Available online at www.sciencedirect.com

ScienceDirect

Journal of Economic Theory 193 (2021) 105196

JOURNAL OF Economic Theory

www.elsevier.com/locate/jet

A theory of dynamic contracting with financial constraints ☆

Ilia Krasikov [a], Rohit Lamba [b,*]

[a] *Higher School of Economics Moscow, Russian Federation*
[b] *Pennsylvania State University, United States of America*

Received 16 February 2019; final version received 7 January 2021; accepted 10 January 2021
Available online 18 February 2021

Abstract

Financial constraints preclude many surplus producing economic transactions, and inhibit the growth of many others. This paper models financial constraints as the interaction of two forces: the agent has persistent private information and is strapped for cash. The wedge between the optimal and efficient allocation, termed distortion, increases over time with each successive "bad shock" and decreases with each "good shock". At any point in the contract, an endogenous number of "good shocks" are required for the principal to provide some liquidity and then eventually for the contract to become efficient. Efficiency is reached almost surely. The average rate at which contract becomes efficient is decreasing in persistence of shocks; in particular, the iid model predicts a quick dissolution of financial constraints. This speaks to the relevance of modeling persistence in dynamic models of agency. The problem is solved recursively, and building on the literature, a technical tool of finding the minimal subset of the recursive domain that houses the optimal contract is further developed.

JEL classification: D80; D82; G32

☆ We are indebted to Nageeb Ali, 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. We would also like to thank the editor, Alessandro Pavan, the associate editor, and anonymous referees for detailed suggestions that helped improve the paper. We dedicate this paper to our beloved friend and colleague Edward James Green. His paper Green (1987) was a pioneering contribution in dynamic contracts. Ed passed away on October 26, 2019.

* Corresponding author.
E-mail addresses: krasikovis.main@gmail.com (I. Krasikov), rlamba@psu.edu (R. Lamba).

https://doi.org/10.1016/j.jet.2021.105196

Keywords: Dynamic mechanism design; Financial constraints; Asymmetric information

1. Introduction

Long term economic transactions are often marred by financial constraints. A sizeable body of empirical work documents the wide prevalence of financial constraints, their micro impact on firms size and growth, and macro impact on the misallocation of capital in an economy.[1] With an aim to provide theoretical constructs to these empirical regularities, Kiyotaki (2012), in an elegant note, advocates "a mechanism design approach to illustrate how different environments of private information and limited commitment generate different financial frictions." 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.

We study a dynamic screening problem with Markovian shocks where the principal offers history dependent allocations and transfers to the agent. If the agent has a cash reserve or pledgeable assets, the principal will ask the agent to post a bond or deposit collateral. Broadly, optimal distortions are then *frontloaded*, going from maximal to zero, and optimal payments are *backloaded*, maximally delayed to the extent possible. However, in many real situations, e.g. in supply contracts, managerial compensation, provision of public goods and regulation, the agent may not have the requisite cash to post a bond or collateralize existing assets. This has implication for both the optimal structure of distortions, and the sequential breakup of payments.[2]

Taking inspiration from the literature on financial contracting, we model the aforementioned situation by restricting the stage (or per-period) utility to be positive. The idea being that the agent requires, in the least, the amount of cash that covers the consumption/production decisions in every period. The economic force generated by the interaction of this stronger feasibility restriction and private information is termed as the *financial constraint*. This is because if there is no private information, the efficient allocation is implementable, and in the presence of a bond or collateral, efficiency is achievable much more easily through maximal backloading of payoffs. So, it is the interaction of the two forces together that produces financial constraints. Further, we show that *persistence* in private information, an empirically relevant feature of the model, makes this interaction even richer in terms of constraining the optimal allocation.[3]

The big picture question is: when do these financial constraints bind and when they do, what dynamic distortions do they generate? In asking and then trying to answer this question, we provide a deeper understanding of the role of financial constraints in dynamic mechanism design,

[1] For example, Campello et al. (2010) conduct 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. Moll (2014) studies credit constraints as a channel of misallocation of capital in firms and aggregate productivity losses, and Midrigan and Xu (2014) provide evidence for financing as source of misallocation in plant level data in Columbia and Korea.

[2] Distortions here are defined as the distance between the optimal and efficient allocation. In a dynamic contract these evolve over time as a function of the exogenous parameters and endogenous realization of shocks.

[3] 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.

and a deeper understanding of the role of persistence in agency frictions in dynamic financial contracting.[4]

The rest of this introduction is divided into three parts. First is the structure of the optimal contract and a plausible mechanism that implements it. Second is unpacking the economic content of the novel elements – (i) cash versus no cash constraints, (ii) interpretation of positivity of stage utility, and (iii) the role of persistence in agency frictions in generating financial constraints. Third, is an overview of the literature. 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. Preferences are quasi-linear. The small firm requires a constant cash flow to cover its costs of production, hence the stage utility must be positive: we say that the agent is thus *strapped for cash*. The big firm is tasked with designing a contract which sets supply of inputs by the small firm, and payments for its production. Both parties can commit to a dynamic contract.

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. Fig. 1a depicts a typical sequence of technology shocks. For a history of cost 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: $q(\theta_L|h^t) = q^e(\theta_L)$.[5] On the other hand, each "bad shock" increases optimal distortions: $q(\theta_H|h^t, \theta_H) < q(\theta_H|h^t) < q^e(\theta_H)$.[6] 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, this 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.[7]

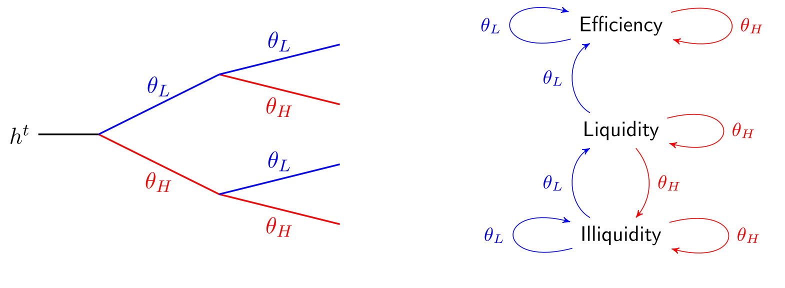
With reference to Fig. 1a, the expected utilities of both the low and high cost types go up after a "good shock" and go down after a "bad shock". That is, as long as the contract is inefficient: $\big(U(\theta_L|h^t, \theta_H), U(\theta_H|h^t, \theta_H)\big) \leq \big(U(\theta_L|h^t), U(\theta_H|h^t)\big)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

[4] 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."

[5] A low cost realization is better for economic surplus than a high cost realization, thus $q^e(\theta_L) > q^e(\theta_H)$.

[6] This is in 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 et al. (2018)).

[7] 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. This provides intuition for the structure of the optimal dynamic contract with highly persistent agency frictions.



(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

Fig. 1. The optimal contract.

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 cash-strapped constraint becomes slack. 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 into the absorbing state of efficiency.

At a technical level, we use the recursive approach to characterize the optimal contract. More specifically, we construct a "shell", the *minimal subset* of the recursive domain which houses the optimal constrained contract. The recursive domain is too large to make crisp predictions about the exact structure of dynamic distortions. We show that as long as the optimal contract is inefficient, the expected utility of the agent must always lie in this shell. 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.

A combination of *working capital* and eventual *take-over* implements the optimal contract. 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.[8]

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

We also reconsider the interpretation of the positivity of stage utility as a limited liability constraint for small businesses. It is clear that the cash strapped constraint is welfare reducing

[8] In the corporate finance view of our model, the Modigilliani-Miller Theorem (Modigliani and Miller (1958)) 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 Modigilliani-Miller theorem holds asymptotically.

from the perspective of total welfare (or surplus), but is it "beneficial" for the agent? 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 perfect persistently limit the ranking can reverse for certain parameters. This points to a cautious interpretation of the positivity of the stage utility as a *limited liability* constraint, which is the standard in the literature.

Finally, we take this model as representative of firm dynamics in an economy with financial constraints and numerically show how persistence in agency frictions makes a marked difference to the substantive predictions of the 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; in particular, 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.[9]

Related literature. This paper sits at the intersection of at least two strands of theoretical models: dynamic mechanism design with serially correlated information (see surveys by Vohra (2012), Krähmer and Strausz (2015a), Pavan (2016), and Bergemann and Välimäki (2019)) and dynamic financial contracting with iid technologies (see surveys by Biais et al. (2013), and Sannikov (2013). Three ingredients interact to determine the structure of dynamic inefficiencies: correlation in agency frictions, strength of feasibility restrictions, and permissibility of termination. The overarching role of each combination of ingredients is to create frictions in dynamic contracting that lead to realistic qualitative predictions.

Tables 1 and 2 enlist the most *closely related papers*; Table 1 features screening and Table 2 features cash flow diversion as the underlying agency friction. Within each table, papers are classified along inclusion/exclusion of the three aforementioned modeling ingredients. In terms of long-term predictions, once the recursive problem is appropriately set up, it can be shown that in each of the papers the optimal contract converges to the efficient allocation in the absence of the termination clause, and it converges either to efficiency or termination in the presence of the termination clause.[10] The key economic force that leads to this result is backloading of information rents to the extent possible.[11]

At a high level, ours is the first paper to precisely characterize the short-run predictions in terms of the monotonic nature of dynamic distortions: "good shocks" monotonically push the allocation towards efficiency and "bad shocks" take it away from it. As noted in Table 1, it is the first paper to analyze a dynamic screening model (as opposed to a cash flow diversion or moral hazard problem) with both persistence in private information and financial constraints, nudging the literature on dynamic mechanism design to explicitly incorporate financial constraints. Further it (i) identifies the minimal subset of the recursive domain that houses the optimal contract;

[9] We believe this an important point to consider for the literature on firm dynamics, as a step towards bridging the reality of the wide and prolonged prevalence of financial constraints, especially in developing countries, and the limited impact of agency frictions in the predictions of the iid model.

[10] The proof follows from a straightforward application of the martingale convergence theorem.

[11] Note that an important assumption in all the papers we discuss is that the principal can commit to the dynamic contract, thus a simple version of the revelation principle is applicable. See Bester and Strausz (2001), Maestri (2017) and Doval and Skerta (2019) for the limited commitment case with no financial constraints.

Table 1
Dynamic screening.

	IR	CS w/o termination	CS w/ termination
IID private information	trivial	KLT & this paper	KLT & this paper
Markovian private information	Battaglini (2005)	this paper	this paper (partially)

IR means individual rationality, CS means cash-strapped, w/o means without, w/ means with, and KLT stands for Krishna et al. (2013).

Table 2
Dynamic cash flow diversion.

	IR	CS w/o termination	CS w/ termination
IID technology	trivial	CH & FK	CH & Biais et al. (2007)
Markovian technology	{∅}	Fu and Krishna (2019)	{∅}

CH and FK stand for Clementi and Hopenhayn (2006) and Fu and Krishna (2019) respectively.

(ii) clarifies the connection between limited liability and being strapped for cash; (iii) provides an explicit characterization of the optimal contract in the perfectly persistent limit which shows "good shocks" have a stronger effect on distortions than "bad shocks", this fact underlies the long-term efficiency result; (iv) solves for the optimal contract in continuous time, which seeks to unify the literatures on cash flow diversion and screening, since the models converge to the same limit in continuous time; and (v) explores the implications of persistence in agency frictions on firm dynamics.

The two most closely related papers are Battaglini (2005) and Krishna et al. (2013). Battaglini (2005) studies a similar screening model, but where the agent has cash to post a bond. More specifically, only the total expected utility over time is required to be positive in every period. The structure of short-run distortions are thus quite different – the contract becomes efficient forever as soon as the agent assumes a "good shock", and it has decreasing distortions along the history of constant "bad shocks". In a departure from that paper, and more generally the literature on dynamic mechanism design, our paper explores the implications of cash constraints for the agent with persistent private information.

