DISCUSSION PAPER SERIES

No. 10956

OPTIMAL DYNAMIC CONTRACTING: THE FIRST-ORDER APPROACH AND BEYOND

Marco Battaglini and Rohit Lamba

INDUSTRIAL ORGANIZATION

ISSN 0265-8003

OPTIMAL DYNAMIC CONTRACTING: THE FIRST-ORDER APPROACH AND BEYOND

Marco Battaglini and Rohit Lamba

Discussion Paper No. 10956
November 2015
Submitted 19 November 2015

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: (44 20) 7183 8801
www.cepr.org

OPTIMAL DYNAMIC CONTRACTING: THE FIRST-ORDER APPROACH AND BEYOND[†]

Abstract

We study a dynamic principal-agent model in which the agent's types are serially correlated. In these models, the standard approach consists of first solving a relaxed version in which only local incentive compatibility constraints are considered, and then in proving that the local constraints are sufficient for implementability. We explore the conditions under which this approach is valid and can be used to characterize the profit maximizing contract. We show that the approach works when the optimal allocation in the relaxed problem is monotonic in the types, a condition that is satisfied in most solved examples. Contrary to the static model, however, monotonicity is generally violated in many interesting economic environments. Moreover, when the time horizon is long enough and serial correlation is sufficiently high, global incentive compatibility constraints are generically binding. By fully characterizing a simple two period example, we uncover a number of interesting features of the optimal contract that cannot be observed in spatial environments in which the standard approach works. Finally, we show that even in complex environments, approximately optimal allocations can be easily characterized by focusing on a particular class of contracts in which the allocation is forced to be monotonic.

JEL Classification: D86
Keywords: contract theory and dynamic contracts

Marco Battaglini battaglini@cornell.edu
Cornell University, EIEF and CEPR

Rohit Lamba rlamba@psu.edu
Pennsylvania State University

[†] For useful comments and discussions we thank Dilip Abreu, Dirk Bergemann, Carlo Cabrera, Sylvain Chassang, Stephen Morris, Wolfgang Pesendorfer, Roland Strausz, Balazs Szentes, Juuso Valimaki and seminar participants at the Einaudi Institute for Economics and Finance, New York University, Princeton University, University of Toronto, Yale University, University of Cambridge, the 9th Csef-Igier Symposium on Economics and Institutions, the Tenth World Congress of the Econometric Society. We thank Nemanja Antic and Edoardo Grillo for outstanding research assistance. Battaglini gratefully acknowledges the hospitality of the Einaudi Institute for Economics and Finance, and Lamba gratefully acknowledges the time spent at Cambridge-INET institute while working on this paper. An earlier version of this paper forms Chapter 3 of Lamba's PhD dissertation at Princeton, and he is especially grateful to Stephen Morris for his guidance.

1 Introduction

Most contractual relationships have a dynamic nature, involving long-term, non-anonymous interactions between a principal and an agent. Examples of these contractual relationships include income taxation, regulation, managerial compensation or a monopolist repeatedly selling a non-durable good to a buyer. In these environments contracts can be made contingent on past realizations of the agent's type, allowing the principal to use the agent's revealed preferences to screen future types' realizations. This may be particularly useful in limiting asymmetric information and agency problems when the agent's type is persistent over time.

Despite recent advances in contract theory, there is still a limited understanding about how to use this information to design optimal screening contracts. Dynamic contracts are difficult to study because they involve a large number of incentive compatibility constraints. The analysis of optimal dynamic contracts has therefore been limited to economic environments in which a form of the "first-order approach" can be applied: environments in which the optimal contract can be fully characterized using only the necessary conditions implied by local incentive compatibility constraints. While the first-order approach can be generally applied in static environments under mild regularity assumptions, in dynamic models local incentive compatibility constraints have been shown to be sufficient only in certain specific economic environments.[1]

This leaves three sets of open questions. First, what is the general applicability of the first-order approach and what are its implications? Second, in environments in which the first-order approach does not hold, what does the optimal contract look like? Are there phenomena associated with dynamic contracts that we are ignoring by focusing on environments in which solving the contract is easy? Finally, if characterizing the optimal contract is complicated, can we approximate the optimal contracts with simpler contracts which guarantee a minimal loss in profits?

To address these questions, we consider a simple principal-agent model in which a monopolist repeatedly sells a non durable good to a buyer. The "type" of the buyer that parametrizes his utility is private information, and it evolves over time according to a general N-state Markov process**.** Higher types are assumed to have higher marginal valuations and their associated conditional distribution on future types first-order stochastically dominates the distribution of lower types.

We present four sets of results. We start by exploring the applicability of the first-order approach. We show that if we ignore global constraints, necessary local incentive compatibility constraints allow us to state a "dynamic envelope theorem" with discrete types through which the agent's equilibrium rent can be expressed just as a function of the expected allocation. The dynamic envelope theorem allows a simple characterization of the profit maximizing contract. In keeping with the terminology from the static literature, this contract is referred to as the first-order optimal contract, or FO-optimal contract. We also show that the envelope formula and a simple

[1] We will discuss the literature in greater detail in Section 8.

form of monotonicity of the allocation are sufficient for implementability.[2] The monotonicity condition invoked in these results is only sufficient and quite strong, but it is verified for virtually all environments in which the optimal dynamic contract has been characterized in the existing literature.[3] Although various characterizations of the envelope conditions have been presented over time,[4] this paper is the first to provide a general characterization of the formula and its implications for discrete types. This approach has two advantages. First, most applied works using numerical methods to study dynamic contracts rely either on the discrete type assumption or discrete approximations of the continuous type model. The formula with discrete types presented in this paper, thus, allows an exact characterization. Second, focusing on discrete types allows us to avoid the measure theoretic complications of the case with continuous types which may obscure otherwise simple economic intuitions.

Second, we show that in general for a large class of economically interesting parameters the dynamic envelope formula is not sufficient to characterize the optimal dynamic contract. In particular, when types' persistence is high and hence private information has sufficient bite, the first-order optimal contract is generically non-monotonic. As a consequence global incentive constraints are generically binding if the time horizon is sufficiently important (that is when number of periods T and the discount factor δ are high enough).

As an example, consider a simple two period model with no discounting, and ex ante uniformly distributed $N+1$ equally spaced agent types. For Markov evolution of types governed by a renewal model with persistence probability α, a second period global incentive compatibility constraint binds for any $\alpha \geq \alpha^*(N)$ and a first period global incentive compatibility constraint binds for any $\alpha \geq \alpha^{**}(N)$, where $\alpha^*(N)$ and $\alpha^{**}(N)$ are both strictly decreasing functions of N.[5] To put things in perspective, for $N = 5 : \alpha^*(5) = 0.43$ and $\alpha^{**}(5) = 0.63$. Numerical calculations presented in the paper show that in many natural examples, the level of persistence in types needed for the failure of the first-order approach is in fact quite low.

These findings on the limits of the first-order approach have important implications for applied work. Recent empirical evidence has shown that in many applications of dynamic principal-agent models (including the study of optimal taxation), the key variable for which agents have private information is highly persistent.[6] These are precisely the environments where our results establish that the use of the first-order approach is particularly problematic: either it does not work, due to

[2] An allocation is implementable if there exist transfers such that the contract is incentive compatible. Monotonicity requires that if $h^t \succeq \widehat{h}^t$, then $q(h^t) \geq q(\widehat{h}^t)$, where $q(h^t)$ (resp., $q(\widehat{h}^t)$) is the quantity allocated following a history h^t (resp., $\widehat{h}^t$). A history is a vector of reports $h^t = (h_1^t, ..., h_t^t)$, so $h^t \succeq \widehat{h}^t$ if $h_j^t \geq \widehat{h}_j^t \ \forall j \leq t$.

[3] Necessary and sufficient conditions for the optimality of the FO-contract can easily be stated, see Section 4. But, these tend to be less intuitive.

[4] See, among others, Baron and Besanko [1984], Besanko [1985], Laffont and Tirole [1996], Courty and Li [2000], Battaglini [2005], and more recently by Pavan, Segal and Toikka [2014].

[5] In a renewal model, in the second period, the agent has the same type (as in the first period) with some probability α, and with probability $1 - \alpha$ types change uniformly.

[6] Using a recent large data set Guvenen et al. [2014] and [2015] show that individual income in the U.S. is very persistent and the empirical distribution of income changes has extremely high kurtosis. Therefore, in all applications where income is the key variable it is appropriate to assume that types are highly persistent.

the violation of some global incentive constraint; or it works only because a non-generic stochastic process has been assumed. A theory that holds only if types are "not too correlated" is clearly unsatisfactory: correlation in types is the *raison d'etre* of dynamic contracts.

Our third contribution is to fully characterize the optimal contract in a simple environment with three types and two periods. The characterization shows that the seller typically finds it optimal to offer a continuation utility in the second period that is not monotonic in the revealed first period type. The optimal contract is characterized by separation of types even when separation is not optimal in static contracts. It is also characterized by what we call *dynamic pooling*: strategic state contingent treatment of types in which types may be initially separated, to be then pooled conditioned on particular histories.

In our final contribution, we make a first step in addressing the problem of designing optimal contracts in complex environments with large T and N. We identify a particular class of allocations for which the optimal implementable contract, which we term *monotonic contracts*, can be easily characterized. Quantities in monotonic contracts *are forced* to be non-decreasing in types (a restriction, following Roger Myerson's original terminology, we call *ironing*). Restricting to this subset of contracts is not optimal in general. However, we show that in addition to being incentive compatible, the optimal monotonic contract converges in probability to the optimal contract as types become highly persistent. And, for an infinitely repeated model, as the types' persistence and discounting converge to one, independent of the order of these limits the expected profit in the optimal monotonic contract converges to the optimal profits. In these cases, the loss in the monopolist's profit goes to zero. Further, numerical calculations show that the optimal monotonic contract performs very well, ensuring a minimal loss in objective, even with lower levels of persistence.

We proceed as follows. In Section 2 we present the model. In Section 3 we present the dynamic envelope formula and the first-order optimal contract. In Section 4 we characterize the validity of the first-order approach. In Section 5 we establish the limits of the first-order approach in the form of an impossibility result for a general class of dynamic models. In Section 6 we completely characterize a three type, two period model. In Section 7 we introduce and explore monotonic contracts. In Section 8 we provide an overview of the literature. Finally, conclusions are presented in Section 9. Proofs can be found in the appendices.

2 Model

There are two players, a buyer (or consumer) and a seller (or monopolist). The buyer repeatedly buys a non-durable good from the seller. Consumer of type θ^t enjoys a per-period utility $u(\theta^t, q) - p$ for q units of the good bought at a price p. In every period, the seller produces the good with a cost function $c(q)$. The utility and cost functions satisfy the usual conditions. The utility function $u(\theta^t, q)$ is increasing and differentiable in both arguments, with $u(\theta^t, 0) = 0$; it is concave in q; and it satisfies the single crossing condition:

Assumption 1. *$u_{\theta q}(\theta, q) > 0$ for any θ and q.*

The cost function $c(q)$ is increasing, convex and differentiable with $c'(0) = 0$ and $\lim_{q\to\infty} c'(q) = \infty$. For future reference, let $s(\theta, q) = u(\theta, q) - c(q)$ be the instantaneous surplus generated by a contract that supplies quantity q to a buyer of type θ. In what follows, $s_q(\theta, q)$ and $u_q(\theta, q)$ denote the derivatives with respect to q. To illustrate some of the results, we will repeatedly use the classic version of this model proposed by Mussa and Rosen [1978] in which $u(\theta^t, q) = \theta^t q$ and $c(q) = (1/2)q^2$.

The type θ^t evolves over time according to a Markov process. There are $N+1$ possible types, $\Theta = \{\theta_0, \theta_1, ..., \theta_N\}$, with $\theta_i - \theta_{i+1} = \Delta\theta > 0$ for any $i = 0, ..., N-1$. Let $\mathcal{N} = \{0, 1, 2, ..., N\}$ denote the set of all indices of types, noting that the indices uniquely identify the types. The probability that type k is reached next period if the agent's current type is i is given by $f(\theta_k|\theta_i) = \alpha_{ik}$. Let F be the conditional CDF, defined $F(\theta_j|\theta_i) = \sum_{k=0}^{N-j} \alpha_{i(j+k)}$. The distribution of types conditional on being type i is denoted $\boldsymbol{\alpha}_i = (\alpha_{i0}, \alpha_{i1}, ..., \alpha_{iN})$, where we assume that $\boldsymbol{\alpha}_i$ has full support (so $\alpha_{ij} > 0$ for any i, j), and $\boldsymbol{\alpha}_i$ first-order stochastically dominates $\boldsymbol{\alpha}_j$ for any i and any $j > i$. Given that higher indices imply lower values, first-order stochastic dominance can be stated as:

Assumption 2. *$F(\theta_j|\theta_i) \leq F(\theta_j|\theta_k)$ for any j and $i \leq k$.*

In each period the consumer observes the realization of his own type; the seller, in contrast, can only observe past allocations. At date 0 the seller has a prior $\mu = (\mu_0, ..., \mu_N)$ on the agent's type. For convenience in most of what follows we assume the prior has full support, so $\mu_i > 0$ for any i. This assumption is made only to simplify notation and is not necessary for the results.

In static models, standard concavity assumptions on the objectives and distributional assumptions like monotone hazard rate on the prior ensure the validity of the first-order approach, see for example Stole [2001]. We require the former assumption, but we do not need the latter. Define:

$$\Phi(\theta_i, q) = s(\theta_i, q) - \frac{1 - \sum_{k=i}^{N} \mu_k}{\mu_i} \cdot [u(\theta_{i-1}, q) - u(\theta_i, q)] .$$

Assumption 3. *Φ is concave and has a a unique interior maximum over q for all i.*

This assumption rules out situations in which even in the static model the optimal solution is the zero supply corner solution.

We assume that time is discrete and the relationship between the buyer and the seller lasts for $T \geq 2$ periods. In period 1 the seller offers a supply contract to the buyer. The buyer can reject the offer or accept it; in the latter case the buyer can walk away from the relationship at any time $t \geq 1$ if the expected continuation utility offered by the contract falls below the reservation value $\underline{U} = 0$. In line with the standard model of price discrimination, the monopolist commits to the contract that is offered. The common discount factor is $\delta \in (0, 1)$.

It is easy to show that in this environment a form of the revelation principle is valid, which allows us to consider, without loss of generality, only contracts that depend on the history of type revelations, i.e., the contract can be written as $\langle \mathbf{p}, \mathbf{q} \rangle = \left(p\left(\theta^t \left| h^{t-1} \right.\right), q\left(\theta^t \left| h^{t-1} \right.\right) \right)_{t=1}^{T}$, where h^{t-1} and θ^t are, respectively, the public history up to period $t-1$ and the type revealed at time t.[7]

[7] Note that the superscript of θ signifies time period and subscript the type: θ_i^t. Often we will write just one of them and the other will be clear from the context.

In general, h^t can be defined recursively as $h^t = \left\{h^{t-1}, \theta^t\right\}$, $h^0 = \emptyset$. The set of possible histories at time t is denoted H^t (for simplicity $H = H^T$). Let κ_t be the mapping that projects the first t elements of a vector. The set of full histories that follow h^t till time t is given by $H(h^t) = \{h \in H | \kappa_t(h) = h^t\}$. It is also useful to define the set $\widehat{H}(h^t) = \{h \in H(h^t) | h_\tau < \theta_0 \ \forall \tau > t\}$. This is the set of histories following h^t in which all realizations after t are lower than θ_0, the highest type.

A strategy for a seller consists of offering a direct mechanism $\langle \mathbf{p}, \mathbf{q} \rangle$ as described above. The strategy of a consumer is, at least potentially, contingent on a richer history $h_A^t = \left\{h_A^{t-1}, \theta^t, \widehat{\theta}^{t-1}\right\}$, where θ^t is the actual type every period and $\widehat{\theta}^t$ is the revealed type. Note that $h_A^0 = \theta^1$. For a given contract, a strategy for the consumer is simply a function that maps a history h_A^t into a revealed type: $h_A^t \mapsto s(h_A^t)$.

3 The first-order approach and the dynamic envelope formula

In this section we characterize the seller's problem and discuss the standard approach that has been used in the literature to solve it: the so called first-order approach. The seller's problem consists of choosing a contract $\langle \mathbf{p}, \mathbf{q} \rangle$ that maximizes profits under two sets of constraints: incentive compatibility constraints, which guarantee that an agent of type i does not want to report being a type j after any history h^t, and individual rationality constraints, which guarantee that all types expect to receive at least their reservation utility $\underline{U} = 0$ after any history h^t. Since the choice of prices and quantities corresponds to the choices of utilities and quantities for the buyer, this problem can be conveniently represented as a choice of $\langle \mathbf{U}, \mathbf{q} \rangle = \left(U\left(\theta^t \left| h^{t-1}\right.\right), q\left(\theta^t \left| h^{t-1}\right.\right)\right)_{t=1}^T$, where $U\left(\theta^t \left| h^{t-1}\right.\right)$ is the expected utility of a type θ^t after history h^{t-1}.

The general incentive compatibility constraint $IC_{i,j}(h^{t-1})$ requires $U(\theta_i | h^{t-1}) \geq U(\theta_j; \theta_i | h^{t-1})$, where $U(\theta_j; \theta_i | h^{t-1})$ is the expected utility of a type θ_i reporting to be a type θ_j at time t after history h^{t-1}. This constraint can be easily rewritten in terms of $\langle \mathbf{U}, \mathbf{q} \rangle$ as:

$$\begin{aligned} U(\theta_i | h^{t-1}) \quad \geq \quad & U(\theta_j | h^{t-1}) + \delta \sum_{k=0}^{N} (\alpha_{ik} - \alpha_{jk}) U(\theta_k | h^{t-1}, \theta_j) \\ & + u(\theta_i, q(\theta_j | h^{t-1})) - u(\theta_j, q(\theta_j | h^{t-1})). \end{aligned} \tag{1}$$

The individual rationality constraint for type i at history h^{t-1}, $IR_i(h^{t-1})$, is a simple non-negativity constraint:

$$U(\theta_i | h^{t-1}) \geq 0. \tag{2}$$

For future reference, we call *local downward constraints* the incentive constraints that are associated with a deviation to a contiguous lower type (i.e. $IC_{i,i+1}(h^{t-1})$), and *local upward constraints* the incentive constraints that are associated with a deviation to a contiguous higher type (i.e. $IC_{i+1,i}(h^{t-1})$). We refer to all the other constraints as *global*. A contract that satisfies all incentive and individual rationality constraints is said to be *implementable*.

Let $\mathbb{E}[S(\mathbf{q})]$ denote the expected value of $s(\theta, q)$ across time and types. The monopolist's problem is to maximize expected surplus net of the buyer's expected equilibrium rents:

$$\max_{\langle \mathbf{U}, \mathbf{q} \rangle} \left\{ \begin{array}{c} \mathbb{E}[S(\mathbf{q})] - \sum_{i=0}^{N} \mu_i U(\theta_i | h^0) \\ \text{s.t. } \mathbf{q} \geq 0 \text{ and } IR_i(h^{t-1}),\ IC_{i,j}(h^{t-1}) \\ \text{for any } i, j,\ t \text{ and } h^{t-1} \in H^{t-1}. \end{array} \right\} \tag{3}$$

This is a standard maximization problem of a concave function under a system of non-linear constraints. As T and N increase the number of variables and constraints becomes prohibitively large making (3) analytically intractable.

The typical approach in the literature is to first study a relaxed problem in which only individual rationality constraint of the lowest type and the local downward constraints $IC_{i,i+1}(h^t)$ are considered. The remaining constraints can be verified ex-post after the solution of the relaxed problem has been characterized.

Definition 1. *A contract is first-order optimal if and only if it maximizes profits under the following constraints:* $IR_N(h^{t-1})$ *and* $IC_{i,i+1}(h^{t-1})$, $\forall i \in \mathcal{N} \backslash \{N\}$, $\forall h^{t-1} \in H^{t-1}, \forall t$.

Interest in FO-optimal contracts is based on the fact that in many environments they coincide with the optimal contracts. Under standard assumptions, the FO-optimal contract coincides with the optimal contract in a static environment, both with finite and continuous type spaces (Stole [2001]).[8]

This approach has also been used in all papers that have extended the principal-agent model to dynamic environments: for example, the first-order autoregressive environment (Besanko [1985]) and the Markov environment with two types (Battaglini [2005]).[9] However, in the absence of generally applicable conditions (on primitives) for the validity of local incentive constraints as being sufficient for implementability, and arguably a limited understanding of the genericity of these hypothetical conditions if they existed; the applied literature often focuses on the FO-optimal contract and numerically checks for its optimality under a sample of global incentive constraints.[10]

It is easy to show that when we consider the relaxed problem with only local downward constraints, the incentive compatibility constraints can be assumed to hold as equalities.[11] This allows us to eliminate utilities from the optimization problem and drastically simplify the constraint set. Let us define:

[8] A sufficient condition for the FO-optimal contract to be optimal in a static environment is that the prior $\boldsymbol{\mu}$ satisfies the monotone hazard rate condition and $u_\theta(\theta, q)$ is not increasing in θ- conditions satisfied, for example, by a uniform prior and $u(\theta, q) = \theta q$. See Stole [2001] for discussion of these results.

[9] Another class of models in which the allocation takes place only in the final period, called sequential screening, has produced interesting cases in which the FO-optimal contract is indeed optimal. See Courty and Li [2000] and the discussion in Section 8.

[10] See Section 8 for a discussion of this literature.

