to overcome the legacy of additive distortions from previous binding constraints.

5.2 Credit constraint versus limited liability

Should the positivity of stage utility be seen as a limited liability or a credit constraint? The answer can be found in a comparison to the benchmark model. It is clear that both the profit of the principal and the total value of economic surplus is higher in the benchmark model than in our model, cash-strapped is a stronger restriction than individual rationality. The ambiguity lies in the ex ante value of the agent's utility.

We ask- in the principal profit maximizing contract, when is the agent better off? The ex ante expected utility of the agent in the two models is given by

$$\underline{v}^{\#} = \mu_L U^{\#}(\theta_L) + \mu_H.0 = \mathbb{P}(\theta_1 = \theta_L)\Delta\theta \sum_{t=1}^{T} \delta^{t-1} (\alpha_L - \alpha_H)^{t-1} q^{\#}(\theta_H|\theta_H^{t-1})$$

$$\underline{v}^{*} = \mu_L U^{*}(\theta_L) + \mu_H U^{*}(\mu_H) = \Delta\theta \sum_{t=1}^{T} \mathbb{P}(\theta_t = \theta_L) q^{*}(\theta_H|\theta_H^{t-1})$$

A careful look at the two formulas would reveal that there is no obvious mathematical way of ranking $\underline{v}^{\#}$ and $\underline{v}^{*}$. We take a two pronged approach. First we theoretically evaluate the ranking between $\underline{v}^{\#}$ and $\underline{v}^{*}$ for the iid and perfectly persistent limits.

Proposition 5. *For $\alpha \approx \frac{1}{2}$, $\underline{v}^{\#} < \underline{v}^{*}$. And $\exists$ $D^{\#}$ and D^{*}, functions of $\Gamma\backslash\{\alpha, v_0\}$, such that for $\alpha \approx 1$, $v^{*} \gtrless v^{\#}$ if and only if $D^{\#} \lesseqgtr D^{*}$.*

Therefore, the iid model would predict that agent always does better with financial constraints, whereas with persistence, the answer depends on the underlying economic environment. Second, we numerically evaluate both values for a large class of parameters and find that $\underline{v}^{*}$ is mostly higher than $\underline{v}^{\#}$ except for very low values of δ and high values of μ_H.[32]

Conceptually, in the absence of (PK), the level of persistence in private information matters for the interpretation of constraining the stage utility to be positive. When the agent is patient, he gains more from signing a contract in which he is "protected" by a rule of law that he cannot be forced to pay large amounts of cash or liquidate high valued assets at any stage of the contract.

[32]The code has been made available at rohitlamba.com/research.

With a binding ex ante promise keeping constraint, (PK), we argue for the interpretation of the cash-strapped constraint as a credit constraint. Since the ex ante value of the agent is fixed, the cash-strapped constraint simply shrinks the size of the total pie.

5.3 Persistence

One key difference of our model from the standard dynamic financial contracting literature is persistence of the underlying agency friction, in our case the private information of the agent's technology. Here we present comparative statics with respect to persistence that the modeler would otherwise miss in an iid setting. We start with the long-run distribution of the economic surplus, that is, the value of the "meta firm". For the general asymmetric Markov chain, the invariant distribution is given by $\mu^* = \left(\mu_L^*, \mu_H^*\right)$, where $\mu_L^* = \frac{\alpha_H}{1-\alpha_L+\alpha_H}$ and $\mu_H^* = 1 - \mu_L^*$. The meta firm's value converges in distribution to the random variable which takes value Q_j^e with the probability μ_j^*, $j = L, H$, where

$$Q_j^e = \alpha_j[s(\theta_L, q^e(\theta_L)) + \delta Q_L^e] + (1-\alpha_j)[s(\theta_H, q^e(\theta_H)) + \delta Q_H^e]$$

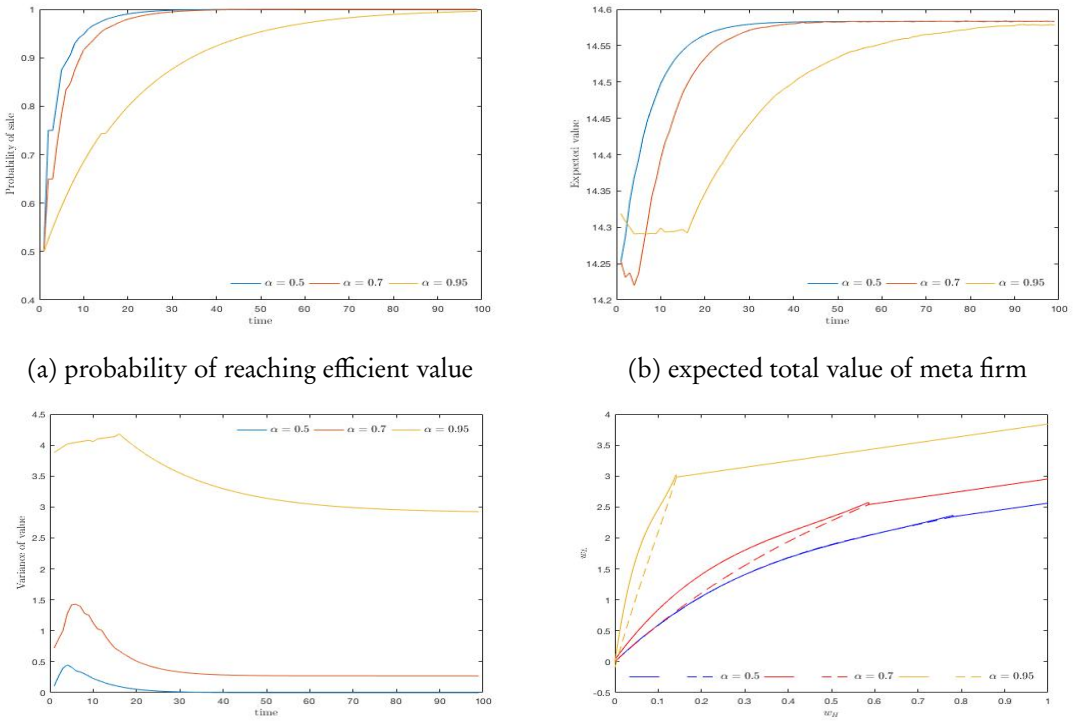
So, the mean and variance of ex ante surplus converge respectively to

$$\mathbb{E}\left[\text{Economic Surplus}\right] \to \mu_L^* \frac{s(\theta_L, q^e(\theta_L))}{1-\delta} + \mu_H^* \frac{s(\theta_H, q^e(\theta_H))}{1-\delta} \text{ as } t \to \infty$$

$$\mathbb{V}\left[\text{Economic Surplus}\right] \to \mu_L^* \mu_H^* \left(\frac{(\alpha_L - \alpha_H)[s(\theta_L, q^e(\theta_L)) - s(\theta_H, q^e(\theta_H))]}{1-\delta(\alpha_L - \alpha_H)}\right)^2 \text{ as } t \to \infty$$

For the symmetric Markov chain, $\alpha_L = 1 - \alpha_H = \alpha$, it is easy to see that $\mathbb{E}$ [Economic Surplus] is independent of α and $\mathbb{V}$ [Economic Surplus] is an increasing function of α. More importantly though, we are interested in the path towards efficiency, that is the size of firms that are not yet mature. A simulation of a large number of firms is documented in Figure 7. First, we look at the average time it takes for a firm to reach its efficient value. Figure 7a shows that rate of convergence is decreasing in the level of persistence- higher the persistence of technology shocks, smaller is the fraction of firms that are efficient at any given point in time. Figure 7b shows the average size of the firm as a function of time. This value is clearly decreasing in persistence, and so is the average time it takes for a firm to converge to its efficient value. Therefore, an iid model would predict too many mature firms, and too few financially constraints firms while analyzing a cross-section of firms in an economy.

Moreover, Figure 7c documents that variance in firm value is increasing in persistence even in the short-run. We have simulated this model for a large number of parameters and we find the relationships to be uniform- the hierarchy in values runs across the entire length of time. Why does this relationship persist robustly even in the short-run? The intuition comes from Figure 7d. It plots the shell that houses the set of expected utilities of the agent in the optimal contract (numerical counterpart of the theoretical pictures in Figure 4). Two factors here determine the evolution of variance in firm value over time as a function of persistence- the Lebesgue measure of the shell and the time it takes to reach its north east corner, that is efficiency. For the iid model, the shell collapses to a line. As we increases the value of persistence the shell first expands and then contracts



(a) probability of reaching efficient value

(b) expected total value of meta firm

(c) variance of total surplus

(d) the shell

Figure 7: Comparative statics with respect to persistence

towards the y-axis. Even as the first factor changes non-monotonically with persistence, the second factor dominates and results in a monotonic relationship. For example, for the iid model most firms converge quickly to the efficient value which results is small variance even in the short-run. For very high persistence, even though the shell is shrinking, the time required to reach efficiency is large and hence so is the variance of firm values.[33]

The larger message here is the following. If we were to take our model as a description of firm behavior in the economy, then in comparison to the iid model, persistence would (i) predict a larger number of firms that are financially constrained, (ii) result in a slower average rate of convergence to the state of being unconstrained, and (iii) produce a larger variance in the values of both the financially constrained and "mature" firms.

6 The model in continuous time

In this section, time in continuous and we let θ_t follow a continuous time Markov chain on $\Theta = \{\theta_L, \theta_H\}$ with transition rates $0 < \lambda_L, \lambda_H < \infty$, respectively:

$$\mathbb{P}\left(\theta_{t+dt} = \theta_i | \theta_t = \theta_i\right) = 1 - \lambda_i dt + o(dt) \quad \text{for } i = L, H$$

[33] It is also interesting to note that irrespective of the level of persistence, variance is a non-monotonic function of time. At the inception surplus takes one of two values depending on the first period shock. But as time grows the possible sequence of shocks increases exponentially that allow the surplus to take any value in the shell. However, on further passage of time, more and more of these contracts become efficient which collapses the variance to its long-run steady state value.

For each t and a history up to this time (h^{t^-}, θ_t), a contract specifies agent's instantaneous utility $u(\theta_t|h^{t^-}) \geq 0$ and a quantity $q(\theta_t|h^{t^-}) \geq 0$. A contract is assumed to be progressively measurable with respect to the natural filtration. In addition, it is assumed that the process of promised utilities defined as

$$U(\theta_t|h^{t^-}) = \mathbb{E}_t\left[\int_t^{+\infty} e^{-r(s-t)}u(h^s)ds\right]$$

is uniformly bounded.[34] With some restrictions on the agent's strategy space, described in the appendix, we can recursify the problem as in the discrete time model. Let $w_j = U(\theta_j|h^{t^-})$ and $z_{jk} = U(\theta_k|h^{t^-}, \theta_j^{[t,t+dt)})$ for $j,k = L,H$. Then, incentive compatibility says that for any small $dt > 0$,

$$w_L - w_H \geq \Delta\theta q_H dt + (1 - rdt)[1 - (\lambda_L + \lambda_H)dt](z_{HL} - z_{HH})$$

where q_H is an average quantity from (h^{t^-}, θ_H) to $(h^{t^-}, \theta_H^{[t,t+dt)})$.[35] There are two cash-strapped constraints that the contract must satisfy. For any small $dt > 0$,

$$w_L \leq (1 - rdt)[(1 - \lambda_L dt)z_{LL} + \lambda_L dt z_{LH}]$$
$$w_H \leq (1 - rdt)[(1 - \lambda_H dt)z_{HH} + \lambda_H dt z_{HL}]$$

The recursive domain is $W = \mathbb{R}^2_+$ and akin to the discrete time model, the efficiency set $E \subseteq W$ is self generating. The recursive problem uses $\mathbf{w} = (w_L, w_H)$ and "last period" shock as the state variables. We take expected surplus to be the objective, defined as $Q_L(\mathbf{w})$ or $Q_H(\mathbf{w})$. Assuming that these functions are continuously differentiable, we get two HJB equations that characterize the optimum. Much of the remaining work then involves proving the existence of the shell and showing that the evolution of the optimal contract follows the properties of Theorem 1. We show that except points 5 and 8, all other properties hold in the continuous time model as well. The two exceptions arise because there is no notion of a "time period" in continuous time, so the liquidity and efficiency regions are synonymous. All the formal details are provided in Section 9.11 in the appendix.

7 Related literature

This paper sits at the intersection of mechanism design and financial economics, using the tools of the former to generate insights referred to in the latter. We take up each in turn. The last decade has seen a burgeoning literature in mechanism design where agents meet or transact in the market repeatedly. A key challenge therein is the modeling of dynamic incentives. This literature has been pioneered by papers such as Courty and Li [2000], Battaglini [2005] and Esö and Szentes [2007] with a focus on optimal contracts, and Athey and Segal [2007], Athey and Segal [2013] and Bergemann and Välimäki [2010] that focus on implementing the efficient allocation. Pavan, Segal, and Toikka [2014] provide a unified treatment of dynamic incentives through what has come to be referred as the dynamic envelope formula.

Most, if not all, of this literature does not ask the question of feasibility or the extent of commitment beyond the usual individual rationality constraint. Esö and Szentes [2017] point out that

[34] U is well-defined because of our measurability assumption and non-negativity of the instantaneous utility process. Moreover, the uniform boundedness assumption can be weakened without affecting our results.