Krishna et al. (2013) study the same model as ours, repeated screening with the cash-strapped constraint, but where the agent's types follow an iid process. Since theirs is a special case of our model, all our results also hold in their setup. However, the focus of the paper is on long-term efficiency. We build upon their work in at least three ways. First, the monotonicity of allocation rule, even for the iid model is novel to our paper. Second, the Markov model is technically much harder to solve, as has already been noted in simpler dynamic mechanism design models without financial constraints.[12] Third, persistence adds greater empirical relevance to the analysis, as is evident from the applications of standard dynamic mechanism design models to public finance (see Stantcheva (2020)).[13]

[12] Describing the important distinguishing feature of dynamic mechanism design, Bergemann and Välimäki (2019) state: "In all of the above applications, the types of some agents and/or the set of allocations available change in a non-trivial manner across periods. For us, this is the distinguishing feature of dynamic mechanism design."

[13] To the best of our knowledge, the earliest reference of modeling financial constraints directly in dynamic mechanisms is Sen (1996). The paper looks at a two period screening model with persistent private information and limits on the agent's liability, but where the principal can fire (and replace) the agent after the first period. It shows that limits on liability restrict the principal's ability to screen the agent, and termination offers greater flexibility.

Clementi and Hopenhayn (2006) and Fu and Krishna (2019) 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: 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. While these models produce similar long-term predictions, the short-run structure of the optimal contract here is quite different than the screening literature.[14]

Our paper is also related to the recent work by Guo and Hörner (2018): They consider a dynamic principal-agent model with persistent private information where preferences are perfectly aligned, transfers are not allowed and the principal wants to maximize efficiency. The optimal contract converges almost surely either to permanent allocation (efficiency) or permanent non-allocation (immiseration), driven by the fact that both front-loading and backloading of payoffs can occur at the optimum. In our framework, preferences are misaligned, the expected utility is continuously backloaded, and the optimal contract always converges to efficiency. A technical aspect we share with Guo and Hörner (2018) is the characterization of a subset of the recursive domain that houses the optimal contract, which allows us to make precise statements about the short and long run properties.[15,16]

Financial constraints have also been explored in the sequential screening literature pioneered by Courty and Li (2000). For example, Krähmer and Strausz (2015b) consider a sequential screening model with ex post participation constraints. They show that with these additional constraints the optimal contract is static and does not illicit the agent's information sequentially. One way to map their framework into ours would be to consider the two period version of our model, and require the first period allocation to be (exogenously) zero. Then, the cash-strapped constraints require that no payments can be charged in the first period. As a consequence the optimal contract replicates the "static optimum". In contrast our model highlights that multi-period interactions can extract private information in an incentive compatible fashion, even with stronger feasibility restrictions.[17]

Finally, the cash-strapped constraint breaks the linearity of transfers across time. The spirit of this exercise is shared by other related works: Amador et al. (2006), and Halac and Yared (2014)

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

[15] Zhang (2009) was the first paper to exploit the construction of a "shell" to characterize the optimal contract in a two-types dynamic adverse selection model. This approach has also been used by Fu and Krishna (2019). The identification of the minimal subset of the recursive domain that houses the optimal contract separates us from these papers. This allows us to show the direction of allocative distortions over time.

[16] See also Li et al. (2017) and Lipnowski and Ramos (2019) for a repeated model of allocation sans transfers, but without commitment on the side of the principal.

[17] In a similar vein, Ashlagi et al. (2019) 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 where all payoffs cannot be delayed to the last period.

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. Krasikov et al. (2019) analyze a dynamic screening model with individual rationality, but where the principal is more patient than the agent. In all these papers, there is a cost to moving transfers or payments across time.

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.[18] The principal values the final good at $V(q)$, where $V:\mathbb{R}_+ \to \mathbb{R}_+$ satisfies the Inada conditions.[19] 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.[20] 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, often 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: $\mu >> \mathbf{0}$ and $f >> \mathbf{0}$. 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.

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 cost 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 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: $m = \langle \mathbf{p}, \mathbf{q} \rangle = \big(p(\hat{\theta}_t|h^{t-1}), q(\hat{\theta}_t|h^{t-1})\big)_{t=1}^{T}$, where h^{t-1} and $\hat{\theta}_t$ are, respectively, the history of reports up to $t-1$ and current report at time t.[21] The reported history h^t is recursively defined as $h^t = \{h^{t-1}, \hat{\theta}_t\}$ starting with $h^0 = \emptyset$. The set of possible histories at time t is denoted by H^t. Define the private history of the agent to be $h_A^t = \{h_A^{t-1}, \theta_t, \hat{\theta}_{t-1}\}$, 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: $h_A^t \mapsto \sigma_t(h_A^t) \in \Theta$.

[18] 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: $F = 0$.

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

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

[21] 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^{∞}.

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

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

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, $\forall\, h^{t-1} \in H^{t-1}\ \forall\, t$:

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

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.

A stronger notion of feasibility is invoked in the paper than the standard dynamic mechanism deign models. A 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, the more permissive feasibility criterion would demand the following:

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

Under individual rationality the agent can be asked to forgo payments or deposit upfront capital with a promise of being compensated for it later. We shall refer to the framework which invokes the individual rationality constraint as the "benchmark model". The cash-strapped constraint precludes contracts with such delayed promises; therefore, it acts as a credit constraint for the agent. As a consequence the principal cannot maximally backload the agent's payoffs. The main idea behind this constraint is that the agent has to be compensated for the production costs in every period.[22]

3. Optimal contract

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 $\overline{S} = \sum_{t=1}^{T} \delta^{t-1}\mathbb{E}[s(\tilde{\theta}_t)]$ 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 $\mathbf{q} \geq 0$,

[22] Two aspects of the contractual space can be noted here. First, if a contract is strapped for cash it necessarily satisfies individual rationality while the opposite is not necessarily true. Second, just as the incentive constraint, the cash-strapped constraint holds both "on" and "off" path. Even if the agent may have misreported in the past, the principal delivers a non-negative stage utility to him if he is truthful today.

$$\text{(PK):}\quad \mu_L U(\theta_L) + \mu_H U(\theta_H) \geq v_0, \text{ and}$$
$$IC_L(h^{t-1}), IC_H(h^{t-1}), \mathcal{C}_L(h^{t-1}), \mathcal{C}_H(h^{t-1}) \; \forall\, h^{t-1} \in H^{t-1} \; \forall\, t,$$

where (PK) is the ex ante promise keeping constraint, and $IC_i(h^{t-1})$ and $\mathcal{C}_i(h^{t-1})$ are the incentive and cash-strapped constraints, respectively, for type θ_i in period t after history h^{t-1}. Since quantity is always non-negative at the optimum, we shall drop that constraint.

Note that v_0 parameterizes the bargaining power of the agent, and maps the Pareto frontier. In the absence of (PK), the solution to $(\mathcal{P}^*)$ will choose the principal-optimal contract which will deliver the agent an ex ante utility, say $\underline{v}$. Moreover, (PK) binds if and only if $v_0 \geq \underline{v}$.

We consider a relaxed problem where we ignore $IC_H(h^{t-1})$ for all histories. A justification of this is provided in Section 6.10, 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

$$IC_L(h^{t-1}), \mathcal{C}_L(h^{t-1}), \mathcal{C}_H(h^{t-1}) \; \forall\, h^{t-1} \in H^{t-1} \; \forall\, t$$

where $R^*(v_0)$ is the value of the objective – the principal's profit at the constrained optimum. Also, we shall denote the ex ante economic surplus generated by the optimal contract by $S^*(v_0) = R^*(v_0) + \max\left\{\underline{v}, v_0\right\}$.

Following Myerson, we 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:

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

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 $\mathcal{Q}: \mathbb{R}_+ \to \mathbb{R}_+$:

$$V'(\mathcal{Q}(x)) = \theta_H + x\Delta\theta \tag{1}$$

where $\Delta\theta = \theta_L - \theta_H$ and $x(h^{t-1})\Delta\theta = \frac{r(h^{t-1})}{\mathbb{P}(h^{t-1}, \theta_H)}$ is the optimal distortion. This equation can instead be written succinctly as $q(\theta_H|h^{t-1}) = q^e(\theta_H) - d(\theta_H|h^{t-1})$, and it is easy to see that d is increasing in x.[23] We shall characterize r and hence x, which in turn pins down the evolution of optimal quantities and expected utilities.

In what follows, we first solve the two period model to provide some basic intuition, and then present the main result on the characterization of the optimal contract. In addition, a complete solution of the optimal limit contract when persistence converges to unity is stated, followed by a sketch of the proof using the recursive approach, and finally a brief discussion on implementation of the optimal contract is provided.

[23] The exact relationship is as follows: $\mathcal{Q}(x) = \left(V'\right)^{-1}(\theta_H + x\Delta\theta) = \theta_H - d$.

3.1. Two period model

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

$$IC_L(\emptyset), \mathcal{C}_L(\emptyset), \mathcal{C}_H(\emptyset), \text{ and } IC_L(\theta_i), \mathcal{C}_L(\theta_i), \mathcal{C}_H(\theta_i) \text{ for } i = L, H$$

It can be shown that (i) $\mathcal{C}_L(\theta_L)$ and $\mathcal{C}_L(\theta_H)$ are implied by the other constraints, (ii) $IC_L(\emptyset)$, $IC_L(\theta_H)$, $\mathcal{C}_H(\emptyset)$ and $\mathcal{C}_H(\theta_H)$ all bind at the optimum, (iii) $\mathcal{C}_L(\emptyset)$ can bind sometimes, and (iv) $IC_L(\theta_L)$ and $\mathcal{C}_H(\theta_L)$ can be assumed to hold as an equality, they bind if $\mathcal{C}_L(\emptyset)$ 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. $IC_L(\emptyset)$ 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 elegantly derived for continuous type spaces in Pavan et al. (2014).[24] 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. The rents to the high cost type have a straightforward expression[25]:

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

Additively, the total information rent to be paid to the agent is given as follows:

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

where $\mathbb{P}(\theta_t = \theta_L)$ is the ex ante probability of being the low cost type in period t. Define the 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 $\underline{v}$.[26]

Finally, $\mathcal{C}_L(\emptyset)$ can be expressed as follows[27]

$$\mathcal{C}_L(\emptyset): \quad q(\theta_H) + \delta\alpha q(\theta_H|\theta_H) \geq \delta\alpha q(\theta_H|\theta_L) \tag{6}$$

---

[24] Battaglini and Lamba (2019) derive the same formula for a general discrete type space. Esö and Szentes (2017) also have a derivation of the result for continuous types.

[25] Equation (3) is generated by the binding $\mathcal{C}_H(\emptyset)$, $\mathcal{C}_H(\theta_H)$ and $IC_L(\theta_H)$ constraints: $U(\theta_H) = u(\theta_H) + \delta\left[(1-\alpha)u(\theta_L|\theta_H) + \alpha u(\theta_H|\theta_H)\right] = \delta(1-\alpha)\Delta\theta q(\theta_H|\theta_H)$.

[26] $\underline{v}$ refers to the agent's expected utility on the principal profit maximizing point of the Pareto frontier.

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

The principal chooses $\mathbf{q}$ to maximize $\overline{S} - [\mu_L U(\theta_L) + \mu_H U(\theta_H)]$ subject to (PK) and $\mathcal{C}_L(\emptyset)$, where $\mu_L U(\theta_L) + \mu_H U(\theta_H)$ is given by equation (4). The precise closed form solution is provided 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* $d(\theta_H|\theta_L) > 0 \Leftrightarrow \mathcal{C}_L$ *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 is 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 with individual rationality as opposed to the cash-strapped constraint, $r_2(h) = 0 \ \forall \ h$. Positivity of r_2 here ensures that dynamic distortions accumulate to have a distinct form than in the benchmark model. Specifically, distortions can persist even after a low cost type has been realized in the first period, and for consecutive high costs, distortions actually increase over time.