[11] The details of the statements made in this section are formally proven in the appendix.

$$\Delta F\left(\theta_j\left|\theta_i\right.\right)=F\left(\theta_j\left|\theta_i\right.\right)-F\left(\theta_j\left|\theta_{i-1}\right.\right).$$

It denotes the effect on the conditional distribution of a marginal change in type in the previous period. It is important to note that first-order stochastic dominance implies $\Delta F\left(\theta_j\left|\theta_i\right.\right)\geq 0$, for all i and j. Recalling that $\widehat{H}\left(h^t\right)$ is the set of histories following h^t in which all realizations after t are lower than θ_0, and representing by h_k the kth element of history h, we have the following characterization of the agent's utility only as a function of $\mathbf{q}$:[12]

Lemma 1. *Corresponding to a FO-optimal contract, we have:*

$$\begin{aligned}\frac{U(\theta_i|h^{t-1})-U(\theta_{i+1}|h^{t-1})}{\Delta\theta} &= \frac{\int_{\theta_{i+1}}^{\theta_i}u_\theta(x,q(\theta_{i+1}|h^{t-1}))dx}{\Delta\theta}\\ &+\sum_{\hat{h}\in\widehat{H}(h^{t-1},\theta_{i+1})}\sum_{\tau>t}\delta^{\tau-t}\left[\begin{array}{c}\prod_{k=t+1}^{\tau}\Delta F\left(\hat{h}_k\left|\hat{h}_{k-1}\right.\right)\\ .\frac{\int_{\hat{h}_\tau}^{\hat{h}_\tau+\Delta\theta}u_\theta(x,q(\hat{h}_\tau|\hat{h}^{\tau-1}))dx}{\Delta\theta}\end{array}\right] \qquad (4)\end{aligned}$$

for any $i\in\mathcal{N}\backslash\{N\}$, $h^{t-1}\in H^{t-1}$ *and* $t=1,...,T$.

Lemma 1 presents a straightforward dynamic extension of the envelope formula introduced by Myerson [1981]. This can be seen by taking δ to zero, in which case the second term on the right hand side vanishes and (4) coincides with the classic static formula. The formula in (4) allows us to express the marginal rent of a type exclusively as a function of the allocation $\mathbf{q}$.[13] Although the formula is a complicated function of the conditional probabilities and the allocation, in specific environments it is quite tractable.

Example 1. When types are i.i.d. we have $f\left(\theta_i\left|\theta_j\right.\right)=f\left(\theta_i\left|\theta_k\right.\right)$ for all i,j,k, so for all histories $\Delta F\left(\hat{h}_k\left|\hat{h}_{k-1}\right.\right)=0$. It follows that $U(\theta_i|h^{t-1})-U(\theta_{i+1}|h^{t-1})=\int_{\theta_{i+1}}^{\theta_i}u_\theta(x,q(\theta_{i+1}|h^{t-1}))dx$. If we assume $u\left(\theta,q\right)=\theta q$, then $u_\theta(x,q(\theta_{i+1}|h^{t-1}))=q(\theta_{i+1}|h^{t-1})$. It follows that

$$\left[U(\theta_i|h^{t-1})-U(\theta_{i+1}|h^{t-1})\right]/\Delta\theta=q(\theta_{i+1}|h^{t-1}).$$

In particular this holds for the null history, h^0. Thus, the expected rent at $t=1$ depends only on quantities in the first period and is same as in the static model. The agent has no private information about future realizations beyond period 1 when the contract is signed. So he or she is unable to extract any rents for $t\geq 2$.

Example 2. Assume $u\left(\theta,q\right)=\theta q$ and, as in Baron and Besanko [1984], that types are constant, i.e. $f\left(\theta_i\left|\theta_i\right.\right)=1$ for all $i=0,...,N$. In this case, after $\left(h^{t-1},\theta_{i+1}\right)$, only history

[12] To interpret (4), note that, given a history $\hat{h}$, $\hat{h}^{\tau-1}=(\hat{h}_1,...,\hat{h}_{\tau-1})$ and $\hat{h}_\tau$ is the realization of the type at time τ. It follows that $q(\hat{h}_\tau|\hat{h}^{\tau-1})$ is the quantity at time τ when the realized history is $\hat{h}^{\tau-1}$.

[13] A continuous type version of the formula is presented in Baron and Besanko [1984] for the case in which $T=3$ and in Besanko [1985] for an infinite horizon model with first-order autoregressive types in which shocks have independent realizations. Battaglini [2005] states the formula for a Markov process with two states: (4) is a direct, but more involved extension of this result for the case with $|\Theta|\geq 2$. Pavan, Segal and Toikka [2014] present a general version of the formula for a continuous type space and other stochastic processes.

$\hat{h} = \{h^{t-1}, \theta_{i+1}..., \theta_{i+1}\}$ (in which the type remains equal to θ_{i+1}) has positive probability and $\Delta F(\theta_i|\theta_i) = 1$ for all i. Applying (4), it follows that:

$$\left[U(\theta_i|h^{t-1}) - U(\theta_{i+1}|h^{t-1})\right]/\Delta\theta = \sum_{\tau \geq t} \delta^{\tau-t} \cdot q(\hat{h}_\tau|\hat{h}^{\tau-1})$$

for all $i \in \mathcal{N}\backslash\{N\}$, where $\hat{h} \in H(h^{t-1}, \theta_{i+1})$ is the history that has all realizations following period t equal to θ_{i+1}. The expected rents are thus a discounted sum of quantities along the constant histories.

Example 3. Assume $u(\theta, q) = \theta q$ and two types, $\theta_0 = \theta_H$ and $\theta_1 = \theta_L$ that are imperfectly correlated. In this case all histories except the "lowest history" (in which all the types' realizations are always θ_L) disappear from (4). Given this, we obtain:

$$\left[U(\theta_H|h^{t-1}) - U(\theta_L|h^{t-1})\right]/\Delta\theta = \sum_{\tau \geq t} \delta^{\tau-t} \cdot \left[F(\theta_L|\theta_L) - F(\theta_L|\theta_H)\right]^{\tau-t} \cdot q(\widehat{h}_\tau|\widehat{h}^{\tau-1})),$$

where $\widehat{h} = \{h^{t-1}, \theta_L..., \theta_L\}$ is the history following h^{t-1} in which all realization after $t-1$ are θ_L. In this case the rent of the agent at $t = 1$ depends only on the quantities in the lowest history, in which the realizations are always θ_L. This is the envelope formula derived in Battaglini [2005].

Example 4. Another example that will prove useful in the remainder of the paper is when $T = 2$. Assuming the usual utility $u(\theta, q) = \theta q$, the rents at $t = 2$ are given by $U(\theta_i|h^1) - U(\theta_{i+1}|h^1)/\Delta\theta = q(\theta_{i+1}|h^1)$ and those at $t = 1$ by $\left[U(\theta_i) - U(\theta_{i+1})\right]/\Delta\theta = q(\theta_{i+1}) + \sum_{k=1}^{N} \delta\Delta F(\theta_k|\theta_{i+1}) \cdot q(\theta_k|\theta_{i+1})$.

Returning to the general model, we can express the utility vector solely as a function of $\mathbf{q}$ using Lemma 1. Define:

$$U^*(\theta_i|h^{t-1}; \mathbf{q}) = \sum_{n=1}^{N-i} \left[\begin{array}{c} \int_{\theta_{i+n}}^{\theta_{i+n-1}} u_\theta(x, q(\theta_{i+n}|h^{t-1}))dx \\ + \sum_{\hat{h} \in \widehat{H}(h^{t-1}, \theta_{i+n})} \sum_{\tau > t} \delta^{\tau-t} \left[\begin{array}{c} \prod_{k=t+1}^{\tau} \Delta F\left(\hat{h}_k \left| \hat{h}_{k-1}\right.\right) \\ \cdot \int_{\hat{h}_\tau}^{\hat{h}_\tau + \Delta\theta} u_\theta(x, q(\hat{h}_\tau|\hat{h}^{\tau-1}))dx \end{array} \right] \end{array} \right] \quad (5)$$

for any $i < N$, and $U^*(\theta_N|h^{t-1}; \mathbf{q}) = 0$. Corollary 1, thus, immediately follows from (4):

Corollary 1. *Corresponding to a FO-optimal contract, we have* $U(\theta_i|h^{t-1}) = U^*(\theta_i|h^{t-1}; \mathbf{q})$ *for any* $i \in \mathcal{N}$, $h^{t-1} \in H^{t-1}, \forall t$.

The FO-optimal contract can now be characterized as the solution of the following program:

$$\max_{\mathbf{q} \geq 0} \left\{ \mathbb{E}\left[S(\mathbf{q})\right] - \sum_{i=0}^{N} \mu_i U^*(\theta_i|h^0; \mathbf{q}) \right\} \quad (6)$$

This problem can be solved to obtain the closed form solution. Let $D(h^t)$ be equal to 1 at $t=1$, and for $t>1$, define:

$$D(h^t)=\begin{cases} 0 & \text{if } h^t_\tau=\theta_0 \text{ for any } \tau\leq t \\ \prod_{\tau=1}^{t-1}\left(\frac{\Delta F(h^t_{\tau+1}|h^t_\tau)}{f(h^t_{\tau+1}|h^t_\tau)}\right) & \text{else}\end{cases} \tag{7}$$

These are the dynamic distortions associated with the FO-optimal contract.[14] Recall that for any θ_i, $s(\theta_i,q)$ is the per period surplus (i.e. $u(\theta_i,q)-c(q)$), and $s_q(\theta_i,q)$ its derivative with respect to q. From the first-order necessary conditions of (6) we can easily characterize the FO-optimal contract as follows.[15]

Proposition 1. *Corresponding to a FO-optimal contract we have:*

$$s_q(\theta_i,q^*\left(\theta_i|h^{t-1}\right))\leq\frac{1-\sum_{k=j}^N\mu_k}{\mu_j}\cdot D(h^{t-1},\theta_i)\cdot\int_{\theta_i}^{\theta_{i-1}}u_{\theta q}(x,q(\theta_i|h^{t-1}))dx \tag{8}$$

for any $i\in\mathcal{N}$, $h^{t-1}\in H^{t-1}$ *and* t, *where* $\theta_j=h^t_1$, *and the above is satisfied with equality if* $q^*\left(\theta_i|h^{t-1}\right)>0$.

It is customary in the literature to assume that the objective function in (6) is concave (see Stole [2001] for example): in this case (8) is necessary and sufficient and so it uniquely defines the FO-optimal contract. Although this assumption is not required for the following results, it is always verified if we assume preferences a' la Mussa and Rosen [1978], when $u(\theta,q)=\theta q$ and $c(q)=(1/2)q^2$. In this case, at an interior solution, we have:

$$q^*\left(\theta_i|h^{t-1}\right)=\theta_i-\frac{1-\sum_{k=j}^N\mu_k}{\mu_j}D(h^{t-1},\theta_i)\Delta\theta \tag{9}$$

where $\theta_j=h^t_1$.

We can now apply (8) to the examples discussed above.

Example 1 (cont.). From (7) and (8) we can see that when types are i.i.d., it is optimal to offer the optimal static contract in the first period and the efficient contract in all following periods since the quantities offered after $t=1$ do not affect rents. For the standard model, we have $q^*(\theta_i)=\theta_i-\frac{1-\sum_{k=i}^N\mu_k}{\mu_i}\Delta\theta$ in the first period and $q^*\left(\theta_i|h^{t-1}\right)=\theta_i$ in the following periods.

Example 2 (cont.). From (9), it follows that when types are constant it is optimal to offer the same quantities $q^*(\theta_i)=\theta_i-\frac{1-\sum_{k=i}^N\mu_k}{\mu_i}\Delta\theta$ in all periods, irrespective of the history of types' realizations. To see this, note that on histories in which types remain constant we have $D(h^{t-1},\theta_i)=1$, so (9) is equal to $\theta_i-\frac{1-\sum_{k=i}^N\mu_k}{\mu_i}\Delta\theta$. On histories in which types are not constant, any quantity is optimal. Since these quantities neither affect the surplus nor the rents of the agent

[14] In some parts of the literature dynamic distortions have been referred to as an impulse response.

[15] Note that, in the following expression, $D(h^{t-1},\theta_i)$ corresponds to $D(h^t)$ for $h^t=\left\{h^{t-1},\theta_i\right\}$. Also, θ_{-1} is any dummy type.

they do not enter the objective function, (6).[16] The quantity $q^*(\theta_i)$ is equal to the optimal quantity that would be offered in a static model with $T = 1$. This observation was first made by Baron and Besanko [1984]. For future reference, note that this is only one of the possible solutions.

Example 3 (cont.). With two types, (9) implies that $q^*\left(\theta_i|h^{t-1}\right) = \theta_i$ if $\theta_i = \theta_H$ and/or θ_H is a realization in h^{t-1}. For the remaining history, $\widetilde{h}^{t-1}$, in which the type is always θ_L, we have $q^*\left(\theta_L|\widetilde{h}^{t-1}\right) = \theta_L - \frac{\mu_H}{\mu_L}\left(\frac{F(\theta_L|\theta_L)-F(\theta_L|\theta_H)}{F(\theta_L|\theta_L)}\right)^{t-1}\Delta\theta$. In this case the FO-optimal contract is efficient for all histories except the lowest in which the type is θ_L. Along the lowest history in which quantities are distorted, the distortion is proportional to $\left(\frac{F(\theta_L|\theta_L)-F(\theta_L|\theta_H)}{F(\theta_L|\theta_L)}\right)^{t-1}$, which is less than 1, and so it vanishes as $t \to \infty$.

Example 4 (cont.). In the first period, we have $q^*(\theta_i) = \theta_i - \frac{1-\sum_{k=i}^{N}\mu_k}{\mu_i}\Delta\theta$, as in the static model, and in period 2, $q^*(\theta_i|\theta_j) = \theta_i - \frac{1-\sum_{k=j}^{N}\mu_k}{\mu_j}\frac{F(\theta_i|\theta_j)-F(\theta_i|\theta_{j-1})}{f(\theta_i|\theta_j)}\Delta\theta$.

Some distinct characteristics easily emerge from (8) even without assuming that it admits a unique solution. Since the right hand side of (8) is non-negative, the contract is always distorted downward, at least weakly: so, analogous to the static case, we never have overprovision, but we can have underprovision. Moreover, the right hand side becomes zero when the type becomes θ_0, the highest type (since $D(h^{t-1},\theta_i) = 0$). In this case, $s_q(\theta_i, q^*\left(\theta_i|h^{t-1}\right)) = 0$ and the contract is efficient in all following periods, a phenomenon that has been called "Generalized No-Distortion at the Top". For any other history, the quantities are distorted strictly below the efficient level. The distortion is exactly equal to $\left[\sum_{k=0}^{j-1}\mu_k/\mu_i\right]D(h^{t-1},\theta_i)\Delta\theta$: this formula is complicated because the wedge is state contingent and it depends on the entire history.

4 When does the first-order approach work?

Given the (relatively) simple characterization of Proposition 1, the imperative question is: when is it without loss of generality to focus on the first-order approach? From previous work we know that there are a number of cases in which the first-order approach works where the optimal contract coincides with (8). Should these be seen as the standard, or do they constitute a specialized class of models? In the remainder of this section we explore when the first-order approach is valid.

To verify the validity of the FO-approach we need to establish that the solution of (6) satisfies the full set of constraints in (3). Corollary 1 tells us that the agent's rents are functions only of the quantities, so the set of constraints also depends only on q. We let $C(\mathbf{q})$ denote this set of constraints. The first-order optimal contract is defined by $\mathbf{q}(\Theta,\mu,F)$, function only of the fundamentals. It follows that a necessary and sufficient condition for the validity of the first-order approach is that the set of fundamentals satisfy the family of inequalities defined by $C(\mathbf{q}(\Theta,\mu,F))$.

[16] In the rest of the paper we assume that types have full support so (7) is always well defined. With perfect persistence, for histories in which types change, $D(h^t)$ is indeterminate: in this cases both the numerator and the denominator of $D(h^t)$ are zero. These histories occur with zero probability, so the associated quantities are irrelevant.

The key question is whether these conditions define reasonably interesting economic environments for which the FO-approach works. The following result provides a unified framework to interpret existing "possibility results" for the FO-approach. Let $q(h^t) = q\left(h_t^t|h^{t-1}\right)$ be an allocation after history h^t, and let $h^t \succeq \widehat{h}^t$ if $h_j^t \geq \widehat{h}_j^t \ \forall j \leq t$. We have:

Definition 2. *An allocation is monotonic if* $q(h^t) \geq q(\widehat{h}^t)$ *for any* $h^t \succeq \widehat{h}^t$.

A simple sufficient condition for the validity of the first-order approach can now be stated.[17]

Proposition 2. *The envelope formula (5) and monotonicity of the FO-optimal contract are sufficient for implementability.*

Proposition 2 directly parallels the well known results in static environments that show that local incentive compatibility (i.e. the envelope formula) and monotonicity of the allocation are necessary and sufficient for implementability. The result is however weaker for two reasons: first the monotonicity condition is stronger than in a static environment, since it involves all histories following a report; second, the result is only sufficient. The are a number of applications in which the FO-optimal contract is indeed monotonic.

Example 1 and 2 (cont.). When types are i.i.d., the contract is history independent and monotonic in all periods $t > 1$ (since it coincides with the efficient allocation). The contract is also monotonic in the type at $t = 1$ if the optimal static contract is monotonic: this is always the case if, for example, the prior satisfies the monotone hazard rate condition and $u_{q\theta\theta} \leq 0$, a condition satisfied, for example, by a uniform prior and $u(\theta, q) = \theta q$. When types are constant, the repetition of the first-order optimal static contract is a FO-optimal contract (although it is not unique), which is monotonic if the first-order optimal static contract is monotonic. It follows that under standard assumptions that guarantee monotonicity in θ of the first-order optimal static contract, FO-optimal contract is an optimal dynamic contract both when types are constant and when they are i.i.d .

Example 3 (cont.). As discussed in the previous section, with two types the FO-optimal contract is efficient in all histories except the history in which types' realizations are all θ_L. This history is also the "lowest history" according to the order $\succeq$. It follows that the contract is monotonic according to Definition 1, and so the FO-optimal contract is optimal.[18]

Example 5: AR(1) model. Besanko [1985] and more recently Pavan, Segal and Toikka [2014] assume an $AR(1)$ model in which $\theta^t = \gamma\theta^{t-1} + \varepsilon_t$, where ε_t is the realization of an i.i.d. random variable and $\gamma \in (0, 1)$. The Markovian framework developed above can be easily adapted to generalize this environment to non i.i.d. shocks. Here we present a two period model to drive

[17] This result generalizes, to an environment with N types, the method used in Battaglini [2005] to establish the sufficiency of (5) for $N = 2$. The proof of Claim 2 in Battaglini [2005] employs a weaker monotonicity condition: the marginal of expected utilities are non-decreasing in the current type, and shows that it is sufficient for implementability. Analogous monotonicity results for continuous types and more general stochastic processes are presented by Pavan, Segal and Toikka [2014].

[18] Boleslavsky and Said [2013] generalize this two type model assuming continuous types at $t = 0$ with binary shocks in the following periods. They also consider a version with a continuum of shocks, but in this case they directly assume that the quantities are monotonic in the reported values.

home the point in a simple fashion. In both periods, the "shocks" have support $\theta_0, ...\theta_N$, with $\theta_k - \theta_{k+1} = \Delta\theta$; in the first period the realization is $\theta^1 = \theta_i$ with prior probability μ_i, in the second period $\varepsilon_2 = \theta_j$ with probability α_{ij} when $\theta^1 = \theta_i$. When $\alpha_{ij} = \alpha_{kj}$ for any i, j, k, we have i.i.d. shocks and the model is equivalent to the models presented in Besanko [1985] and Pavan, Segal and Toikka [2014].

When we consider only local incentive constraints, it is easy to show that they must hold as equalities. In period 2, we have $U(\theta_k^2|\theta_i) = U(\theta_{k+1}^2|\theta_i) + \Delta\theta q(\theta_{k+1}^2|\theta_i)$, where $U(\theta_k^2|\theta_i)$ and $q(\theta_k^2|\theta_i)$ are respectively the second period utility and quantity of the agent when the realization at $t = 1$ and $t = 2$ are θ_i and θ_k^2.[19] Without loss of generality we can set $U(\theta_N^2|\theta_i) = 0$, so that $U(\theta_k^2|\theta_i) = \Delta\theta \sum_{l=k+1}^{N} q(\theta_l^2|\theta_i)$. Similarly, in the first period, we have:

$$\begin{aligned} U(\theta_i) &= U(\theta_{i+1}) + \left[\Delta\theta q(\theta_{i+1}|h^0) + \delta\gamma\Delta\theta \sum_{k=0}^{N} \alpha_{ik} q(\theta_k^2|\theta_{i+1})\right] \\ &\quad + \delta \sum_{k=0}^{N} \left(\alpha_{ik} - \alpha_{(i+1)k}\right) U(\theta_k^2|\theta_{i+1}) \end{aligned} \tag{10}$$

The expected utility of the agent with type θ_i is equal to the utility of type θ_{i+1} plus an *informational rent*. The informational rent can be decomposed into two parts. First, we have a deterministic part $\Delta\theta q(\theta_{i+1}|h^0) + \delta\gamma\Delta\theta \sum_{k=0}^{N} \alpha_{ik} q(\theta_k^2|\theta_{i+1})$: type at time 1 affects rents at $t = 1$ (i.e., $\Delta\theta q(\theta_{i+1}|h^0)$), but it effects rents at $t = 2$ as well, in a way that is proportional to γ (i.e., $\delta\gamma\Delta\theta \sum_{k=0}^{N} \alpha_{ik} q(\theta_k^2|\theta_{i+1})$). Second, we have the stochastic part $\delta \sum_{k=0}^{N} \left(\alpha_{ik} - \alpha_{(i+1)k}\right) U(\theta_k^2|\theta_{i+1})$. This term depends only on the fact that types i and $i + 1$ have different expectations on the probability of the shock ε_t at $t = 2$. With i.i.d. shocks the stochastic term of the information rent is zero. The distortions are then *exclusively* deterministic. The first-order optimal quantities are given by $\theta_1 - \Delta\theta \left(1 - \sum_{k=i}^{N} \mu_k\right) / \mu_i$, in period 1, and $\theta_2 - \gamma\Delta\theta \left(1 - \sum_{k=i}^{N} \mu_k\right) / \mu_i$ in period 2 (when θ_i is the realization in the first period). It is immediate to observe that, under a monotone hazard rate assumption on the prior, quantities are monotonic in the sense of Definition 2. It follows from Proposition 2 that the first-order approach works and these quantities describe the optimal contract.