[35] This quantity is well-defined, because the allocation process is uniformly bounded.

these modelling choices lead to a kind of irrelevance of dynamics. Since incentives and feasibility interact only at the start of the contract, distortions are akin to the static model and are only augmented by marginal "innovations" to agent types. What would happen if this unbridled backloading instrument is not available? How do incentives interact with a stronger notion of feasibility? Our paper places a standard mechanism design problem in a dynamic framework where the agent is constrained on how much credit he can borrow- producing markedly different time structure of information rents, and hence distortions.[36]

Krähmer and Strausz [2015b] consider a (two-period) sequential screening model with ex post participation constraints. This comes close to our framework of demanding stronger notions of feasibility. They show that with these additional constraints the optimal contract is static and does not illicit the agent's information sequentially. Our model is multi-period, the optimal contract is not "static", and efficiency (and overcoming of private information) is attained eventually.[37]

Krishna, Lopomo, and Taylor [2013] analyze a model similar to ours, but with iid types. Their theoretical results are limited to documenting the long-term efficiency properties, point 9 of Theorem 1; the rest of the theoretical results are novel here. Other than the fact that the model with persistent types is technically harder to analyze, in Section 5.3 we describe the conceptual and empirical importance of the model with persistent types over the iid framework.

The second strand of literature that our paper connects to is dynamic financial contracting. Agency frictions here are motivated in two forms: moral hazard and cash flow diversion. Our paper is the most closely related to Clementi and Hopenhayn [2006] and Fu and Krishna [2017]. Both study the problem of cash flow diversion by the agent in a repeated setting, the former looks at an iid technology and the latter at a Markovian one.

A simple way to map their framework into ours would be to change the time structure of the supply of inputs and payments. At the start of every period the agent commits to a production plan after which his cost type is realized. The type is reported, agreed upon input quantity is supplied, and the agent is compensated for by the principal. The interpretation here is that agent does not know whether his cost would be low or high when he makes the production decision. Despite being a low cost type, he can misreport to be a high cost type, supply some portion of the produced quantity and sell the rest in the black market- a diversion of the economic surplus.

The space of contracts is now defined by $m = \left(p(\theta_t|h^{t-1}), q(h^{t-1})\right)_{t=1}^{T}$. The efficient contract is Markov:

$$V'(q^e(h^{t-1}, \theta_j)) = \alpha_j \theta_L + (1 - \alpha_j)\theta_H$$

Stage utility in this case has a different form: $u(\theta_t|h^{t-1}) = p(\theta_t|h^{t-1}) - \theta_t q(h^{t-1})$. The incentive

[36]There are other papers that seek to temper the flexibility allowed by the linearity of transfers in dynamic models. Guo and Hörner [2017] analyze a dynamic screening model without transfers (see Lipnowski and Ramos [2015] for a related model without commitment). Amador et al. [2006], and Halac and Yared [2014] study models of delegation. Thomas and Worrall [1990], Garrett and Pavan [2015], Luz [2015], and Arve and Martimort [2016] consider dynamic models of private information where the agent is risk averse. Our paper presents a tractable model of restricting the feasible set of transfers which in turn interacts with incentives to produce distinctly different allocation rules.

[37]In a similar vein, Ashlagi et al. [2016] consider a framework where a monopolist wants to sell k goods in k periods, valuations are iid over time, and the mechanism must satisfy ex post individual rationality. They provide an implementation of the optimal mechanism through delayed payments where all the utility is paid in the last period. We look at a different model, solve for the optimal allocation rule, and provide a plausible implementation.

constraint is given by:

$$U(\theta_t|h^{t-1}) \geq p(\hat{\theta}_t|h^{t-1}) - \theta_t q(h^{t-1}) + \delta \mathbb{E}\left[U(\tilde{\theta}_{t+1}|h^{t-1}, \hat{\theta}_t) \mid \theta_t\right]$$

and cash-strapped constraint simply by $p(\theta_t|h^{t-1}) - \theta_t q(h^{t-1}) \geq 0$. An immediate consequence of the time structure is that any finite horizon version of the modified model has a trivial solution. In the last period the principal does not have enough instruments to screen types, so pools both at the best contract for the high cost type. By backward induction the same is true for all periods.

Drawing on Clementi and Hopenhayn [2006], we can show that in the modified model, points 6 and 9 of Theorem 1 will continue to hold for iid types. Fu and Krishna [2017] show that points 5 and 8 also hold in the Markov extension of Clementi and Hopenhayn [2006]. Neither paper though can explicitly characterize the optimal distortions, that is, points 1, 2 and 3 in Theorem 1 are unique to our paper. As a consequence, points 4 and 7 too are novel.[38]

The literature on dynamic financial contracting is also seeped in plausible implementations of the optimal allocation rules. DeMarzo and Fishman [2007] and Biais et al. [2007] are leading references, and Golosov, Tsyvinski, and Werquin [2016] provide an excellent survey of the techniques with an emphasis on the applications to macroeconomics and finance. We show that such implementations have natural interpretation in the corresponding screening and adverse selection models through the dynamic information operator.

8 Conclusion

This paper motivates the study of financial constraints in dynamic contracting through the interaction between persistent asymmetric information and cash or liquidity constraints. The agent has access to a viable technology marred by agency frictions, and is strapped for cash. The paper situates itself in between the literatures on dynamic mechanism design and dynamic financial contracting.

In the appendix, we discuss in detail how our results extend to the model where types evolve according to an asymmetric Markov chain and general iid distributions. Theoretically speaking, the paper is limited to the two-types model because it is difficult to pin down the optimal allocation rule with more than two Markov types. Global incentive constraints are generically binding for high persistence, even for the benchmark model (see Battaglini and Lamba [2017]). Looking for approximate optimal and easily characterizable contracts is a promising approach going forward. Moreover, a missing link from our analysis, present in Clementi and Hopenhayn [2006], is liquidation or scrapping of the contract between the principal and agent. There are sequences of types for which the value of the "meta firm" could become very low, and perhaps the fire sale value of assets at that point is greater than expected value of continuation. The evolution of liquidation decisions in dynamic financial contracting with persistent agency frictions is a worthwhile question for future research.

Finally, the ideas developed in the paper potentially hold promise for other economically meaningful questions such as optimal taxation and double auctions. In optimal taxation, the agent is the

[38]The continuous time limit of both models is the same. Therefore, the analysis in Section 6 suggests that the continuous time limit of the cash flow diversion model will also inherit the properties of Theorem 1, except of course liquidity as a penultimate zone towards efficiency.

citizen with privately observed labor productivity. The principal is the government seeking to maximizing a Pareto-weighted welfare function. Presumable the government cannot force the citizens to consume below a certain threshold in any given period. Similarly, in repeated transactions in financial markets, a double auction with liquidity constraints involving buyers and sellers with privately observed values seems like a reasonable baseline model which could generate attractive properties.

9 Appendix

We divide the appendix into eleven subsections- the benchmark model, two period model, the sequential approach, followed by the recursive approach, then the main theorem, the optimal limit contract, interpretation of cash strapped as limited liability, dynamics of payments, general IID model, sufficiency conditions for global optimality, and finally the model in continuous time. Throughout we will invoke the general model where $f(\theta_L|\theta_i) = \alpha_i$ for $i = L, H$. We shall assume the following, their role aptly explained by their title:

(A1) Persistence: $\alpha_L, 1 - \alpha_H \geq 1/2$.

(A2) Limited asymmetry: $1 - \alpha_L \geq \alpha_H \geq (1 - \alpha_L)\alpha_L$.

(A3) Ranking of prior: $\alpha_H \leq \mu_L \leq \alpha_L$.

(A1) is assumed throughout. (A2) is used in constructing the shell in the recursive approach. (A3) ensures that the optimal contract starts in the shell.

9.1 Benchmark: dynamic model without financial constraints

Consider a relaxed problem where the principal chooses to maximize $\overline{S} - [\mu_L U(\theta_L) + \mu_H U(\theta_H)]$ subject to $IC_L(h^{t-1})$ and $IR_H(h^{t-1})$ $\forall$ h^{t-1}, $\forall$ t. All constraints can be assumed to hold as an equality, and it can be easily shown that the solution to the relaxed problem is globally optimal (see Battaglini [2005]).

Inductively applying the binding $IC_L(\theta_H^{t-1})$ gives us

$$U(\theta_L) = U(\theta_H) + \Delta\theta \sum_{t=1}^{T} \delta^{t-1}(\alpha_L - \alpha_H)^{t-1} q(\theta_H|\theta_H^{t-1})$$

Substituting back into the objective function, we get that

$$q^{\#}(\theta_L|h^{t-1}) = q^e(\theta_L)\ \forall h^{t-1},\ \forall\ t, \text{ and } q^{\#}(\theta_H|h^{t-1}) = q^e(\theta_H)\ \forall h^{t-1} \neq \theta_H^{t-1},\ \forall\ t$$

$$q^{\#}(\theta_H|\theta_H^{t-1}) = Q\left((1-\lambda)\frac{\mu_L}{\mu_H}\left(\frac{\alpha_L - \alpha_H}{1-\alpha_H}\right)^{t-1}\right)$$

where λ is the Lagrange multiplier on (PK) satisfying $(1 - \lambda)U^{\#}(\theta_H) = 0$ with complementary slackness. It is routine to show that $\lambda \in [0, 1]$, and $\lambda < 1$ if and only if $v_0 \leq \overline{v} := \Delta\theta \sum_{t=1}^{T} \delta^{t-1}(\alpha_L - \alpha_H)^{t-1} q^e(\theta_H)$.

9.2 Two period model

In this section, we prove Proposition 1 and its corollaries. We first establish the set of binding constraints for the relaxed problem. For the mechanism $\langle \mathbf{U}, \mathbf{q} \rangle$, the constraints can be written as:

$$\begin{aligned}
IC_L: &\quad U(\theta_L) \geq U(\theta_H) + \Delta\theta q(\theta_H) + \delta(\alpha_L - \alpha_H)\left[u(\theta_L|\theta_H) - u(\theta_H|\theta_H)\right] \\
C_i: &\quad U(\theta_i) = u(\theta_i) + \delta\left[(\alpha_i u(\theta_L|\theta_i) + (1-\alpha_i)u(\theta_H|\theta_i)\right] \quad \text{and } u(\theta_i) \geq 0 \quad \text{for } i = L, H \\
IC_L(\theta_i): &\quad u(\theta_L|\theta_i) \geq u(\theta_H|\theta_i) + \Delta\theta q(\theta_H|\theta_i) \quad \text{for } i = L, H \\
C_H(\theta_i): &\quad u(\theta_H|\theta_i) \geq 0 \quad \text{for } i = L, H
\end{aligned}$$

Notice that the first-period constraints are relaxed when $u(\theta_H|\theta_i)$ and $u(\theta_L|\theta_i) - u(\theta_H|\theta_i)$ are decreased. Therefore, the second-period constraints can be assumed to hold as an equality.

Let $\mu_L\beta$ and $\mu_L\eta$ be the Lagrange multipliers on IC_L and C_L, respectively. As in the benchmark model, λ is the Lagrange multiplier on (PK). Since the problem is concave, the optimal allocation is described by

$$\begin{aligned}
&q^*(\theta_L) = q^e(\theta_L), \text{ and } q^*(\theta_H) = Q\left(\beta\frac{\mu_L}{\mu_H}\right) \\
&q^*(\theta_L|\theta_L) = q^e(\theta_L), \text{ and } q^*(\theta_H|\theta_L) = Q\left(\eta\frac{\alpha_L}{1-\alpha_L}\right) \\
&q^*(\theta_L|\theta_H) = q^e(\theta_L), \text{ and } q^*(\theta_H|\theta_H) = Q\left(\beta\frac{\mu_L}{\mu_H}\left(\frac{\alpha_L}{1-\alpha_H}\right) + (1-\lambda)\frac{\alpha_H}{1-\alpha_H}\right)
\end{aligned}$$

such that $\beta + \eta = 1 - \lambda$ and $u^*(\theta_H)[\mu_L\beta + \mu_H(1-\lambda)] = 0$. Clearly, $\lambda \in [0, 1]$, and $\lambda < 1$ if and only if $v_0 \leq \overline{v} = \Delta\theta \sum_{t=1}^{2} \delta^{t-1}\mathbb{P}(\theta_t = \theta_L)q^e(\theta_H)$.

We claim that IC_L can be assumed hold as an equality. Suppose that IC_L is satisfied as a strict inequality at the optimum, then $\beta = 0$ and $q^*(\theta_H) = q^e(\theta_H) \geq q^*(\theta_H|\theta_L)$. It follows that C_L is also satisfied as a strict inequality, $\eta = 1 - \lambda = 0$. Therefore, the optimal contract is simply the efficient contract. Since transfers are not uniquely pinned down, we can pick one possible implementation of the efficient contract where IC_L holds as an equality.