When does $\mathcal{C}_L$ bind? Equation (6) clearly establishes that low values of $q(\theta_H)$ and $q(\theta_H|\theta_H)$ would violate $\mathcal{C}_L$. 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. Fig. 2a shows the parametric range for which $\mathcal{C}_L$ binds. In the $\mu_H \times \alpha$ rectangle, it plots $q^*(\theta_H|\theta_L)$ – shades represent numerical values as shown on the vertical key on the right. The lightest region is the efficient quantity, and darker the shade, greater is the optimal distortion. It is clear that low values of μ_H and high values of α correspond to the largest liquidity crunch.[28]

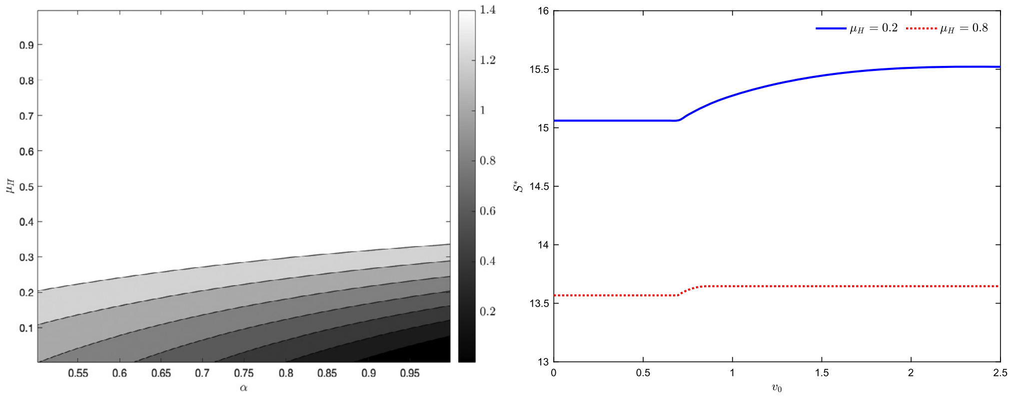
Finally, the following result establishes the dynamics of utility, and value of total surplus. Let $S^*(v_0)$ be the total economic surplus generated by the optimal allocation in Proposition 1.

Corollary 1. *In the optimal contract:*

1. $(u^*(\theta_L|\theta_L), u^*(\theta_H|\theta_L)) \geqslant (u^*(\theta_L|\theta_H), u^*(\theta_H|\theta_H))$;
2. $S^*(v_0)$ *is increasing in* v_0, *and strictly so for* $v_0 \in [\underline{v}, \overline{v}]$.

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. For a "good shock" the next period's utility is larger for both types than that for a "bad shock". Moreover, the optimal value of economic surplus is strictly

[28] 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 $\delta = 1$.



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

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

increasing in the interval $[\underline{v}, \overline{v}]$; this is because a higher v_0 shifts bargaining power towards the possessor of private information, increasing the total size of the pie.[29]

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? What role does persistence play in short and long run?

3.2. The main result

Now we state the main result that enlists an exhaustive characterization of the optimal contract. It will be followed by a brief discussion of the result. We assume that the prior is first-order stochastically ranked around its Markov evolution: $1-\alpha \le \mu_L \le \alpha$. This is strictly more general than assuming a "seed type", that is $\mu_L = \alpha$ or $\mu_L = 1-\alpha$ which is a standard in the recursive contracting literature.[30]

Using binding incentive and cash-strapped constraints, the total expected utility for the agent can be expressed as:

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

This is direct generalization of equation (4) from the two period model. We denote 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 $\underline{v}$.[31] Further, (for infinite horizon) define $E \subset \mathbb{R}^2_+$ to be the largest set of expected utility values, $\big(U(\theta_L|h^t), U(\theta_H|h^t)\big)$, that simultaneously satisfies the all

[29] For $v_0 < \underline{v}$, (PK) does not bind, and for $v_0 > \overline{v}$ the efficient contract is the optimum, so S^* is constant in both regions.

[30] The assumption $1-\alpha \le \mu_L \le \alpha$ is made to ensure that the optimal contract starts in shell defined in Section 3.4. If the 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.

[31] It is easy to see that optimal contract is efficient for $v_0 \ge \overline{v}$ and selects the principal optimal contract for $v_0 < \underline{v}$.

constraints in ($\mathcal{RP}^*$) such that the allocation is efficient: $q(\theta_i|h^t) = q^e(\theta_i)$. It is easy to see (and will be proven later) that this set is non-empty; it takes the shape of cone; and it has a "lowest" point $(\underline{w}^e_L, \underline{w}^e_H)$ such that $\overline{v} = \mu_L \underline{w}^e_L + \mu_H \underline{w}^e_H$.

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 and within period monotonic: $q^*(\theta_L|h^{t-1}) = q^e(\theta_L)$, $q^*(\theta_H|h^{t-1}) = q^e(\theta_H) - d(\theta_H|h^{t-1})$, where $d(\theta_H|h^{t-1}) \geq 0$, and $q^*(\theta_L|h^{t-1}) > q^*(\theta_H|h^{t-1})$.*
2. *Distortions are strictly increasing for consecutive "bad shocks": $d(\theta_H|h^{t-1}, \theta^s_H)$ is strictly increasing in s for $q^*(\theta_H|h^{t-1}) < q^e(\theta_H)$.*
3. *Distortions are muted after "good shock": $d(\theta_H|h^{t-1}, \theta_L) < d(\theta_H|h^{t-1})$ for $q^*(\theta_H|h^{t-1}) < q^e(\theta_H)$.*

B **Expected utility***:*

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

C **Liquidity***:*

5. *The contract becomes liquid above a fixed threshold which is below the efficient level: $\exists\ \underline{w}^{liq}_L < \underline{w}^e_L$ such that for $U^*(\theta_L|h^{t-1}) \geq \underline{w}^{liq}_L$, $\mathcal{C}_L(h^{t-1})$ 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 $\big(U^*(\theta_L|h^{t-1}), U^*(\theta_H|h^{t-1})\big) \in E \Rightarrow \big(U^*(\theta_L|h^{t+s-1}), U^*(\theta_H|h^{t+s-1})\big) \in E\ \forall\ h^{t+s-1} \in H^{t+s-1}|_{h^{t-1}}$.*
7. *An endogenous and monotonic number of "good shocks" are required for efficiency: $\exists\ n^*(h^{t-1}) \in \mathbb{N}$ such that $d(\theta_H|h^{t-1}, \theta^{n^*}_L) = 0$, 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: $\mathcal{C}_L(h^{t-1})$ binds $\Leftrightarrow n^*(h^{t-1}) \geq 2$, and $\mathcal{C}_L(h^{t-1})$ 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 $\big(U^*(\theta_L|h^{t+s-1}), U^*(\theta_H|h^{t+s-1})\big) \xrightarrow[s\to\infty]{} \mathbf{w} \in E$ 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 $r(h^{t-1})$, viz. the marginal cost of incentive provision, determines the optimal distortions d, which drives the rest of the theorem.

In the inefficient region, every further high cost realization strictly increases dynamic distortions, thereby strictly decreasing the optimal quantity. In Fig. 1a, quantity along the history (h^{t-1}, θ^2_H) is less than the quantity along (h^{t-1}, θ_H). This is in contrast to the results in dynamic mechanism design (without financial constraints) that emphasize decreasing distortions over time (see Section 6.1). Moreover, a "good shock" reduces the optimal distortions: quantity along $(h^{t-1}, \theta_L, \theta_H)$ is greater than that along (h^{t-1}, θ_H). These rankings of optimal distortions form the bedrock of our analysis.[32]

Part B tracks the optimal path of expected utility. For the inefficient contract, expected utility strictly increases along both dimensions after a "good shock" and reduces after a "bad shock".

Part C characterizes liquidity. The interval $\left[\underline{w}_L^{liq}, \underline{w}_L^e\right)$ for $U^*(\theta_L|h^{t-1})$ witnesses slacking of $\mathcal{C}_L(h^{t-1})$. The region can only be attained through a "good shock" (given the contract does not start in this region). 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, $\left[0, \underline{w}_L^{liq}\right)$.

The path to efficiency is completed in part D. Efficiency is an absorbing state (point 6), that is, once the distortion for the high cost type reduces to zero, it stays zero: $q(\theta_H|h^{t-1}) = q^e(\theta_H)$ implies $q(\theta_H|h^{t-1}, h^s) = q^e(\theta_H)$. The efficient region is reached through an endogenous number of "good shocks". Moreover, on the realization of a high cost type, the number of consecutive low cost types required to reach the efficient region increases (point 7). 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 $\left[0, \underline{w}_L^{liq}\right)$), the efficient region is achieved through a penultimate liquid region - $\left[\underline{w}_L^{liq}, \underline{w}_L^e\right)$ (point 8). Once the cash-strapped constraint is slack, efficiency is attained through one more low cost type. Thus, the liquid region is strictly larger than the efficient region.

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. In fact, at any point in the inefficient region, it requires infinitely many bad shocks to get to zero allocation, and finitely many good shocks to get to efficiency, the martingale convergence theorem does the rest.

In terms of screening literature (Table 1 in the introduction), Krishna et al. (2013) study the iid version of this model, and focus on long-term efficiency: only parts 1, 4, 6 and 9 of Theorem 1 are established in the paper. The short-run results on monotonicity of the allocation rule over time are especially missing. In terms of the cash flow diversion models (Table 2 in the introduction), Clementi and Hopenhayn (2006) show that points 6 and 9 of Theorem 1 will continue to hold for iid types; and Fu and Krishna (2019) 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.

3.3. Optimal limit contract

While Theorem 1 provides a fairly precise characterization, the optimal contract can in fact 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 of the structure of dynamic contracts with highly persistent agency frictions. Recall the function $\mathcal{Q}$ defined in equation (1).

[32] In the appendix we prove Theorem 1 for 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.

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

1. *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: $\mathcal{Q}\left((1-\lambda)\frac{\mu_L}{\mu_H}\right) \leq \delta^N q^e(\theta_H)$.*
2. *Specifically, $\lim_{\alpha \to 1} q^*(\theta_H | h^{t-1}) = \mathcal{Q}(d_n)$ such that for n draws of θ_L in h^{t-1},*

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

If $\alpha = 1$, then the optimal contract is a repetition of the static optimum, distortions are positive and constant. However, if $\alpha < 1$, then efficiency is achieved almost surely. Intuitively, Proposition 2 describes the path to efficiency for α close to 1. In particular, distortions start at the level of the static optimum, strictly decrease for every low cost type, and remain constant for a high cost: so distortions for a fourth period θ_H are the same for two sequences of histories $\{\theta_H, \theta_L, \theta_L\}$ and $\{\theta_L, \theta_H, \theta_L\}$. Therefore by continuity, 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". This provides some intuition for why the optimal contract converges almost surely to efficiency in the long-run. Moreover, the Proposition explicitly pins down the exact distortions for the limit contract, which decrease multiplicative for every low cost type as a function of the discount factor.[33]

3.4. *Sketch of the proof through the recursive formulation*

A natural way to solve the infinite horizon problem would be to extend the sequential approach from the two period model. However, this quickly runs into the curse of dimensionality problem reminiscent of the literature on repeated games.[34] We therefore turn to the recursive approach to reduce the dimensionality and provide a precise characterization of the optimal contract.

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.

Let $S(h^{t-1})$ be the expected total surplus generated by the sequential contract from period t onwards:

[33] It is also interesting to note that when $\alpha = 1$, both the benchmark model with individual rationality and the model with financial constraints give the same prediction. However, for $\alpha < 1$, the predictions are markedly different.

[34] 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 $\mathcal{C}_L(h^{t-1})$, where $\mu_L U(\theta_L) + \mu_H U(\theta_H)$ is given by equation (7). Introducing Lagrange multipliers for (PK) and all the $\mathcal{C}_L(h^{t-1})$ constraints, we can then write down the optimal allocation rule, see Section 6.3 in the appendix. It is, however, hard to derive general arguments about the nature of dynamic distortions because the Lagrange multipliers are endogenous and jointly determined at the optimum.

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

where $H^t|_{h^\tau}$ for $\tau \le t$ 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

$$IC_L(h^{t+s}), \mathcal{C}_L(h^{t+s}), \mathcal{C}_H(h^{t+s}) \; \forall \, h^{t+s} \in H^{t+s}|_{(h^{t-1}, \theta_j)} \quad \forall \, s.$$

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 exactly equal to $\mathbf{w}$. It is standard practice to show that all the history dependence is encoded in the two-dimensional expected utility $\mathbf{w}$ and last period's type j; hence the simple expression $Q_j^*(\mathbf{w})$. 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}.[35]

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

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

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

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

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 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.