Example 6: AR(k) model and its variations. When the shock ε_t is i.i.d. we can easily generalize the analysis to T periods following the same steps as in Example 5. In this case we can verify that the quantity at time t is

$$\theta^t - \frac{1 - \sum_{k=i}^{N} \mu_k}{\mu_i} \cdot \gamma^{t-1}\Delta\theta \tag{11}$$

when θ_i is the realization in the first period. Distortions, therefore, are history dependent- they depend only on time through γ^{t-1}. Note that θ^t is the first best efficient quantity: the optimal contract is characterized by deterministic distortions that are independent of the Markov process

[19] So, with realization θ_i in period 1, we have that $\theta_k^2 \in \{\gamma\theta_i + \theta_0, .., \gamma\theta_i + \theta_N\}$.

governing the evolution of types. Given this, it is easy to extend the analysis to the k-order autocorrelation case: $\theta^t = \sum_{j=0}^{k} \gamma_j \theta^{t-j} + \varepsilon_t$. The examples can also be extended to a non-linear case in which $\theta^t = l_1(h^{t-1}, \mathbf{q^{t-1}}) + l_2(h^{t-1}, \mathbf{q^{t-1}})\varepsilon_t$, where $l_i(h^{t-1}, \mathbf{q^{t-1}})$ $i = 1, 2$ are both functions of the sequences of types and quantities up to $t-1$. To see this point, note that at time t the terms $\sum_{j=0}^{k} \gamma_j \theta^{t-j}$ or $l_1(h^{t-1}, q^{\mathbf{t-1}})$ are just constants for all types, so they do not have any effect on incentives to reveal the true type and $l_2(h^{t-1}, \mathbf{q^{t-1}})$ disappears since ε_t is an i.i.d. random variable. In all these cases, the key assumption is that the shock is an independent linear addition to the agent's type.[20]

The examples presented above show that the first-order approach can be extended to study quite complex dynamic environments. All the examples, however, can be reconducted to two basic assumptions. The environment studied in Besanko [1985] allows for many possible types (in fact a continuum), but assumes that types change because of linearly additive stochastic shocks uncorrelated with the agent's type. In this environment the shocks are irrelevant for the equilibrium distortions, which are independent of the history of realized types (except for the first).[21] The environment studied in Battaglini [2005] allows the conditional distributions of the types to depend on the type, but limits the analysis to two types only. In this case the optimal contract is history dependent. These two environments have a common feature: in all these cases the FO-optimal allocation is monotonic. In the next section, however, we show that this is not a general property of FO-optimal contracts.

5 The limits of the first-order approach

In static environments monotonicity only requires that the quantity is non-decreasing in the type. This condition is satisfied under standard regularity conditions. Notably, a sufficient condition for the monotonicity of the optimal contract is that the prior satisfies the monotone hazard rate condition and that $u_{\theta q}$ is non-decreasing in θ. The examples in the previous section may suggest that monotonicity of the optimal contract is a feature of dynamic contracts as well. Dynamic environments, however, are different and monotonicity should not be expected even in the simplest examples.

To see this, consider an example with two periods ($T = 2$), Mussa and Rosen [1978] preferences, three types: $\theta_0 = \theta_H$, $\theta_1 = \theta_M$ and $\theta_2 = \theta_L$ with $\theta_H - \theta_M = \theta_M - \theta_L = \Delta\theta > 0$, and transition probabilities $f(\theta_i \,|\theta_i\,) = \alpha$, $f(\theta_i \,|\theta_j\,) = (1-\alpha)/2$ for $i \neq j$. These simple transition probabilities satisfy first-order stochastic dominance, so they preserve "order" in the stochastic evolution of types. From Example 4 in Section 3 we have:

$$\begin{aligned} q^*\left(\theta_M \,|\theta_M\,\right) &= \theta_M - \frac{1-\sum_{k=L,M}\mu_k}{\mu_M}\frac{F(\theta_M|\theta_M)-F(\theta_M|\theta_H)}{f(\theta_M|\theta_M)}\Delta\theta \\ &= \theta_M - \frac{\mu_H}{\mu_M}\frac{3\alpha-1}{2\alpha}\Delta\theta < \theta_M. \end{aligned}$$

[20] Multiplicative independent shocks also share a similar structure, see Coutry and Li [2000].

[21] We will return on the importance of these assumptions for the first-order approach to work in Section 5.1.

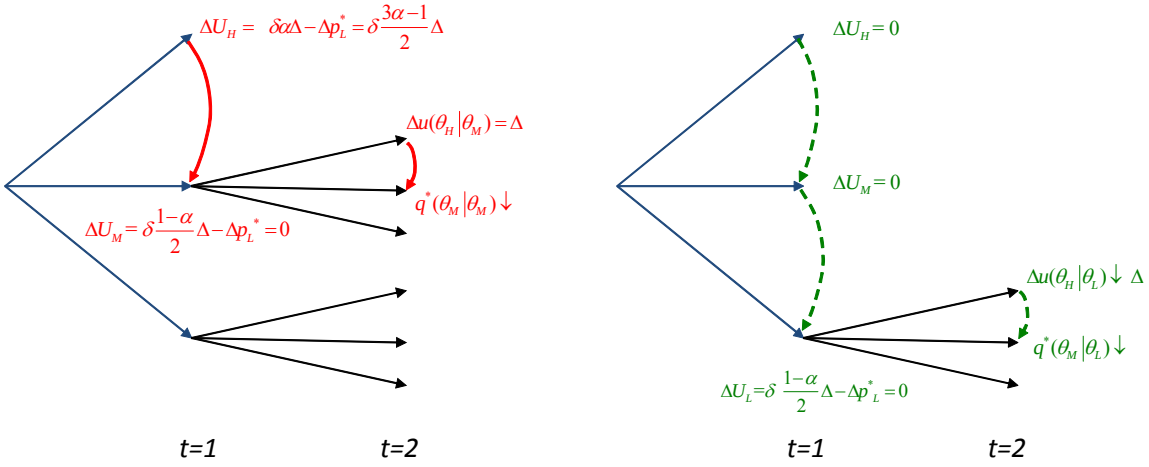

Figure 1: Dynamic screening and optimal non-monotonic allocations.

Monotonicity would require $q^*(\theta_M|\theta_M) > q^*(\theta_M|\theta_L)$. On the contrary, we have:

$$q^*(\theta_M|\theta_L) = \theta_M - \frac{1-\mu_L}{\mu_L}\frac{F(\theta_M|\theta_L)-F(\theta_M|\theta_M)}{f(\theta_M|\theta_L)}\Delta\theta = \theta_M,$$

since $F(\theta_M|\theta_M) = F(\theta_M|\theta_L) = \alpha + \frac{1-\alpha}{2}$. So $q^*(\theta_M|\theta_L) > q^*(\theta_M|\theta_M)$ and the FO-optimal contract is not monotonic with respect to the realization at $t=1$.

To understand why the seller finds it first-order optimal to offer a non-monotonic contract in this example, consider the role of distortions in the screening problem. The planner distorts quantities at $t=2$ in order to reduce the agent's rent at $t=1$. By decreasing $q^*(\theta_M|\theta_L)$, the planner reduces the rent of type H at $t=2$ by say Δ; this in turn reduces the expected utility of L at $t=1$ by $\delta(1-\alpha)/2\cdot\Delta$ (see the right panel of Figure 1). This reduction, however, does not benefit the monopolist directly. Since type L at $t=1$ is receiving the reservation utility: any reduction in rents at $t=2$ must be compensated by an equivalent increase of rents at $t=1$, otherwise type L's reservation utility would be violated. This compensation is achieved by decreasing p_L^*, the price paid by L at $t=1$, by exactly $\delta(1-\alpha)/2\cdot\Delta$. The change in $q^*(\theta_M|\theta_L)$, therefore, reduces surplus and leaves the rents extracted by L untouched. Reducing $q^*(\theta_M|\theta_L)$ could be beneficial only if it helps reducing some other type's rent. The change in $q^*(\theta_M|\theta_L)$ changes the outside option of H and M types when they consider misreporting to be an L type in period 1. The first effect is ignored by assumption in a first-order optimal contract since only local deviations are considered, and the second effect in this example is exactly zero. To see this, note that if the M type reports to be a L type, he benefits from a reduction in p_L^* by $(1-\alpha)/2\cdot\Delta$ at $t=1$ and he suffers a reduction in expected rent if he becomes a H type at $t=2$. The expected value of this loss however is the same as the expected value for L, $(1-\alpha)/2\cdot\Delta$, since both L and M types become a H type at $t=2$ with the same probability $(1-\alpha)/2$. Reducing $q^*(\theta_M|\theta_L)$ therefore does not help the monopolist: it makes surplus smaller and it has no apparent benefit

(assuming that only local incentive constraints are binding).

The same can not be said of a reduction in $q^*(\theta_M|\theta_M)$ (see the left panel of Figure 1) As before, this reduction has no direct effect on the rents of type M (or L),[22] it does however reduce the rent of type H both at time $t=1$ and $t=2$. By standard arguments, reducing $q^*(\theta_M|\theta_M)$ induces a reduction at $t=2$ of type H's rent by, say, Δ. In addition, type H is more likely to remain H than a M type is to become H. This implies that the net expected rent of a high type at $t=1$ is reduced by $(3\alpha-1)/2\cdot\Delta>0$.[23] Distorting $q^*(\theta_M|\theta_M)$ downward now yields a stritly positive payoff: this is the reason why $q^*(\theta_M|\theta_L)>q^*(\theta_M|\theta_M)$.

Is the failure of monotonicity a general phenomenon? In the example presented above we have assumed a particular transition function $f_\alpha(\theta_j|\theta_i)$ in which the probability of persistence is the same for all types ($f_\alpha(\theta_i|\theta_i)=\alpha$) and all deviations are equally likely ($f_\alpha(\theta_j|\theta_i)=(1-\alpha)/N$ for $i\neq j$). Does the phenomenon illustrated by the example extend to general transition functions? To address this question consider the general set Λ of all possible transition functions $f_\alpha(\theta_j|\theta_i)$ satisfying Assumption 2 and parametrized by $\alpha\in[0,1]$ such that for all i, $f_\alpha(\theta_i|\theta_i)\to 1$ as $\alpha\to 1$.[24]

Definition 3. *We say that a property holds for a* generic *transition function in* Λ *if and only if it holds for an open and dense set of functions in* Λ.

This is the standard definition of genericity in this environment.[25] Our first result proves that, for a generic transition function, the optimal contract is non-monotonic when types are sufficiently persistent.

Proposition 3. *For any* μ, δ, $|\Theta|>2$, $T>2$, *and a generic transition function in* Λ, *there exists an* $\alpha^*<1$ *such that the FO-optimal contract is not monotonic for any* $\alpha>\alpha^*$.

To grasp the intuition behind this result, consider the FO-optimal contract when $u(\theta,q)=\theta q$. Consider the two histories $h_i=\{\theta_i,\theta_i\}$ and $h_{i+1}=\{\theta_{i+1},\theta_{i+1}\}$ where $i\in(0,N)$, i.e. neither the highest nor the lowest type. From the general formula presented in (9), we have:

$$\begin{aligned} q^*(\theta_i|h_i) &= \theta_i-\tfrac{1-\sum_{k=i}^N\mu_k}{\mu_i}D(h^{t-1},\theta_i)\Delta\theta \\ &= \theta_i-\tfrac{1-\sum_{k=i}^N\mu_k}{\mu_i}\left[\tfrac{\Delta F_\alpha(\theta_i|\theta_i)}{f_\alpha(\theta_i|\theta_i)}\cdot\tfrac{\Delta F_\alpha(\theta_i|\theta_i)}{f_\alpha(\theta_i|\theta_i)}\right]\Delta\theta \end{aligned} \tag{12}$$

It is easy to verify that, as types become persistent, the term in the square parenthesis converges to one: so $q^*(\theta_i|h_i)$, converges to $\theta_i+\left(1-\sum_{k=i}^N\mu_k\right)/\mu_i\cdot\Delta\theta$, the optimal static contract for type

[22] Type M's rent at $t=1$ is equal to his outside option, that is the utility of reporting to be a L type. This rent, in turn, depends only on the quantities that follow a L report at $t=1$. Type L's rent is equal to the reservation utility.

[23] The expected change in H's rent at $t=1$ is $\alpha\Delta-\Delta p_M^*$ where $\Delta p_M^*=(1-\alpha)/2\cdot\Delta$ to keep M's rent constant.

[24] Note that since we impose no additional restrictions, the probabilities of persistence of different types may be different for $\alpha<1$ and they can even converge to one at different speeds: α is just an index of the level of persistence of the stochastic process.

[25] Endowed with a sup norm, the space of transition functions Λ is a complete metric space. The complement of an open and dense set in Λ is a set of first category. The Baire Category Theorem guarantees that these sets have empty interior and therefore are topologically small (Royden [1988], ch.7.8).

θ_i. By a similar argument, since this conclusion does not depend on the specific i, we also have that, as types become persistent, $q^*(\theta_{i+1}|h_{i+1})$ converges to $\theta_{i+1} + \left(1 - \sum_{k=i+1}^{N} \mu_k\right)/\mu_{i+1} \cdot \Delta\theta$.

What happens after a "mixed" history $h_i' = \{\theta_i, \theta_{i+1}\}$? Again (9) tells us that:

$$q^*(\theta_i|h_i') = \theta_i - \frac{1-\sum_{k=i}^{N}\mu_k}{\mu_i}\left[\frac{\Delta F_\alpha(\theta_i|\theta_{i+1})}{f_\alpha(\theta_i|\theta_{i+1})} \cdot \frac{\Delta F_\alpha(\theta_{i+1}|\theta_i)}{f_\alpha(\theta_{i+1}|\theta_i)}\right]\Delta\theta \tag{13}$$

The only difference between (12) and (13) is in the terms in the square parenthesis. Now, as types become persistent, it is not any more the case that the square parenthesis in (13) converges to 1. In the example presented at the beginning of this section (in which there are 3 types and $f(\theta_i|\theta_i) = \alpha$ and $f(\theta_j|\theta_i) = (1-\alpha)/2$), we have $\Delta F_\alpha(\theta_M|\theta_L) = 0$ and $\Delta F_\alpha(\theta_L|\theta_M) = 0$ for any α, so the term in the square parenthesis is zero for $i = M$ and $q^*(\theta_i|h_i') = \theta_i$. For a generic process, however,

$$\frac{\Delta F_\alpha(\theta_i|\theta_{i+1})}{f_\alpha(\theta_i|\theta_{i+1})} \cdot \frac{\Delta F_\alpha(\theta_{i+1}|\theta_i)}{f_\alpha(\theta_{i+1}|\theta_i)} \tag{14}$$

can either be larger or smaller than one (as formally proven in the appendix, the case in which it is *exactly* one is a non-generic knife-hedge case).

When it is smaller than one, we have $q^*(\theta_i|h_i') > q^*(\theta_i|h_i)$, proving the failure of monotonicity. And when the term is larger than one? If this is the case, consider a history $h_i'' = \{\theta_{i+1}, \theta_i\}$ in which the type realizations are switched. Applying again general formula (9), we have:

$$\begin{aligned} q^*(\theta_{i+1}|h_i'') &= \theta_{i+1} - \frac{1-\sum_{k=i+1}^{N}\mu_k}{\mu_{i+1}}\left[\frac{\Delta F(\theta_{i+1}|\theta_i)}{f(\theta_{i+1}|\theta_i)} \cdot \frac{\Delta F(\theta_i|\theta_{i+1})}{f(\theta_i|\theta_{i+1})}\right]\Delta\theta \\ &= \theta_{i+1} - \frac{1-\sum_{k=i+1}^{N}\mu_k}{\mu_{i+1}}\left[\frac{\Delta F(\theta_i|\theta_{i+1})}{f(\theta_i|\theta_{i+1})} \cdot \frac{\Delta F(\theta_{i+1}|\theta_i)}{f(\theta_{i+1}|\theta_i)}\right]\Delta\theta \\ &< \theta_{i+1} - \frac{1-\sum_{k=i+1}^{N}\mu_k}{\mu_{i+1}}\Delta\theta \end{aligned}$$

where the second equality follows by switching the terms in the square parenthesis, the final inequality follows from the assumption that (14) is larger than one. Since $q^*(\theta_{i+1}|h_{i+1})$ converges to the optimal static contract $\theta_{i+1} + \frac{1-\sum_{k=i+1}^{N}\mu_k}{\mu_{i+1}}$, it follows that $q^*(\theta_{i+1}|h_{i+1}) > q^*(\theta_{i+1}|h_i'')$ when types are sufficiently persistent. Since $h_i'' \succ h_{i+1}$, we end up with a non-monotonic contract, no matter how we pick the distribution of types.

Does Proposition 3 imply that the FO-approach generically fails when types are highly persistent? It is easy to see that a failure of monotonicity can not alone be sufficient for the first-order approach to fail. When δ is small, the future becomes irrelevant and the problem is essentially static. What happens when the future is sufficiently important? In the next result we use Proposition 3 to show that when types are highly persistent and the future is sufficiently important, then even a small failure of monotonicity is sufficient to make the FO-approach invalid. We have:

Proposition 4. *For any μ, $|\Theta| > 2$ and a generic transitionfunction in Λ, there exists an $\alpha^* < 1$, $T^* > 2$, and $\delta^* < 1$ such that the first-order approach fails to be verified for any $\alpha > \alpha^*$, $T \geq T^*$ and $\delta > \delta^*$.*

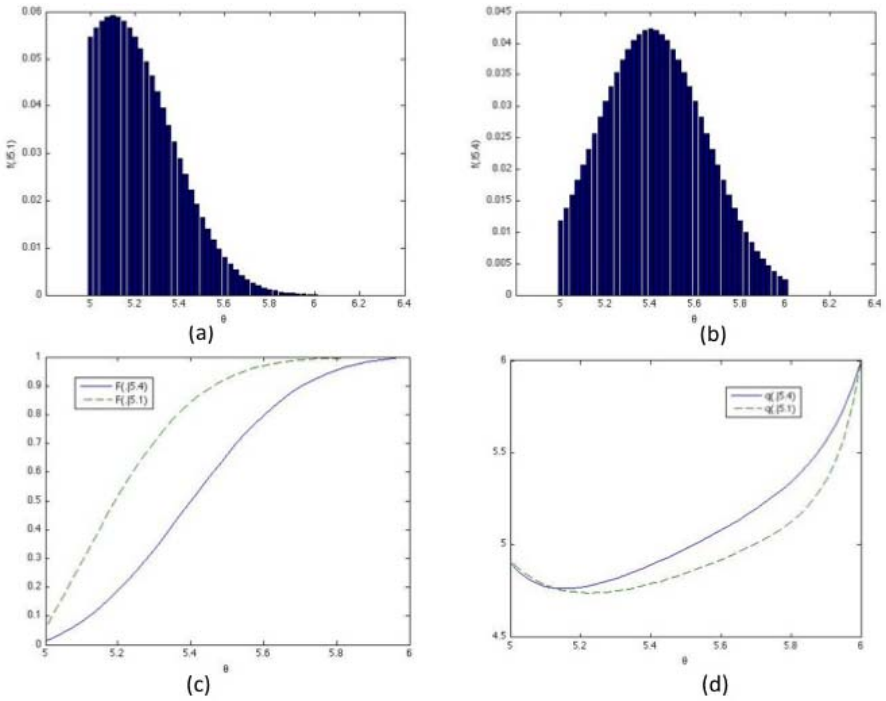

Figure 2: Example 7, the truncated normal case. $\sigma = 0.25, \Delta\theta = 0.025$.

Proposition 4 shows that with sufficiently high persistence in types and a sufficiently long horizon, the effect of even a small failure of monotonicity is highly magnified as $\delta \to 1$, to the point of inducing a violation of global incentive constraints.

To see the intuition behind this result, consider again the example presented at the beginning of this section. As we said, $q^*(\theta_M|\theta_L)$ is left undistorted in the FO-optimal contract because its reduction affects neither type L's rents nor the incentives for M to falsely report to be a type L. The incentive for type H to report to be a type L, however, are ignored in the FO-optimal contract and this is a problem. When we consider type H's incentives to report L, there are two effects. As in static models, type H's rent is equal to the utility received by reporting to be a type M. The first effect of reporting L is that at $t = 1$ type H is offered a quantity that is lower than the quantity offered by reporting M: this make the deviation to L less appealing. In the second period, however, $q^*(\theta_M|\theta_L) > q^*(\theta_M|\theta_M)$ implies that H receives a higher utility by reporting L rather than M at $t = 2$: this makes the deviation to L more appealing. When the discount factor is sufficiently small and/or types are not sufficiently correlated, the first effect dominates and global incentive compatibility constraints can be ignored. However, with sufficient correlation in types and a reasonably large continuation utility, the second effect dominates. As we formally show in the proof of Proposition 4, with more than two periods the seller not only finds it optimal to have a smaller distortion at $t = 2$ after type L than after type M at $t = 1$, she

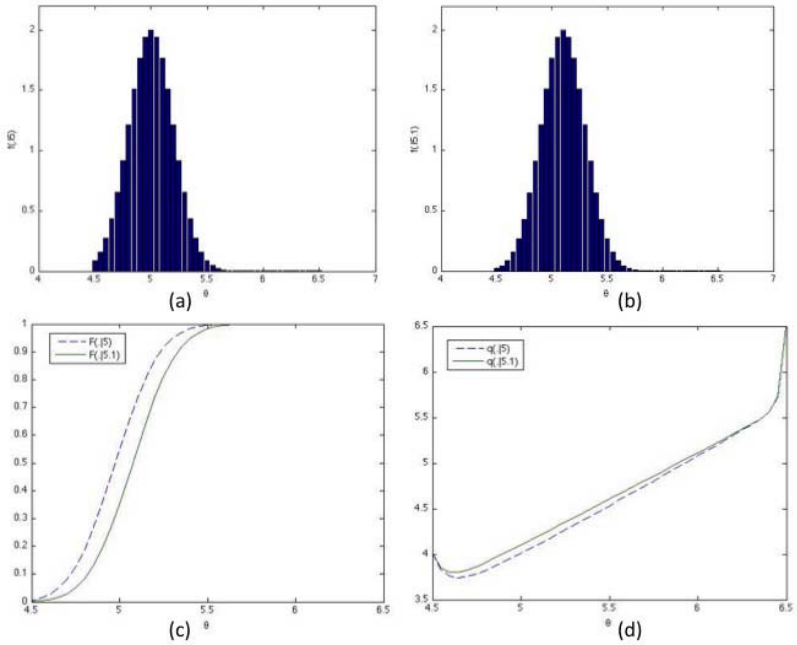

Figure 3: Example 8, the exponential case. $\alpha = 2, \Delta\theta = 0.05$.

also finds it optimal to have smaller expected distortions for all periods following L. This implies that by reporting L, type H suffers a loss at $t = 1$, but gains in all the following periods. When T and δ are sufficiently high, the dynamic effect on future expected utility dominates the static effect in the first period.