Finally, we prove two corollaries. Corollary 1 is equivalent to $q^*(\theta_H|\theta_L) \geq q^*(\theta_H|\theta_H)$ which is satisfied trivially no matter if C_L binds or not. Corollary 2 says that $S^*(v_0) = R^*(v_0) + \max\{\underline{v}, v_0\}$ is non-decreasing, and it is strictly increasing on $[\underline{v}, \overline{v}]$ where $\underline{v}$ is total expected utility for the agent when (PK) is ignored. For $v_0 \leq \underline{v}$, (PK) does not bind and $S^*(v_0) = S^*(\underline{v})$. On other hand, (PK) binds whenever $v_0 > \underline{v}^*$. Since R^* is concave, the set of its sub-differentials is non-empty and consists of all $-\lambda$ supporting the optimum. Recall that $\lambda < 1$ for $v_0 < \overline{v}$ implying that S^* is strictly increasing on $[\underline{v}, \overline{v}]$.

9.3 Sequential approach with $T = \infty$

The set of constraints in $(\mathcal{RP}^*)$ can be enlisted as follows:

$$\begin{aligned}
IC_L(h^{t-1}): &\quad U(\theta_L|h^{t-1}) \geq U(\theta_H|h^{t-1}) + \Delta\theta q(\theta_H|h^{t-1}) + \delta(\alpha_L - \alpha_H)\left[U(\theta_L|h^{t-1}, \theta_H) - U(\theta_H|h^{t-1}, \theta_H)\right] \\
C_i(h^{t-1}): &\quad U(\theta_i|h^{t-1}) \geq \delta\left[\alpha_i U(\theta_L|h^{t-1}, \theta_i) + (1-\alpha_i)U(\theta_H|h^{t-1}, \theta_i)\right] \quad \text{for } i = L, H
\end{aligned}$$

As in the two period problem, each $C_H(h^{t-1})$ with $h^{t-1} \neq \emptyset$ can be assumed to hold as an equality. In addition, C_H binds for $v_0 < \overline{v} = \Delta\theta \sum_{t=1}^{\infty} \delta^{t-1}\mathbb{P}(\theta_t = \theta_L)q^e(\theta_H)$, and the optimal allocation is efficient for $v_0 \geq \overline{v}$.

Finally, the incentive compatibility constraints hold as an equality at the optimum. We show this within the recursive approach introduced in the next section, see the proof of Claim 5, Part 3. Using the set of binding constraints, we can write $C_L(h^{t-1})$ in terms of quantities:

$$C_L(h^{t-1}): \quad \sum_{s=0}^{\infty} \delta^s p_s q(\theta_H|h^{t-1}, \theta_H^s) \geq \sum_{s=1}^{\infty} \delta^s p_s q(\theta_H|h^{t-1}, \theta_L, \theta_H^{s-1})$$

where $p_s = \mathbb{P}(\theta_{t+s} = \theta_L|\theta_t = \theta_L) = \frac{\alpha_H + (1-\alpha_L)(\alpha_L - \alpha_H)^s}{1-\alpha_L+\alpha_H}$.

In Claim 4, we show that the optimal contract is interior, and therefore it can be characterized using the Lagrangian method with multipliers in l^1. So, letting $\delta^{t-1}\zeta(h^{t-1})$ be the multiplier on $C_L(h^{t-1})$, we get the optimal allocation for $v_0 \leq \overline{v}$:

$$q^*(\theta_L|h^{t-1}) = q^e(\theta_L) \quad \forall h^{t-1}, \ \forall\, t$$

$$q^*(\theta_H|h^{t-1}, \theta_L, \theta_H^{s-1}) = Q\left(\zeta(h^{t-1})\frac{p_s}{\mathbb{P}(h^{t-1}\theta_L\theta_H^s)} - \sum_{\tau=0}^{s-1}\zeta(h^{t-1}, \theta_L, \theta_H^\tau)\frac{p_{s-1-\tau}}{\mathbb{P}(h^{t-1}\theta_L\theta_H^s)}\right) \quad \forall h^{t-1}, \ \forall\, t$$

$$q^*(\theta_H|\theta_H^{t-1}) = Q\left((1-\lambda)\frac{\mathbb{P}(\theta_t = \theta_L)}{\mathbb{P}(\theta_H^t)} - \sum_{\tau=0}^{t-1}\zeta(\theta_H^\tau)\frac{p_{t-1-\tau}}{\mathbb{P}(\theta_H^t)}\right)$$

where λ is the multiplier on (PK).

One can immediately note that for positive values of η, distortions are pervasive. A binding $C_L(h^{t-1})$ leaves a legacy of distortions on all high cost quantities that follow- $q^*(\theta_H|h^{t-1}, \theta_L, \theta_H^{s-1})$. It is also important to note that distortions are a function of shadow prices as measured from the last time a low cost type was realized. However, it is hard to drive home general arguments about the nature of dynamic distortions because ηs are endogenous and jointly determined at the optimum.

9.4 Recursive approach

In this section, we convert $(\mathcal{RP}^*)$ into its recursive avatar. The recursive formulations have been defined as $(\mathcal{RF})$ and $(\mathcal{RF}_0)$ in Section 3.4. First, $(\mathcal{RF})$ can be restated for the general Markovian framework as follows:

$$(\mathcal{RF}) \qquad Q_j^*(\mathbf{w}) = \max_{\langle \mathbf{z}_L, \mathbf{z}_H, \mathbf{q}\rangle} \quad \alpha_j[s(\theta_L, q_L) + \delta Q_L^*(\mathbf{z_L})] + (1-\alpha_j)[s(\theta_H, q_H) + \delta Q_H^*(\mathbf{z_H})]$$

subject to $\langle \mathbf{z_L}, \mathbf{z_H}, \mathbf{q}\rangle \in W^2 \times \mathbb{R}_+^2$, and

$$w_L - w_H \geq \Delta\theta q_H + \delta(\alpha_L - \alpha_H)(z_{HL} - z_{HH})$$
$$w_L \geq \delta[\alpha_L z_{LL} + (1-\alpha_L)z_{LH}]$$
$$w_H \geq \delta[\alpha_H z_{HL} + (1-\alpha_H)z_{HH}]$$

$(\mathcal{RF}_0)$ can similarly be rewritten for the general model.

The rest of the section is divided into six claims. In Claims 1 and 2 we describe the recursive domain and the efficiency set, respectively. Next in Claim 3, we show that the optimal contract exists and the recursive formulation can be used to obtain it. Then, we discuss several standard properties of the value function in Claim 4. Finally, claims5 and 6 form the core of our analysis; the former constructs the shell introduced in the main text and the latter describes the behavior of the optimal contract within the shell, and its transition to the efficiency set.

Claim 1 (Recursive domain)**.** $W = \{\mathbf{w} \in \mathbb{R}^2_+ : w_L \geq w_H\}$.

Proof. The cash-strapped constraint implies $W \subseteq \mathbb{R}^2_+$. In addition, it is easy to see that $\{\mathbf{w} \in \mathbb{R}^2_+ : w_L \geq w_H\} \subseteq W$. We prove the converse inclusion by iterative approximations of W.

Relax the constraints set in $\boxed{\star}$ ignoring the low cost cash-strapped constraints, and let $\tilde{W}$ be the set of $\mathbf{w} \in \mathbb{R}^2_+$ such that this constraints set is non-empty. Of course, $W \subseteq \tilde{W}$ and $\tilde{W}$ does not depend on $h^t = (h^{t-1}, \theta_j)$. Notice that it suffices to consider $\mathbf{q} = 0$ in order to determine $\tilde{W}$.

Denote $w_i = U(\theta_i|h^t)$, $z_{ik} = U(\theta_k|h^t, \theta_i)$ for $i, k = L, H$. First, ignore all constraints at date $t+2$ and later, but $z_i \in \mathbb{R}^2_+$. So, we are left only with $\mathbf{z}_i \in \mathbb{R}^2_+$ and two date $t+1$ constraints:

$$\begin{aligned} w_L - w_H &\geq \delta(\alpha_L - \alpha_H)(z_{HL} - z_{HH}) \\ w_H &\geq \delta[\alpha_H z_{HL} + (1-\alpha_H) z_{HH}] \end{aligned} \tag{9}$$

Let $\tilde{W}_0$ be the set of $\mathbf{w} \in \mathbb{R}^2_+$ such that there exist $\mathbf{z_H}, \mathbf{z_L} \in \mathbb{R}^2_+$ satisfying Equation (9). Then, define recursively $\tilde{W}_l$ as a set of $\mathbf{w} \in \mathbb{R}^2_+$ such that there exist $\mathbf{z_H}, \mathbf{z_L} \in \tilde{W}_{l-1}$ satisfying Equation (9). In other words, $\tilde{W}_l$ is found by ignoring all constraints at date $t+l+2$ and later, but $U(\theta_{t+l+2}|h^{t+l+1}) \geq 0$ $\forall \theta_{t+l+2}$ and $h^{t+l+1} \in H^{t+l+1}|_{h^t}$. We claim that $\tilde{W}_l \subseteq \tilde{W}_{l-1}$ for any l and $\tilde{W} \subseteq \bigcap_{l=0}^{+\infty} \tilde{W}_l = \{\mathbf{w} \in \mathbb{R}^2_+ : w_L \geq w_H\}$.

Fix $a \in [0, 1)$, then let $\tilde{W}^{old}_a = \{\mathbf{w} \in \mathbb{R}^2_+ : w_L \geq a w_H\}$ and define $\tilde{W}^{new}_a = \{\mathbf{w} \in \mathbb{R}^2_+ : \exists(\mathbf{z_H}, \mathbf{z_L}) \in \tilde{W}^{old}_a \times \tilde{W}^{old}_a \text{ s.t. } (9)\}$. Notice that $\mathbf{w}' \in \tilde{W}^{new}_a$ if and only if there exists $\mathbf{w}$ with $w'_L = w_L$, $w'_H \geq w_H = \delta[\alpha_H a + (1-\alpha_H)] z_{HH}$ and $w_L - w_H \geq \delta(\alpha_L - \alpha_H)(a-1) z_{HH}$. It follows that $\tilde{W}^{new}_a = \{w \in \mathbb{R}^2_+ : w_L \geq [1 - \frac{(\alpha_L - \alpha_H)(1-a)}{\alpha_H a + (1-\alpha_H)}] w_H\}$.

So, define $a_0 = 0$, $a_l = 1 - \frac{(\alpha_L - \alpha_H)(1 - a_{l-1})}{\alpha_H a_{l-1} + (1-\alpha_H)}$, then $\tilde{W}_l = \tilde{W}^{old}_{a_l}$. The claim follows from $a_{l-1} < a_l < 1$ for any l and $a_l \to_{l \to \infty} 1$. □

Now, we look at the efficiency set. Formally, for $j = L, H$, $Q_j(\mathbf{w}) < Q^e_j$ $\forall \mathbf{w} \in W - E$ and $Q_j(\mathbf{w}) = Q^e_j$ $\forall \mathbf{w} \in E$ where $\mathbf{Q}^e$ solves

$$Q^e_j = \alpha_j[s(\theta_L, q^e(\theta_L)) + \delta Q^e_L] + (1-\alpha_j)[s(\theta_H, q^e(\theta_H)) + \delta Q^e_H]$$

To characterize the set, define $\kappa = \frac{\Delta\theta q^e(\theta_H)}{1 - \delta(\alpha_L - \alpha_H)}$, $(1-\delta)\underline{w}^e_H = \delta\alpha_H\kappa$ and $\underline{w}^e_L = \underline{w}^e_H + \kappa$. Clearly, $\{\mathbf{w} \in W : w_H \geq \underline{w}^e_H \text{ and } w_L \geq w_H + \kappa\} \subseteq E$.

Claim 2 (Efficiency set)**.** $E = \{\mathbf{w} \in W : w_H \geq \underline{w}^e_H \text{ and } w_L \geq w_H + \kappa\}$.

Proof. The proof is similar to Claim 1, see the online appendix for details. □

Remark 1 (Liquidity)**.** *Clearly, if there is some $\mathbf{z}_i \in E$ satisfying the constraints, then this $\mathbf{z}_i$ is optimal. It is easy to see that $\mathbf{z}_H \in E$ if only if $\mathbf{w} \in E$. Moreover, $\mathbf{z}_L \in E$ if and only if $w_L \geq \underline{w}^{liq}_L =$*

$\delta[\alpha_L w_L^e + (1-\alpha_L)w_H^e] < \underline{w}_L^e$, *because* $\mathbf{w}^e$ *is the smallest point of the efficiency set. So, a transition to efficiency is possible from outside of* E*, and it requires a "good shock" provided that* $w_L \geq \underline{w}_L^{liq}$*. We shall show in Claim 6 that the low cost type is liquid if and only if it is possible to transit to efficiency.*

To make our next claim formal, we need several auxiliary objects. Let $(\mathcal{Z}_L(\mathbf{w}), \mathcal{Z}_H(\mathbf{w}), \mathcal{Q}(\mathbf{w}))$ be the set of maximizers in the problem $(\mathcal{RF})$ given $\mathbf{w}$ and θ_j. Importantly, this set is independent of θ_j, because of the structure of the problem. The policy correspondence is a correspondence which maps $\mathbf{w}$ into $(\mathcal{Z}_L(\mathbf{w}), \mathcal{Z}_H(\mathbf{w}), \mathcal{Q}(\mathbf{w}))$. We say that a contract is generated from the policy correspondence when for $i, k = L, H$ and $\forall h^t, \forall t$

$$U(\theta_k|h^t, \theta_i) \in \mathcal{Z}_{ik}\left(U(\theta_L|h^t), U(\theta_H|h^t)\right)$$
$$q(\theta_i|h^t) \in \mathcal{Q}_i\left(U(\theta_L|h^t), U(\theta_H|h^t)\right)$$

Claim 3 (Validity of the recursive approach)**.**

1. *There exists a unique continuous bounded function satisfying the Bellman equation in* $(\mathcal{RF})$*.*
2. *The policy correspondence is non-empty, compact-valued and upper hemicontinuous.*
3. *A contract is generated from the policy correspondence if and only if it solves* $\boxed{\star}$ $\forall h^t, \forall t$*, with* $\mathbf{w} = (U(\theta_L|h^t), U(\theta_H|h^t))$*.*
4. *Value functions in* $\boxed{\star}$ *and* $(\mathcal{RF})$*, and in* $(\mathcal{RP}^*)$ *and* $(\mathcal{RF}_0)$ *coincide.*

Proof. See Exercises 9.4, 9.5 and 9.7 of Stokey et al. [1989]. □

We continue this section with the claim showing standard properties of the value function such as concavity, supermodularity and differentiability.