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

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

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

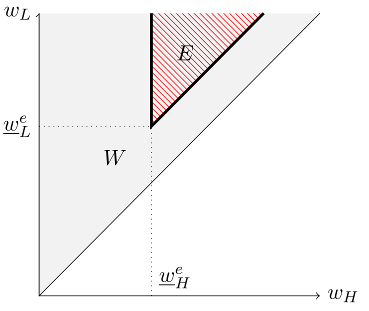


Fig. 3. Recursive domain W and the efficient set E.

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

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

The recursive domain W is the set of all possible expected utilities that generate themselves in an incentive compatible and feasible manner; here it turns out to be the positive orthant above the 45 degree line.[36] Fig. 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. Fig. 3 also plots the efficient set E; it is characterized by its lowest point $\underline{\mathbf{w}}^{\mathbf{e}}$ and two rays.[37] Intuitively, one can note that $\overline{v} = \mu_L \underline{w}_L^e + \mu_H \underline{w}_H^e$.

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.

3.4.1. *Shape of the optimal contract*

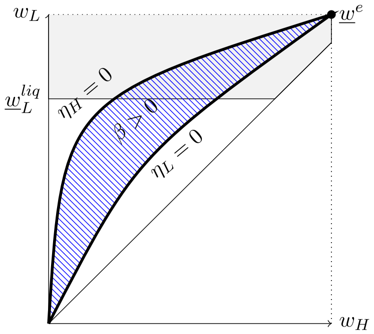
The optimal contract is characterized by the first-order and envelope conditions. These are provided in the appendix. Here we geometrically explain the structure of expected utilities that arise as part of the optimal contract. First, there exists a threshold $\underline{w}_L^{liq}$ on the expected utility 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 $w_L \geq \underline{w}_L^{liq}$. This threshold lies below the efficient level: $\underline{w}_L^{liq} < \underline{w}_L^e$, see Fig. 4a. We also show that the for any w such that $w_L \geq \underline{w}_L^{liq}$, $\mathbf{z_L} \in E$, so one more low cost type is required to move from the liquid region to efficiency.[38]

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 ($\mathcal{RF}$):

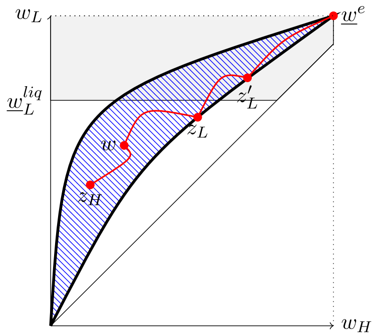
[36] Self generation identifies the largest possible set such that given an expected utility vector $\mathbf{w}$ in that set, there exists some feasible policy choice q, z_L, z_H such that $\mathbf{z_L}, \mathbf{z_H}$ also lie in the set (see Abreu et al. (1990)).

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

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



(a) Liquidity threshold and the Shell, B



(b) Evolution of the optimal contract

Fig. 4. Constructing the space wherein the optimal contract lies.

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

This problem can essentially be thought of as the "maximization of a utility function subject to a budget set", where Q^* represents the utility function, w_L the income, $\alpha/(1-\alpha)$ the relative prices and z_{LL} and z_{LH} the consumption bundles. The point of tangency for an interior optimum is then given by the condition "marginal rate of substitution = relative prices", which in our two dimensional setup is the following:

$$\frac{D_L Q_L^*(\mathbf{z_L})}{\alpha} = \frac{D_H Q_L^*(\mathbf{z_L})}{1-\alpha} \tag{8}$$

where D_i is the directional derivative of Q_j^* for $i = L, H$. Similarly, we cull the following sub-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 the other point of tangency:

$$\frac{D_H Q_H^*(\mathbf{z_H})}{1-\alpha} = \frac{D_L Q_H^*(\mathbf{z_H})}{\alpha} \tag{9}$$

Now, as we vary the "incomes", w_L and w_H, respectively in both sub-problems, we get what is knows as the income offer curve, the locus of all points of tangency as the level of the budget is changed. These two loci are denoted by $\eta_L(w_L) = 0$ and $\eta_H(w_H) - 0$ respectively. Fig. 4a plots the curves. The positivity of the directional derivatives implies that the both commodities are "normal goods" and hence the income offer curves are upward sloping. We show that both curves join the origin and $\underline{\mathbf{w}}^{\mathbf{e}}$, and that $\eta_H = 0$ lies above $\eta_L = 0$. The optimal constrained contract resides on or in the interior of the curves, a space we call the *shell*. The shell is characterized by binding incentive constraints, $\beta > 0$.

Fig. 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 the realization of a high cost type. In fact

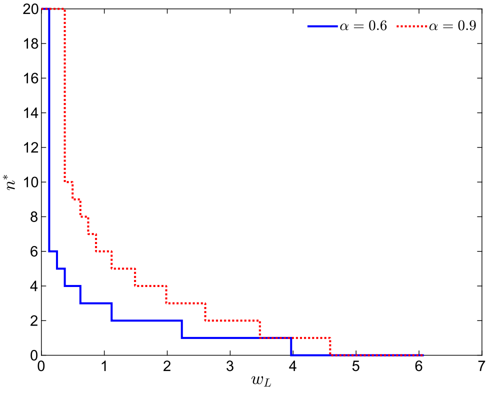


Fig. 5. Number of consecutive low cost types required to reach E.

the example is chosen so that on the realization of two consecutive low cost types, the contract becomes liquid (at $\mathbf{z}'_{\mathbf{L}}$), and therefore a third "good shock" makes it efficient.[39]

Finally, the number of "good shocks" required to make the contract efficient is monotonic in the promised utility. Fig. 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 $w_L \to 0$.

3.5. *Implementation*

There are two 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.

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.[40] We define the mechanism formally in the next proposition. For $j = L, H$, let $v^e_j = \alpha_j \underline{w}^e_L + (1-\alpha_j)\underline{w}^e_H$ be the promised utility offered to the agent at the lowest point of the efficiency set.

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

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

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.

---

[39] As Fig. 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 $\mathbf{z}'_{\mathbf{L}}$ lie *on* the curve. Whereas, realization of a high cost type chooses a point in the southwest direction in the *interior* of the shell.

[40] $\mathbf{U}^*$ 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, $\mathbf{u}^*$ (and hence $\mathbf{p}^*$) is determined through definitional identities.

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

$$I(h^{t-1}) = \Delta\theta \sum_{s=1}^{\infty} \delta^{s-1}(2\alpha - 1)^{s-1} q^*(\theta_H|h^{t-1}, \theta_H^{s-1}) \tag{10}$$

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 takeover is 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.

4. Role of financial constraints and persistence in private information

There are at least 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, and (iii) the impact of persistence in agency frictions on the evolution of the optimal contract and economic surplus.

4.1. Interaction of incentives and stronger notion of feasibility

The elegance of capturing real economic frictions in much of mechanism design is embedded in the interaction of various incentive and feasibility constraints.[41] Fig. 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. Dynamic distortions are a sum of two components: backloading of payoffs to the extent possible and illiquidity due to financial constraints; the latter increases with each "bad shock", overturning the standard result of decreasing distortions.[42]

[41] 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).

[42] In the benchmark model (described formally in Section 6.1) 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.

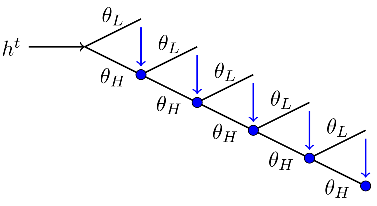


Fig. 6. Propagation of distortions.

Moreover, since the cash strapped constraint of the low cost type binds in the illiquid region, distortions propagate along consecutive sequence of "bad shocks" even when "good shocks" have been realized before it. The cash strapped constraint of the low cost type too interacts with the incentive constraints to sustain distortions. It requires an endogenous number of "good shock" realizations to overcome the legacy of additive distortions from previous binding constraints.

4.2. *Credit constraint versus limited liability*

If (PK) binds at the optimum, which it does if $v_0 \geq \underline{v}$, then the expected utility of the agent is fixed at v_0, and from an ex ante perspective the agent is indifferent between the benchmark model and the model with the cash strapped constraint. The cash strapped constraint is then simply a credit constraint, which shrinks the total surplus and hence the principal's profit. On the other hand if we look at model without (PK), so that all the bargaining power rests with the principal, then it is less clear whether the agent "prefers" a situation with or without the cash strapped constraint.

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

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

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

A careful look at the two formulas would reveal that there is no obvious mathematical way of ranking $\underline{v}^{\#}$ and $\underline{v}^{*}$. The next result theoretically evaluates the ranking between $\underline{v}^{\#}$ and $\underline{v}^{*}$ for the iid and perfectly persistent limits.

Proposition 4. *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 $D^{\#} \lessgtr D^{*}$.*

Therefore, the iid model would predict that agent is always does better with financial constraints, whereas with persistence, the answer depends on the underlying economic environment. We also numerically evaluate both values for a large class of parameters and find that $\underline{v}^{*}$ is higher than $\underline{v}^{\#}$, but not always.[43] In a related two period model, Grillo and Ortner (2020) show that with serial correlation, the cash-strapped constraint would make the agent worse off.

[43] For low values of δ and high values of α and μ_H, $\underline{v}^{\#}$ is in fact greater than $\underline{v}^{*}$. The code is available on request.

In the literature, the positivity of stage utility has been regarded as limited liability, under the assumption that it would always make the agent better off. Arguably, this interpretation needs a more careful analysis with persistent agency frictions.

4.3. Persistence

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. 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 value of total surplus converges in distribution to the random variable which takes value Q_j^e with the probability μ_j^*, $j = L, H$, where

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

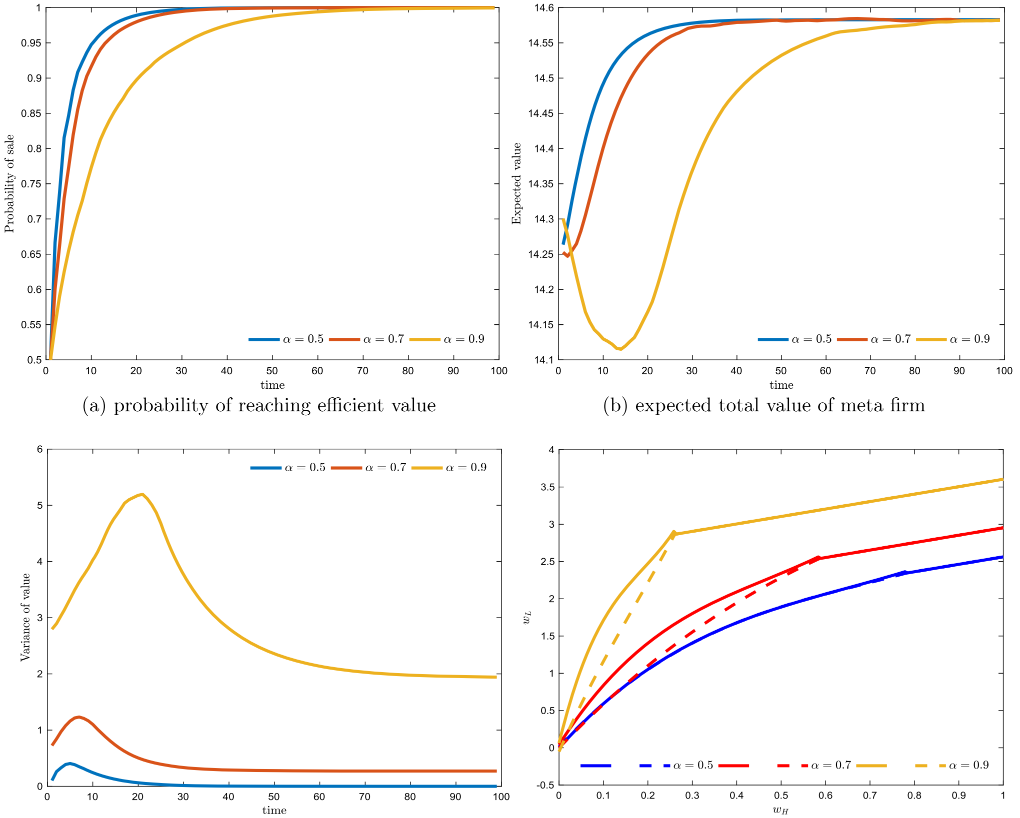
So, the mean and variance of surplus converge respectively to

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

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

For the symmetric Markov chain, $\alpha_L = 1 - \alpha_H = \alpha$, it is easy to see that $\mathbb{E}\left[\text{Economic Surplus}\right]$ is independent of α and $\mathbb{V}\left[\text{Economic Surplus}\right]$ 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 Fig. 7. First, we look at the average time it takes for a firm to reach its efficient value. Fig. 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. Fig. 7b shows the average size of the firm as a function of time. This value is 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, Fig. 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 Fig. 7d. It plots the shell that houses all values of expected utility of the agent in the optimal contract (numerical counterpart of the theoretical pictures in Fig. 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 increase the value of persistence the shell first expands and then contracts 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



(a) probability of reaching efficient value (b) expected total value of meta firm

(c) variance of total surplus (d) the shell

Fig. 7. Comparative statics with respect to persistence.

in the short-run. For very high persistence, though the shell is shrinking, the time to efficiency in large and hence so is the variance in firm values.[44]

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.