The fact that the first-order approach fails only if types are sufficiently persistent and expected payoffs are sufficiently important should not be surprising. As we have seen in Example 1, the first-order approach always works when types are sufficiently serially uncorrelated: in this case types have small private information about the future, so there is no point in imposing distortions on future quantities. Similarly, if agents are impatient and the time horizon is short, the model is close to being static.

Interestingly, it is easy to compute standard examples in which very limited serial correlation is sufficient to induce a failure of the first-order approach. To illustrate this point, we conclude this section with three such examples.

Examples 7. Assume that the type in the first period, θ_1, is uniformly distributed on $\Theta = [5, 6]$ and the distribution in the second period is a (truncated) normal $f_\alpha(\theta_2 | \theta_1) = \frac{A(\theta_1)}{\sigma}\Phi(\frac{\theta_2 - \theta_1}{\sigma})\Delta\theta$ where Φ is a standard normal density with variance σ and $A(\theta_1)$ is chosen so that the distribution assigns probability one on Θ.[26] (The specific values chosen for the support are obviously irrelevant

[26] To obtain a discrete density that can be applied to any size $\Delta\theta$ (even arbitrarily small), we discretize a

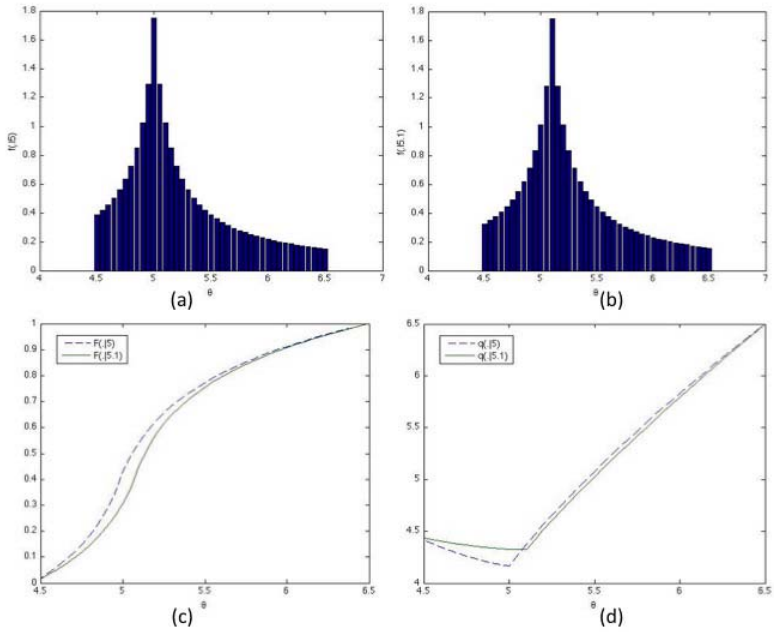

Figure 4: Example 9, the hyperbolic case. $\alpha = 1.75, \Delta\theta = 0.05$.

and chosen only as examples). In this case, the probability that a type remains constant is $f_\alpha(\theta_i|\theta_i) = \frac{A(\theta_i)}{\sigma}\Delta\theta\Phi(0)$, a function of the first period realization. It is easy to verify the probability of persistence converges to one as $\sigma \to 0$ (in the notation used above, the process can therefore be parametrized by $\alpha = 1 - \sigma$). The top panels of Figure 2 illustrate $f_\alpha(\theta_j|\theta_i)$ for two values: $\theta_i = 5.1$ (top left panel) and $\theta_i = 5.4$ (top right panel) The bottom right and left panel of the figure shows the FO-optimal contracts at $t = 2$ after histories $\theta_i = 5.1$ and $\theta_i = 5.4$, respectively. The contract is not monotonic: it is not monotonic with respect to the realization at $t = 2$ (this can be seen from the fact that the lines are not non-decreasing); and it is not monotonic in the realization at $t = 1$ (this can be seen form the fact that the contracts intersect at $t = 2$). It can also be verified that the FO-optimal contract is not incentive compatible.

Examples 8 and 9. Assume that the type in the first period, θ_1, is uniformly distributed and consider now the transition probabilities $f_\alpha(\theta_j|\theta_i) = \alpha e^{-\frac{(\theta_j-\theta_i)^2}{\sigma_i(\alpha\Delta\theta)}}\Delta\theta$, and $f_\alpha(\theta_j|\theta_i) = \frac{\alpha\Delta\theta}{1+\sigma_i|\theta_j-\theta_i|}$, where σ_i chosen so that the probabilities sum to one. In this case, $f_\alpha(\theta_i|\theta_i) = \alpha\Delta\theta$ so the probability of persistence is identical for all types; it is the variance of the distribution that is adjusted so that f_α assigns probability one on Θ_2. As it is straightforward to verify, we have $\sigma_i \to 0$ as $\alpha\Delta\theta \to 1$. Figures 3 and 4 illustrate $f_\alpha(\theta_j|\theta_i)$ for two values: $\theta_i = 5$ (top left panel) and $\theta_i = 5.1$ (top right panel) and compares the two implied distribution functions (bottom left panel)

continous density $f(\theta)$. As standard, in this case the probability of type θ_i is equal to $f(\theta_i)\Delta\theta$, i.e. the "histogram" approximation of the continuous density.

assuming $\Theta_1 = [5, 6]$ and $\Theta_2 = [4.5, 6.5]$. The bottom right panel of the figures illustrates the FO-optimal contracts at $t = 2$ after histories $\theta_i = 5$ and $\theta_i = 5.1$, respectively. As in Example 7, the contract is not monotonic and the FO-optimal contract associated to this case is not incentive compatible.

It is interesting to note that both in Examples 7, 8 and 9 the contract is not monotonic despite the fact that the transition probabilities have moderate levels of persistence. Finally, the discussion above, Propositions 3 and 4, and non-monotonicity in the examples are all valid independent of the prior μ, re-affirming our assertion that the failure of the first-order approach is not a technical irregularity, but a consequence of the added structure that dynamics present to the economic problem of contracting.

5.1 Discussion

We conclude this section with a few remarks on Propositions 3 and 4.

Perfectly persistent shocks. As we have seen in the previous sections, the first-order approach *always* works when types are perfectly persistent; Proposition 4, however, shows that the FO-approach does not generically work when types are highly persistent. How is this possible? The key to understanding this apparent contradiction is to realize that when types are constant, the repetition of the optimal static contract is only one of the many possible solutions: in histories which occur with *exactly* zero probability, the quantities are irrelevant and so they can be set to any arbitrary number, for example, equal to the static optimum. On the contrary, when types are highly persistent, but probabilities off the main diagonal are not exactly zero, quantities can not be set arbitrarily in these histories. The effect of these histories on the agent's rents is small, but so is the effect of these quantities on the surplus. Typically, the quantities are uniquely defined along all histories. As persistence converges to one, these quantities along the non-constant histories converge to values that are different from the static optimum and that are non-monotonic. In the example presented at the beginning of Section 5, $q_M(M) = \theta_M - \frac{\mu_H}{\mu_M}\frac{3\alpha-1}{2\alpha}$ that converges to $\theta_M - \frac{\mu_H}{\mu_M}$ in the limit; $q_M(L)$, on the contrary, is equal to θ_M for any α: in the limit, therefore, $q_M(M) < q_M(L)$. The problem is that there is a lack of lowerhemicontinuity at the limit with constant types, and some of the limit solutions (including the repetition of the static optimum) can not be seen as the limit of solutions as persistence converges to one.

AR(k) models. To see why monotonicity is a fragile property in $AR(k)$ models consider (11). In this formula the terms $D(h^{t-1}, \theta_i)$ are all identically equal to γ^{t-1} and independent of h^{t-1}: trivially, therefore, we have $q^*\left(\theta_i|h^{t-1}\right) = q^*\left(\theta_i|\widehat{h}^{t-1}\right)$ for *any* two histories with $h^{t-1} \succeq \widehat{h}^{t-1}$. This however is not a generic property: it follows from the fact that the shocks ε_t are assumed to be i.i.d and linearly additive. If we assume that the distribution of ε_t depends on the past realization, even if the effect of the past realization is very small, then $D(h^{t-1}, \theta_i)$ is history dependent and it is no longer the case that $q^*\left(\theta_i|h^{t-1}\right) = q^*\left(\theta_i|\widehat{h}^{t-1}\right)$.

In addition to this, even assuming a constant γ, to make the $AR(k)$ model conceivable we need to assume that the support shifts with the type as in Examples 4 and 5 or alternatively that the

type support is unbounded above and below. If we assume a given constant and bounded support then a perfect horizontal translation of the distribution is obviously impossible. Again, it is not generally true that $q^*\left(\theta_i|h^{t-1}\right) = q^*\left(\theta_i|\widehat{h}^{t-1}\right)$ for *any* two histories with $h^{t-1} \succeq \widehat{h}^{t-1}$. Example 7 can be seen as an $AR(k)$ model with bounded support in which the shock follows a truncated normal. As evident from Figure 2, neither monotonicity nor the FO-approach works in this case.

On serially independent shocks. There is a hope that the *AR(k)* model can be seen as an example of a more general class of environments for which the first-order approach works. This suggestion is based on an observation by Eso and Szentes [2007] and judiciously extended by Pavan, Segal and Toikka [2014], that any model with correlated and continuous types can be transformed into an equivalent model with i.i.d. shocks. To see this, note that if the cumulative distribution is $F(\theta^t|\theta^{t-1})$, then assuming that the agent observes θ^t is equivalent to assuming that he or she observes the variable $v_t = F(\theta^t|\theta^{t-1})$ (since $F(\theta^t|\theta^{t-1})$ is increasing and invertible in θ^t): v_t is a random variable with a uniform distribution on $[0,1]$. Eso and Szentes' [2007] observation is insightful in interpreting the screening contract and useful to derive the envelope formula (see Pavan, Segal and Toikka [2014], and Eso and Szentes [2013]). Unfortunately, however, it does not generalize the insights from the *AR(1)* models in that it does not help to solve the problems of the FO-approach elucidated above. It is useful to illustrate why transforming the stochastic process to an i.i.d. shock does not simplify the set of binding constraints. Assume the utility is $u(\theta, q) = \theta q$. To make the equivalent transformation, we need to substitute $\theta^t = F^{-1}(v_t; \theta^{t-1})$, so we have: $u(v_1, q_1) = F^{-1}(\mathbf{v}_1) \cdot q_1$, $u(\mathbf{v}^2, q_2) = F^{-1}(v_2; F^{-1}(v_1)) \cdot q_2$ and iterating:

$$u(\mathbf{v}^t, q_t) = F^{-1}(v_t; F^{-1}(v_{t-1}; F^{-1}(v_{t-2}; [...]))) \cdot q_t \tag{15}$$

where $\mathbf{v}^t = (v_1, ..., v_t)$. It is clear from (15) that, even starting from the simplest utility function, the per period utility of the equivalent transformation is a very complicated, time inseparable function of the entire history of the shocks $\mathbf{v}^t$. The change of variables, from θ^t to v_t, allows one to get rid of serial correlation in the types; the correlation however, does not disappear: it must be incorporated in a transformed utility function. All the problems that induce a failure of the first-order approach in the original problem are just shifted from the distribution function to the transformed per period utility function. The benefit of having independent shocks is compensated by the complications of having these per period utilities.

From discrete to continuous types. While we have focused the analysis on the case with discrete types, there is a strict connection between models with discrete and continuos types; and the same issues discussed above arise in continuous type models as well. Consider a continuous types model with type set $\Theta = \left[\underline{\theta}, \overline{\theta}\right] \subset \mathbb{R}^+$, prior distribution $\Gamma(\theta)$ and transition distribution $F(\theta' | \theta)$. We can define an associated discrete model by defining the type space as $\Theta^N = \{\theta_0, ..., \theta_N\}$ with $\theta_0 = \overline{\theta}$, $\theta_N = \underline{\theta}$ and $\theta_i = \theta_{i+1} + \Delta\theta_N$, the prior as $\Gamma^N(\theta_i) = \Gamma(\theta_i)$ and the transition matrix as $F^N(\theta_j | \theta_i) = F(\theta_j | \theta_i)$. In the online appendix we show that the envelope formula and the FO-optimal contracts of the continuous model can be obtained as limits of the discrete formulas (4) and (9). We also present a number of solved continuous type examples

(including continuous type versions of Examples 8 and 9) to illustrate the limits of the FO-approach.

6 What does the optimal contract look like when the first-order approach is invalid?

As we have seen in Section 5, even with two periods and three types the FO-optimal contract fails to be monotonic and the FO-approach can not be generally applied. In this section we fully characterize the optimal contract in the motivating example of Section 4, in which $f(\theta|\theta) = \alpha$ and $f(\theta|\theta') = \frac{1-\alpha}{2}$ for any $\theta, \theta' \in \{\theta_H, \theta_M, \theta_L\}, \theta \neq \theta'$ and $\alpha > 1/3$. The goal of this section is twofold- to elucidate the structure of optimal contracts that is otherwise elusive in models where the FO-approach can be applied, and to illustrate the trade-offs between rent and efficiency in a dynamic model.

To characterize the optimal contract we focus on a *weakly relaxed program* that constitutes problem (3) with $|\Theta| = 3$ and $T = 2$, with the following subset of constraints:

$$IR_L, IC_{HM}, IC_{ML}, IC_{HL}, \tag{16}$$

$$IC_{HM}(M), IC_{ML}(M), IC_{LM}(M), IC_{HM}(L), IC_{ML}(L), IC_{LM}(L)$$

where IR_L is the individual rationality constraint of type L at $t = 0$, $IC_{i,j}$ is incentive compatibility constraint requiring that type i doesn't want to misreport being a type j in period 1, and $IC_{i,j}(k)$ is the incentive compatibility constraint requiring that type i doesn't want to misreport being a type j in period 2, after the agent reports to be a type k in period 1. In contrast to the FO-approach, this problem has two key differences. First, now we are ignoring all the individual rationality constraints of the lowest type in period 2 and incentive compatibility constraints after history H. Second, and most importantly, we are adding three new constraints: the global downward constraint IC_{HL}, and the local upward constraints $IC_{LM}(M), IC_{LM}(L)$ in period 2. The constraint set of the problem is illustrated in the relevant history tree in Figure 5. In the following we will refer to this program as the *WR-program*.

Since this is a three type and two period model we simplify notation. Let U_i be the expected utility of type i in the first period and $u_i(h)$ be the expected utility of type i after history h in the second period. Note that since the second period is the terminal period, the expected utility and stage utility are the same. Similarly, we define q_i and $q_i(h)$ to be the first and second period allocations respectively. The following lemma allows to simplify the constraint set:[27]

Lemma 2. *In the WR-program, constraints IR_L, IC_{HM}, IC_{ML} bind at the optimum.*

[27] When only the usual local downward incentive compatibility constraints are considered, the following result is immediate. If, for example IC_{HM} were not binding, the principal could simply raise the price that type θ_H is paying. In the *WR-program* the proof of the result is complicated by the additional constraints: reducing type θ_H's rent at $t = 0$ may conflict with IC_{HL}.

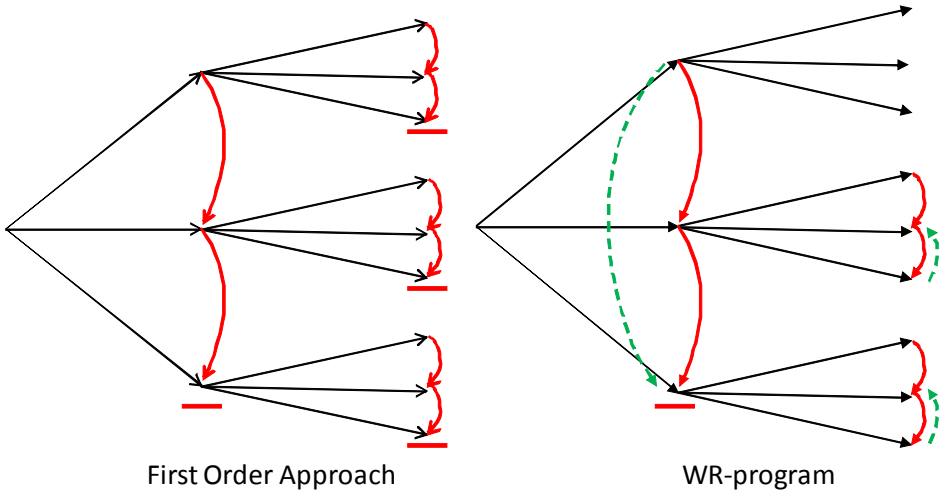

Figure 5: The dashed arrows are the constraints in the *WR-program* that are ignored in the first-order approach.

We can now use the equalities implied by Lemma 2 to reduce the number of free variables in the optimization problem. In particular we can eliminate the period 1 utility vectors. Define $\omega_{HM}(i) = u_H(i) - u_M(i)$ and $\omega_{ML}(i) = u_M(i) - u_L(i)$ for $i = M, L$. The variable $\omega_{kl}(i)$ is the net utility of reporting to be type k rather than a type l after history i. Using this notation, we can rewrite the *WR-program* as a maximization problem in which the control variables are the quantities $\mathbf{q}$ and second period marginal utilities $\boldsymbol{\omega}$:

$$\max_{\langle \boldsymbol{\omega}, \mathbf{q} \rangle} \left\{ \begin{array}{c} \sum\limits_{i=H,M,L} \mu_i \left[\theta_i q_i - \frac{1}{2} q_i^2 + \delta \sum\limits_{k=H,M,L} \mu\left(k \,|\, i\right) \left(\theta_k q_k(i) - \frac{1}{2} q_k(i)^2 \right) \right] \\ -\mu_H \left[\Delta\theta q_M + \delta \frac{3\alpha - 1}{2} \omega_{HM}(M) \right] \\ -(\mu_H + \mu_M) \left[\Delta\theta q_L + \delta \frac{3\alpha - 1}{2} \omega_{ML}(L) \right] \end{array} \right\} \tag{17}$$

subject to

$$\begin{array}{rcl} [\lambda]: & \Delta\theta q_M + \delta \dfrac{3\alpha - 1}{2} \omega_{HM}(M) \geq \Delta\theta q_L + \delta \dfrac{3\alpha - 1}{2} \omega_{HM}(L) & \\ [\lambda_{HM}(M)]: & \omega_{HM}(M) \geq \Delta\theta q_M(M) & | \quad [\lambda_{HM}(L)]: \quad \omega_{HM}(L) \geq \Delta\theta q_M(L) \\ [\lambda_{ML}(M)]: & \omega_{ML}(M) \geq \Delta\theta q_L(M) & | \quad [\lambda_{ML}(L)]: \quad \omega_{ML}(L) \geq \Delta\theta q_L(L) \\ [\lambda_{LM}(M)]: & \omega_{ML}(M) \leq \Delta\theta q_M(M) & | \quad [\lambda_{LM}(L)]: \quad \omega_{ML}(L) \leq \Delta\theta q_M(L) \end{array}$$

where the variables in the square brackets on the left are the Lagrange multipliers associated with the constraints. Program (17) is a standard maximization problem, but it is complicated by a still significantly large number of constraints. The difference between (17) and the problem of

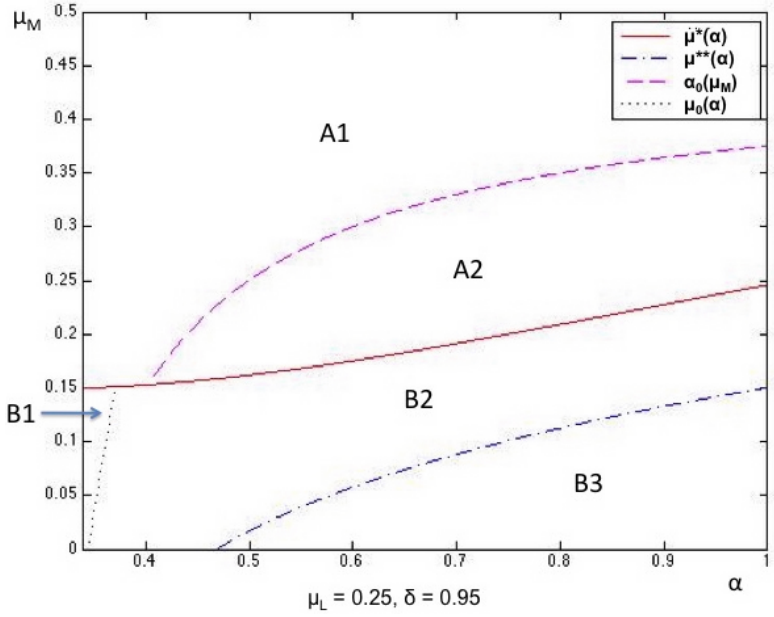

Figure 6: Fully characterized contract

the first-order approach (6) is the global constraint IC_{HL} and the presence of the local upward constraints $IC_{LM}(M)$ and $IC_{LM}(L)$. The latter are essentially *monotonicity conditions* requiring $q_M(h) \geq q_L(h)$ for $h = M, L$.[28] We cannot ignore any of these three constraints. Moreover now we cannot assume without loss of generality that all local downward incentive constraints are binding at $t = 2$: so the envelope formula (4) cannot be directly applied. Hence, we still have utilities in the objective function. The next lemma validates our focus on problem (17) :

Lemma 3. *A contract is optimal if and only if it solves the WR-program.*

The analysis can be divided into two cases: first the case in which the global constraint can be ignored and so it is sufficient to look at local constraints, i.e. $\lambda = 0$; second, the case in which the global constraint is binding, i.e. $\lambda > 0$.