Claim 4 (Properties of the value function)**.**

1. *Each* Q_j^* *is concave.*
2. *Each* Q_j^* *is supermodular.*
3. *Each* Q_j^* *is continuously differentiable on* $int(W)$ *with*

$$\lim_{w_L \to w_H} D_L Q_j^*(\mathbf{w}) = \infty \; \forall \; w_H \text{ and } \lim_{w_H \to 0} D_H Q_j^*(\mathbf{w}) = \infty \; \forall \; w_L \neq 0$$

4. *Each* Q_j^* *is strictly concave in* w_L *and* w_H *on*

$$H = \{\mathbf{w} \in int(W) : DQ_L^*(\mathbf{w}) \gg 0 \text{ and } DQ_H^*(\mathbf{w}) \gg 0\}$$

Proof. See the online appendix. □

Now, we derive the optimality conditions which turn out to be useful for our characterization of the optimal contract. Let $(1-\alpha_j)\beta$, $\alpha_j \rho_L$ and $(1-\alpha_j)\rho_H$ be the Lagrange multipliers for the

respective constraints in $(\mathcal{RF})$. Let $w \in int(W)$. Since the optimum is interior by Claim 4, it is characterized by the following first-order conditions:

$$\begin{aligned} &D_L Q_L^*(\mathbf{z}_L) = \alpha_L \rho_L \\ &D_H Q_L^*(\mathbf{z}_L) = (1-\alpha_L)\rho_L \\ &D_L Q_H^*(\mathbf{z}_H) = \alpha_H \rho_H + (\alpha_L - \alpha_H)\beta \\ &D_H Q_H^*(\mathbf{z}_H) = (1-\alpha_H)\rho_H - (\alpha_L - \alpha_H)\beta \\ &D_q s(\theta_H, q_H) = \Delta\theta\beta \end{aligned} \tag{10}$$

In addition, the following envelope conditions are satisfied:

$$\begin{aligned} &D_L Q_L^*(\mathbf{w}) = \alpha_L \rho_L + (1-\alpha_L)\beta \\ &D_H Q_L^*(\mathbf{w}) = (1-\alpha_L)(\rho_H - \beta) \\ &D_L Q_H^*(\mathbf{w}) = \alpha_H \rho_L + (1-\alpha_H)\beta \\ &D_H Q_H^*(\mathbf{w}) = (1-\alpha_H)(\rho_H - \beta) \end{aligned} \tag{11}$$

At the initial date, the problem is different. Let λ be the multiplier on (PK) in $(\mathcal{RF}_0)$, and $\mu_H\beta$, $\mu_L\rho_L$ and $\mu_H\rho_H$ be the other multipliers. The first-order conditions with respect to $\mathbf{z}_L$, $\mathbf{z}_H$ and q_H are the same as in Equation (10). The extra first-order conditions are

$$\begin{aligned} &\mu_L\rho_L + \mu_H\beta = \mu_L\lambda \\ &\mu_H(\rho_H - \beta) = \mu_H\lambda \end{aligned} \tag{12}$$

We proceed by characterizing the shell, the optimal contract and its dynamics. The shell is extremely important, because the optimal contract always lies in this set (Claim 6), so we start with it. As in the main text, define $\eta_j(\mathbf{w}) = (1-\alpha_j)D_L Q_j^*(\mathbf{w}) - \alpha_j D_H Q_j^*(\mathbf{w})$ for $j = L, H$ and $\mathbf{w} \in int(W)$. Formally, the shell is defined as:

$$B = \{\mathbf{w} \in W \cap (0, \underline{w}_L^e) \times (0, \underline{w}_H^e) : \eta_L(\mathbf{w}) \leq 0 \leq \eta_H(\mathbf{w})\}$$

We focus on the case with $\alpha_L \neq \alpha_H$, the generalized IID model is discussed in the next section. The following claim establishes that the shell looks like the shaded area in Figure 4. It is the intersection of epigraph and hypograph of two strictly increasing, continuous functions connecting 0 and $\underline{\mathbf{w}}^e$. The shell has a non-empty interior and it lies above the line connecting 0 and $\underline{\mathbf{w}}^e$.

Claim 5 (Shape of the shell)**.**

1. *For $j = L, H$ and $\forall w_H \in (0, \underline{w}_H^e)$, $\exists$ unique $w_L^j(w_H) \in (0, \underline{w}_L^e)$ such that $\eta_j(w_L^j(w_H), w_H) = 0$.*
2. *Each w_L^j is continuous and strictly increasing with*

$$\lim_{w_H \to 0} w_L^j(w_H) = 0 \quad \textit{and} \quad \lim_{w_H \to \underline{w}_H^e} w_L^j(w_H) = \underline{w}_L^e.$$

3. *$w_L^H > w_L^L$ on $(0, \underline{w}_H^e)$.*

4. $w_L^L > \frac{\delta\alpha_H w_H}{1-\delta(1-\alpha_H)}$ *on* $(0, \underline{w}_H^e)$.

Proof.

Parts 1 and 2. First, $\{w_L \in (w_H, \underline{w}_L^e] : \eta_j(\mathbf{w}) = 0\} \neq \emptyset$ for any $w_H \in (0, \underline{w}_H^e)$, because η_j is continuous in w_L with $\lim_{w_L \to w_H} \eta_j(\mathbf{w}) = +\infty$ and $\eta_j(\underline{w}_L^e, w_H) \leq 0$ by Claim 4.

Next, we want to show that $B \subseteq H$. If $\mathbf{w} \in B$ with $D_L Q_L^*(\mathbf{w}) > 0$, then $D_H Q_j^*(\mathbf{w}) > 0$ for $j = L, H$ by the definition of η_L. Consider $D_L Q_L^*(\mathbf{w}) = 0$ which imply $\rho_L = \beta = 0$. Clearly, all the Lagrange multipliers can not be zero at the same time for $\mathbf{w} \notin E$ as it implies $DQ_j^*(\mathbf{w}) = 0$ for $j = L, H$. Then, $DQ_j(\mathbf{z}_j) = 0$ for $j = L, H$. Iterating forward, conclude that $\mathbf{w}$ must be in E, a contradiction. So, $\rho_H > 0$ and $\eta_H(\mathbf{w}) < 0$, as a result $\mathbf{w} \notin B$.

Given that $\mathbf{w} \in H$, each Q_j is strictly concave in its coordinates. Uniqueness, continuity and strict monotonicity of w_L^j is due to strict concavity and supermodularity of the value function.

Part 3. Notice that $\eta_j(\mathbf{w}) = \alpha_j(1-\alpha_j)(\rho_L - \rho_H) + (1-\alpha_j)\beta$ by Equation 11, thus

$$\frac{\eta_H(\mathbf{w})}{\alpha_H(1-\alpha_H)} = \frac{\eta_L(\mathbf{w})}{\alpha_L(1-\alpha_L)} + \frac{(\alpha_L - \alpha_H)\beta}{\alpha_L \alpha_H} \tag{13}$$

So, $\eta_H(\mathbf{w}) > \eta_L(\mathbf{w})$ whenever $\beta > 0$, and it suffices to establish $\beta > 0$ in B to prove this part of the claim.

For any $\mathbf{w} \in B$ with $w_L \geq \underline{w}_L^{liq} = \delta[\alpha_L w_L^e + (1-\alpha_L) w_H^e]$, there exists $\mathbf{z}_L \in E$ satisfying the cash-strapped constraint of the low cost type. So, $\rho_L = 0$ and $\beta > 0$, because $\mathbf{w} \in B \subseteq H$ by the first part of the claim above.

It remains to look at $\mathbf{w} \in B$ with $w_L < \underline{w}_L^{liq}$. Consider $\mathbf{w}^0$ such that $\eta_H(\mathbf{w}^0) = 0$ and $w_L^0 \geq \underline{w}_L^{liq}$. Then, $\eta_L(\mathbf{w}^0) < 0$ by Equation (13). There exists $\mathbf{w}^1$ such that $\eta_L(\mathbf{w}^1) = 0$ and $w_H^1 = w_H^0$, $w_L^1 < w_L^0$.

By Lemma in the online appendix, $\beta(w_L', w_H) \geq \beta(w_L, w_H)$ for any $w_L > w_L' > w_H > 0$. In particular $\beta(\mathbf{w}^1) \geq \beta(\mathbf{w}^0) > 0$. Notice that $\eta_H(w^1) > 0$ by Equation (13). Thus, there exists w^2 such that $\eta_H(\mathbf{w}^2) = 0$ and $w_H^2 < w_H^1$, $w_L^2 = w_L^2$. By strict concavity on H, $\eta_L(\mathbf{w}^2) < 0$, implying that $\beta(\mathbf{w}^2) > 0$ by Equation (13). Iteratiing, get that $\beta > 0$ on $\{\mathbf{w} \in (0, \underline{w}_L^e) \times (0, \underline{w}_H^e) : \eta_L(\mathbf{w}) = 0\} \subseteq B$ which implies $\beta > 0$ on B by Lemma in the online appendix.

Part 4. Finally, we argue that $w_L^L > \frac{\delta\alpha_H w_H}{1-\delta(1-\alpha_H)}$ when $\alpha_L \neq \alpha_H$ and (A2) holds.

Take $\mathbf{w} \in W$ with $\eta_L(\mathbf{w}) \leq 0$, then $\alpha_L(\rho_L - \rho_H) + \beta \leq 0$ and (A2) implies that $D_L Q_H^*(\mathbf{z}_H) - D_L Q_H^*(\mathbf{w}) = \alpha_H(\rho_H - \rho_L) - (1-\alpha_L)\beta \geq 0$. In addition, assume that $\beta > 0$, then that $D_H Q_H^*(\mathbf{z}_H) - D_H Q_H^*(\mathbf{w}) = (1-\alpha_L)\beta > 0$. So, $\mathbf{w} \neq \mathbf{z}_H$ and they are ordered by strict concavity and supermodularity. Clearly, the cash strapped constraint for the high type can be assumed to hold as an equality which implies that $w_H \neq \delta[\alpha_H w_L + (1-\alpha_H) w_H]\}$. By the previous part of the claim, our argument implies that B and $\{\mathbf{w} \in (0, \mathbf{w}_L^e) \times (0, \mathbf{w}_H^e) : w_H = \delta[\alpha_H w_L + (1-\alpha_H) w_H]\}$ do not intersect. Even more, if the shell lies below the line connecting 0 and $\underline{\mathbf{w}}^e$, then $\beta = 0$ on this line.

Next, suppose that $w_L^H < \frac{\delta\alpha_H w_H}{1-\delta(1-\alpha_H)}$ on $(0, \underline{w}_H^e)$. And, take $\mathbf{w}$ on the line, $w_H = \delta[\alpha_H w_L + (1-\alpha_H) w_H]$. Since $\beta = 0$, it must be the case that $q_H = q^e(\theta_H)$ and

$$w_L - w_H \geq \Delta\theta q^e(\theta_H) + \delta(\alpha_L - \alpha_H)(z_{HL} - z_{HH})$$

As $\mathbf{w}$ tends to 0 along the line, $\mathbf{z}_H$ also tends to 0 by the cash strapped constraint. The above equation must be violated at some $\mathbf{w}$ close to 0, therefore the shell can not lie below the line. □

Our last claim points out the optimal contract is initialized in the shell, and it stays within the shell until it reaches E. Moreover, while in the shell, a good/bad shock strictly increases/decreases $\mathbf{w}$ in each coordinate, the allocation is monotonically decreasing along the sequence of θ_H's.

Claim 6 (Optimal contract)**.**

1. *The optimal contract is initialized at* $\mathbf{w} \in B$ *such that* $w_L = w_I^L(w_H)$ *where* $w_I^L : (0, \underline{w}_H^e) \to (0, \underline{w}_H^e)$ *is a continuous, strictly increasing function* $w_I^L : (0, \underline{w}_H^e) \to (0, \underline{w}_L^e)$ *such that* $w_L^H \geq w_L^I \geq w_L^L$ *on* $(0, \underline{w}_H^e)$.

2. $\forall \mathbf{w} \in B$, $\mathbf{z}_L \gg \mathbf{w} \geq \mathbf{z}_H$ *with* $w_H > z_{HH}$ *and* $w_H > z_{HL}$ *if* $\eta_L(\mathbf{w}) < 0$.