5. Final remarks

This paper motivates the study of financial constraints in dynamic contracting through the interaction between persistent private information and cash or liquidity constraints. The agent

[44] 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.

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 a number of extensions and other results not covered in the main text. First, we offer the complete proof of Theorem 1 building on the recursive approach. Second, we extend the model to incorporate asymmetric Markov chains and general iid distributions for the agent's type. Third, we look at the continuous time version of our model, and show that except points 5 and 8, all other properties from Theorem 1 continue to hold. The two exceptions arise because there is no notion of a "time period" in continuous time, so the liquidity and efficiency regions are synonymous. Fourth, we document sufficiency conditions for the validity of the relaxed problem approach and offer numerical results for the same. And, fifth, we incorporate termination into the model, allowing the principal to end the contract before the start of every period.

Theoretically speaking, the paper is limited to the two-types model because it is difficult to determine the optimal allocation for more than two Markov types. Global incentive constraints generically bind for high persistence, even for the benchmark model (see Battaglini and Lamba (2019)). Looking for approximate optimal and easily characterizable contracts is a promising approach going forward.

What if the agent is allowed to borrow or save? We conjecture that the set of allocation that satisfy incentive compatibility and cash strapped constraints would remain the same, and hence the predictions would be unchanged. Whether the agent is allowed to borrow from the principal or a third party, as long as this money has to be paid back within the life of contract, our conjecture should hold. An analogous result should hold if the agent is allowed to save and draw from those savings at any point. Of course, both these observations rely on the linearity of transfers across time.[45]

Now, an alternate way to express the cash-strapped constraint could 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 $p_t \geq \theta_t q_t$, both for tractability and because it is the standard in the literature.[46] Our preliminary results indicate that the nature of dynamic distortions would differ under this alternate modeling of cash constraints. It is potentially a good 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 citizen with privately observed labor productivity. The principal is the government seeking to maximize a Pareto-weighted welfare function. Presumably 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.

6. Appendix

We divide the appendix into twelve subsections - the benchmark model, two period model, the sequential approach, followed by the recursive approach, proof of the main theorem, opti-

[45] Notes on a proof of the conjecture for a two-type two-period model are available from the authors upon request.

[46] When the constraint binds, the principal only supplies payments that covers the cost of production.

mal limit contract, conceptual interpretation of cash-strapped as limited liability, dynamics of payments, general IID model, sufficiency conditions, introducing termination, and 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.

6.1. Benchmark: dynamic model without financial constraints

The dynamic mechanism design literature considers models akin to the one studied here with the key difference being the agent has access to deep pockets. The problem looks exactly the same as $(\mathcal{P}^*)$, except that $\mathcal{C}_i(h^{t-1})$ 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 et al. (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.

Formally, 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.

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

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

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

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 $v_0 \le \overline{v}^{\#} := \Delta\theta \sum_{t=1}^{T} \delta^{t-1}(\alpha_L - \alpha_H)^{t-1} q^e(\theta_H)$.

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 implement the optimal allocation with a restriction that first period expected utilities, $U(\theta_L)$ and $U(\theta_H)$, are uniquely pinned down. Esö and Szentes (2017) point out that these modeling 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. Our paper shows that when the agent is constrained on cash, information rents and hence distortions have markedly different time structure.[47]

6.2. 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:

$$\begin{aligned}
IC_L: &\quad U(\theta_L) \ge U(\theta_H) + \Delta\theta q(\theta_H) + \delta(\alpha_L - \alpha_H)[u(\theta_L|\theta_H) - u(\theta_H|\theta_H)] \\
\mathcal{C}_i: &\quad U(\theta_i) = u(\theta_i) + \delta[(\alpha_i u(\theta_L|\theta_i) + (1-\alpha_i)u(\theta_H|\theta_i)] \\
&\qquad\qquad\qquad\qquad \text{and } u(\theta_i) \ge 0 \quad \text{for } i = L, H \\
IC_L(\theta_i): &\quad u(\theta_L|\theta_i) \ge u(\theta_H|\theta_i) + \Delta\theta q(\theta_H|\theta_i) \quad \text{for } i = L, H \\
\mathcal{C}_H(\theta_i): &\quad u(\theta_H|\theta_i) \ge 0 \quad \text{for } i = L, H
\end{aligned}$$

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 $\mathcal{C}_L$, respectively. As in the benchmark model, λ is the Lagrange multiplier on (PK). Since the problem is concave, the optimal allocation is described by

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

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 $v_0 \le \overline{v}^* = \Delta\theta \sum_{t=1}^{2} \delta^{t-1}\mathbb{P}(\theta_t = \theta_L) q^e(\theta_H)$.

[47] It is possible for distortions to increase over time in dynamic mechanism design without financial constraints when the private information is regarding the parametrization of the Markov process itself, see Boleslavsky and Said (2013).

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 $\mathcal{C}_L$ 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 Corollary 1. Part 1 is equivalent to $q^*(\theta_H|\theta_L) \geq q^*(\theta_H|\theta_H)$ which is satisfied trivially no matter if $\mathcal{C}_L$ binds or not. Part 2 says that $S^*(v_0) = R^*(v_0) + \max\{\underline{v}^*, v_0\}$ is non-decreasing, and 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 $[\underline{v}^*, \overline{v}^*]$.

6.3. Sequential approach with $T = \infty$

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

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

As in the two period problem, each $\mathcal{C}_H(h^{t-1})$ with $h^{t-1} \neq \emptyset$ can be assumed to hold as an equality. In addition, $\mathcal{C}_H$ 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 $\mathcal{C}_L(h^{t-1})$ in terms of quantities:

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

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 $\delta^{t-1}\eta(h^{t-1})$ be the multiplier on $\mathcal{C}_L(h^{t-1})$, we get the optimal allocation for $v_0 \leq \overline{v}^*$:

$$
\begin{aligned}
& q^*(\theta_L|h^{t-1}) = q^e(\theta_L) \quad \forall h^{t-1},\ \forall\, t \\
& q^*(\theta_H|h^{t-1}, \theta_L, \theta_H^{s-1}) \\
& \quad = \mathcal{Q}\left(\eta(h^{t-1})\frac{p_s}{\mathbb{P}(h^{t-1}\theta_L\theta_H^s)} - \sum_{\tau=0}^{s-1}\eta(h^{t-1}, \theta_L, \theta_H^\tau)\frac{p_{s-1-\tau}}{\mathbb{P}(h^{t-1}\theta_L\theta_H^s)}\right) \quad \forall h^{t-1},\ \forall\, t \\
& q^*(\theta_H|\theta_H^{t-1}) = \mathcal{Q}\left((1-\lambda)\frac{\mathbb{P}(\theta_t = \theta_L)}{\mathbb{P}(\theta_H^t)} - \sum_{\tau=0}^{t-1}\eta(\theta_H^\tau)\frac{p_{t-1-\tau}}{\mathbb{P}(\theta_H^t)}\right)
\end{aligned}
$$

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 η s are endogenous and jointly determined at the optimum.

6.4. 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:

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

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

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

$(\mathcal{RF}_0)$ 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, Claims 5 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 $W \subseteq \mathbb{R}_+^2$. It is easy to see that $\{\mathbf{w} \in \mathbb{R}_+^2 : w_L \geq w_H\} \subseteq W$: fix $\mathbf{w}$ such that $w_L \geq w_H \geq 0$, and let $z_{LL} = z_{LH} = w_L$, $z_{HL} = z_{HH} = w_H$ and $q_H = q_L = 0.$, then $\mathbf{w} \in W$. We prove the converse 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:

$$\begin{aligned}
&w_L - w_H \geq \delta(\alpha_L - \alpha_H)(z_{HL} - z_{HH}) \\
&w_H \geq \delta[\alpha_H z_{HL} + (1-\alpha_H) z_{HH}]
\end{aligned} \tag{11}$$

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 (11). 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 (11). 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 $\tilde{W}_a^{old} = \{\mathbf{w} \in \mathbb{R}_+^2 : w_L \geq a w_H\}$ and define $\tilde{W}_a^{new} = \{\mathbf{w} \in \mathbb{R}_+^2 : \exists(\mathbf{z_H}, \mathbf{z_L}) \in \tilde{W}_a^{old} \times \tilde{W}_a^{old}$ s.t. (11)$\}$. Notice that $\mathbf{w}' \in \tilde{W}_a^{new}$ if and only if there exists $\mathbf{w}$ with $w'_L = w_L$, $w'_H \geq w_H = \delta\big[\alpha_H a + (1-\alpha_H)\big]z_{HH}$ and $w_L - w_H \geq \delta(\alpha_L - \alpha_H)(a-1)z_{HH}$. It follows that $\tilde{W}_a^{new} = \big\{w \in \mathbb{R}_+^2 : w_L \geq \big[1 - \frac{(\alpha_L-\alpha_H)(1-a)}{\alpha_H a + (1-\alpha_H)}\big]w_H\big\}$.

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 $\tilde{W}_l = \tilde{W}_{a_l}^{old}$. The claim follows from $a_{l-1} < a_l < 1$ for any l and $a_l \to_{l\to\infty} 1$. $\square$

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

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

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

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

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

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 $w_L \geq \underline{w}_L^{liq} = \delta[\alpha_L w_L^e + (1-\alpha_L)w_H^e] < \underline{w}_L^e$, 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 $\forall h^t$, $\forall t$

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

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 $\mathbf{w} = \big(U(\theta_L|h^t), U(\theta_H|h^t)\big)$.*
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). $\square$

The next claim establishes 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:

$$\begin{aligned}
&D_L Q^*_L(\mathbf{z}_L) = \alpha_L\rho_L \\
&D_H Q^*_L(\mathbf{z}_L) = (1-\alpha_L)\rho_L \\
&D_L Q^*_H(\mathbf{z}_H) = \alpha_H\rho_H + (\alpha_L - \alpha_H)\beta \\
&D_H Q^*_H(\mathbf{z}_H) = (1-\alpha_H)\rho_H - (\alpha_L - \alpha_H)\beta \\
&D_q s(\theta_H, q_H) = \Delta\theta\beta
\end{aligned} \tag{12}$$

In addition, the following envelope conditions are satisfied:

$$\begin{aligned}
&D_L Q^*_L(\mathbf{w}) = \alpha_L\rho_L + (1-\alpha_L)\beta \\
&D_H Q^*_L(\mathbf{w}) = (1-\alpha_L)(\rho_H - \beta) \\
&D_L Q^*_H(\mathbf{w}) = \alpha_H\rho_L + (1-\alpha_H)\beta \\
&D_H Q^*_H(\mathbf{w}) = (1-\alpha_H)(\rho_H - \beta)
\end{aligned} \tag{13}$$

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 (12). The extra first-order conditions are

$$\begin{aligned}
&\mu_L\rho_L + \mu_H\beta = \mu_L\lambda \\
&\mu_H(\rho_H - \beta) = \mu_H\lambda
\end{aligned} \tag{14}$$

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:

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

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 Fig. 4. It is the intersection of epigraph and hypograph of two strictly increasing, continuous functions connecting $\mathbf{0}$ and $\underline{\mathbf{w}}^e$. The shell has a non-empty interior and it lies above the line connecting $\mathbf{0}$ and $\underline{\mathbf{w}}^e$.