6.1 Case 1: Local IC is sufficient

The following result characterizes the necessary and sufficient condition for $\lambda = 0$. For a given μ_L and δ, the environment is fully described by two parameters, μ_M, α, and therefore it can be represented in the two dimensional box $(\mu_M, \alpha) \in E(\mu_L) = (0, 1 - \mu_L) \times (1/3, 1)$.[29] In the rest of the analysis we will fix μ_L and δ and study how the equilibrium changes as we change μ_M, α.

[28] To see this note that given $IC_{ML}(h)$, $q_M(h) \geq q_L(h)$ if and only if $IC_{LM}(h)$ is satisfied.

[29] The thresholds defined below do not depend on the types θ.

This approach is without loss of generality and it allows for simpler statements (and a graphical representation) of the relevant cases. We have:

Lemma 4. *There exists a threshold $\mu^*(\alpha)$ such that the global incentive constraint IC_{HL} can be ignored if and only if $\mu_M \geq \mu^*(\alpha)$.*

Within the two regions defined by $\mu^*(\alpha)$, the particular shape of the optimal contract depends on the remaining set of binding constraints. Explicit solutions of the optimal quantities for all feasible parameters are presented in Table 1 in the appendix. The following proposition describes what the optimal contract looks like for $\mu_M \geq \mu^*(\alpha)$, when the global constraint can be ignored:

Proposition 6. *Assume $\mu_M \geq \mu^*(\alpha)$. There is a threshold $\alpha_0(\mu_M)$, such that:*

- **Case A1.** *If $\alpha < \alpha_0(\mu_M)$, the optimal contract is fully separating and first-order optimal.*
- **Case A2.** *If $\alpha \geq \alpha_0(\mu_M)$, the optimal contract is fully separating after all histories except M. After this history types M and L are pooled: $q_M(M) = q_L(M)$.*

Regions *A1* and *A2* are illustrated in Figure 6 in a simple parametric example, where the threshold $\alpha_0(\mu_M)$ is represented by a dashed line.[30] In region $A1$, the envelope formula is sufficient to characterize the optimal contract. In this case the FO-optimal contract is not monotonic (as in Definition 2), but this lack of monotonicity is not sufficient to cause a failure of incentive compatibility. The contract is not monotonic because $q_M(M) < q_M(L)$. However, given any h, $q_\theta(h)$ is monotonic in θ. In region A2, even though the global constraints can be ignored, the envelope formula is not sufficient to determine the contract since at $t = 2$ we have pooling after history M. It is interesting to note that although pooling makes sure that $q_\theta(h)$ is monotonic in h for all θ, the optimal contract remains non-monotonic with respect to the realization at $t = 1$ (since $q_M(M) < q_M(L)$ in A2 as well). For reasons akin to the discussion of Figure 1, when types are positively correlated across time, it is optimal to offer non-monotonic contracts by having higher distortions at constant histories.

Figure 7 illustrates what happens when we make the payoffs in the second period more important by increasing δ. We know from Proposition 4 that as the future becomes more important, the first-order approach is never valid for high levels of α. A similar phenomenon occurs here: as δ increases, $\mu^*(\alpha)$ shifts up and the region in which local constraints are sufficient shrinks. These higher values of δ should be seen as representative of dynamic models with longer time horizons.

6.2 Case 2: Local IC is not sufficient

When $\mu_M < \mu^*(\alpha)$ *both* the global constraint IC_{HL} *and* the local constraints IC_{HM} and IC_{ML} are simultaneously binding in the first period. There are three relevant cases. The following result characterizes the optimal contract in these situations:

Proposition 7. *There exists a threshold $\mu^{**}(\alpha)$ such that:*

[30] In Figure 6 we assume $\mu_L = 0.25$ and $\delta = 0.95$.

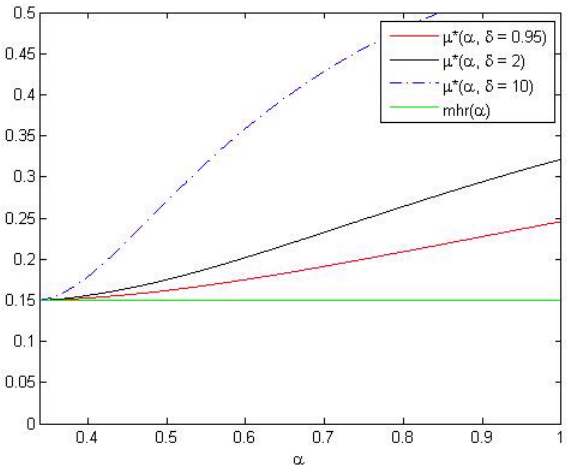

Figure 7: $\mu^*(\alpha)$ for $\delta = 0.95, 2$ and 10, and "monotone hazard rate" condition for static model.

- **Case B1&B2**. *Assume $\mu_M \in [\mu^{**}(\alpha), \mu^*(\alpha))$. The optimal contract is fully separating at $t = 1$. There exists a threshold $\mu_0(\alpha)$, such that the optimal contract is fully separating at $t = 2$ as well if $\mu_M > \mu_0(\alpha)$ (case B1). If $\mu_M \leq \mu_0(\alpha)$, types M and L are pooled after history M : $q_M(M) = q_L(M)$ (case B2).*

- **Case B3**. *If $\mu_M < \mu^{**}(\alpha)$, the optimal contract pools types M and L in the first period: $q_M = q_L$. In the second period, after history H, the contract is separating and efficient. After histories M and L, types M and L are pooled across both histories: $q_M(j) = q_L(j)$ and $q_j(M) = q_j(L)$ for $j = M, L$.*

Propositions 6 and 7 provide a full characterization of the optimal contract that can be used to gain new insights on how types are optimally screened in dynamic environments that are not apparent in the models discussed in Section 4.[31]

How does the possibility of repeated interactions affect the structure of the optimal contract? It is imperative to note that, in contrast to the static model, binding global constraints are no longer synonymous with pooling alone. In regions $B1$ and $B2$, even though the global incentive constraint binds, there is complete separation of types in period 1. Region $B2$ interestingly, like in $A2$, has a strategic separation in period 1 followed by history dependent pooling in period 2, which we term dynamic pooling. Region $B3$ has pooling in period 1 and in period 2 after histories θ_M and θ_L (but not θ_H). The contract captures a loss of history in region $B3$- types are always

[31] Table 1, presented in the appendix, enlists closed form solutions of the optimal quantities for the entire parameter space. Note that Assumption 3 ensures that the corner solution of zero quantity never arises; a sufficient condition for this is a small value of $\Delta\theta$.

pooled in period 2 after being pooled in period 1; it is as if we are in a two-type model following the pooled histories.

Another lesson is that pooling of types at $t = 1$ is always lower in the dynamic model than in a static optimal contract. It is easy to verify that in a static model types are pooled only if $\mu_M \leq \underline{\mu}_M = \frac{\mu_L(1-\mu_L)}{1+\mu_L}$, which in the example presented above is equal to 0.15- the horizontal line in Figure 7. This condition, however, is irrelevant in a dynamic model. We have pooling at $t = 1$ only if $\mu_M < \mu^{**}(\alpha)$, and $\mu^{**}(\alpha) < \underline{\mu}_M$ for all $\alpha \in (1/3, 1)$. The reason that the region for pooling at $t = 1$ is strictly smaller with two periods than with one is fairly intuitive. Separation is efficient, and the principal would always like to separate types as long as the information rents required for incentives do not outweigh the benefit of separation. In a dynamic environment the principal has an added instrument in the form of continuation value to screen the agent's types. The burden of inefficiency in the form of distortions due to information rents can therefore be pushed into the future. Yet full separation in the static model does not imply full separation over time in the dynamic model as is evident in region $A2$, with $\mu_M > \underline{\mu}_M$, but pooling in period 2.

7 Approximate optimality and Implementability

The results of the previous sections make clear that in order to solve for an optimal contract, the principal cannot generally use the first-order approach and limit the analysis to local incentive compatibility constraints. Without the first-order approach, we have no systematic way of simplifying the constraint set. This may make the analysis extremely complicated even from a numerical point of view. What does the optimal contract look like in general environments with large T and N? What kind of advice can we give to a seller who needs to design an optimal contract? In this section we show that there is a class of contracts that is relatively easy to characterize, and that induces a minimal loss (if any) on the principal's payoff precisely when the first-order approach fails, that is, when the agent's types are highly persistent. This class consists of contracts that are monotonic in the sense of Definition 2.

In static environments the envelope formula plus monotonicity are necessary and sufficient for a contract to be implementable: if we ignore the monotonicity constraint, then the contract *must* be ironed out to make it monotonic, otherwise implementability fails (see Myerson [1981]). In a dynamic environment monotonicity is not necessary: it follows that if we impose monotonicity in the seller's problem, we guarantee implementability even if we ignore the global constraints, but we may obtain a suboptimal contract.[32] We show that as types' persistence converges to one, the optimal monotonic contract converges in probability to the optimal contract, and so the loss from focusing on this class of contracts converges to zero. Further, the main result in this section establishes that for an infinitely repeated model, as the types' persistence and discounting converge to one, the expected profit in the optimal monotonic contract converges to the optimal profit, that is, the loss in profit from using a monotonic contract converges to zero *independent* of

[32] Note that in all the cases presented in section 4 in which the FO-optimal contract coincides with the optimal contract, the optimal monotonic contract is exactly optimal.

δ = 0.95	α						
	0.38	0.48	0.58	0.68	0.78	0.88	0.98
$\mu_H = 0.5$ $\mu_M = 0.1$	**0.01** 11.00	**0.01** 9.87	**0.02** 8.49	**0.02** 6.87	**0.01** 4.98	**0.01** 2.86	**0.00** 0.51
$\mu_H = 0.5$ $\mu_M = 0.2$	**0.01** 10.70	**0.02** 9.62	**0.04** 8.32	**0.06** 6.77	**0.06** 4.96	**0.04** 2.87	**0.01** 0.51
$\mu_H = 0.5$ $\mu_M = 0.3$	**0.01** 10.01	**0.01** 9.87	**0.02** 8.51	**0.03** 6.91	**0.03** 5.06	**0.02** 3.93	**0.01** 0.52
$\mu_H = 0.3$ $\mu_M = 0.1$	**0.01** 10.75	**0.01** 9.73	**0.01** 8.45	**0.02** 6.91	**0.02** 5.08	**0.01** 2.95	**0.00** 0.53
$\mu_H = 0.3$ $\mu_M = 0.2$	**0.01** 10.61	**0.01** 9.61	**0.01** 8.37	**0.03** 6.87	**0.04** 5.08	**0.03** 3.98	**0.01** 0.54
$\mu_H = 0.3$ $\mu_M = 0.3$	**0.01** 10.41	**0.01** 9.42	**0.01** 8.20	**0.02** 6.72	**0.02** 4.97	**0.02** 2.92	**0.01** 0.53

Figure 8: Percentage loss of optimal objective (monopolist's profit) by using monotonic contracts (in bold) and repetition of the static optimum.

the order of these limits.[33]

Define $\mathcal{M}$ as the set of *monotonic contracts*:

$$\mathcal{M} = \left\{ \mathbf{q} \;\middle|\; \begin{array}{c} q(\theta_i|h^{t-1}) \geq q(\theta_{i+1}|h^{t-1}), i < N, \text{ and } q(\theta_i|h^{t-1}) \geq q(\theta_i|\widehat{h}^{t-1}), \\ i = 1, ..., N, \forall h^{t-1} \text{ and } h^{t-1} \succeq \widehat{h}^{t-1} \end{array} \right\} \tag{18}$$

where, as before, $h^t \succeq \widehat{h}^t$ if $h^t_j \geq \widehat{h}^t_j \ \forall j \leq t$. It follows immediately from Proposition 2 that the optimal monotonic contract can be characterized by solving the following program:

$$\max_{\mathbf{q} \in \mathcal{M}} \left\{ \mathbb{E}\left[S(\mathbf{q})\right] - \sum_{i=0}^{N} \mu_i U^*(\theta_i, h^0; \mathbf{q}) \right\} \tag{19}$$

where $U^*(\theta_i, h^0; \mathbf{q})$ is given by the envelope formula (5). Problem (19), moreover, is sufficiently tractable to allow a partial characterization of the properties of its solution.

Proposition 8. *In the optimal monotonic allocation,* $q\left(\theta^t \left| h^{t-1} \right.\right) \leq \theta^t$ for any θ^t and h^{t-1}. *Moreover, for any arbitrarily small* $\varepsilon_1, \varepsilon_2 > 0$ *we have* $\Pr\left(\left| q\left(\theta^t \left| h^{t-1} \right.\right) - \theta^t \right| > \varepsilon_1\right) \leq \varepsilon_2$ *for* t *and* T *sufficiently large.*

The first part of the proposition establishes that, analogous to the static model, the optimal monotonic contract is uniformly downward distorted. The second part states that the contract converges to an efficient contract in probability.

[33] All limit results in this section for persistence going to one are also valid for persistence going to zero, that is the stochastic process converging to an iid process. At both limits the optimal contract is monotonic.

How good is the optimal monotonic contract as an approximation of the optimal contract? Let $\mathbf{q} = \{q(h^t)\}_{h^t \in H}$ be an allocation and let $\mathbf{q}^{**} = \{q^{**}(h^t)\}_{h^t \in H}$ be the optimal allocation. Let I be the identity matrix that describes the transition matrix when types are perfectly correlated. As types become perfectly persistent, we must have that the transition matrix converges to I, i.e. $\boldsymbol{\alpha} \to I$. We say that $\mathbf{q}$ converges in probability to the optimal allocation as types become perfectly persistent if $\lim_{\boldsymbol{\alpha} \to I} \Pr(|q(h^t) - q^{**}(h^t)| \geq \varepsilon) = 0$ for any $\varepsilon > 0$.

Proposition 9. *For all μ, δ, T, and transition matrices, the optimal monotonic contract converges in probability to a contract that maximizes the seller's profits as types become perfectly persistent.*

The intuition of this result is as follows. As persistence converges to one, the optimal contract converges to the optimal static contract on the constant histories. The constant histories are histories along which types remain constant. In the proof of Proposition 9 we show that the optimal monotonic contract also converges to the optimal static contract along these histories. The optimal contract and the optimal monotonic contract, therefore, can differ at most only along histories in which the types change over time. As persistence converges to one, however, the probability of these histories goes to zero.[34]

This result implies that for any δ and T, as types become increasingly persistent the profit associated with the optimal monotonic contract converges to the profit in the optimal contract. The table in Figure 8 illustrates the loss of profits associated with the optimal monotonic contract in an example with 3 periods, 3 types and the Markov matrix used in Section 6. The loss is expressed as a percentage of the profit in the optimal contract. As can be seen, the approximation is quite good for all cases, with a loss of profit that is never higher than 0.06%. It is interesting to note the inverse-U relationship between losses and the level of persistence. As persistence increases, losses increase, peak and then come down again. The reason is simple. At $\alpha = 1/3$, the model is akin to the i.i.d. shock framework, where we know that the optimal contract is monotonic. At the other extreme, $\alpha = 1$, the optimal contract constitutes repetition of the static optimum which too is monotonic. As we increase α, the distortions vary and the probability of non-constant histories decreases. Thus, the loss in using monotonic contracts increases with the non-monotonicities only to be suppressed in probability by the increasing weight of constant histories along which the optimal monotonic allocation converges to the optimal allocation.

When simultaneously types' persistence, the discount factor and the length of the contract are high, Proposition 9 may not be sufficient to guarantee that the optimal monotonic contract is a good approximation for the seller. Even a contract that converges to the efficient contract as $\boldsymbol{\alpha} \to I$ may perform very poorly as $\delta \to 1$ and $T \to \infty$ as well. For example, the repetition of the optimal static contract converges in probability to an optimal contract (as shown in Example 2): for any given $\boldsymbol{\alpha}$ (even arbitrarily close to I), however, the difference in profits between this contract and the optimal contract becomes arbitrarily large as $\delta \to 1$ and $T \to \infty$.[35] The problem

[34] It is useful here to note that the fact that optimal monotonic contract converges to the optimal contract has nothing to say on the FO-optimal contract. The FO-optimal contract also converges in probability to the optimal contract. The problem with the FO-optimal contract, however, is that it is not incentive compatible.

[35] As proven in Battaglini [2005], the optimal (dynamic) contract becomes efficient and the seller appropriates

is that the contract may not converge to the efficient contract fast enough in α. Therefore, in general, the order of limits may matter when we allow both the probability of persistence and the discount factor to converge to one.

The following result shows that for the optimal monotonic contract, profits converges to the optimal level independently of the order of limits. Define $\pi_m(\boldsymbol{\alpha}, \delta, T)$ and $\pi^*(\boldsymbol{\alpha}, \delta, T)$ to be the expected average discounted profits corresponding to the optimal monotonic contract and the optimal contract.[36] Moreover, let $\pi_m(\boldsymbol{\alpha}, \delta) = \lim_{T\to\infty} \pi_m(\boldsymbol{\alpha}, \delta, T)$ and $\pi^*(\boldsymbol{\alpha}, \delta) = \lim_{T\to\infty} \pi^*(\boldsymbol{\alpha}, \delta)$ be the limit expected average profits as $T \to \infty$.[37]

Proposition 10. *When $\boldsymbol{\alpha} \to I$ and $\delta \to 1$, the profits of the optimal monotonic contract converges to the profits of the optimal contract independent of the order of limits:* $\lim_{\delta\to 1}\lim_{\boldsymbol{\alpha}\to I} \pi_m(\boldsymbol{\alpha}, \delta) = \lim_{\delta\to 1}\lim_{\boldsymbol{\alpha}\to I} \pi^*(\boldsymbol{\alpha}, \delta)$ and $\lim_{\boldsymbol{\alpha}\to I}\lim_{\delta\to 1} \pi_m(\boldsymbol{\alpha}, \delta) = \lim_{\boldsymbol{\alpha}\to I}\lim_{\delta\to 1} \pi^*(\boldsymbol{\alpha}, \delta)$.

The table in Figure 8 illustrates this point comparing the loss of profits of the optimal monotonic contract with the loss of profits obtained with the repetition of the optimal static contract: the loss can be higher than 10% of the optimal profits, even in this simple example with only 3 periods. Naturally larger losses should be expected with longer horizons.

The results of this section may be useful in applied work. As mentioned in the introduction, many works in the applied literature postulate that the first-order approach works. The risk is that the contracts thus characterized are not incentive compatible. Further in the most natural environments, this risk can not be fully resolved by numerical methods. To the extent that it is not possible to check *all* the incentive compatibility constraints, studying optimal monotonic contracts may be a more robust option, since it guarantees implementability and it is equal to the true optimal contract with high probability when types are highly persistent.

8 Related literature

Our paper is related to four main literatures. First, we have the traditional literature studying dynamic principal-agent models when the agent's type follows a stochastic process and the allocation is chosen in every period. The first paper to use the first-order approach to study dynamic models and state an associated "envelope formula" is Baron and Besanko [1984].[38] Their paper states the formula in general terms and shows it to be sufficient in two benchmark cases: when types are constant over time, in which case the optimal dynamic contract corresponds to a repetition of the static optimum; and when types' realizations are independently distributed over

all the surplus as $\delta \to 1$ when types are imperfectly persistent. With the repetition of the optimal static contract, however, per period surplus is below the efficient level and only a fraction is appropriated by the seller. The difference in discounted profits, therefore, becomes arbitrarily large as $\delta \to 1$ and $T \to \infty$.

[36] If $\Pi_m(\boldsymbol{\alpha}, \delta, T)$ and $\Pi^*(\boldsymbol{\alpha}, \delta, T)$ are the expected discounted profits correspondening to the optimal monotonic contract, then $\pi_m(\boldsymbol{\alpha}, \delta, T) = (1-\delta)\Pi_m(\boldsymbol{\alpha}, \delta, T)$ and $\pi^*(\boldsymbol{\alpha}, \delta, T) = (1-\delta)\Pi^*(\boldsymbol{\alpha}, \delta, T)$.

[37] This limit exists without loss of generality since $\pi_m(\boldsymbol{\alpha}, \delta, T)$ and $\pi^*(\boldsymbol{\alpha}, \delta, T)$ are bounded for any $\boldsymbol{\alpha}$ and δ.

[38] See Section 3 for a discussion of the first-order approach and envelope formula. See Stole [2001], Laffont and Martimort [2002], Milgrom (2004), and Bolton and Dewatripont [2005] for general discussions of the envelope formula in the static case.

time, in which case the optimal contract is efficient starting from period 2.[39] Extensions of this approach to environments with imperfect correlation of types are presented by Besanko [1985], Laffont and Tirole [1990] and Battaglini [2005]. Besanko [1985] extends the analysis to an infinite horizon with continuous types following a AR(1) process; Laffont and Tirole [1990] focus on a two periods environment with two types. Battaglini [2005] extends the two types model to an infinite horizon.[40] The main contributions of these papers is in showing that the first-order approach is sufficient in their respective environments. Laffont and Tirole [1996], and more recently, Pavan, Segal and Toikka [2014] and Eso and Szentes [2013] have derived "envelope formulas" for continuous types applicable to more complex environments. Both of the latter papers build on Eso and Szentes [2007], where the principal-agent problem is transformed in to a problem in which the shocks are i.i.d. through an appropriate change in utility. Contrary to the previous literature, these papers are not focused on finding specific environments in which these envelope formulas are sufficient for incentive compatibility, leaving open the question of the general applicability of the first-order approach.[41]

The second literature to which our paper is related is on sequential screening started by Courty and Li [2000]. This literature studies environments in which the agent receives information gradually over time, but the allocation is determined only in the last period. The models in this literature have 2 stages: in the beginning of period 1, the agent receives an informative signal and the contract is signed at the end of this period, but no allocation is made; in the second period the type is revealed to the agent and the allocation takes place. Courty and Li is one of the first papers to clearly discuss the limitations of the first-order approach in dynamic environments: one of their main achievements is to identify environments in which the first-order approach can be applied in the class of problems that they study. More recently, Courty and Li's work has

[39] See also Townsend [1982] and Clementi and Hopenhayn [2006], amongst others, for dynamic principal-agent models in which types are serially uncorrelated.