3. $\mathbf{z}_L \in B$ *whenever* $\mathbf{w} \in B$, $w_L < \underline{w}_L^{liq}$, *and* $\mathbf{z}_L \in E$ *whenever* $\mathbf{w} \in B$, $w_L \geq \underline{w}_L^{liq}$.

4. $z_L(\mathbf{w}) \geq z_L(\mathbf{w}')$ *whenever* $\mathbf{w}, \mathbf{w}' \in B$ *and* $w_L' \leq w_L$, $w_L' \leq \underline{w}_L^{liq}$.

5. $\mathbf{z}_H \in B$ *whenever* $\mathbf{w} \in B$.

6. $\forall \mathbf{w} \in B$, $q_H(w) \geq q_H(\mathbf{z}_H)$ *with a strict inequality when* $\eta_L(\mathbf{w}) < 0$.

7. $\forall \mathbf{w} \in B$, $q_H(\mathbf{z}_L) \geq q_H(\mathbf{w})$.

Proof.

Part 1. By equation (12) the contract is initialized at $\mu_L(\rho_L - \rho_H) + \beta = 0$. Existence of w_I^L and its properties can be easily seen by the same argument as in the first two parts of Claim 5. Then, by Assumption (A3) and Equation (13), $\alpha_L(\rho_L - \rho_H) + \beta \leq 0 \leq \alpha_H(\rho_L - \rho_H) + \beta$ which implies that $w_L^H \geq w_L^I \geq w_L^L$ on $(0, \underline{w}_H^e)$.

Part 3. Clearly, $\mathbf{z}_L \in H$ is such that $\eta_L(\mathbf{z}_L) = 0$ if $\mathbf{w} \in B$, but $w_L < \underline{w}_L^{liq}$. And, $\mathbf{z}_L \in B$, because $H \subseteq (0, \underline{w}_L^e) \times (0, \underline{w}_H^e)$. Notice that $\underline{\mathbf{w}}^e \in int(W)$ and $DQ_j^*(\underline{\mathbf{w}}^e) = 0$ by construction. For $\mathbf{w} \geq \underline{\mathbf{w}}^e$ and $\mathbf{w} \notin E$, there exists $(w_L', w_H) \in E$ with $w_L' > w_L$. Hence, each $D_L Q_j^*(\mathbf{w}) \leq 0 = D_L Q_j^*(w_L', w_H)$ by concavity and supermodularity. And, if $w_L \geq \underline{w}_L^e$ and $w_H \in (0, \underline{w}_H^e)$, then $D_L Q_j^*(\mathbf{w}) \leq 0 = D_L Q_j^*(\underline{\mathbf{w}}^e)$. The case with $w_H \geq \underline{w}_H^e$ and $w_L \in (0, \underline{w}_L^e)$ is similar.

On the other hand, there exists $\mathbf{z}_L \in E$ satisfying the cash-strapped constraint of the low cost type $\forall \mathbf{w} \in B$ with $w_L \geq \underline{w}_L^{liq}$.

Part 2. By the previous part of the claim, $\mathbf{z}_L \in E$ whenever $\mathbf{w} \in B$, $w_L \geq \underline{w}_L^{liq}$, thus $\mathbf{z}_L \gg \mathbf{w}$. Now, Take any $\mathbf{w} \in B$ and $w_L < \underline{w}_L^{liq}$. In this case, $\rho_L > 0$ and the cash strapped constraint for the low cost type holds as an equality:

$$w_L = \delta[\alpha_L z_{LL} + (1 - \alpha_L) z_{LH}] < z_{LL}$$

where we used $\mathbf{z}_L \in int(W)$ and $\delta < 1$. Given that w_L^L is strictly increasing (see Claim 5), $z_{LH} \gg w_H$ must hold as well.

Next, we show that $\mathbf{w} \geq \mathbf{z}_H$ and $w_H > z_{HH}$. In the proof of Part 4 of Claim 5, we argue that if $\mathbf{w} \in B$, $D_L Q_H^*(\mathbf{z}_H) \geq D_L Q_H^*(\mathbf{w})$ and $D_H Q_H^*(\mathbf{z}_H) > D_H Q_H^*(\mathbf{w})$. This implies that $\mathbf{w} \neq \mathbf{z}_H$ are

ordered by strict concavity and supermodularity. Using the fact that $\mathbf{w}$ lies above the line connecting 0 and $\underline{\mathbf{w}}^e$, and the cash strapped constraint:

$$0 \leq \alpha_H(w_L - z_{HL}) + (1 - \alpha_H)(w_H - z_{HH})$$

Clearly, $\mathbf{w} \geq \mathbf{z}_H$ with $w_H > z_{HH}$ is the only possible choice. The assertion could be strengthen to $\mathbf{w} \gg \mathbf{z}_H$ when $\eta_L(\mathbf{z}_H) < 0$ as $D_L Q_H^*(\mathbf{z}_H) > D_L Q_H^*(\mathbf{w})$.

Part 4. Take $\mathbf{w}, \mathbf{w}' \in B$ with $w'_L \leq w_L$, $w'_L \leq \underline{w}_L^{liq}$ and suppose that $\mathbf{z}_L < \mathbf{z}'_L$. By the third part of this claim, the low type cash-strapped constraint binds for both $\mathbf{w}$, $\mathbf{w}'$. Also, Equation 10 yields $\eta_L(\mathbf{z}_L) = \eta_L(\mathbf{z}'_L) = 0$ implying that $\mathbf{z}_L > \mathbf{z}'_L$, because w_L^L is strictly increasing as shown in Claim **??**. This is a clear contradiction.

Part 5. First, we argue that the level curves of η_H in $(0, \underline{w}_L^e) \times (0, \underline{w}_H^e)$ cross $\{\mathbf{w} \in B : \eta_L(\mathbf{w}) = 0\}$ at most once. Suppose not, namely for some $\mathbf{w} \neq \mathbf{w}'$ within the square, $\eta_H(\mathbf{w}) = \eta_H(\mathbf{w}')$. Then, $\beta(\mathbf{w}) = \beta(\mathbf{w}') = \rho_L(\mathbf{w}) - \rho_H(\mathbf{w}) = \rho_L(\mathbf{w}') - \rho_H(\mathbf{w}') = 0$, which is a contradiction to the last part of Claim 5.

Now, take $\mathbf{w} \in B$ with $\eta_L(\mathbf{w}) = 0$. By Equation (13), and the second assumption, $\eta_H(\mathbf{w}) = \frac{1-\alpha_H}{\alpha_L}\eta_H(\mathbf{z}_H) \geq \eta_H(\mathbf{z}_H) = (\alpha_L - \alpha_H)\beta \geq 0$. From $\eta_H(\mathbf{w}) \geq \eta_H(\mathbf{z}_H)$, the fact that the level curves of η_H cross at most once and $w \leq \mathbf{z}_H$, conclude $\mathbf{z}_H \in B$. The general case is implied by monotonicity of β (Lemma in the online appendix) and the previous result.

Part 6. We have shown above that the optimal contract lies in the shell with $\frac{1-\alpha_H}{\alpha_L}\eta_H(\mathbf{z}_H) \geq \eta_H(\mathbf{w})$ and $\eta_H(\mathbf{z}_H) = (\alpha_L - \alpha_H)\beta \geq 0$. Iterating forward on θ_H, $\frac{1-\alpha_H}{\alpha_L}\beta(\mathbf{z}_H) \geq \beta(\mathbf{w})$. Using the first-order condition $D_q s(\theta_H, q_H) = \Delta\theta\beta$, and strict concavity of $s(\theta_H, \cdot)$, conclude that $q_H(\mathbf{z}_H) \leq q_H(\mathbf{w})$. For $\eta_L(\mathbf{w}) < 0$, the result can be easily strengthen to $q_H(\mathbf{z}_H) < q_H(\mathbf{w})$, because $\frac{1-\alpha_H}{\alpha_L}\eta_H(\mathbf{z}_H) > \eta_H(\mathbf{w})$.

Part 7. Clearly, a level curve of η_H can be described by an increasing, continuous function starting at 0 (see the first part of Claim 5). This function must be strictly increasing on B.

In the previous part, we showed that the level curves of η_H cross $\{\mathbf{w} \in B : \eta_L(\mathbf{w}) = 0\}$ at most once. Now, as we are moving along $\{\mathbf{w} \in B : \eta_L(\mathbf{w}) = 0\}$ from 0 to $\underline{\mathbf{w}}^e$, η_H must be strictly decreasing. To see this, take $\mathbf{w} \gg \mathbf{w}' \in B$ with $\eta_L(\mathbf{w}) = \eta_L(\mathbf{w}') = 0$. Then, $\{\tilde{\mathbf{w}} \in B : \eta_H(\tilde{\mathbf{w}}) = \eta_H(\mathbf{w})\}$ is to the left of $\{\tilde{\mathbf{w}}' \in B : \eta_H(\tilde{\mathbf{w}}') = \eta_H(\mathbf{w}')\}$. Consider the section at w'_L, strict concavity yield the desired result.

We have shown that $\eta_H w$ is strictly decreasing as we are move towards $\underline{\mathbf{w}}^e$. By equation (13), β is strictly decreasing, thus q_H is strictly increasing (see equation (10)). □

9.5 Main result

Our main result is a translation of our recursive characterization into sequential notations. By the Claim 3, the contract is optimal if it is generated from the police correspondence and the first period choice of $\mathbf{w}$ solves $(\mathcal{RF}_0)$. Parts A-D of Theorem 1 completely established in Claims in the recursive approach. We briefly describe the connection. The optimal quantity is downward distortion from equation (10): $D_q s(\theta_H, q_H) = \Delta\beta$ and fact that $\beta > 0$ in B, which establishes part 1. Parts 2 and 3 follow from 6 and 7 of Claim 6. Part 4 follows from 2 of Claim 6. Part 5 follows from 3 of Claim 6.

Part 6 of follows from the way set E is constructed in Claim 2. Parts 7 and 8 are implied by 2, 3 and 4 of Claim 6.

It is left to be shown that the optimal contract converges to the efficient allocation, that is part E. The first part of Claim 6 says that the optimal contract is initialized within $B \subseteq int(W)$. Let $D^* = (D_L + D_H)$, by Equations (10) and (11), the stochastic process $D^*Q_j^*$ is a non-negative martingale, namely

$$D^*Q_L^*(\mathbf{z}) = \alpha_L D^*Q_L^*(\mathbf{z}_L) + (1-\alpha_L)D^*Q_H^*(\mathbf{z}_H) \geq 0$$
$$D^*Q_H^*(\mathbf{z}) = \alpha_H D^*Q_L^*(\mathbf{z}_L) + (1-\alpha_H)D^*Q_H^*(\mathbf{z}_H) \geq 0$$

So, the Martingale convergence theorem implies that $D^*Q_j^*$ converges almost surely. Therefore, the Lagrange multipliers are uniquely pinned by the limits through Equation (11). Clearly, w converges to a point in E, hence $q^*(\theta_i|h^{t-1})$ converges to $q^e(\theta_i)$ almost surely.

9.6 Optimal limit contract

In this section, we prove Proposition 2. We shall invoke the sequential approach discussed in Section 9.3. Since we are interested in the limit result as the Markov matrix approaches the identity matrix, we will simply consider the symmetric Markov chain: $\alpha_H = 1 - \alpha_L = \alpha$.

Recall that $C_L(h^{t-1})$ could be expressed as:

$$\sum_{s=0}^{\infty} \delta^s p_s q(\theta_H|h^{t-1}, \theta_H^s) \geq \sum_{s=1}^{\infty} \delta^s p_s q(\theta_H|h^{t-1}, \theta_L, \theta_H^{s-1})$$

where $p_s = \frac{1+(2\alpha-1)^s}{2}$. Let $\delta^{t-1}\mathbb{P}(h^{t-1}\theta_L)(1-\alpha)\zeta(h^{t-1})$ be the Lagrange multiplier on this constraint, then the optimal allocation takes the following form

$$q^*(\theta_H|\theta_H^s) = Q\left((1-\lambda)\frac{1+(\mu_L-\mu_H)(2\alpha-1)^s}{2\mu_H\alpha^s} - \frac{p_s}{\alpha_s}\frac{\mu_L}{\mu_H}(1-\alpha)\zeta - \sum_{j=1}^{s}\frac{p_{s-j}}{\alpha^{s-j}}\frac{(1-\alpha)^2}{\alpha}\zeta(\theta_H^j)\right), \quad s \geq 0$$
$$q^*(\theta_H|h^{t-1}, \theta_L, \theta_H^{s-1}) = Q\left(\frac{p_s}{\alpha_s}\alpha\zeta(h^{t-1}) - \frac{p_{s-1}}{\alpha_{s-1}}\alpha\zeta(h^{t-1}, \theta_L) - \sum_{j=1}^{s-1}\frac{p_{s-1-j}}{\alpha^{s-1-j}}\frac{(1-\alpha)^2}{\alpha}\zeta(h^{t-1}, \theta_L, \theta_H^j)\right), \quad s \geq 1$$

where λ is the Lagrange multiplier on (PK).