Claim 5 *(Shape of the shell).*

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

$$\lim_{w_H \to 0} w^j_L(w_H) = 0 \quad \text{and} \quad \lim_{w_H \to \underline{w}^e_H} w^j_L(w_H) = \underline{w}^e_L.$$

3. *$w^H_L > w^L_L$ on $(0, \underline{w}^e_H)$.*
4. *$w^L_L > \frac{\delta \alpha_H w_H}{1-\delta(1-\alpha_H)}$ on $(0, \underline{w}^e_H)$.*

Proof. Parts 1 and 2. First, $\{w_L \in (w_H, \underline{w}^e_L] : \eta_j(\mathbf{w}) = 0\} \neq \emptyset$ for any $w_H \in (0, \underline{w}^e_H)$, because η_j is continuous in w_L with $\lim_{w_L \to w_H} \eta_j(\mathbf{w}) = +\infty$ and $\eta_j(\underline{w}^e_L, w_H) \leq 0$ 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 $DQ^*_j(\mathbf{w}) = \mathbf{0}$ for $j = L, H$. Then, $DQ_j(\mathbf{z}_j) = \mathbf{0}$ 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^j_L 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 (13), thus

$$\frac{\eta_H(\mathbf{w})}{\alpha_H(1-\alpha_H)} = \frac{\eta_L(\mathbf{w})}{\alpha_L(1-\alpha_L)} + \frac{(\alpha_L - \alpha_H)\beta}{\alpha_L \alpha_H} \tag{15}$$

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 $w_L \geq \underline{w}^{liq}_L = \delta[\alpha_L w^e_L + (1-\alpha_L) w^e_H]$, 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 $w_L < \underline{w}^{liq}_L$. Consider $\mathbf{w}^0$ such that $\eta_H(\mathbf{w}^0) = 0$ and $w^0_L \geq \underline{w}^{liq}_L$. Then, $\eta_L(\mathbf{w}^0) < 0$ by Equation (15). There exists $\mathbf{w}^1$ such that $\eta_L(\mathbf{w}^1) = 0$ and $w^1_H = w^0_H$, $w^1_L < w^0_L$.

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 (15). Thus, there exists w^2 such that $\eta_H(\mathbf{w}^2) = 0$ and $w^2_H < w^1_H$, $w^2_L = w^2_L$. By strict concavity on H, $\eta_L(\mathbf{w}^2) < 0$, implying that $\beta(\mathbf{w}^2) > 0$ by Equation (15). Iterating, get that $\beta > 0$ on $\{\mathbf{w} \in (0, \underline{w}^e_L) \times (0, \underline{w}^e_H) : \eta_L(\mathbf{w}) = 0\} \subseteq B$ 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 $\mathbf{0}$ 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 $\mathbf{0}$ 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 $\mathbf{0}$, therefore the shell can not lie below the line. $\square$

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 (14) 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 (15), $\alpha_L(\rho_L - \rho_H) + \beta \leq 0 \leq \alpha_H(\rho_L - \rho_H) + \beta$ 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 $DQ_j^*(\underline{\mathbf{w}}^e) = \mathbf{0}$ 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 $\mathbf{0}$ 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 (12) 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 5. 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 (15), 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 $\mathbf{0}$ (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 $\mathbf{0}$ 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 (15), β is strictly decreasing, thus q_H is strictly increasing (see equation (12)). □

6.5. Main result

Theorem 1 translates the recursive characterization into sequential notations. By 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 are completely established in the Claims above. We briefly describe the connection. The optimal quantity is downward distortion from equation (12): $D_q s(\theta_H, q_H) = \Delta\beta$ 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 in $B \subseteq int(W)$.

Let $D^* = (D_L + D_H)$; equations (12) and (13) imply that the stochastic process $D^* Q_j^*$ is a non-negative martingale:

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

So, the Martingale convergence theorem delivers that $D^* Q_j^*$ converges almost surely. Therefore, the Lagrange multipliers are uniquely pinned by the limits through Equation (13). Clearly, w converges to a point in E, hence $q^*(\theta_i|h^{t-1})$ converges to $q^e(\theta_i)$ almost surely.

6.6. Optimal limit contract

In this section, we prove Proposition 2. We shall invoke the sequential approach discussed in Section 6.3. Since we are interested in the limit result as the Markov matrix approaches the identity matrix, we will consider the symmetric Markov chain: $\alpha_H = 1 - \alpha_L = \alpha$.

Recall that $\mathcal{C}_L(h^{t-1})$ 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 $\delta^{t-1}\mathbb{P}(h^{t-1}\theta_L)(1-\alpha)\eta(h^{t-1})$ be the Lagrange multiplier on this constraint, then the optimal allocation takes the following form

$$q^*(\theta_H|\theta_H^s) = \mathcal{Q}\Bigg((1-\lambda)\frac{1+(\mu_L-\mu_H)(2\alpha-1)^s}{2\mu_H\alpha^s} - \frac{p_s}{\alpha_s}\frac{\mu_L}{\mu_H}(1-\alpha)\eta$$
$$- \sum_{j=1}^{s} \frac{p_{s-j}}{\alpha^{s-j}}\frac{(1-\alpha)^2}{\alpha}\eta(\theta_H^j)\Bigg), \quad s \geq 0$$

$$q^*(\theta_H|h^{t-1}, \theta_L, \theta_H^{s-1}) = \mathcal{Q}\Bigg(\frac{p_s}{\alpha_s}\alpha\eta(h^{t-1}) - \frac{p_{s-1}}{\alpha_{s-1}}\alpha\eta(h^{t-1}, \theta_L)$$
$$- \sum_{j=1}^{s-1} \frac{p_{s-1-j}}{\alpha^{s-1-j}}\frac{(1-\alpha)^2}{\alpha}\eta(h^{t-1}, \theta_L, \theta_H^j)\Bigg), \quad s \geq 1$$

where λ is the Lagrange multiplier on (PK).

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

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

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

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

$$\mathcal{Q}\left(\eta(h^{t-1}) - \eta(h^{t-1}, \theta_L)\right) \geq \delta \mathcal{Q}\left(\eta(h^{t-1}, \theta_L, \theta_H^{s-1}) - \eta(h^{t-1}, \theta_L, \theta_H^{s-1}, \theta_L)\right), \quad s \geq 1$$

First of all, we argue that $\zeta(h^{t-1}) := \eta(h^{t-1}) - \eta(h^{t-1}, \theta_L) \geq 0$, and it holds with as an equality if and only if $\eta(h^{t-1}) = 0$. 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 $\zeta(h^{t-1}) = 0$. Then, $\mathcal{Q}(0) > \delta \mathcal{Q}\left(\zeta(h^{t-1}, \theta_L)\right)$ which leads to $\eta(h^{t-1}, \theta_L) = 0$ by the complimentary slackness. So, we could use η interchangeably with ζ as our set of Lagrange multipliers.

Next, we explicitly solve for ζ. To begin, notice that $\zeta(\theta_H^s)$ appears only in $\mathcal{Q}\left((1-\lambda)\frac{\mu_L}{\mu_H}\right) \geq \delta \mathcal{Q}\left(\zeta(\theta_H^s)\right)$ and $\mathcal{Q}\left(\zeta(\theta_H^s)\right) \geq \delta \mathcal{Q}\left(\zeta(\theta_H^s, \theta_L)\right)$. Since higher $\zeta(\theta_H^s)$ relaxes the latter constraint, $\mathcal{Q}\left(\zeta(\theta_H^s)\right) = \min\left\{q^e(\theta_H), \frac{1}{\delta}\mathcal{Q}\left((1-\lambda)\frac{\mu_L}{\mu_H}\right)\right\}$. By induction, $\zeta(h^{t-1})$ 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.

6.7. *Credit constraint versus limited liability*

Now, we show Proposition 4. 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 6.3 and 6.5, $\underline{v}^\# = \Delta\theta\mu_L q^\#(\theta_H) \leq \Delta\theta\mu_L q^*(\theta_H) < \underline{v}^*$, 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 $D^\#$:

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

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

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

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

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

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.

6.8. 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 utility spread 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.[48] Solving the promised utility identity and the "envelope formula" together:

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

gives

$$U^*(\theta_L) = v_0 + \mu_H I \quad \text{and} \quad U^*(\theta_H) = v_0 - \mu_L I \tag{16}$$

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

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

Solving the two equation gives us

$$\begin{aligned}
U^*(\theta_L|h^{t-1}, \theta_j) &= v^*(\theta_j|h^{t-1}) + (1-\alpha_j)I(h^{t-1}, \theta_j) \\
U^*(\theta_H|h^{t-1}, \theta_j) &= v^*(\theta_j|h^{t-1}) - \alpha_j I(h^{t-1}, \theta_j)
\end{aligned} \tag{17}$$

Starting from promised utility v_0 and choosing per period transfers optimally, equations (16) and (17) inductively define future expected and promised utilities. The proof of Proposition 3 then simply follows from this induction.

6.9. 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 (15) 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

$$Q^*(w) = \max_{\mathbf{z} \in W} Q_j^*(\mathbf{z}) \text{ s.t. } w = \mu_L z_L + \mu_H z_H \tag{18}$$

This definition is based on the problem $(\mathcal{RF})$, the problem $(\mathcal{RF}_0)$ is trivially modified. Importantly, that the value function in equation (18) solves the simpler Bellman equation $(\mathcal{RF}')$.

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

[48] In case (PK) is not binding, replace v_0 with $\underline{v}$ in equation (16).

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

$$w = \mu_L(u_L + \delta z_L) + \mu_H(u_H + \delta z_H)$$
$$u_L + \delta z_L \geq \Delta\theta q_H + u_H + \delta z_H$$

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 $Q^e = \mu_L Q^e_L + \mu_H Q^e_H$, then $Q^* \leq Q^e$ and $Q^* = Q^e$ if and only if $w \geq \underline{w}^e = \mu_L \underline{w}^e_L + \mu_H \underline{w}^e_H$. 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 $(0, \underline{w}^e)$.

6.10. Sufficiency conditions and global optimality

We say that the first-order approach is valid if the solution to $(\mathcal{RP}^*)$ defined in Section 3 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 3.1 the "upward" incentive constraint, IC_H, never binds. It is possible, however as we argue largely implausible that 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:

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

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

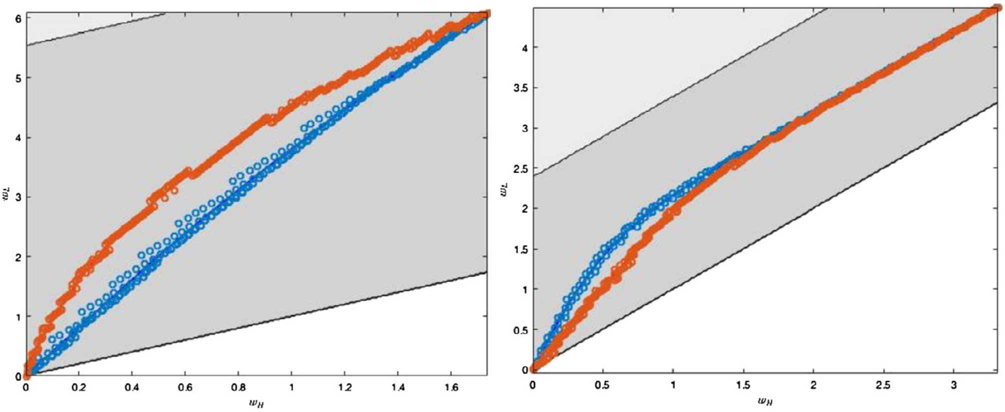
where $a_s = \mathbb{P}(\theta_{t+s} = \theta_L | \theta_t = \theta_L) = \frac{1}{1-\alpha_L+\alpha_H} \{\alpha_H + (1-\alpha_L)(\alpha_L - \alpha_H)^s\}$.