[40] Other important contributions in the dynamic contracting literature are Dewatripont [1989], Hart and Tirole [1988], Rey and Salanie (1990), Rustichini and Wolinsky [1995], Battaglini [2007], Williams [2011], Bergeman and Valimaki [2010], Strulovici [2011], Garrett and Pavan [2012], Athey and Segal [2013], Boleslavsky and Said [2013], Maestri [2013]. These papers however focus on different aspects of the problem and limit the analysis to environment that are quite different from ours. Hart and Tirole (1988) assumes that supply can have two values, zero or one. Rustichini and Wolinsky (1995) assume consumers are not strategic and ignore that future prices depend on their current actions. Dewatripont (1989), Rey and Salanie (1990), Battaglini [2007], Maestri [2013] and Strulovici [2011] focus on renegotiation. Garrett and Pavan [2012] look at managerial compensation when allocations are ex post monotone. Bergemann and Valimaki [2010] and Athey and Segal [2013] study implementation of efficient allocations extending the pivot mechanism to dynamic environments. See also Bergemann and Said [2011] for a short survey.

[41] The problems encountered in studying incentive constraints in dynamic settings are related to the problems encountered in multidimensional environments: in both cases it is hard to reduce the constraint set by ordering the types in an ex ante obvious manner. Rochet [1987] shows that cyclical monotonicity (a monotonicity chain amongst all finite sequence of types) is both necessary and sufficient for implementability in static models. In the single dimension model this condition simplifies to monotonicity of the allocation rule. A similar condition can stated for dynamic contracts, however just as in the static multidimensional environment, in dynamic contracts cyclical monotonicity has no immediate economic interpretation, and the restriction it imposes on the primitives of the model is less than well understood. There is however a crucial difference between multidimensional and dynamic contracts: in the former all information is endowed to the agent at the same time, and revealed to the principal also in one block; in the latter the multidimensional information is gradually released to the agent and hence through the mechanism gradually reported to the principal. See Rochet and Stole [2003] for an elegant survey on multidimensional screening.

been extended in many directions. Eso and Szentes [2007] consider the case in which the seller can choose to voluntarily disclose information in the first period. They show that the agent does not receive private rents for the disclosure of information. Li and Shi [2013] show that discriminatory disclosure of information can be optimal when the amount of additional private information that the buyer can learn depends on his type. Krahmer and Strausz [2015] argue that in this class of models the benefit of sequential screening is due to the joint relaxation of incentive and participation constraints. To solve their model, the authors propose an original approach to deal with global constraints that works in their environment with N types.[42] In all these papers the key question is whether the contract must depend on the interim informative signal, or if it can depend only on the type revealed in the last stage. In our model, because the allocation is chosen in all periods, information must be disclosed in all periods.

Third, our paper is related to a recent literature devoted to the study of approximately optimal mechanisms in environments in which fully optimal mechanisms are hard to characterize (see Madarasz and Prat [2014], Chassang [2013] for recent contributions and Hartline [2012] for a summary of the computer science approach). While parts of this literature deal with more general environments than ours, the approach we adopt in Section 7 takes full advantage of the dynamic structure of the framework we study; this allows us to obtain an approximately optimal contract that guarantees incentive compatibility for all types at all histories.

Finally, there is a large and growing literature using the first-order approach to solve dynamic contracts in complex environments using numerical methods. Understanding the conditions for the applicability of the first-order approach with discrete types seems particularly important in these exercises. Even when using models with continuous types, these papers typically compute the equilibrium policies and verify incentive compatibility using discretized approximations.[43] When discrete approximations are not used to construct the first-order optimal contract, incentive compatibility is verified numerically on a grid of points.[44] The envelope formula presented in our paper provides an exact formula for discrete types that can be used to compute the first-order optimal contract and to verify incentive compatibility directly without approximations.

9 Conclusion

In this paper we have studied a simple principal-agent model in which the agent's type is private information and follows a Markov process. We have presented four sets of results. First, following the standard approach in the literature, we have studied the optimal contract when only local

[42] In another paper, Krahmer and Strausz [2011] argue that in sequential screening models a restriction to deterministic contracts again makes monotonicity a necessary condition. It would be interesting to pursue this idea in the context of our model in future research.

[43] This is the case, for example, in Kapicka [2013], Farhi and Werning [2013], and Golosov et al. [2013] who study models of intertemporal consumption smoothing using numerical methods.

[44] Exceptions are Zhang [2009] and Williams [2011] who use continuous time methods. Zhang [2009] and Williams [2011] verify that the conditions for the first order approach are satisfied in their model. Zhang [2009] however, limits the analysis to a two types model; and Williams [2011] limits the set of possible deviations available to the agent (who can report only incomes lower or equal to the true income).

incentive constraints are considered. We have shown that the agent's equilibrium rents can be represented purely as a function of the allocation through a dynamic version of the so called "envelope formula." Moreover, as in the static model, the envelope formula and a natural monotonicity condition on the allocation guarantee that the contract is implementable. Although this condition is only sufficient and quite strong, it is verified for virtually all the natural environments in which the optimal dynamic contract has been characterized in the existing literature.

Second, and most importantly, we have shown that the environments for which the envelope formula is sufficient to characterize the optimal dynamic contract are special. In general, even in the simplest examples, the allocation is not monotonic. Thus, for high persistence and sufficiently long time horizons global incentive constraints generically bind. Moreover, numerical examples show that moderate levels of persistence are sufficient to violate the first-order approach.

Third, to gain insight on how the optimal contract looks like when the first-order approach doesn't work, we have characterized it in a simple case with three types and two periods. We show that the optimal contract is characterized by *dynamic pooling*: strategic, state contingent treatment of types in which types may be initially separated, but then be pooled conditioned on particular histories. And, once types are pooled in the first period, the model operates like a two-type model by pooling the same types in the second period across histories.

Finally, we have shown that some insights in general environments with many types and periods can be gained by studying a simple class of incentive compatible and robust contracts: monotonic contracts, in which non-monotonicities in the allocation are "ironed" out. The appeal of optimal monotonic contracts is dual: it is always incentive compatible, and the loss in profit (with respect to the optimal contract) from using these contracts converges to zero as the persistence of types and discounting converge to one, independent of the order of these limits, precisely when the first-order approach tends to fail.

The analysis suggests a number of important research questions. The characterization of the optimal contract with three types and two periods suggests that state dependent pooling of types plays an important role in dynamic screening. The example suggests a number of features that one naturally expects to hold in more general environments as well. The analysis in Section 7, moreover, suggests that even when it is not possible to fully characterize the optimal contract, useful insights can be gained by studying contracts that are approximately optimal. We leave the further development of these ideas for future research.

References

Athey, S. and I. Segal (2013), "An Efficient Dynamic Mechanism," *Econometrica*, 81 (6), 2463-2485.

Baron, D. and D. Besanko (1984), "Regulation and Information in a Continuing Relationship," *Information Economics and Policy*, 1(3), 267-302.

Battaglini, M. (2005), "Long-term Contracting with Markovian Consumers," *American Economic Review*, 95, 637–658.

Battaglini, M. (2007), "Optimality and Renegotiation in Dynamic Contracting," *Games and Economic Behavior*, 60 (2), 213-246.

Bergemann D. and J. Valimaki (2010). "The Dynamic Pivot Mechanism," *Econometrica*, 78(2), 771-789.

Bergemann D. and M. Said (2011). "Dynamic Auctions: A Survey," *Wiley Encyclopedia of Operations Research and Management Science*, 2, 1511-1522.

Besanko, D. (1985), "Multiperiod Contracts between Principal and Agent with Adverse selection," *Economic Letters*, 17, 33-37.

Bolton, P. and M. Dewatripont (2005). *Contract Theory*. The MIT Press, Cambridge, MA.

Chassang, S. (2013), "Calibrating Incentive Contracts," *Econometrica*, 81 (5), 1935-1971.

Clementi, G. and H. Hopenhayn (2006), "A Theory of Financing Constraints and Firm Dynamics," *Quarterly Journal of Economics*, 121 (1), 229-265.

Courty, P. and H. Li (2000), "Sequential Screening," *Review of Economic Studies*, 67 (4), 697-718.

Dewatripont, M. (1989), "Renegotiation and Information Revelation over Time: The Case of Optimal Labor Contracts," *Quarterly Journal of Economics*, 104 (3), 589–619.

Eso, P. and B. Szentes (2007), "Optimal Information Disclosure in Auctions and the Handicap Auction," *Review of Economic Studies*, 74 (3), 705-731.

Eso, P. and B. Szentes (2013), "Dynamic Contracting: An Irrelevance Result," *working paper*.

Farhi, E. and I. Werning (2013), "Insurance and Taxation over the Life Cycle," *Review of Economic Studies*, 80, 596-635.

Garrett, D. and A. Pavan (2012), "Managerial Turnover in a Changing World," *Journal of Political Economy*, 120 (5), 879-925.

Golosov, M., M. Troshkin, and A. Tsyvinski, (2013), "Redistribution and Social Insurance," *working paper*.

Guvenen, F., S. Ozkan, and J. Song, (2014), "The Nature of Countercyclical Income Risk," *Journal of Political Economy*, 122 (3), 621-660.

Guvenen, F., F. Karahan, S. Ozkan, and J. Song, (2015), "What Do Data on Millions of U.S. Workers Reveal about Life-Cycle Earnings Risk?" *working paper*.

Hartline, J. (2012), "Approximation in Mechanism Design," *American Economic Review P&P*, 102 (3), 330-36.

Kapicka, M. (2013), "Efficient Allocations in Dynamic Private Information Economies with Persistent Shocks: A First-Order Approach," *Review of Economic Studies,* 80 (3), 1027-1054.

Krahmer, D. and R. Strausz (2011), "Optimal Procurement Contracts with Pre-Project Planning," *Review of Economic Studies,* 78 (3), 1015-1041.

Krahmer, D. and R. Strausz (2015), "Optimal Sales Contracts with Withdrawal Rights," *Review of Economic Studies (forthcoming).*

Laffont, J.J. and D. Martimort (2002), *The Theory of Incentives,* Princeton University Press, Princeton, NJ.

Laffont J.J. and J. Tirole (1990) "Adverse Selection and Renegotiation in Procurement," *Review of Economic Studies*, 57 (4), 597–625.

Laffont J.J. and J. Tirole (1996), "Pollution Permits and Compliance Strategies," *Journal of Public Economics*, 62 (1–2), 85–125.

Madarasz, K. and A. Prat (2014), "Sellers with Misspecified Models," *working paper.*

Maestri, L. (2013), "Dynamic Contracting under Adverse Selection and Renegotiation," *working paper.*

Milgrom, P. (2004), *Putting Auction theory to Work,* Cambridge University Press.

Mussa M. and S. Rosen (1978), "Monopoly and Product Quality," *Journal of Economic Theory*, 18 (2), 301–317.

Myerson R. (1981), "Optimal Auction Design," *Mathematics of Operations Research*, 6 (1), 58-73.

Pavan, A., I. Segal, and J. Toikka (2014), "Dynamic Mechanism Design: A Myersonian Approach," *Econometrica*, 82 (2), 601-653.

Rey, P. and Salanie, B. (1990), "Long-Term, Short-Term and Renegotiation: On the Value of Commitment in Contracting," *Econometrica*, 58 (3), 597–619.

Rochet, J.C. (1987), "A Necessary and Sufficient Condition for Rationalizability in a Quasi-linear Context," *Journal of Mathematical Economics*, 16 (2), 191-200.

Rochet, J.C. and L. Stole (2003), "The Economics of Multidimensional Screening," *Advances in Economics and Econometrics, Eight World Congress*, Cambridge University Press, Vol. 1, 150-197.

Stole, L. (2001), "Lectures on the Theory of Contracts and Organizations," *mimeo*, The University of Chicago.

Royden, H. (1988), *Real Analysis*, Prentice Hall, Third Edition.

Rustichini, A. and A. Wolinsky.(1995), "Learning about Variable Demand in the Long Run," *Journal of Economic Dynamics and Control*, 19 (5–7), 1283–92.

Strulovici, B. (2011), "Contracts, Information Persistence, and Renegotiation," *mimeo.*

Townsend, R. M. (1982), "Optimal Multiperiod Contracts and the Gain from Enduring Relationships under Private Information," *Journal of Political Economy*, 90 (6), 1166–86.

Williams, N. (2011), "Persistent Private Information," *Econometrica,* 79 (4), 1233–1275.

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

10 Appendix

10.1 Proof of Lemma 1 and Corollary 1

We first show that all the constraints in the relaxed problem can be assumed to hold as equalities.

Lemma A1. *In a FO-relaxed problem: $IR_N(h^{t-1})$ can be assumed to hold as equality for all $h^{t-1} \in H^{t-1}$; $IC_{i,i+1}(h^{t-1})$ can be assumed to hold as an equality for all $h^{t-1} \in H^{t-1}$ and $i = 0, 1, ..., N-1$.*

Proof. The proof of this result is in the on line appendix. The appendix is available for download at http://www.mbattaglini.com/dyncontractingfoa. ■

We can now prove Lemma 1 and Corollary 1 together. We shall proceed by (backward) induction on t. Note that at $t = T$, Lemma A1 implies:

$$U(\theta_N|h^{T-1}) = 0 \text{ and } U(\theta_i|h^{T-1}) = \sum_{l=1}^{N-i} \Delta u(\theta_{i+l}|h^{T-1}; \mathbf{q}) \ \forall i \leq N-1. \tag{20}$$

where $\Delta u(\theta_{i+1}|h^{t-1}; \mathbf{q})$ is defined by: $\Delta u(\theta_{i+1}|h^{t-1}; \mathbf{q}) = u(\theta_i, q(\theta_{i+1}|h^{t-1})) - u(\theta_{i+1}, q(\theta_{i+1}|h^{t-1}))$. Similarly, for $t = T-1$, we have for $i \leq N-1$:

$$\begin{aligned} U(\theta_i|h^{T-2}) &= \Delta u(\theta_{i+1}|h^{T-2}; \mathbf{q}) + U(\theta_{i+1}|h^{T-2}) + \delta \sum_{k=0}^{N} (\alpha_{ik} - \alpha_{(i+1)k}) U(\theta_k|h^{T-2}, \theta_{i+1}) \\ &= \sum_{n=1}^{N-i} \left[\Delta u(\theta_{i+n}|h^{T-2}; \mathbf{q}) + \delta \sum_{k=0}^{N} (\alpha_{(i+n-1)k} - \alpha_{(i+n)k}) U(\theta_k|h^{T-2}, \theta_{i+n}) \right] \\ &= \sum_{n=1}^{N-i} \left[\Delta u(\theta_{i+n}|h^{T-2}; \mathbf{q}) + \delta \sum_{k=0}^{N-1} (\alpha_{(i+n-1)k} - \alpha_{(i+n)k}) \sum_{l=1}^{N-k} \Delta u(\theta_{k+l}|h^{T-2}, \theta_{i+n}; \mathbf{q})) \right] \end{aligned}$$

Now, let

$$\sum_{k=0}^{N-1} (\alpha_{(i+n-1)k} - \alpha_{(i+n)k}) \sum_{l=1}^{N-k} \Delta u(\theta_{k+l}|h^{T-2}, \theta_{i+n}; \mathbf{q})) = \sum_{k=0}^{N-1} \left(\alpha_{(i+n-1)k} - \alpha_{(i+n)k}\right) \sum_{l=1}^{N-k} Q_{k+l}, \tag{21}$$

where $Q_j = \Delta u(\theta_j|h^{T-2}, \theta_{i+n}; \mathbf{q}))$ for for any type θ_j. The right hand side of (21) can be written as:

$$\sum_{k=0}^{N-1} \left(\alpha_{(i+n-1)k} - \alpha_{(i+n)k}\right) \sum_{l=1}^{N-k} Q_{k+l} = \begin{bmatrix} \left(\alpha_{(i+n-1)0} - \alpha_{(i+n)0}\right) (Q_1 + ... + Q_N) \\ + \left(\alpha_{(i+n-1)1} - \alpha_{(i+n)1}\right) (Q_2 + ... + Q_N) \\ +... + \left(\alpha_{(i+n-1)(N-1)} - \alpha_{(i+n)(N-1)}\right) Q_N \end{bmatrix}$$

Rearranging the terms, we have:

$$\begin{aligned}
&\sum_{k=0}^{N-1}\left(\alpha_{(i+n-1)k}-\alpha_{(i+n)k}\right)\sum_{l=1}^{N-k}Q_{k+l}\\
=&\left[\begin{array}{c}\left(\alpha_{(i+n-1)0}-\alpha_{(i+n)0}\right)Q_1\\ +\left(\left(\alpha_{(i+n-1)0}+\alpha_{(i+n-1)1}\right)-\left(\alpha_{(i+n)0}+\alpha_{(i+n)1}\right)\right)Q_2\\ +...+\left(\left(\alpha_{(i+n-1)0}+...+\alpha_{(i+n-1)(N-1)}\right)-\left(\alpha_{(i+n)0}+...+\alpha_{(i+n)(N-1)}\right)\right)Q_N\end{array}\right]\\
=&\sum_{k=1}^{N}\Delta F(\theta_k|\theta_{i+n})Q_k
\end{aligned}$$

where, we recall, $\Delta F\left(\theta_j\left|\theta_i\right.\right)=F\left(\theta_j\left|\theta_i\right.\right)-F\left(\theta_j\left|\theta_{i-1}\right.\right)$. This implies that:

$$\sum_{k=0}^{N-1}(\alpha_{(i+n-1)k}-\alpha_{(i+n)k})\sum_{l=1}^{N-k}\Delta u(\theta_{k+l}|h^{T-2},\theta_{i+n};\mathbf{q}))=\sum_{k=1}^{N}\Delta F(\theta_k\left|\theta_{i+n}\right.)\Delta u(\theta_k|h^{T-2},\theta_{i+n};\mathbf{q}))$$

It follows that we can write:

$$\begin{aligned}
U(\theta_i|h^{T-2}) &= \sum_{n=1}^{N-i}\left[\Delta u(\theta_{i+n}|h^{T-2};\mathbf{q})+\delta\sum_{k=1}^{N}\Delta F(\theta_k\left|\theta_{i+n}\right.)\Delta u(\theta_k|h^{T-2},\theta_{i+n};\mathbf{q}))\right]\\
&= \sum_{n=1}^{N-i}\left[\begin{array}{c}\Delta u(\theta_{i+n}|h^{T-2};\mathbf{q})\\ +\sum_{\hat{h}\in\widehat{H}(h^{T-2},\theta_{i+n})}\sum_{\tau>T-1}\delta^{\tau-T-1}\prod_{k=T}^{\tau}\Delta F\left(\hat{h}_k\left|\hat{h}_{k-1}\right.\right)\Delta u(\hat{h}_\tau|\hat{h}^{\tau-1};\mathbf{q}))\end{array}\right] \qquad (22)
\end{aligned}$$

where, we recall, $\widehat{H}(h^t)$ is the set of histories following h^t in which all realizations after t are lower than θ_0.