Fixing λ, $q^*(\theta_H|\theta_H^s) \xrightarrow[\alpha\to 1]{} Q\left((1-\lambda)\frac{\mu_L}{\mu_H}\right)$ and $q^*(\theta_H|h^{t-1}, \theta_L, \theta_H^{s-1}) \xrightarrow[\alpha\to 1]{} Q\left(\zeta(h^{t-1}) - \zeta(h^{t-1}, \theta_L)\right)$. This implies that the agent's ex ante utility converges to $\frac{\Delta\theta\mu_L}{1-\delta}Q\left((1-\lambda)\frac{\mu_L}{\mu_H}\right)$. Therefore, $\lambda = 0$ for any $v_0 \leq \underline{v} = \frac{\Delta\theta\mu_L}{1-\delta}Q\left(\frac{\mu_L}{\mu_H}\right)$. And, for $v_0 \in (\underline{v}, \overline{v}]$, λ is uniquely pinned down by the (PK) where $\overline{v} = \frac{\Delta\theta\mu_L}{1-\delta}q^e(\theta_H)$:

$$\lambda = 1 - \frac{\mu_H}{\mu_L}Q^{-1}\left(\frac{(1-\delta)v^0}{\Delta\theta\mu_L}\right)$$

Notice that the limiting allocation does not depend on the number of θ_H's since the last θ_L.

Using the cash-strapped constraint, obtain that

$$Q\left((1-\lambda)\frac{\mu_L}{\mu_H}\right) \geq \delta Q\left(\zeta(\theta_H^s) - \zeta(\theta_H^s, \theta_L)\right), \quad s \geq 0$$
$$Q\left(\zeta(h^{t-1}) - \zeta(h^{t-1}, \theta_L)\right) \geq \delta Q\left(\zeta(h^{t-1}, \theta_L, \theta_H^{s-1}) - \zeta(h^{t-1}, \theta_L, \theta_H^{s-1}, \theta_L)\right), \quad s \geq 1$$

First of all , we argue that $x(h^{t-1}) := \zeta(h^{t-1}) - \zeta(h^{t-1}, \theta_L) \geq 0$, and it holds with as an equality if and only if $\zeta(h^{t-1}) = 0$. The first part and the "if" direction follow from the main theorem which says that the optimal distortions are downwards. Considering the "only if" direction, let $x(h^{t-1}) = 0$. Then, $Q(0) > \delta Q\left(x(h^{t-1}, \theta_L)\right)$ which leads to $\zeta(h^{t-1}, \theta_L) = 0$ by the complimentary slackness. So, we could use ζ interchangebly with x as our set of Lagrange multipliers.

Next, we explicitly solve for x. To begin, notice that $x(\theta_H^s)$ appears only in $Q\left((1-\lambda)\frac{\mu_L}{\mu_H}\right) \geq \delta Q\left(x(\theta_H^s)\right)$ and $Q\left(x(\theta_H^s)\right) \geq \delta Q\left(x(\theta_H^s, \theta_L)\right)$. Since higher $x(\theta_H^s)$ relaxes the latter constraint, $Q\left(x(\theta_H^s)\right) = \min\left\{q^e(\theta_H), \frac{1}{\delta} Q\left((1-\lambda)\frac{\mu_L}{\mu_H}\right)\right\}$. By induction, $x(h^{t-1})$ is constant on h^{t-1} with the same number of θ_L's. Then, the exact expression of d_n are obtained using complimentary slackness.

9.7 Credit constraint versus limited liability

Now, we show Proposition 5. Our argument is based on calculations done in the previous section. Since (PK) is ignored, it is akin to assuming $v_0 = 0$. For $\alpha = 0.5$, using the results of Sections 9.3 and 9.5, $\underline{v}^{\#} = \Delta\theta\mu_L q^{\#}(\theta_H) \leq \Delta\theta\mu_L q^*(\theta_H) < \underline{v}^*$, because $\mathbf{q}^* \gg 0$.

The case of $\alpha \approx 1$ is more delicate. Clearly, $\lim_{\alpha\to 1} \underline{v}^{\#} = \lim_{\alpha\to 1} \underline{v}^*$. Thus, we need to compare the rates at which $\underline{v}^{\#}$ and $\underline{v}^*$ converge. Direct calculations yield that the derivate of $\underline{v}^{\#}$ evaluated at $\alpha = 1$ is proportional to $D^{\#}$:

$$D^{\#} := \frac{d}{d\alpha}\frac{\underline{v}^{\#}}{\Delta\theta\mu_H}\bigg|_{\alpha=1} = \frac{\delta}{(1-\delta)^2}\left[2\left(\frac{\mu_L}{\mu_H}\right)Q\left(\frac{\mu_L}{\mu_H}\right) + \left(\frac{\mu_L}{\mu_H}\right)^2 Q'\left(\frac{\mu_L}{\mu_H}\right)\right]$$

By the implicit function theorem, the derivative of $\mathbf{q}^*$ at $\alpha = 1$ is well-defined. Using the expression for $\mathbf{q}^*$:

$$\frac{d}{d\alpha} q^*(\theta_H|\theta_H^s)\bigg|_{\alpha=1} = Q'\left(\frac{\mu_L}{\mu_H}\right)\left(\frac{\mu_L}{\mu_H} d_1 - s\right), \quad s \geq 0$$

Totally differentiating $\underline{v}^*$, obtain that its derivate evaluated at $\alpha = 1$ is proportional to D^*:

$$D^* := \frac{d}{d\alpha}\frac{\underline{v}^*}{\Delta\theta\mu_H}\bigg|_{\alpha=1} = \frac{\delta}{(1-\delta)^2}\left[\left(\frac{\mu_L - \mu_H}{\mu_H}\right)Q\left(\frac{\mu_L}{\mu_H}\right) - \left(\frac{\mu_L}{\mu_H}\right)Q'\left(\frac{\mu_L}{\mu_H}\right)\right] + \frac{1}{1-\delta}\left(\frac{\mu_L}{\mu_H}\right)^2 Q'\left(\frac{\mu_L}{\mu_H}\right) d_1$$

Clearly, for $\alpha \approx 1$, $\underline{v}^* > \underline{v}^{\#}$ if and only if $\underline{v}^{\#}$ is strictly steeper than $\underline{v}^*$ that is $D^* < D^{\#}$. The case of $\underline{v}^* < \underline{v}^{\#}$ is similar.

9.8 Dynamics of payments

Define the promised utility of the agent to be:

$$v^*(\theta_j|h^{t-1}) = \frac{1}{\delta}\left[U^*(\theta_j|h^{t-1}) - u^*(\theta_j|h^{t-1})\right]$$

and the dynamic information operator as:

$$I(h^{t-1}) = \Delta\theta \sum_{s=1}^{\infty} \delta^{s-1}(\alpha_L - \alpha_H)^{s-1} q^*(\theta_H|h^{t-1}, \theta_H^{s-1})$$

The dynamics of payments are as follows. Fix the optimal allocation rule and initial promised utility v_0.[39] Solving the promised utility identity and the "envelope formula" together:

$$\mu_L U^*(\theta_L) + \mu_H U^*(\theta_H) = v_0$$

$$U^*(\theta_L) = U^*(\theta_H) + I$$

gives

$$U^*(\theta_L) = v_0 + \mu_H I \quad \text{and} \quad U^*(\theta_H) = v_0 - \mu_L I \tag{14}$$

Now, $U^*(\theta_i) = u^*(\theta_i) + \delta v^*(\theta_i)$. Choosing $u(\theta_i)$ automatically determines $v(\theta_i)$. Proceeding inductively, we have:

$$\alpha_j U^*(\theta_L|h^{t-1}, \theta_j) + (1-\alpha_j) U^*(\theta_H|h^{t-1}, \theta_j) = v^*(\theta_j|h^{t-1})$$

$$U^*(\theta_L|h^{t-1}, \theta_j) = U^*(\theta_H|h^{t-1}, \theta_j) + I(h^{t-1}, \theta_j)$$

Solving the two equation gives us

$$\begin{aligned} U^*(\theta_L|h^{t-1}, \theta_j) &= v^*(\theta_j|h^{t-1}) + (1-\alpha_j) I(h^{t-1}, \theta_j) \\ U^*(\theta_H|h^{t-1}, \theta_j) &= v^*(\theta_j|h^{t-1}) - \alpha_j I(h^{t-1}, \theta_j) \end{aligned} \tag{15}$$

Starting from promised utility v_0 and choosing per period transfers optimally, equations (14) and (15) inductively define future expected and promised utilities. The proof of Proposition 3 then simply follows from this induction.

To study the relationship between total economic surplus, and promised utility and utility spread, we introduce a new recursive representation. Define promised utility utility spread respectively by

$$\begin{aligned} v(\theta_j|h^{t-1}) &= \alpha_j U(\theta_L|h^{t-1}, \theta_j) + (1-\alpha_j) U(\theta_H|h^{t-1}, \theta_j) \\ U^s(\theta_j|h^{t-1}) &= U(\theta_L|h^{t-1}, \theta_j) - U(\theta_H|h^{t-1}, \theta_j) \end{aligned} \tag{16}$$

Our goal is to show how that expected surplus increases with expected utility and the utility spread.

Denote $w^e = v(\theta_j|h^{t-1})$ and $w^s = U^e(\theta_j|h^{t-1})$ to be the expected utility and utility spread, respectively, that the principal is committed to deliver to the agent after history (h^{t-1}, θ_j). Then,

[39] In case (PK) is not binding, replace v_0 with $\underline{v}$ in equation (14).

$\boxed{\star}$ can be restated in terms of the new state variables as:

$$Q_j^*(\mathbf{w}) = \max_{(v_L, v_H, q)} \alpha_j[s(\theta_L, q_L)) + \delta\tilde{Q}_L(z_L)] + (1-\alpha_j)[s(\theta_H, q_H) + \delta\tilde{Q}_H(z_H)]$$

subject to $\langle \mathbf{z_L}, \mathbf{z_H}, \mathbf{q}\rangle \in W^2 \times \mathbb{R}^2_+$, and

$$\begin{aligned} w^s &\geq \Delta\theta q_H + \delta(\alpha_L - \alpha_H) z_H^s \\ w^e + (1-\alpha_j) w^s &\geq \delta z_L^e \\ w^e - \alpha_j w^s &\geq \delta z_H^e \end{aligned}$$

This problem is recursive, and analogous to $(\mathcal{RF})$. However, there are some key differences. The recursive domain W_j now depends on θ_j. Indeed, using the transformations in Equation (16), and the previous recursive domain, one could show that $W_j = \{w \in \mathbb{R}^2_+ : w^e \geq \alpha_j w^s\}$. In addition, $\tilde{Q}_j$ is well-defined, bounded, continuously differentiable and concave as it is obtained from Q_j by the linear transformation of variables

$$\tilde{Q}_j(\mathbf{w}) = Q_j[w^e + (1-\alpha_j)w^s, w^e - \alpha_j w^s]$$

Let $(1-\alpha_j)\beta$, $\alpha_j\rho_L$ and $(1-\alpha_j)\rho_H$ be the Lagrange multipliers. Then the first-order and envelope conditions are given by

$$\begin{aligned} D_e\tilde{Q}_L(v_L) &= \rho_L \\ D_s\tilde{Q}_L(v_L) &= 0 \\ D_e\tilde{Q}_H(v_H) &= \rho_H \\ D_s\tilde{Q}_H(v_H) &= (\alpha_L - \alpha_H)\beta \\ D_q s(q_H, \theta_H) &= \Delta\theta\beta \end{aligned} \tag{17}$$

$$\begin{aligned} D_e\tilde{Q}_L(w) &= \alpha_L\rho_L + (1-\alpha_L)\rho_H \\ D_s\tilde{Q}_L(w) &= \alpha_L(1-\alpha_L)(\rho_L - \rho_H) + (1-\alpha_L)\beta \\ D_e\tilde{Q}_H(w) &= \alpha_H\rho_L + (1-\alpha_H)\rho_H \\ D_s\tilde{Q}_H(w) &= \alpha_H(1-\alpha_H)(\rho_L - \rho_H) + (1-\alpha_H)\beta \end{aligned} \tag{18}$$

In principle, the multipliers and the solution depend on θ_j, but we omit this dependence to ease notation. It follows from Equations (17) and (18) that the expected surplus is non-decreasing in the expected utility (globally) and it is non-decreasing in the utility spread (for the optimal contract): $D\tilde{Q}_j(w) \geq 0$. Finally, it is straightforward to show that the ex ante value of economic surplus is increasing in the initial promised utility, v_0. This proves Proposition 4.

9.9 General IID model

We show how to solve the model with the independent types. Suppose that $\alpha_L = 1 - \alpha_H = \mu_L$, then $\eta_L = \eta_H$ by equation (13) implying that the optimal contract lives on a one-dimensional curve.