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 as for $\alpha_L = \alpha_H$ and $\alpha_H = 0$.*

Proof. $IC_H(h^{t-1})$ trivially holds when $\alpha_L = \alpha_H$, and for $\alpha_H = 0$, $\mathcal{C}_L(h^{t-1})$ 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 $\mathcal{C}_L(h^{t-1})$ is



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

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

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 $\mathcal{C}_L(h^{t-1})$ bind? It binds when quantities on the left hand side of $\mathcal{C}_L(h^{t-1})$, 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 $\mathcal{C}_L(h^{t-1})$, 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.[49] Two such examples are presented in Fig. 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,

[49] We have used the parametric setting: $V(q) = 10\sqrt{q}$, $\delta = 0.8$, $\theta_L = 3, \theta_H = 4$, $v_0 = 0$. The code is available on request.

wherein the optimal contract resides, lies within the darker shaded area. Hence the first-order approach is valid.

6.11. Introducing termination

Suppose that at the start of every period the principal can terminate the contract with some probability say λ. Upon termination, the principal gets a scrap value Φ and the agent gets his outside option which has been normalized to zero. If the principal chooses to continue the relationship, then the new type is realized and reported by the agent in return for endogenous supply of quantity and payment. We first analyze the general Markov model stated, and then show that in special case of the iid model, which is analogous to the cash flow diversion model in Clementi and Hopenhayn (2006) and the dynamic screening model in Krishna et al. (2013), more results can be established.

As before, we study the problem recursively using a two-dimensional vector of promised utilities as a state variable. There are two recursive problems to deal with: Let $\hat{Q}^*_j(\mathbf{w})$ be the maximal surplus the principal can achieve *before termination* and $Q^*_j(\mathbf{w})$ be the same value *upon continuation*.

The former problem where the principal has to decide whether to terminate or not reads as follows:

$$\hat{Q}^*_j(\mathbf{w}) = \max_{\lambda\in[0,1]} \quad \lambda\Phi + (1-\lambda)Q^*_j\left(\frac{\mathbf{w}}{1-\lambda}\right)$$

where λ is the probability of termination of the contract. The latter problem is explored separately for the Markov model and the iid model, where as expected, the iid model allows for a us more precise characterization.

General Markov model. The problem is exactly the same as $(\mathcal{RF})$, but $Q^*_j(\mathbf{z_j})$ is replaced on the right hand side of the objective by $\hat{Q}^*_j(\mathbf{z_j})$ for $j = L, H$:

$$Q^*_j(\mathbf{w}) = \max_{\langle \mathbf{z_L}, \mathbf{z_H}, \mathbf{q}\rangle} \quad \alpha_j\left[s(\theta_L, q_L) + \delta\hat{Q}^*_L(\mathbf{z_L})\right] + (1-\alpha_j)\left[s(\theta_H, q_H) + \delta\hat{Q}^*_H(\mathbf{z_H})\right]$$

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

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

As before the problem at $t = 1$ will be different and analogous to $(\mathcal{RF}_0)$.

We suppose that the termination payoff is not too large, so the termination is inefficient absent information frictions: $\Phi < Q^e_j$ for $j = L, H$, where

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

Because of this assumption the scrap value is not too large, the efficient set remains the same as before, when the model does not have termination:

$$E = \left\{\mathbf{w} \in \mathbb{R}^2_+ : w_L \ge w_H + \frac{\Delta\theta q^e(\theta_H)}{1-\delta(\alpha_L - \alpha_H)}, w_H \ge 0\right\}$$

It can be show, as before, that E is an absorbing set. Moreover, the value function Q^* posses exactly the same properties as the earlier model, namely Q^* is well-defined, strictly concave, strictly increasing on $\left[0, \frac{\mu_L \Delta\theta q^e(\theta_H)}{1-\delta}\right]$ and it is constant on E. Finally, it is continuously differentiable with $\lim\limits_{w \downarrow 0} DQ^*(w)$.

The principal's decision of whether to terminate or not is summarized in the following proposition.

Proposition 5. *Fix the expected utility of the two types to be $w_L > w_H > 0$, and let $r = \frac{w_L}{w_H}$. There exists a unique value $\hat{v}_j(r) \geq 0$ such that*

(a) the contract is continued with probability one whenever $w_H \geq \hat{v}_j(r)$;
(b) the contract is continued with probability $\frac{w_H}{\hat{v}_j(r)}$ whenever $w_H \leq \hat{v}_j(r)$.

Proof. Rewrite the problem as

$$\hat{Q}_j^*(\mathbf{w}) = \max_{v \geq w_H} \quad \left(\frac{w_H}{v}\right)\Phi + \left(1 - \frac{w_H}{v}\right) Q_j^*(rv, v)$$

The FOC is as follows:

$$\Phi - Q_j^*(rv, v) - (rv, v) \cdot D_j^*(rv, v) \begin{cases} = 0, & v > w_H \\ \leq 0, & v = w_H \end{cases}$$

Concavity of Q_J^* implies that the function $v \mapsto \Phi - Q_j^*(rv, v) - (rv, v) \cdot D_j^*(rv, v)$ is non-increasing for any $r \in (1, \infty)$. Moreover, for sufficiently large v, $(rv, v) \in E$, thus the FOC holds as "<". So, there exists unique $\hat{v}_j(r) \geq 0$ such that

$$\begin{cases} \Phi - Q_j^*(\hat{v}_j(r)) + \big(r\hat{v}_j(r), \hat{v}_j(r)\big) \cdot DQ_j^*\big(r\hat{v}_j(r), \hat{v}_j(r)\big) = 0, & \hat{v}_j(r) > 0 \\ \Phi - Q_j^*(\hat{v}_j(r)) \leq 0, & \hat{v}_j(r) = 0 \end{cases}$$

Then, $w_H = (1 - \lambda)\max\{w_H, \hat{v}_j(r)\}$ which concludes the proof. □

Proposition 5 establishes that a simple termination policy is in fact optimal. For r $w_H \leq \hat{v}_j(r)$, the contract is either termination or it is continued with the promised utility of $\hat{v}_j(r)$, and the exact mixing probability is chosen to satisfy the promise-keeping constraint. The "slope" of the expected utilities, the line connecting $\mathbf{0}$ and $\mathbf{w}$, essentially pins down the threshold $\hat{v}_j(r)$ for w_H above which the contract is continued with probability one, and below which it is terminated with a unique probability.

The uniqueness of the termination threshold and the fact of efficiency as an absorbing state, in conjunction with the martingale convergence theorem, gives the following long-term termination-efficiency result.

Proposition 6. *Consider the model with correlated types and possibility of termination. Then the optimal contract is almost surely either terminated or becomes efficient.*

Proof. Proposition 5 implies that the contract is not terminated whenever $w_H > \hat{v}_j(w_L/w_H)$ for $j = L, H$, thus $\hat{Q}_j^*(\mathbf{w}) = Q_j^*(\mathbf{w})$. Then, by the same reasoning as in the model without termination:

$$(D_L + D_H)\hat{Q}_j^*(\mathbf{w}) = \alpha_j (D_L + D_H)\hat{Q}_L^*(\mathbf{z}_L) + (1-\alpha_j)(D_L + D_H)\hat{Q}_H^*(\mathbf{z}_H) \quad j = L, H$$

Conditional on no termination, the process of promised utilities is bounded from below. Moreover, $(D_L + Q_H)\hat{Q}_j^*$ is a non-negative martingale whenever termination is suboptimal. Conclude that $(D_L + Q_H)\hat{Q}_j^*$ must converge almost surely (provided the contract is not terminated). And, a point of convergence to must be such that $(D_L + Q_H)\hat{Q}_j^* = 0$, that is the contract is efficient. □

IID model. The value function before the decision of termination is taken reads the same as before:

$$\hat{Q}^*(w) = \max_{\lambda\in[0,1]} \quad \lambda\Phi + (1-\lambda)Q^*\left(\frac{w}{1-\lambda}\right)$$

The latter problem to be solved upon the decision to continue the contract is now simpler than before:

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

subject to $\langle \mathbf{z}, \mathbf{q}\rangle \in \mathbb{R}_+^4$, and

$$w = \mu_L(u_L + \delta z_L) + \mu_H(u_H + \delta z_H)$$
$$u_L + \delta z_L \geq \Delta\theta q_H + u_H + \delta z_H$$

As before, to make the problem interesting we suppose that termination is inefficient when information frictions are "small": $\Phi < \lim_{w\to\infty} Q^*(w)$, or in terms of primitives:

$$\Phi < \frac{\mu_L s(\theta_L, q^e(\theta_L)) + \mu_H s(\theta_H, q^e(\theta_H))}{1-\delta}$$

This ensures that the efficiency set is the same as before in the iid model without termination

$$E = \left[\frac{\mu_L \Delta\theta q^e(\theta_H)}{1-\delta}, \infty\right)$$

Our first result characterizes the optimal termination policy.

Proposition 7. *There exists a unique value $\hat{v} \in \left[0, \frac{\mu_L\Delta\theta q^e(\theta_H)}{1-\delta}\right)$ such that*

1. *the contract is continued with probability one whenever $w \geq \hat{v}$;*
2. *the contract is continued with probability $\frac{w}{\hat{v}}$ whenever $w \leq \hat{v}$.*

Proof. The first-order condition with respect to λ can be written as:

$$\Phi - Q^*\left(\frac{w}{1-\lambda}\right) + \left(\frac{w}{1-\lambda}\right) DQ^*\left(\frac{w}{1-\lambda}\right)\begin{cases} = 0, & \lambda \in (0,1) \\ \leq 0, & \lambda = 0\end{cases}$$

Moreover, concavity of Q^* implies that this FOC is sufficient and determines the unique value of λ. Our assumption on the level of Φ implies that the FOC holds as "<" whenever λ is large enough. So, there exists unique $\hat{v}$ satisfying the following:

$$\begin{cases} \Phi - Q^*(\hat{v}) + \hat{v} D Q^*(\hat{v}) = 0, & \hat{v} > 0 \\ \Phi - Q^*(0) \leq 0, & \hat{v} = 0 \end{cases}$$

Then, $w = (1 - \lambda) \max\{w, \hat{v}\}$ which concludes the proof. □

Next, we summarize dynamic properties of the optimum.

Proposition 8. *Consider the iid model with the possibility of termination, then the optimal contract satisfies the following properties:*

1. *Efficiency is achieved above a threshold and it is an absorbing set: For $w \in E \Rightarrow z_L, z_H \in E$.*
2. *Promised utility increases after a low cost and decreases after a high cost: $w \notin E \Rightarrow z_L > w > z_H$;*
3. *Almost surely, the contract is either terminated or becomes efficient.*

Proof. Part (1) follows from the structure of E.

First show we Part (2). The incentive constraint must bind and $u_H = 0$, thus $w = \mu_L \Delta\theta q_H + \delta z_H$ and $\delta z_L \geq w + \mu_H \Delta\theta q_H$. It is immediate that $z_L > w$. We shall show that $z_H < w$ whenever $w < \frac{\mu_L \Delta\theta q^e(\theta_H)}{1-\delta}$. By the Envelope theorem applied to $\hat{Q}^*$:

$$D\hat{Q}^*(w) = DQ^* \left(\max\{\hat{v}, w\}\right) \leq DQ^*(w)$$

where the last inequality follows from concavity of Q^*. By the Envelope theorem applied to Q^*:

$$DQ^*(w) = \mu_L D\hat{Q}^*(z_L) + \mu_H D\hat{Q}^*(z_H)$$

Combining these two results, obtain that

$$DQ^*(w) \leq \mu_L D\hat{Q}^*(z_L) + \mu_H D\hat{Q}^*(z_H)$$

Strict concavity of Q^* and $z_L > w$ implies that $z_H < w$.