It is easy to see that (20) and (22) prove the statement in Corollary 1 and in Lemma 1 respectively for $t=T$ and $t=T-1$. We therefore conclude that our hypothesis holds for $t\geq T-1$. Next, suppose it holds for $t+1$ where $t\geq T-2$. We want to show that it holds for t. We have,

$$\begin{aligned}
U(\theta_i|h^{t-1}) &= \Delta u\left(\theta_{i+1}|h^{t-1};\mathbf{q}\right)+U\left(\theta_{i+1}|h^{t-1}\right)+\delta\sum_{k=0}^{N}(\alpha_{ik}-\alpha_{(i+1)k})U(\theta_k|h^{t-1},\theta_{i+1}) \qquad (23)\\
&= \sum_{n=1}^{N-i}\left[\Delta u\left(\theta_{i+n}|h^{t-1};\mathbf{q}\right)+\delta\sum_{k=0}^{N}(\alpha_{(i+n-1)k}-\alpha_{(i+n)k})U\left(\theta_k|h^{t-1},\theta_{i+n}\right)\right]\\
&= \sum_{n=1}^{N-i}\left[\Delta u\left(\theta_{i+n}|h^{t-1};\mathbf{q}\right)+\delta\sum_{k=0}^{N-1}(\alpha_{(i+n-1)k}-\alpha_{(i+n)k})\sum_{m=1}^{N-k}\left(\Delta u\left(\theta_{k+m}|h^{t-1},\theta_{i+n};\mathbf{q}\right)+\right.\right.\\
&\left.\left.\sum_{\hat{h}\in\widehat{H}(h^{t-1},\theta_{i+n},\theta_{k+m})}\sum_{\tau>t+1}\delta^{\tau-(t+1)}\prod_{\iota=t+2}^{\tau}\Delta F\left(\hat{h}_\iota\left|\hat{h}_{\iota-1}\right.\right)\Delta u\left(\hat{h}_\tau|\hat{h}^{\tau-1};\mathbf{q}\right)\right)\right],
\end{aligned}$$

where the third equality follows from the induction hypothesis. Now,

$$\sum_{k=0}^{N-1}(\alpha_{(i+n-1)k}-\alpha_{(i+n)k})\sum_{m=1}^{N-k}\Delta u(\theta_{k+m}|h^{t-1},\theta_{i+n};\mathbf{q})=\sum_{k=1}^{N}\Delta F(\theta_k\,|\theta_{i+n})\Delta u(\theta_k|h^{t-1},\theta_{i+n};\mathbf{q}), \tag{24}$$

and,

$$\delta\sum_{k=0}^{N-1}(\alpha_{(i+n-1)k}-\alpha_{(i+n)k})\sum_{m=1}^{N-k}\sum_{\hat{h}\in\widehat{H}(h^{t-1},\theta_{i+n},\theta_{k+m})}\sum_{\tau>t+1}\delta^{\tau-(t+1)}\prod_{\iota=t+2}^{\tau}\Delta F\left(\hat{h}_\iota\left|\hat{h}_{\iota-1}\right.\right)\Delta u\left(\hat{h}_\tau|\hat{h}^{\tau-1};\mathbf{q}\right)$$

$$=\delta\sum_{k=0}^{N-1}(\alpha_{(i+n-1)k}-\alpha_{(i+n)k})\sum_{m=1}^{N-k}Q_{k+m},$$

where, $Q_l=\sum_{\hat{h}\in\widehat{H}(h^{t-1},\theta_{i+n},\theta_l)}\sum_{\tau>t+1}\delta^{\tau-(t+1)}\prod_{\iota=t+2}^{\tau}\Delta F\left(\hat{h}_\iota\left|\hat{h}_{\iota-1}\right.\right)\Delta u\left(\hat{h}_\tau|\hat{h}^{\tau-1};\mathbf{q}\right)$. As before, after some algebraic manipulation, this becomes:

$$\delta\sum_{k=0}^{N-1}(\alpha_{(i+n-1)k}-\alpha_{(i+n)k})\sum_{m=1}^{N-k}Q_{k+m}=\delta\sum_{k=1}^{N}\Delta F\left(\theta_k|\theta_{i+n}\right)Q_k \tag{25}$$

Combining (24) and (25) we obtain:

$$\delta\sum_{k=1}^{N}\Delta F\left(\theta_k|\theta_{i+n}\right)\left[\Delta u(\theta_k|h^{t-1},\theta_{i+n};\mathbf{q})+Q_k\right] \tag{26}$$

$$=\delta\sum_{k=1}^{N}\Delta F\left(\theta_k|\theta_{i+n}\right)\left[\begin{array}{c}\Delta u(\theta_k|h^{t-1},\theta_{i+n};\mathbf{q})+\\ \sum_{\hat{h}\in\widehat{H}(h^{t-1},\theta_{i+n},\theta_k)}\sum_{\tau>t+1}\delta^{\tau-(t+1)}\prod_{\iota=t+2}^{\tau}\Delta F\left(\hat{h}_\iota\left|\hat{h}_{\iota-1}\right.\right)\Delta u\left(\hat{h}_\tau|\hat{h}^{\tau-1};\mathbf{q}\right)\end{array}\right]$$

$$=\sum_{\hat{h}\in\widehat{H}(h^{t-1},\theta_{i+n})}\sum_{\tau>t}\delta^{\tau-t}\prod_{\iota=t+1}^{\tau}\Delta F\left(\hat{h}_\iota\left|\hat{h}_{\iota-1}\right.\right)\Delta u\left(\hat{h}_\tau|\hat{h}^{\tau-1};\mathbf{q}\right)$$

Combining (23) and (26), we obtain:

$$U\left(\theta_i|h^{t-1}\right)=\sum_{n=1}^{N-i}\left[\Delta u\left(\theta_{i+n}|h^{t-1};\mathbf{q}\right)+\sum_{\hat{h}\in\widehat{H}(h^{t-1},\theta_{i+n})}\sum_{\tau>t}\delta^{\tau-t}\prod_{\iota=t+1}^{\tau}\Delta F\left(\hat{h}_\iota\left|\hat{h}_{\iota-1}\right.\right)\Delta u(\hat{h}_\tau|\hat{h}^{\tau-1};\mathbf{q})\right].$$

Note that:

$$\Delta u(\theta_{i+1}|h^{t-1};\mathbf{q})=u(\theta_i,q(\theta_{i+1}|h^{t-1}))-u(\theta_{i+1},q(\theta_{i+1}|h^{t-1}))=\int_{\theta_{i+1}}^{\theta_i}u_\theta(x,q(\theta_{i+1}|h^{t-1}))dx$$

It follows that we have:

$$U\left(\theta_i|h^{t-1}\right)=\sum_{n=1}^{N-i}\left[\begin{array}{c}\int_{\theta_{i+n}}^{\theta_{i+n-1}}u_\theta(x,q(\theta_{i+n}|h^{t-1}))dx+\\ \sum_{\hat{h}\in\widehat{H}(h^{t-1},\theta_{i+n})}\sum_{\tau>t}\delta^{\tau-t}\prod_{\iota=t+1}^{\tau}\Delta F\left(\hat{h}_\iota\left|\hat{h}_{\iota-1}\right.\right)\int_{\hat{h}_\tau}^{\hat{h}_\tau+\Delta\theta}u_\theta(x,q(\hat{h}_\tau|\hat{h}^{\tau-1}))dx\end{array}\right].$$

This proves Corollary 1. Subtracting $U(\theta_{i+1}|h^{t-1})$ and dividing by $\Delta\theta$ from the above expression gives us Lemma 1.

10.2 Proof of Proposition 2

Recall that $\Delta U(\theta_k \left| h^{t-1}, \theta_i \right.) = U(\theta_k \left| h^{t-1}, \theta_i \right.) - U(\theta_k \left| h^{t-1}, \theta_{i+1} \right.)$. We start with some useful lemmas.

Lemma A2. *If $q(\theta_i|h^{t-1})$ and $\Delta U(\theta_k \left| h^{t-1} \right.)$ are non increasing in, respectively, i and k for any h^{t-1}, then (5) implies that local upward incentive compatibility constraints are satisfied.*

Proof. The proof of this result is in the online appendix. ■

Lemma A3. *If $q(\theta_i|h^{t-1})$ and $\Delta U(\theta_k \left| h^{t-1} \right.)$ are non increasing in, respectively, i and k for any h^{t-1} and (5) holds, then the local incentive compatibility constraints imply the global incentive compatibility constraints.*

Proof. The proof of this result is in the online appendix. ■

Given the lemmas presented above, Proposition 2 is proven if we establish that when the allocation is monotonic as defined in Definition 2, then $q(\theta_i|h^{t-1})$ and $\Delta U(\theta_k \left| h^{t-1}, \theta_i \right.)$ are non increasing in i for any h^{t-1}. The fact that $q(\theta_i|h^{t-1})$ is non increasing in i for any h^{t-1} is an immediate consequence of the monotonicity. The fact that $\Delta U(\theta_k \left| h^{t-1}, \theta_i \right.)$ is non increasing in i for any h^{t-1} is established by the following result.

Lemma A4. *If the allocation is monotonic as in Definition 2, then $\Delta U(\theta_k \left| h^{t-1} \right.)$ is non increasing in k $\forall h^{t-1}$.*

Proof. Note first that $U(\theta_N \left| h^{t-1}, \theta_i \right.) = U(\theta_N \left| h^{t-1}, \theta_{i+1} \right.) = 0$, so $\Delta U(\theta_N \left| h^{t-1}, \theta_i \right.) = 0$. By Lemma 1, we have:

$$\begin{aligned} U(\theta_{N-1}|h^{t-1},\theta_i) &= \int_{\theta_N}^{\theta_{N-1}} u_\theta(x, q(\theta_N|h^{t-1},\theta_i))dx \\ &+ \sum_{\hat{h}\in\widehat{H}(h^{t-1},\theta_i,\theta_{N-1})} \sum_{\tau>t+1} \delta^{\tau-t-1} \left[\begin{array}{c} \prod_{\iota=t+2}^{\tau} \Delta F\left(\hat{h}_\iota \left| \hat{h}_{\iota-1}\right.\right) \\ \cdot \int_{\hat{h}_\tau}^{\hat{h}_\tau+\Delta\theta} u_\theta(x, q(\hat{h}_\tau|\hat{h}^{\tau-1}))dx \end{array} \right] \end{aligned} \quad (27)$$

It is useful to write this expression with a different notation. Let $\widehat{H}_t(i)$ be set of realizations of length $T-t$ that start with the first element equal to θ_i (we denote ${}_th$ is the typical element of $\widehat{H}_t(i)$, so ${}_th_1 = \theta_i$). A history $h^\tau \in \widehat{H}\left(h^t\right)$ with $(t+1)$-th element equal to θ_i ($h^\tau_{t+1} = \theta_i$) is then $h^\tau = \{h^t, {}_th^{\tau-t}\}$ for ${}_th \in \widehat{H}_t(i)$ (by convention we write $h^t = \left\{h^t, {}_th^0\right\}$). We can then write:

$$\begin{aligned} U(\theta_{N-1} \left| h^{t-1}, \theta_i \right.) &= \int_{\theta_N}^{\theta_{N-1}} u_\theta(x, q(\theta_N|h^{t-1},\theta_i))dx \\ &+ \sum_{{}_th\in\widehat{H}_t(N-1)} \sum_{\tau>t+1} \delta^{\tau-t-1} \left[\begin{array}{c} \prod_{l=t+2}^{\tau} \Delta F\left({}_th_l \left| {}_th_{l-1}\right.\right) \cdot \\ \int_{{}_th_\tau}^{{}_th_\tau+\Delta\theta} u_\theta(x, q({}_th_\tau|h^{t-1},\theta_i,{}_th^{\tau-t-1})) \end{array} \right] \end{aligned} \quad (28)$$

Similarly we can write:

$$U(\theta_{N-1}\left|h^{t-1},\theta_{i+1}\right.) = \int_{\theta_N}^{\theta_{N-1}} u_\theta(x,q(\theta_N|h^{t-1},\theta_{i+1}))dx + \sum_{{}_th\in\widehat{H}_t(N-1)}\sum_{\tau>t+1}\delta^{\tau-t-1}\left[\begin{array}{c}\prod_{l=t+2}^{\tau}\Delta F\left({}_th_l\left|{}_th_{l-1}\right.\right)\cdot \\ \int_{{}_th_\tau}^{{}_th_\tau+\Delta\theta} u_\theta(x,q({}_th_\tau|h^{t-1},\theta_{i+1,t}\,h^{\tau-t-1}))dx\end{array}\right] \quad (29)$$

Therefore we have:

$$\begin{aligned}&\Delta U(\theta_{N-1}\left|h^{t-1},\theta_i\right.) = \\ &\int_{\theta_N}^{\theta_{N-1}}\left[u_\theta(x,q(\theta_N|h^{t-1},\theta_i))-u_\theta(x,q(\theta_N|h^{t-1},\theta_{i+1}))\right]dx \\ &+\sum_{{}_th\in H_t(N-1)}\sum_{\tau>t+1}\delta^{\tau-t-1}\left[\begin{array}{c}\prod_{l=t+2}^{\tau}\Delta F\left({}_th_l\left|{}_th_{l-1}\right.\right) \\ \cdot\int_{{}_th_\tau}^{{}_th_\tau-\Delta\theta}\left[\begin{array}{c}u_\theta(x,q({}_th_\tau|h^{t-1},\theta_{i,t}\,h^{\tau-t-1})) \\ -u_\theta(x,q({}_th_\tau|h^{t-1},\theta_{i+1,t}\,h^{\tau-t-1}))\end{array}\right]\end{array}\right]dx\end{aligned}$$

Note that by monotonicity, we must have $q(\theta_N|h^{t-1},\theta_i)-q(\theta_N|h^{t-1},\theta_{i+1})\geq 0$ and

$$q({}_th_\tau|h^{t-1},\theta_{i,t}\,h^{\tau-t-1})-q({}_th_\tau|h^{t-1},\theta_{i+1,t}\,h^{\tau-t-1})\geq 0$$

The above condition plus the single crossing condition (Assumption 1) imply that $\Delta U(\theta_{N-1}\left|h^{t-1},\theta_i\right.)\geq \Delta U(\theta_N\left|h^{t-1},\theta_i\right.)$. Assume now that $\Delta U(\theta_j\left|h^{t-1},\theta_i\right.)$ is monotonic in j for $j\geq m$. We show below that $\Delta U(\theta_{m-1}\left|h^{t-1},\theta_i\right.)\geq \Delta U(\theta_m\left|h^{t-1},\theta_i\right.)$, the result then follows from induction. Applying Lemma 1 and using the notation developed above, we have:

$$\begin{aligned}&\Delta U(\theta_{m-1}\left|h^{t-1},\theta_i\right.) \\ =\;&\Delta U(\theta_m\left|h^{t-1},\theta_i\right.)+\int_{\theta_N}^{\theta_{N-1}}\left[u_\theta(x,q(\theta_m|h^{t-1},\theta_i))-u_\theta(x,q(\theta_m|h^{t-1},\theta_{i+1}))\right]dx \\ &+\sum_{{}_th\in H_t(m-1)}\sum_{\tau>t+1}\delta^{\tau-t-1}\left[\begin{array}{c}\prod_{l=t+2}^{\tau}\Delta F\left({}_th_l\left|{}_th_{l-1}\right.\right) \\ \cdot\int_{{}_th_\tau}^{{}_th_\tau-\Delta\theta}\left[\begin{array}{c}u_\theta(x,q({}_th_\tau|h^{t-1},\theta_{i,t}\,h^{\tau-t-1})) \\ -u_\theta(x,q({}_th_\tau|h^{t-1},\theta_{i+1,t}\,h^{\tau-t-1}))\end{array}\right]\end{array}\right]dx\end{aligned}$$

Thus, the single crossing condition and monotonicity of the allocation imply $\Delta U(\theta_{m-1}\left|h^{t-1},\theta_i\right.)\geq \Delta U(\theta_m\left|h^{t-1},\theta_i\right.)$. ■

10.3 Proof of Propositions 3

For $0 < i < N$, define $\Psi_i(f_\alpha)$ as:

$$\begin{aligned}\Psi_i(f_\alpha) &= \left[\frac{\Delta F_\alpha(\theta_i|\theta_{i+1})}{f_\alpha(\theta_i|\theta_{i+1})} \cdot \frac{\Delta F_\alpha(\theta_{i+1}|\theta_i)}{f_\alpha(\theta_{i+1}|\theta_i)}\right] \\ &= \frac{\sum_{k=i+1}^{N}[f_\alpha(\theta_k|\theta_i) - f_\alpha(\theta_k|\theta_{i-1})]}{f_\alpha(\theta_{i+1}|\theta_i)} \cdot \frac{\sum_{k=i}^{N}[f_\alpha(\theta_k|\theta_{i+1}) - f_\alpha(\theta_k|\theta_i)]}{f_\alpha(\theta_i|\theta_{i+1})}\end{aligned}$$

In Lemma A5 we prove that if:

$$\lim_{\alpha\to 1}\Psi_i(f_\alpha) \neq 1 \text{ for some } i \in (0,N) \tag{30}$$

then the optimal contract is not monotonic as $\alpha \to 1$. In Lemma A6 we prove that condition (30) is generically satisfied.

Lemma A5. *For any μ, δ, $|\Theta| > 2$, $T > 2$, if $\lim_{\alpha\to 1}\Psi_i(f_\alpha) \neq 1$ for some $i \in (0,N)$, then there is an $\alpha^* < 1$ such that the FO-optimal contract is not monotonic for any $\alpha > \alpha^*$.*

Proof. Suppose first $D = \lim_{\alpha\to 1}\Psi_i(f_\alpha) < 1$. We show that $\lim_{\alpha\to 1} q(\theta_i|\theta_i,\theta_{i+1}) > \lim_{\alpha\to 1} q(\theta_i|\theta_i,\theta_i)$. Note that

$$s_q(\theta_i, q(\theta_i|\theta_i,\theta_i)) \leq \frac{1-\sum_{k=i}^{N}\mu_k}{\mu_i} \cdot \left[\frac{\Delta F_\alpha(\theta_i|\theta_i)}{f_\alpha(\theta_i|\theta_i)} \cdot \frac{\Delta F_\alpha(\theta_i|\theta_i)}{f_\alpha(\theta_i|\theta_i)}\right] \cdot \int_{\theta_i}^{\theta_{i-1}} u_{\theta,q}(x, q(\theta_i|\theta_i,\theta_i)dx$$

and

$$s_q(\theta_i, q(\theta_i|\theta_i,\theta_{i+1})) \leq \frac{1-\sum_{k=i}^{N}\mu_k}{\mu_i} \cdot \left[\frac{\Delta F_\alpha(\theta_i|\theta_{i+1})}{f_\alpha(\theta_i|\theta_{i+1})} \cdot \frac{\Delta F_\alpha(\theta_{i+1}|\theta_i)}{f_\alpha(\theta_{i+1}|\theta_i)}\right] \cdot \int_{\theta_i}^{\theta_{i-1}} u_{\theta,q}(x, q(\theta_i|\theta_i,\theta_{i+1})dx.$$

Let $q_1 = \lim_{\alpha\to 1} q(\theta_i|\theta_i,\theta_i)$. Distortions converge to 1 along constant histories, so $q_1 = q(\theta_i|h^0)$.[45] By Assumption 3, since the static optimum is an interior solution, we have

$$s_q(\theta_i, q_1) = \frac{1-\sum_{k=i}^{N}\mu_k}{\mu_i} \cdot \int_{\theta_i}^{\theta_{i-1}} u_{\theta,q}(x, q_1)dx.$$

Also, letting $q_2 = \lim_{\alpha\to 1} q(\theta_i|\theta_i,\theta_{i+1})$, we have

$$s_q(\theta_i, q_2) \leq \frac{1-\sum_{k=i}^{N}\mu_k}{\mu_i} \cdot D \cdot \int_{\theta_i}^{\theta_{i-1}} u_{\theta,q}(x, q_2)dx.$$

It follows that:

$$\begin{aligned}\Phi_q(\theta_i, q_1) &= s_q(\theta_i, q_1) - \frac{1-\sum_{k=i}^{N}\mu_k}{\mu_i} \cdot \int_{\theta_i}^{\theta_{i-1}} u_{\theta,q}(x, q_1)dx \\ &= 0 \geq s_q(\theta_i, q_2) - \frac{1-\sum_{k=i}^{N}\mu_k}{\mu_i} \cdot D \cdot \int_{\theta_i}^{\theta_{i-1}} u_{\theta,q}(x, q_2)dx \\ &> s_q(\theta_i, q_2) - \frac{1-\sum_{k=i}^{N}\mu_k}{\mu_i} \cdot \int_{\theta_i}^{\theta_{i-1}} u_{\theta,q}(x, q_2)dx = \Phi_q(\theta_i, q_2),\end{aligned}$$

[45] This follows from the fact that $\lim_{\alpha\to 1}\Delta F_\alpha(\theta_i|\theta_i)/f_\alpha(\theta_i|\theta_i) = 1$.

where the strict inequality follows from $D < 1$. Since Φ is concave, we have $q_2 > q_1$.

Next, suppose $D = \lim_{\alpha\to 1} \Psi_i(f_\alpha) > 1$. Then, analogous to the steps above we show that $\lim_{\alpha\to 1} q(\theta_{i+1}|\theta_{i+1}, \theta_{i+1}) > \lim_{\alpha\to 1} q(\theta_{i+1}|\theta_{i+1}, \theta_i)$. Letting $q_3 = \lim_{\alpha\to 1} q(\theta_{i+1}|\theta_{i+1}, \theta_i)$, and $q_4 = \lim_{\alpha\to 1} q(\theta_{i+1}|\theta_{i+1}, \theta_{i+1})$, we get

$$s_q(\theta_i, q_4) = \frac{1 - \sum_{k=i+1}^{N} \mu_k}{\mu_{i+1}} \cdot \int_{\theta_{i+1}}^{\theta_i} u_{\theta,q}(x, q_4)dx$$

and,

$$s_q(\theta_i, q_3) \leq \frac{1 - \sum_{k=i+1}^{N} \mu_k}{\mu_{i+1}} \cdot D \cdot \int_{\theta_{i+1}}^{\theta_i} u_{\theta,q}(x, q_3)dx$$

Thus, using $D > 1$, we obtain $\Phi_q(\theta_{i+1}, q_3) < \Phi_q(\theta_{i+1}, q_4)$, implying $q_4 > q_3$. ■

We now prove that (30) is generically satisfied. Define $\Gamma_i = \{f_\alpha \in \Lambda \mid \lim_{\alpha\to 1} \Psi_i(f_\alpha) \neq 1\}$. We will show that Γ_i is open and dense in Λ, thereby establishing that $\Gamma = \cup_{i=1}^{N-1}\Gamma_i$ is open and dense in Λ, proving our result.

Lemma A6. *(30) is a generic property of* Λ.

Proof. Note that Λ is a space of functions, say χ, from $[0, 1]$ to $[0, 1]^{N+1} \times [0, 1]^{N+1}$, where $\chi(\alpha) = f_\alpha$ is a full support Markov matrix that satisfies Assumption 2 and has all diagonal entries converge to 1 as α converges to 1**.** Endow this space with the sup norm:

$$\|f\| = \sup_{\alpha\in[0,1]} \max_{i\in\{0,..,N\},\ j\in\{0,..,N\}} f_\alpha(\theta_i|\theta_j).$$

Given an $i \in (0, N)$, we proceed in two steps.

Step 1. We first prove that Γ_i is open. Assume not. Then for some $f_\alpha \in \Gamma_i$ and any ε-neighborhood $N_\varepsilon(f_\alpha)$ of f_α we can find a $f'_\alpha \in \overline{\Gamma}_i = \Lambda \backslash \Gamma_i$. It follows that there exists a sequence $(f^n_\alpha) \in \overline{\Gamma}_i$ such that $f^n_\alpha \to f_\alpha$. By definition $\lim_{\alpha\to 1} \Psi_i(f^n_\alpha) = 1$ for all n. Since $\Psi_i(f)$ is continuous in f, this implies

$$\lim_{\alpha\to 1} \Psi(f_\alpha) = \lim_{\alpha\to 1} \lim_{n\to\infty} \Psi(f^n_\alpha) = \lim_{n\to\infty} \lim_{\alpha\to 1} \Psi(f^n_\alpha) = 1$$

proving that $f_\alpha \in \overline{\Gamma}_i$, a contradiction.