To characterize the optimal contract, it suffices to have only one state variable, namely expected promised utility. Notice that $Q_L^* = Q_H^*$, then $\forall w \geq 0$ define Q^* by

$$Q^*(w) = \max_{\mathbf{z} \in W} Q_j^*(\mathbf{z}) \text{ s.t. } w = \mu_L z_L + \mu_H z_H \tag{19}$$

This definition is based on the problem $(\mathcal{RF})$, the problem $(\mathcal{RF}_0)$ is trivially modified. Importantly, that the value function in equation (19) solves the simpler Belman equation $(\mathcal{RF}')$.

$$(\mathcal{RF}') \qquad Q^*(w) = \max_{\langle \mathbf{z_L}, \mathbf{z_H}, \mathbf{q} \rangle} \quad \mu_L[s(\theta_L, q_L) + \delta Q^*(z_L)] + \mu_H[s(\theta_H, q_H) + \delta Q^*(z_H)]$$

subject to $\langle \mathbf{u}, \mathbf{z}, \mathbf{q} \rangle \in \mathbb{R}_+^6$, and

$$\begin{aligned} w &= \mu_L(u_L + \delta z_L) + \mu_H(u_H + \delta z_H) \\ u_L + \delta z_L &\geq \Delta\theta q_H + u_H + \delta z_H \end{aligned}$$

The problem $(\mathcal{RF}')$ inherits many properties of the original problem and it has a simpler structure. In particular, Q^* is well-defined and unique in the space of continuous bounded functions. Let $Q^e = \mu_L Q_L^e + \mu_H Q_H^e$, then $Q^* \leq Q^e$ and $Q^* = Q^e$ if and only if $w \geq \underline{w}^e = \mu_L \underline{w}_L^e + \mu_H \underline{w}_H^e$. In addition, Q^* is continuously differentiable on $(0, +\infty)$ with a unbounded right derivative at 0 and strictly increasing, concave on $(0, \underline{w}^e)$.

Consider $w \in (0, \underline{w}^e)$. Given the shape of Q^*, it is easy to see that the constraints in the problem $(\mathcal{RF}')$ could be rewritten as $0 \leq \delta z_H = w - \mu_H \Delta\theta q_H$ and $0 \leq \delta z_L \leq w + \mu_L \Delta\theta q_H$. This implies that $0 < z_H < z_L \leq \underline{w}^e$ and there exists $\underline{w}^{liq} \in (0, \underline{w}^e)$ such that $z_L = \underline{w}^e$ if and only if $w \geq \underline{w}^{liq}$. Finally, z_H is strictly increasing on $(0, \underline{w}^e)$, z_L is also strictly increasing on $(0, \underline{w}^{liq})$ and $0 < q_H < q^e(\theta_H)$ is strictly increasing on $(0, \underline{w}^e)$.

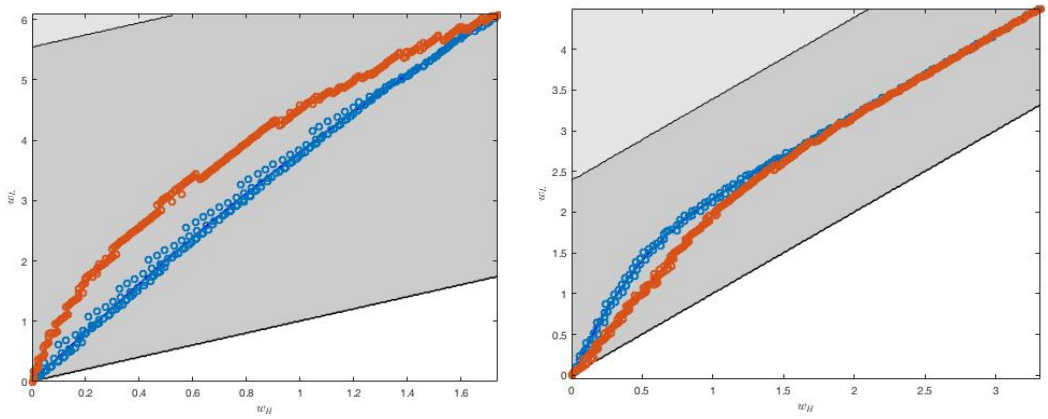
9.10 Sufficiency conditions and global optimality

We say that the first-order approach is valid if the solution to $(\mathcal{RP}^*)$ defined in section 3.3 is incentive compatible, that is the high cost type or "upward" incentive constraints do not bind at the optimum. In the two period model discussed in section 3.2 the "upward" incentive constraint, IC_H, never binds. It is possible, however as we argue largely implausible, that for a long enough time horizon and large enough discount factor the "upward" incentive constraint may bind. In a nutshell, the measure of parameters for which we need to add the "upward" incentive constraint to the relaxed problem after some history is very small and therefore the economic message delivered by our solution worth consideration.

After any history h^{t-1}, using the set of binding constraints in $(\mathcal{RP}^*)$, the "upward" incentive constraint and the cash-strapped constraint can respectively be expressed as:

$$IC_H(h^{t-1}): \quad q^e(\theta_L) + \sum_{s=1}^{\infty} \delta^s (\alpha_L - \alpha_H)^s q(\theta_H | h^{t-1}, \theta_L, \theta_H^{s-1}) \geq \sum_{s=0}^{\infty} \delta^s (\alpha_L - \alpha_H)^s q(\theta_H | h^{t-1}, \theta_H^s)$$

$$C_L(h^{t-1}): \quad \sum_{s=0}^{\infty} \delta^s a_s q(\theta_H | h^{t-1}, \theta_H^s) \geq \sum_{s=1}^{\infty} \delta^s a_s q(\theta_H | h^{t-1}, \theta_L, \theta_H^{s-1})$$



(a) $\alpha = 0.9$ (b) $\alpha = 0.6$

Figure 8: Numerical examples depicting the shell and the region where $\mathcal{RP}^*$ is valid

where $a_s = \mathbb{P}(\theta_{t+s} = \theta_L | \theta_t = \theta_L) = \frac{1}{1-\alpha_L+\alpha_H} \{\alpha_H + (1-\alpha_L)(\alpha_L - \alpha_H)^s\}$.

First, we document that in the neighborhood of both iid types and perfect persistence, "upward" incentive constraints can be safely ignored. Recollect that $\Gamma = \{\Theta, \mu, \alpha_L, \alpha_H, \delta, v_0\}$ is the entire set of parameters.

Claim 7. *For any $\Gamma \setminus \{\alpha_L, \alpha_H\}$, the first-order approach is valid for $\alpha_L = \alpha_H$ and $\alpha_H = 0$.*

Proof. $IC_H(h^{t-1})$ trivially holds when $\alpha_L = \alpha_H$, and for $\alpha_H = 0$, $C_L(h^{t-1})$ implies $IC_H(h^{t-1})$. □

Second, we enlist sufficiency conditions that ensure that the first-order optimal contract is globally optimal. The primary motivation behind them is the following. When $C_L(h^{t-1})$ is slack, $q(\theta_H | h^{t-1}, \theta_L, \theta_H^{s-1})$ are efficient for all $s \geq 1$, therefore, $IC_H(h^{t-1})$ necessarily holds. When does $C_L(h^{t-1})$ bind? It binds when quantities on the left hand side of $C_L(h^{t-1})$, that is $q(\theta_H | h^{t-1}, \theta_H^s)$ for $s \geq 0$, are highly distorted owing to the interaction of binding incentive and cash-strapped constraints in previous periods. But, it is precisely when these quantities are highly distorted that it is easy for $IC_H(h^{t-1})$ to be satisfied for they appear on the right hand side of the constraint. Combining the efficient and inefficient regions, the measure of parameters for which the "upward" incentive constraint may bind after some history is quite small.

Claim 8. *The first-order approach is valid if either of the following condition holds.*

$$(\mathcal{S}_1): \quad q^e(\theta_H) \frac{1}{1 - \delta(\alpha_L - \alpha_H)} \leq q^e(\theta_L)$$

$$(\mathcal{S}_2): \quad \alpha_H q^e(\theta_H) \left(\frac{\delta}{1-\delta} - \frac{\delta(\alpha_L - \alpha_H)}{1 - \delta(\alpha_L - \alpha_H)} \right) \leq (1 - \alpha_L + \alpha_H) q^e(\theta_L)$$

Proof. $(\mathcal{S}_2)$ is derived only from $IC_H(h^{t-1})$. To see $(\mathcal{S}_1)$, note that $p_s \propto \alpha_H + (1-\alpha_L)(\alpha_L - \alpha_H)^s = (1 - \alpha_L + \alpha_H)(\alpha_L - \alpha_H)^s + \alpha_H[1 - (\alpha_L - \alpha_H)^s]$ and quantities are always distorted downward. Use $C_L(h^{t-1})$, which binds, and plug into $IC_H(h^{t-1})$. Finally, bound $q(\theta_H | h^{t-1}, \theta_L, \theta_H^{s-1})$ by $q^e(\theta_H)$ and $q(\theta_H | h^{t-1}, \theta_H^s)$ by 0, because $1 - (\alpha_L - \alpha_H)^s \geq 0$. □

Third, we have numerically calculated the optimal contract for a large range of parameters to show that the first-order approach is indeed valid. The code for these numerical simulations has

been made available online to test any combination of parameter values.[40] Two such examples are presented in Figure 8. The shaded region is the recursive domain for the inefficient contract (easy to see that the efficient contract is first-order optimal). The darkly shaded region is the set of expected utility vectors for which the "upward" constraint is slack at the optimum. The shell, wherein the optimal contract resides, lies within the darker shaded area. Hence the first-order approach is valid.

9.11 The model in continuous time

Now θ_t follows a continuous time Markov chain on $\Theta = \{\theta_L, \theta_H\}$ with transition rates $0 < \lambda_L, \lambda_H < \infty$. We consider contracts which are progressively measurable with respect to the natural filtration with the uniformly bounded process of promised utilities. Agent's strategy is an adapted cadlag process taking values in Θ with at most finitely many jumps in any finite time interval. These structural properties are required for the principal to be not able to detect a deviation from truth-telling.

We adopt the first order approach restricting agent's strategy even further: the agent can not understate his costs. [41] Then, it is with no loss of generality to consider only contracts delivering the efficient quantity to the cost-efficient type. And, it is with out loss of generality to focus on contracts delivering a downward distorted quantity to the cost-inefficient type.

By the revelation principle, it suffices to optimize over the *incentive compatible* contracts. Given uniform-boundedness of U, a contract is incentive compatible if and only if the agent can not gain by misreporting only for a short time interval and being truthful afterwards. As in the main text, let $w_j = U(\theta_j|h^{t-})$ and $z_{jk} = U(\theta_k|h^{t-}, \theta_j^{[t,t+dt)})$ for $j,k = L,H$. Then, the incentive compatibility simply says that for any small $dt > 0$,

$$w_L - w_H \geq \Delta\theta q_H dt + (1 - rdt)[1 - (\lambda_L + \lambda_H)dt](z_{HL} - z_{HH})$$

where q_H is an average quantity from (h^{t-}, θ_H) to $(h^{t-}, \theta_H^{[t,t+dt)})$. Uniform boundedness guarantees that q_H is well-defined.

Two more constraints need to be imposed, because the agent is cash-strapped. For any small $dt > 0$,

$$w_L \leq (1 - rdt)[(1 - \lambda_L dt)z_{LL} + \lambda_L dt z_{LH}]$$
$$w_H \leq (1 - rdt)[(1 - \lambda_H dt)z_{HH} + \lambda_H dt z_{HL}]$$

It immediately follows that, as in the discrete time case, the recursive domain is $W = \mathbb{R}^2_+$ and the

[40] The code can be found at www.rohitlamba.com/research. We have used the parametric setting: $V(q) = 10\sqrt{q}$, $\delta = 0.8$, $\theta_L = 3, \theta_H = 4$, $v_0 = 0$.

[41] The agent can not claim to have a transition to θ_L when no transition happened and the agent has to announce a transition to θ_H if one happened.

efficiency set is $E = \{\mathbf{w} \in W : w_L - w_H \geq \kappa \text{ and } w_H \geq \underline{w}_H^e\}$ such that

$$\underline{w}_L^e - \underline{w}_H^e \geq \kappa = \lim_{dt \to 0} \frac{\Delta\theta q_L^e dt}{1 - (1 - r dt)[1 - (\underline{\lambda} + \overline{\lambda})dt]} = \frac{\Delta\theta q_L^e}{r + \lambda_L + \lambda_H}$$
$$\underline{w}_H^e = \lim_{dt \to 0} \frac{(1 - r dt)\overline{\lambda}}{r}\kappa = \frac{\Delta\theta \lambda_L q_L^e}{r(r + \lambda_L + \lambda_H)}$$

So, we have established continuous time analogs of Claims 1 and 2.

Let $\dot{z}_{jk} = \frac{\dot{z}_{jk} - w_j}{dt} \in [-\infty, +\infty]$ for $j, k = L, H$.[42] Since $\dot{z}_{jk}(dt)^2 \to_{dt \to 0} 0$ almost surely by uniform boundedness, we can rewrite our constraints in the following way:

$$\dot{z}_{LL} \leq (\lambda_L + r)w_L - \lambda_L w_H \tag{20}$$
$$\dot{z}_{HH} \leq (\lambda_H + r)w_H - \lambda_H w_L \tag{21}$$
$$\dot{z}_{HL} - \dot{z}_{HH} \leq (\lambda_L + \lambda_H + r)(w_L - w_H) - \Delta\theta q_H \tag{22}$$

Now, we set up our recursive problem using $\mathbf{w}$ as state variables. Let $Q_H(\mathbf{w})$ be the expected surplus if the "last" type was θ_H and $Q_L(w_L)$ be the expected surplus if the "last" type was θ_L.[43] For now, assume that these functions are continuously differentiable, then the HJB equations are as it follows:

$$(\lambda_L + r)Q_L(w_L) = s(\theta_L, q_L^e) + \max_{\dot{z}_{LL}, w_H \in [0, w_L]} \left\{Q_L'(w_L)\dot{z}_{LL} + \lambda_L Q_H(\mathbf{w})\right\} \text{ s.t. (20)}$$
$$(\lambda_H + r)Q_H(\mathbf{w}) = \lambda_H Q_L(w_L) + \max_{\dot{\mathbf{z}}, q_H \in [0, q_H^e]} \left\{s(\theta_H, q_H) + DQ_L(\mathbf{w}) \cdot \dot{\mathbf{z}}_H\right\} \text{ s.t. (21) and (22)}$$

We shall suppose that the optimal contract exists. We also suppose that both value functions satisfy all the properties which have been established in the discrete case, see Claim 4. For simplicity, we assume each value function is twice differentiable.