Finally, we argue that the contract must either converge to efficiency or get terminated, establishing Part (3). By Proposition 7, there is no termination whenever $w > \hat{v}$, thus $\hat{Q}^*(w) = Q^*(w)$, and

$$D\hat{Q}^*(w) = \mu_L D\hat{Q}^*(z_L) + \mu_H D\hat{Q}^*(z_H)$$

Since the promised utility upon continuation is at least $\hat{w}$ and the derivate of $\hat{Q}^*$ is a non-negative martingale whenever $w \geq \hat{v}$, the process of promised utilities must converge almost surely (conditional on no termination). It is routine to verify it must converge to a point in E, that is $D\hat{Q}^* = 0$. □

The propositions in this subsection can be seen can be seen as generalizations of corresponding results in the iid cash flow diversion model in Clementi and Hopenhayn (2006). The optimal allocation rule, as noted in Proposition 6 in Clementi and Hopenhayn (2006), looses its monotonicity properties upon the introduction of termination. In fact, we numerically find the allocation is decreasing in the recursive domain close to the termination threshold and increasing close to the efficiency region. For low value of expected utility, instead of simply decreasing the allocation on the realization of a "bad shock" the principal prefers to simultaneously increase the allocation and the probability of termination.

6.12. The model in continuous time

We show that in "continuous time", 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.

Time here 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:

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

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.[50] 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.

Following Williams (2011) we adopt the first order approach restricting agent's strategy even further: the agent can not understate his costs.[51] 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 $dt > 0$,

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

where q_H is an average quantity from (h^{t-}, θ_H) to $\left(h^{t-}, \theta_H^{[t,t+dt)}\right)$. 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 $dt > 0$,

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

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

[50] 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.

[51] 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.

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

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

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 $j, k = L, H$.[52] 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:

$$\dot{z}_{LL} \leq (\lambda_L + r)w_L - \lambda_L w_H \tag{19}$$

$$\dot{z}_{HH} \leq (\lambda_H + r)w_H - \lambda_H w_L \tag{20}$$

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

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.[53] For now, assume that these functions are continuously differentiable, then the HJB equations are as it follows:

$$(\lambda_L + r)Q_L(w_L) = s(\theta_L, q_L^e) + \max_{\dot{z}_{LL}, w_H \in [0, w_L]} \left\{ Q_L'(w_L)\dot{z}_{LL} + \lambda_L Q_H(\mathbf{w}) \right\} \text{ s.t. (19)}$$

$$(\lambda_H + r)Q_H(\mathbf{w}) = \lambda_H Q_L(w_L) + \max_{\dot{\mathbf{z}}, q_H \in [0, q_H^e]} \left\{ s(\theta_H, q_H) + DQ_L(\mathbf{w}) \cdot \dot{\mathbf{z}}_H \right\}$$

$$\text{s.t. (20) and (21)}$$

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 supermodularity. And, $DQ_H(\mathbf{w}) = \mathbf{0}$ if and only if $\mathbf{w} \in E$. So, for $\mathbf{w} \notin E$ both cash-strapped constraint must bind.[54] We will look at the set H which is defined as in Claim 4:

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

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 inverses of w_L^i defined in the discrete case, so they have the same properties. To be specific, they continuously connect $\mathbf{0}$ 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

52 $\dot{z}_{jk} \pm \infty$ stays for a discrete jump.

53 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.

54 In Equation (19), w_H is a promised utility after an adjustment.

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

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

$$\begin{aligned} \lambda_H[D_H Q_H(\mathbf{w}) - Q_L'(w_L)] &= D(D_L Q_H)(\mathbf{w}) \cdot \dot{\mathbf{z}}_H \\ \lambda_H D_L Q_H(\mathbf{w}) &= D(D_H Q_H)(\mathbf{w}) \cdot \dot{\mathbf{z}}_H \end{aligned} \tag{22}$$

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 (22) 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 $\dot{z}_{HH} \le (w_H^L)'(w_L)\dot{z}_{HL}$. 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 (22), 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: $q_H(\mathbf{w}) \le q_H(w_L, w_H^L(w_L))$. To see this, notice that in the shell $w_H \le w_H^L(w_L)$. 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.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jet.2021.105196.

References

Abreu, D., Pearce, D., Stacchetti, E., 1990. Toward a theory of discounted repeated games with imperfect monitoring. Econometrica 58 (5), 1041–1063.

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

Arve, M., Martimort, D., 2016. Dynamic procurement under uncertainty: optimal design and implications for incomplete contracts. Am. Econ. Rev. 106 (11), 3238–3274.

Ashlagi, I., Daskalakis, C., Haghpanah, N., 2019. Sequential Mechanisms with Ex-Post Participation Guarantees. Stanford University and Massachusetts Institute of Technology and Pennsylvania State University.

Banerjee, A.V., Duflo, E., 2014. Do firms want to borrow more? Testing credit constraints using a directed lending program. Rev. Econ. Stud. 81 (2), 572–607.

Battaglini, M., 2005. Long-term contracting with Markovian consumers. Am. Econ. Rev. 95 (3), 637–658.

Battaglini, M., Lamba, R., 2019. Optimal dynamic contracting: the first-order approach and beyond. Theor. Econ. 14 (4), 1435–1482.
Bergemann, D., Välimäki, J., 2019. Dynamic mechanism design: an introduction. J. Econ. Lit. 57 (2), 235–274.
Besanko, D., 1985. Multi-period contracts between principal and agent with adverse selection. Econ. Lett. 17 (1–2), 33–37.
Bester, H., Strausz, R., 2001. Contracting with imperfect commitment and the revelation principle: the single agent case. Econometrica 69 (4), 1077–1098.
Biais, B., Mariotti, T., Plantin, G., Rochet, J.-C., 2007. Dynamic security design: convergence to continuous time and asset pricing implications. Rev. Econ. Stud. 74 (2), 345–390.
Biais, B., Mariotti, T., Rochet, J.-C., 2013. Dynamic financial contracting. In: Acemoglu, D., Arellano, M., Dekel, E. (Eds.), Advances in Economics and Econometrics: Tenth World Congress. Cambridge University Press, pp. 125–171. Chapter 9.
Boleslavsky, R., Said, M., 2013. Progressive screening: long-term contracting with a privately known stochastic process. Rev. Econ. Stud. 80 (1), 1–34.
Campello, M., Graham, J.R., Harvey, C.R., 2010. The real effects of financial constraints: evidence from a financial crisis. J. Financ. Econ. 97 (3), 470–487.
Clementi, G.L., Hopenhayn, H.A., 2006. A theory of financing constraints and firm dynamics. Q. J. Econ. 121, 229–265.
Courty, P., Li, H., 2000. Sequential screening. Rev. Econ. Stud. 67 (4), 697–717.
DeMarzo, P.M., Fishman, M.J., 2007. Optimal long-term financial contracting. Rev. Financ. Stud. 20 (6), 2079–2128.
Doval, L., Skerta, V., 2019. Mechanism Design with Limited Commitment. Caltech and UT Austin.
Esö, P., Szentes, B., 2017. Dynamic contracting: an irrelevance theorem. Theor. Econ. 12 (1), 109–139.
Fernandes, A., Phelan, C., 2000. A recursive formulation for repeated agency with history dependence. J. Econ. Theory 91 (2), 223–247.
Fu, S., Krishna, V.R., 2019. Dynamic financial contracting with persistent private information. Rand J. Econ. 50 (2), 418–452.
Garrett, D., Pavan, A., 2015. Dynamic managerial compensation: a variational approach. J. Econ. Theory 159, 775–818.
Garrett, D., Pavan, A., Toikka, J., 2018. Robust Predictions of Dynamic Optimal Contracts (Working Slides). Toulouse School of Economics, Northwestern University and MIT.
Golosov, M., Tsyvinski, A., Werquin, N., 2016. Recursive contracts and endogenously incomplete markets. In: Taylor, J.B., Uhlig, H. (Eds.), Handbook of Macroeconomics, Vol. 2. Elsevier, pp. 725–841. Chapter 10.
Green, E.J., 1987. Lending and the smoothing of uninsurable income. In: Prescott, E.C., Wallace, N. (Eds.), Contractual Arrangements for Intertemporal Trade, Vol. 1. University of Minnesota Press, pp. 3–25. Chapter 1.
Grillo, E., Ortner, J., 2020. Dynamic Contracting with Limited Liability Constraints. Collegio Carlo Alberto and Boston University.
Guo, Y., Hörner, J., 2018. Dynamic Allocation Without Money. Northwestern University and Yale University.
Halac, M., Yared, P., 2014. Fiscal rules and discretion under persistent shocks. Econometrica 82 (5), 1557–1614.
İmrohoroğlu, A., Tüzel, Ş., 2014. Firm-level productivity, risk, and return. Manag. Sci. 60 (8), 2073–2090.
Kiyotaki, N., 2012. A mechanism design approach to financial frictions. In: Allen, F., Aoki, M., Fitoussi, J.-P., Kiyotaki, N., Gordon, R., Stiglitz, J.E. (Eds.), The Global Macro Economy and Finance. Palgrave Macmillan, UK, pp. 177–187. Chapter 9.
Krähmer, D., Strausz, R., 2015a. Dynamic mechanism design. In: An Introduction to the Theory of Mechanism Design by Tilman Börgers. Oxford University Press, pp. 204–234. Chapter 11.
Krähmer, D., Strausz, R., 2015b. Optimal sales contracts with withdrawal rights. Rev. Econ. Stud. 82 (2), 762–790.
Krasikov, I., Lamba, R., Mettral, T., 2019. Implications of Unequal Discounting in Dynamic Contracting. Pennsylvania State University.
Krishna, R.V., Lopomo, G., Taylor, C., 2013. Stairway to heaven or highway to hell: liquidity, sweat equity, and the uncertain path to ownership. Rand J. Econ. 44 (1), 104–127.
Laffont, J.-J., Martimort, D., 2002. The Theory of Incentives: the Principal-Agent Model. Princeton University Press.
Li, J., Matouschek, N., Powell, M., 2017. Power in organizations. Am. Econ. J. Microecon. 9 (1), 217–241.
Lipnowski, E., Ramos, J., 2019. Repeated Delegation. University of Chicago and University of Southern California.
Luz, V.F., 2015. Dynamic Competitive Insurance. University of British Columbia.
Maestri, L., 2017. Dynamic contracting under adverse selection and renegotiation. J. Econ. Theory 171, 136–173.
Midrigan, V., Xu, D.Y., 2014. Finance and misallocation: evidence from plant-level data. Am. Econ. Rev. 104 (2), 422–458.
Modigliani, F., Miller, M.H., 1958. The cost of capital, corporation finance and the theory of investment. Am. Econ. Rev. 48 (3), 261–297.

Moll, B., 2014. Productivity losses from financial frictions: can self-financing undo capital misallocation? Am. Econ. Rev. 104 (10), 3186–3221.

Myerson, R.B., Satterthwaite, M.A., 1983. Efficient mechanisms for bilateral trading. J. Econ. Theory 29 (2), 265–281.

Pavan, A., 2016. Dynamic mechanism design: robustness and endogenous types. In: 2015 World Congress of the Econometric Society.

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

Sannikov, Y., 2013. Dynamic security design and corporate financing. In: Constantinides, G.M., Harris, M., Stulz, R.M. (Eds.), Handbook of the Economics of Finance, vol. 2A. Elsevier, pp. 71–122. Chapter 2.

Sen, A., 1996. Termination clauses in long-term contracts. J. Econ. Manag. Strategy 5 (4), 473–496.

Spear, S.E., Srivastava, S., 1989. On repeated moral hazard with discounting. Rev. Econ. Stud. 54 (4), 599–617.

Stantcheva, S., 2020. Dynamic taxation. Annu. Rev. Econ. 12 (1), 801–831.

Stokey, N.L., Lucas Jr., R.E., Prescott, E., 1989. Recursive Methods in Economic Dynamics. Harvard University Press.

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

Vohra, R.V., 2012. Dynamic mechanism design. Surv. Oper. Res. Manag. Sci. 17 (1), 60–68.

Williams, N., 2011. Persistent private information. Econometrica 79 (4), 1233–1275.

Zhang, Y., 2009. Dynamic contracting with persistent shocks. J. Econ. Theory 144 (2), 635–675.