From the HJB equation, $Q_L'(w_L) \geq 0$, $D_L Q_H(\mathbf{w}) \geq 0$ and $D_L Q_H(\mathbf{w}) + D_H Q_H(\mathbf{w}) \geq 0$. It is easy to see that $Q_L'(w_L) > 0$ if and only if $w_L < \underline{w}_L^e$ by concavity and supernodularity. And, $DQ_H(\mathbf{w}) = 0$ if and only if $\mathbf{w} \in E$. So, for $\mathbf{w} \notin E$ both cash-strapped constraint must bind.[44] We will look at the set H which is defined as in Claim 4:

$$H = \{w \in int(W) : Q_L(w_L) > 0 \text{ and } DQ_H(\mathbf{w}) \gg 0\} \subseteq (0, \underline{w}_L^e) \times (0, \underline{w}_H^e)$$

On this set, $Q_L'' < 0$ and $D_{jj} Q_H < 0$ for $j = L, H$, the incentive constraint is binding.

First, we establish the analogue of Claim 5. The shell can be identified with $B \subseteq W \cap (0, \underline{w}_L^e) \times (0, \underline{w}_H^e) \subseteq H$ such that

$$w_L \in (0, \underline{w}_L^e) \text{ and } w_H^H(w_L) \leq w_H \leq w_H^L(w_L)$$

where $D_H Q_H(w_L, w_H^L(w_L)) = Q_L'(w_L)$ and $D_L Q_H(w_L, w_H^H(w_L)) = 0$. These two functions are

[42] $\dot{z}_{jk} \pm \infty$ stays for a discrete jump.

[43] Q_L depends only on w_L, because z_{LH}' is unrestricted. This means that w_H can be freely adjusted discontinuously when there is a switch from the θ_H to θ_L.

[44] In Equation 20, w_H is a promised utility after an adjustment.

inverses of w_L^i defined in the discrete case, so they have the same properties. To be specific, they continuously connect 0 and $\underline{\mathbf{w}}^e$ and they are strictly increasing. Moreover, $w_H^H(w_L) < w_H^L(w_L)$ on B. To see this differentiate the former HJB equation to obtain that

$$\lambda_L D_L Q_H(w_L, w_H^L(w_L)) = -Q_L''(w_L)\dot{z}_{LL} > 0 = D_L Q_H(w_L, w_H^H(w_L))$$

Next, we show that the shell lies above the line connecting 0 and $\underline{\mathbf{w}}^e$. Differentiate the second HJB equation

$$\begin{aligned} \lambda_H[D_H Q_H(\mathbf{w}) - Q_L'(w_L)] &= D(D_L Q_H)(\mathbf{w}) \cdot \dot{\mathbf{z}}_H \\ \lambda_H D_L Q_H(\mathbf{w}) &= D(D_H Q_H)(\mathbf{w}) \cdot \dot{\mathbf{z}}_H \end{aligned} \tag{23}$$

For any point on the aforementioned line: $\frac{w_H}{w_L} = \frac{\underline{w}_H^e}{\underline{w}_H^e + \kappa}$, it holds that $\dot{\mathbf{z}}_{HH} = 0$. Equation (23) implies $\dot{z}_{HL} > 0$, hence $D_H Q_H(\mathbf{w}) < Q_L'(w_L) = D_H Q_H(w_L, w_H^L(w_L))$.

Now, we establish the analogue of Claim 6. We argue that both boundaries of the shell are reflecting which implies that the shell is absorbing. The claim is trivial for w_H^H and needs to be shown only for w_H^L. To be concrete, we need to show that $\dot{z}_{HH} \leq (w_H^L)'(w_L)\dot{z}_{HL}$. By the implicit function theorem

$$(w_H^L)'(w_L) = \frac{Q_L''(w_L) - D_{LH} Q_H(w_L, w_H^L(w_L))}{D_{HH} Q_H(w_L, w_H^L(w_L))} > \frac{D_{HH} Q_H(w_L, w_H^L(w_L))}{D_{HH} Q_H(w_L, w_H^L(w_L))}$$

Using equation (23), one can obtain the desired result.

We claim that following monotonicity properties are true for the optimal contract in the interior of the shell: $\dot{q}_H < 0$ and $\dot{\mathbf{z}}_H < 0 < \dot{\mathbf{z}}_L$ where $\dot{z}_{LH} = (w_H^L)'(w_L)\dot{z}_{LL}$. First of all, $\dot{z}_{HH} < 0 < \dot{z}_{LL}$ holds trivially in the shell and $\dot{z}_{LH} > 0$ as $(w_H^L)'(w_L) > 0$. And, from the second HJB equation, $D_q s(\theta_H, q_H) = \Delta\theta D_L Q_H(\mathbf{w})$. Totally differentiating with respect to time:

$$\frac{d}{dt} D_L Q_H(\mathbf{w}) = D Q_H(\mathbf{w}) \cdot \dot{\mathbf{z}}_L = \lambda_H[D_H Q_H(\mathbf{w}) - Q_L'(w_L)] > 0$$

implying that $\dot{q}_H < 0$ and $\dot{z}_{HL} < 0$.

Also, the distortions are muted after θ_L: $q_H(\mathbf{w}) \leq q_H(w_L, w_H^L(w_L))$. To see this, notice that in the shell $w_H \leq w_H^L(w_L)$. Then, the first-order condition $D_q s(\theta_H, q_H) = \Delta\theta D_L Q_H(\mathbf{w})$ and supermodularity of Q_H implies the claim. This establishes the continuous time analogue of Claim 6.

References

M. Amador, I. Werning, and G.-M. Angeletos. Commitment vs. flexibility. *Econometrica*, 74(2): 365–396, 2006.

M. Arve and D. Martimort. Dynamic procurement under uncertainty: Optimal design and implications for incomplete contracts. *American Economic Review*, 106(11):3238–3274, 2016.

I. Ashlagi, C. Daskalakis, and N. Haghpanah. Sequential mechanisms with ex-post participation

guarantees. Stanford University and Massachusetts Institute of Technology and Pennsylvania State University, 2016.

S. Athey and I. Segal. Designing efficient mechanisms for dynamic bilateral trading games. *American Economic Review*, 97(2):131–136, 2007.

S. Athey and I. Segal. An efficient dynamic mechanism. *Econometrica*, 81(6):2463–2485, 2013.

A. V. Banerjee and E. Duflo. Do firms want to borrow more? testing credit constraints using a directed lending program. *The Review of Economic Studies*, 81(2):572–607, 2014.

M. Battaglini. Long-term contracting with markovian consumers. *American Economic Review*, 95 (3):637–658, 2005.

M. Battaglini and R. Lamba. Optimal dynamic contracting: the first-order approach and beyond. Cornell University and Pennsylvania State University, 2017.

D. Bergemann and P. Strack. Dynamic revenue maximization: A continuous time approach. *Journal of Economic Theory*, 159:819–853, 2015.

D. Bergemann and J. Välimäki. The dynamic pivot mechanism. *Econometrica*, 78(2):771–789, 2010.

D. Bergemann and J. Välimäki. Dynamic mechanism design: an introduction. Yale University and Aalto University, 2017.

D. Besanko. Multi-period contracts between principal and agent with adverse selection. *Economics Letters*, 17(1-2):33–37, 1985.

B. Biais, T. Mariotti, G. Plantin, and J.-C. Rochet. Dynamic security design: Convergence to continuous time and asset pricing implications. *Review of Economic Studies*, 74(2):345–390, 2007.

B. Biais, T. Mariotti, and J.-C. Rochet. Dynamic financial contracting. In D. Acemoglu, M. Arellano, and E. Dekel, editors, *Advances in Economics and Econometrics: Tenth World Congress*, chapter 9, pages 125–171. Cambridge University Press, 2013.

R. Boleslavsky and M. Said. Progressive screening: Long-term contracting with a privately known stochastic proces. *Review of Economic Studies*, 80(1):1–34, 2013.

M. Campello, J. R. Graham, and C. R. Harvey. The real effects of financial constraints: Evidence from a financial crisis. *Journal of Financial Economics*, 97(3):470–487, 2010.

G. L. Clementi and H. A. Hopenhayn. A theory of financing constraints and firm dynamics. *Quarterly Journal of Economics*, 121:229–265, 2006.

P. Courty and H. Li. Sequential screening. *Review of Economic Studies*, 67(4):697–717, 2000.

P. M. DeMarzo and M. J. Fishman. Optimal long-term financial contracting. *Review of Financial Studies*, 20(6):2079–2128, 2007.

P. Esö and B. Szentes. Optimal information disclosure in auctions and the handicap auction. *Review of Economic Studies*, 74(3):705–731, 2007.

P. Esö and B. Szentes. Dynamic contracting: an irrelevance theorem. *Theoretical Economics*, 12(1): 109–139, 2017.

A. Fernandes and C. Phelan. A recursive formulation for repeated agency with history dependence. *Journal of Economic Theory*, 91(2):223–247, 2000.

S. Fu and V. R. Krishna. Dynamic financial contracting with persistent private information. University of Rochester and Florida State University, 2017.

D. Garrett and A. Pavan. Dynamic managerial compensation: A variational approach. *Journal of Economic Theory*, 159:775–818, 2015.

D. Garrett, A. Pavan, and J. Toikka. Robust predictions of dynamic optimal contracts (working slides). Toulouse School of Economics, Northwestern University and MIT, 2017.

M. Golosov, A. Tsyvinski, and N. Werquin. Recursive contracts and endogenously incomplete markets. In J. B. Taylor and H. Uhlig, editors, *Handbook of Macroeconomics, Volume 2*, chapter 10, pages 725–841. Elsevier, 2016.

E. J. Green. Lending and the smoothing of uninsurable income. In E. C. Prescott and N. Wallace, editors, *Contractual arrangements for intertemporal trade*, volume 1, chapter 1, pages 3–25. University of Minnesota Press, 1987.

Y. Guo and J. Hörner. Dynamic mechanisms without money. Northwestern University and Yale University, 2017.

M. Halac and P. Yared. Fiscal rules and discretion under persistent shocks. *Econometrica*, 82(5): 1557–1614, 2014.

A. İmrohoroğlu and Ş. Tüzel. Firm-level productivity, risk, and return. *Management Science*, 60(8): 2073–2090, 2014.

N. Kiyotaki. A mechanism design approach to financial frictions. In F. Allen, M. Aoki, J.-P. Fitoussi, N. Kiyotaki, R. Gordon, and J. E. Stiglitz, editors, *The Global Macro Economy and Finance*, chapter 9, pages 177–187. Palgrave Macmillan UK, 2012.

D. Krähmer and R. Strausz. Dynamic mechanism design. In *An Introduction to the Theory of Mechanism Design by Tilman Börgers*, chapter 11, pages 204–234. Oxford University Press, 2015a.

D. Krähmer and R. Strausz. Optimal sales contracts with withdrawal rights. *Review of Economic Studies*, 82(2):762–790, 2015b.

R. V. Krishna, G. Lopomo, and C. Taylor. Stairway to heaven or highway to hell: Liquidity, sweat equity, and the uncertain path to ownership. *RAND Journal of Economics*, 44(1):104–127, 2013.

E. Lipnowski and J. Ramos. Repeated delegation. University of Chicago and University of Southern California, 2015.

V. F. Luz. Dynamic competitive insurance. University of British Columbia, 2015.

R. B. Myerson. A model of moral-hazard credit cycles. *Journal of Political Economy*, 120(5):847–878, 2012.

R. B. Myerson and M. A. Satterthwaite. Efficient mechanisms for bilateral trading. *Journal of Economic Theory*, 29(2):265–281, 1983.

A. Pavan, I. Segal, and J. Toikka. Dynamic mechanism design: A myersonian approach. *Econometrica*, 82(2):601–653, 2014.

Y. Sannikov. Dynamic security design and corporate financing. In G. M. Constantinides, M. Harris, and R. M. Stulz, editors, *Handbook of the Economics of Finance*, volume 2A, chapter 2, pages 71–122. Elsevier, 2013.

S. E. Spear and S. Srivastava. On repeated moral hazard with discounting. *Review of Economic Studies*, 54(4):599–617, 1989.

N. L. Stokey, R. E. Lucas Jr, and E. Prescott. *Recursive methods in economic dynamics*. Harvard University Press, 1989.

J. Thomas and T. Worrall. Income fluctuation and asymmetric information: An example of a repeated principal-agent problem. *Journal of Economic Theory*, 51(2):367–390, 1990.

R. V. Vohra. Dynamic mechanism design. *Surveys in Operations Research and Management Science*, 17(1):60–68, 2012.

N. Williams. Persistent private information. *Econometrica*, 79(4):1233–1274, 2011.