Step 2. Next we prove that the Γ_i is dense in Λ. For this we need to show that for any $f_\alpha \in \Lambda$ and $\varepsilon > 0$, there is a function f'_α such that and $\|f - f'\| < \varepsilon$ and $f'_\alpha \in \Gamma_i$. If $f_\alpha \in \Gamma_i$, then the result is immediate. Assume therefore that $f_\alpha \in \overline{\Gamma}_i$. Let $\tilde{f}$ be a given stochastic matrix that satisfies first-order stochastic dominance strictly and with $\tilde{f}(\theta_k|\theta_i) > 0$ for all i**,** k. Fix $\bar{\varepsilon}$, and define

$$f'_\alpha = \varepsilon(\alpha)\tilde{f} + (1 - \varepsilon(\alpha))\, f_\alpha$$

where $\varepsilon(\alpha)$ is a non negative function of α such that $\varepsilon(\alpha) \leq \bar{\varepsilon}\ \forall\alpha$, and $\varepsilon(\alpha) \to 0$ as $\alpha \to 1$. It is

easy to see that $\|f - f'\| < \overline{\varepsilon}$. Moreover,

$$\Psi_i(f'_\alpha) = \frac{\sum_{k=i}^{N} \left[\begin{array}{c} \frac{f_\alpha(\theta_k|\theta_{i+1}) - f_\alpha(\theta_k|\theta_i)}{f_\alpha(\theta_i|\theta_{i+1})} \\ + \frac{\varepsilon(\alpha)}{1-\varepsilon(\alpha)} \frac{\tilde{f}(\theta_k|\theta_{i+1}) - \tilde{f}(\theta_k|\theta_i)}{f_\alpha(\theta_i|\theta_{i+1})} \end{array} \right]}{1 + \frac{\varepsilon(\alpha)}{1-\varepsilon(\alpha)} \frac{\tilde{f}(\theta_i|\theta_{i+1})}{f_\alpha(\theta_i|\theta_{i+1})}} \cdot \frac{\sum_{k=i+1}^{N} \left[\begin{array}{c} \frac{f_\alpha(\theta_k|\theta_i) - f_\alpha(\theta_k|\theta_{i-1})}{f_\alpha(\theta_{i+1}|\theta_i)} \\ + \frac{\varepsilon(\alpha)}{1-\varepsilon(\alpha)} \frac{\tilde{f}(\theta_k|\theta_i) - \tilde{f}(\theta_k|\theta_{i-1})}{f_\alpha(\theta_{i+1}|\theta_i)} \end{array} \right]}{1 + \frac{\varepsilon(\alpha)}{1-\varepsilon(\alpha)} \frac{\tilde{f}(\theta_{i+1}|\theta_i)}{f_\alpha(\theta_{i+1}|\theta_i)}} \tag{31}$$

Since $f_\alpha(\theta_i|\theta_{i+1})$ and $f_\alpha(\theta_{i+1}|\theta_i)$ both converge to zero as $\alpha \to 1$, there are two separate cases to consider: (i) $\lim_{\alpha\to 1} \frac{f_\alpha(\theta_i|\theta_{i+1})}{f_\alpha(\theta_{i+1}|\theta_i)} = \beta$ for some $\beta \geq 0$, and (ii) $\lim_{\alpha\to 1} \frac{f_\alpha(\theta_{i+1}|\theta_i)}{f_\alpha(\theta_i|\theta_{i+1})} = 0$. Consider the first case. Choose $\varepsilon(\alpha) = \min\{\bar{\varepsilon}, f_\alpha(\theta_i|\theta_{i+1})\}$.

Now, if $\lim_{\alpha\to 1} \Psi_i(f'_\alpha) = \infty$, then $f'_\alpha \in \Gamma_i$, and we are done. So, let $\lim_{\alpha\to 1} \Psi_i(f'_\alpha)$ be a finite constant. With some simple algebra, we can write

$$\lim_{\alpha\to 1} \Psi_i(f'_\alpha) = \frac{A + \sum_{k=i}^{N} \left(\tilde{f}(\theta_k|\theta_{i+1}) - \tilde{f}(\theta_k|\theta_i) \right)}{1 + \tilde{f}(\theta_i|\theta_{i+1})} \cdot \frac{B + \beta \sum_{k=i+1}^{N} \left(\tilde{f}(\theta_k|\theta_i) - \tilde{f}(\theta_k|\theta_{i-1}) \right)}{1 + \beta \tilde{f}(\theta_{i+1}|\theta_i)} \tag{32}$$

where A, B are non-negative finite constants.

Since $\tilde{f}$ is a generic transition matrix that satisfies first-order stochastic dominance strictly, we will have $f'_\alpha \in \Gamma_i$ by construction. To see this, assume $\lim_{\alpha\to 1} \Psi_i(f'_\alpha) = 1$ and consider a matrix $\widehat{f}$ that is equal to $\tilde{f}$ except that $\widehat{f}(\theta_i|\theta_{i+1}) = \tilde{f}(\theta_i|\theta_{i+1}) - \epsilon$ and $\widehat{f}(\theta_{i+1}|\theta_{i+1}) = \tilde{f}(\theta_{i+1}|\theta_{i+1}) + \epsilon$ where $\epsilon > 0$ is a arbitrarily small. The new matrix still satisfies first order stochastic dominance strictly, and using equation (32) it is easy to see that

$$\lim_{\alpha\to 1} \Psi_i(\widehat{f}_\alpha) = \frac{1 + \tilde{f}(\theta_i|\theta_{i+1})}{1 + \tilde{f}(\theta_i|\theta_{i+1}) - \epsilon} > 1$$

We can thus repeat the argument presented above choosing $\widehat{f}$ instead of $\tilde{f}$ and obtain $\widehat{f}_\alpha \in \Gamma_i$

Finally, for the second case where $\lim_{\alpha\to 1} \frac{f_\alpha(\theta_{i+1}|\theta_i)}{f_\alpha(\theta_i|\theta_{i+1})} = 0$ we choose $\varepsilon(\alpha) = \min\{\bar{\varepsilon}, f_\alpha(\theta_{i+1}|\theta_i)\}$ and proceed as in the first case to establish denseness of Γ_i in Λ . ■

10.4 Proof of Proposition 4

We prove that for any μ, $|\Theta| > 2$ and a generic transition probability function, there exists an $\alpha^* < 1$, T^*, and $\delta^* < 1$ such that the first-order approach fails to be verified for any $\alpha > \alpha^*$, $T \geq T^*$ and $\delta > \delta^*$ if $\lim_{\alpha\to 1} \Psi_i(f_\alpha) \neq 1$ for some $i \in (0, N)$. Given this, the statement of the proposition follows from Lemma A6.

Note that as $\alpha \to 1$, $\Delta F_\alpha(\theta_k|\theta_k) \to 1$ and $\Delta F_\alpha(\theta_j|\theta_k) \to 0$ for $\forall k \neq j$. We have two cases to consider:

Case 1: $D = \lim_{\alpha\to 1} \Psi_i(f_\alpha) < 1$. We prove the result by showing that the FO-optimal contract violates the second period global incentive constraint $IC_{i-1,i+1}(\theta_i)$. To this end, we first make a useful observation.

Lemma A7. *$IC_{i-1,i+1}(h^{t-1})$ holds if and only if*

$$\int_{\theta_i}^{\theta_{i-1}} \begin{bmatrix} u_\theta(x, q(\theta_i|h^{t-1})) \\ -u_\theta(x, q(\theta_{i+1}|h^{t-1})) \end{bmatrix} dx + \delta \sum_{k=0}^{N} \begin{bmatrix} (f_\alpha(\theta_k|\theta_{i-1}) - f_\alpha(\theta_k|\theta_i)) \cdot \\ \left(U(\theta_k|h^{t-1}, \theta_i) - U(\theta_k|h^{t-1}, \theta_{i+1})\right) \end{bmatrix} \geq 0 \tag{33}$$

where $U(\theta_k|h^{t-1}, \theta_i) = U^*(\theta_k \left| h^{t-1}, \theta_i; \mathbf{q}\right)$, as defined in (5) in the paper.

Proof. The global incentive compatibility constraint $IC_{i-1,i+1}(h^{t-1})$ can be written as:

$$\begin{aligned} U(\theta_{i-1}|h^{t-1}) - U(\theta_{i+1}|h^{t-1}) \quad \geq \quad & u(\theta_{i-1}, q(\theta_{i+1}|h^{t-1})) - u(\theta_{i+1}, q(\theta_{i+1}|h^{t-1})) \\ & + \delta \sum_{k=0}^{N} \left(f_\alpha(\theta_k|\theta_{i-1}) - f_\alpha(\theta_k|\theta_{i+1})\right) U(\theta_k|h^{t-1}, \theta_{i+1}) \end{aligned} \tag{34}$$

Note that

$$U(\theta_{i-1}|h^{t-1}) - U(\theta_{i+1}|h^{t-1}) = \left(U(\theta_{i-1}|h^{t-1}) - U(\theta_i|h^{t-1})\right) + \left(U(\theta_i|h^{t-1}) - U(\theta_{i+1}|h^{t-1})\right).$$

So using $IC_{i-1,i}(h^{t-1})$ and $IC_{i,i+1}(h^{t-1})$, we have:

$$\begin{aligned} & U(\theta_{i-1}|h^{t-1}) - U(\theta_{i+1}|h^{t-1}) - \begin{bmatrix} u(\theta_{i-1}, q(\theta_{i+1}|h^{t-1})) - u(\theta_{i+1}, q(\theta_{i+1}|h^{t-1})) \\ +\delta \sum\limits_{k=0}^{N} \left(f_\alpha(\theta_k|\theta_{i-1}) - f_\alpha(\theta_k|\theta_{i+1})\right) U(\theta_k|h^t, \theta_{i+1}) \end{bmatrix} \\ = & \begin{bmatrix} u(\theta_{i-1}, q(\theta_i|h^{t-1})) - u(\theta_i, q(\theta_i|h^{t-1})) \\ +u(\theta_i, q(\theta_{i+1}|h^{t-1})) - u(\theta_{i-1}, q(\theta_{i+1}|h^{t-1})) \end{bmatrix} + \delta \sum_{k=0}^{N} \begin{pmatrix} (f_\alpha(\theta_k|\theta_{i-1}) - f_\alpha(\theta_k|\theta_i)) \\ \cdot \left(U(\theta_k|h^{t-1}, \theta_i) - U(\theta_k|h^{t-1}, \theta_{i+1})\right) \end{pmatrix} \end{aligned} \tag{35}$$

Using (34) and (35), it follows that that $IC_{i-1,i}(h^{t-1})$ holds if and only if (33) holds. ■

So, $IC_{i-1,i+1}(\theta_i)$ holds if and only if

$$\int_{\theta_i}^{\theta_{i-1}} \begin{bmatrix} u_\theta(x, q(\theta_i|\theta_i)) \\ -u_\theta(x, q(\theta_{i+1}|\theta_i)) \end{bmatrix} dx + \delta \sum_{k=0}^{N} \begin{bmatrix} (f_\alpha(\theta_k|\theta_{i-1}) - f_\alpha(\theta_k|\theta_i)) \cdot \\ (U(\theta_k|\theta_i, \theta_i) - U(\theta_k|\theta_i, \theta_{i+1})) \end{bmatrix} \geq 0$$

We first note that:

$$\begin{aligned} & \sum_{k=0}^{N} \left(f_\alpha(\theta_k|\theta_{i-1}) - f_\alpha(\theta_k|\theta_i)\right) \left[U(\theta_k|\theta_i, \theta_i) - U(\theta_k|\theta_i, \theta_{i+1})\right] \\ = \quad & f_\alpha(\theta_{i-1}|\theta_{i-1}) \begin{bmatrix} U(\theta_{i-1}|\theta_i, \theta_i) \\ -U(\theta_{i-1i}|\theta_i, \theta_{i+1}) \end{bmatrix} - f_\alpha(\theta_i|\theta_i) \begin{bmatrix} U(\theta_i|\theta_i, \theta_i) \\ -U(\theta_i|\theta_i, \theta_{i+1}) \end{bmatrix} + o(\alpha) \end{aligned}$$

where $o(\alpha)$ is such that $o(\alpha) \to 0$ as $\alpha \to 1$. Either the distortions are finite in which case the first-order quantities are finite, or the distortions go to infinity in which case the non-negativity

constraint binds and quantities are zero. Thus, all the quantities along non-constant histories remain finite in the limit and the associated probabilities converge to zero. Hence, $o(\alpha) \to 0$ as $\alpha \to 1$.

Next, let $\tilde{h}^t(\theta)$ be an history in which the realization is θ in every period for t periods. Using (4), we have:

$$\begin{aligned} U(\theta_{i-1}|\theta_i,\theta_i) - U(\theta_i|\theta_i,\theta_i) &= \sum_{t=3}^{T} (\delta \Delta F_\alpha(\theta_i\,|\theta_i))^{t-3} \Delta u(\theta_i|\tilde{h}^{t-1}(\theta_i)) + o(\alpha) \\ &= \sum_{t=3}^{T} \left[\begin{array}{c} (\delta \Delta F_\alpha(\theta_i\,|\theta_i))^{t-3} \\ \cdot \left[\int_{\theta_i}^{\theta_{i-1}} u_\theta(x, q(\theta_i|\tilde{h}^{t-1}(\theta_i)))dx\right] \end{array} \right] + o(\alpha) \end{aligned}$$

Moreover:

$$\begin{aligned} U(\theta_{i-1}|\theta_i,\theta_{i+1}) - U(\theta_i|\theta_i,\theta_{i+1}) &= \sum_{t=3}^{T} (\delta \Delta F_\alpha(\theta_i\,|\theta_i))^{t-3} \Delta u(\theta_i|\theta_i,\theta_{i+1},\tilde{h}^{t-3}(\theta_i)) + o(\alpha) \\ &= \sum_{t=3}^{T} \left[\begin{array}{c} (\delta \Delta F_\alpha(\theta_i\,|\theta_i))^{t-3} \\ \cdot \left[\int_{\theta_i}^{\theta_{i-1}} u_\theta(x, q(\theta_i|\theta_i,\theta_{i+1},\tilde{h}^{t-3}(\theta_i)))dx\right] \end{array} \right] + o(\alpha) \end{aligned}$$

As $\alpha \to 1$, we have by (8), $q(\theta_i|\tilde{h}^{t-1}(\theta_i)) \to q_1$, defined as the unique solution of:

$$s_q\,(\theta_i, q_1) = \frac{1 - \sum_{k=i}^{N} \mu_k}{\mu_i} \cdot \int_{\theta_i}^{\theta_{i-1}} u_{\theta q}(x, q_1) dx$$

and, $q(\theta_i|\theta_i,\theta_{i+1},\tilde{h}^{t-3}(\theta_i)) \to q_2$, where q_2 satisfies

$$s_q(\theta_i, q_2) \leq \frac{1 - \sum_{k=i}^{N} \mu_k}{\mu_i} \cdot D. \int_{\theta_i}^{\theta_{i-1}} u_{\theta,q}(x, q_2) dx$$

As in the proof of Proposition 3, we get $q_2 > q_1$. Next,

$$\begin{aligned} &\lim_{\alpha\to 1} \sum_{k=0}^{N} \left(f_\alpha(\theta_k|\theta_{i-1}) - f_\alpha(\theta_k|\theta_i)\right) \left[U(\theta_k|\theta_i,\theta_i) - U(\theta_k|\theta_i,\theta_{i+1})\right] \\ = &\lim_{\alpha\to 1} \sum_{t=3}^{T} (\delta \Delta F_\alpha(\theta_i\,|\theta_i))^{t-3} \int_{\theta_i}^{\theta_{i-1}} \left[u_\theta(x, q(\theta_i|\tilde{h}^{t-1}(\theta_i))) - u_\theta(x, q(\theta_i|\theta_i,\theta_{i+1},\tilde{h}^{t-3}(\theta_i)))\right] dx \\ = &\frac{1-\delta^{T-2}}{1-\delta} \int_{\theta_i}^{\theta_{i-1}} \left[u_\theta\,(x, q_1) - u_\theta\,(x, q_2)\right] dx \end{aligned}$$

Finally, as $\alpha \to 1$, $IC_{i-1,i+1}(\theta_i)$ holds only if:

$$\begin{aligned} &\lim_{\alpha\to 1} \int_{\theta_i}^{\theta_{i-1}} \left[u_\theta(x, q(\theta_i|\theta_i)) - u_\theta(x, q(\theta_{i+1}|\theta_i))\right] dx \\ \geq\ &\frac{1-\delta^{T-2}}{1-\delta} \int_{\theta_i}^{\theta_{i-1}} \left[u_\theta\,(x, q_2) - u_\theta\,(x, q_1)\right] dx \end{aligned} \tag{36}$$

The left hand side of (36) is clearly bounded for any δ. Since u_θ is strictly increasing in q and $q_2 > q_1$, the right hand side of (36) diverges to ∞ as $\delta \to 1$ and $T \to \infty$. We conclude that there exist thresholds for δ and T above which the inequality does not hold.

Case 2: $D = \lim_{\alpha \to 1} \Psi_i(f_\alpha) > 1$. We prove the result by showing that the FO-optimal contract violates the second period upward local incentive constraint $IC_{i+1,i}(\theta_{i+1})$. To this end, we first make a useful observation. Analogous to the arguments in Lemma A7 above, it is easy to show that $IC_{i+1,i}(\theta_{i+1})$ holds if and only if

$$\int_{\theta_{i+1}}^{\theta_i} \begin{bmatrix} u_\theta(x, q(\theta_i|\theta_{i+1})) \\ -u_\theta(x, q(\theta_{i+1}|\theta_{i+1})) \end{bmatrix} dx + \delta \sum_{k=0}^{N} \left(f_\alpha(\theta_k|\theta_i) - f_\alpha(\theta_k|\theta_{i+1})\right) \cdot \begin{bmatrix} U(\theta_k|\theta_{i+1}, \theta_i) \\ -U(\theta_k|\theta_{i+1}, \theta_{i+1}) \end{bmatrix} \geq 0$$

Now,

$$\begin{aligned} & \sum_{k=0}^{N} \left(f_\alpha(\theta_k|\theta_i) - f_\alpha(\theta_k|\theta_{i+1})\right) \left[U(\theta_k|\theta_{i+1}, \theta_i) - U(\theta_k|\theta_{i+1}, \theta_{i+1})\right] \\ = \quad & f_\alpha(\theta_i|\theta_i) \begin{bmatrix} U(\theta_i|\theta_{i+1}, \theta_i) \\ -U(\theta_i|\theta_{i+1}, \theta_{i+1}) \end{bmatrix} - f_\alpha(\theta_{i+1}|\theta_{i+1}) \begin{bmatrix} U(\theta_{i+1}|\theta_{i+1}, \theta_i) \\ -U(\theta_{i+1}|\theta_{i+1}, \theta_{i+1}) \end{bmatrix} + o(\alpha) \end{aligned}$$

where $o(\alpha)$ is such that $o(\alpha) \to 0$ as $\alpha \to 1$. Then, following the same steps as in case 1, we get that as $\alpha \to 1$, $IC_{i+1,i}(\theta_{i+1})$ holds as if and only if

$$\begin{aligned} & \lim_{\alpha \to 1} \int_{\theta_{i+1}}^{\theta_i} \left[u_\theta(x, q(\theta_i|\theta_{i+1})) - u_\theta(x, q(\theta_{i+1}|\theta_{i+1}))\right] dx \\ \geq \quad & \frac{1 - \delta^{T-2}}{1 - \delta} \int_{\theta_i}^{\theta_{i-1}} \left[u_\theta\left(x, q_4\right) - u_\theta\left(x, q_3\right)\right] dx \end{aligned}$$

where $q_3 = \lim_{\alpha \to 1} q(\theta_{i+1}|\theta_{i+1}, \theta_i)$, and $q_4 = \lim_{\alpha \to 1} q(\theta_{i+1}|\theta_{i+1}, \theta_{i+1})$ are as in the proof of proposition 3 above, and $q_4 > q_3$ gives us the result. ■

10.5 Proof of Lemma 2-4 and Propositions 6-7

The proof of these results are in the online appendix. ■

10.6 Proof of Proposition 8

For simplicity of notation, we present the proof for Mussa and Rosen [1978] preferences: $u(\theta, q) = \theta q$ and $c(q) = (1/2)q^2$. The more general case follows analogously. We proceed in two steps.

Step 1. We say that a quantity $q\left(\theta_i \left| h^{t-1}\right.\right)$ is distorted downward (respectively, upward) if $q\left(\theta_i \left| h^{t-1}\right.\right) \leq \theta_i$ (respectively, $q\left(\theta_i \left| h^{t-1}\right.\right) > \theta_i$). We first show that in the optimal monotonic contract distortions are all downward. Consider the constraint set of (19), as described by $\mathcal{M}$.

Define,

$$\Gamma\left(h^{t-1}\right)=\left\{\begin{array}{c}\widehat{h}^{t-1}|\exists k\leq t-1 \text{ s.t. given } h_k^{t-1}=\theta_l \text{ for some } l=0,1,...,N-1; \\ \text{we have } \widehat{h}_k^{t-1}=\theta_{l+1} \text{ and } h_j^{t-1}=\widehat{h}_j^{t-1}\ \forall j\neq k\end{array}\right\}$$

Thus, $\Gamma\left(h^{t-1}\right)$ is the set of histories that differ from h^{t-1} only once: the type in period k is replaced by the contiguous lower type. It is easy to see that a contract is monotonic if and only if for any history h^{t-1}: 1. $q\left(\theta_i|h^{t-1}\right)\geq q\left(\theta_{i+1}|h^{t-1}\right)$ for all $i<N$; and, 2. $q\left(\theta_i|h^{t-1}\right)\geq q\left(\theta_i|\widehat{h}^{t-1}\right)$ for all i and for all $\widehat{h}^{t-1}\in\Gamma\left(h^{t-1}\right)$.

Next, we introduce the following complete order on the set of all histories at time t. For any two histories h^{t-1} and $\widehat{h}^{t-1}$, let $\tau^*\left(h^{t-1},\widehat{h}^{t-1}\right)$ be the first period in which they diverge: $\tau^*\left(h^{t-1},\widehat{h}^{t-1}\right)=\min_j\left\{0\leq j\leq t-1 \text{ s.t. } h_j^{t-1}\neq\widehat{h}_j^{t-1}\right\}$, with $\tau^*\left(h^{t-1},\widehat{h}^{t-1}\right)=t-1$ if $h^{t-1}=\widehat{h}^{t-1}$. We say that $h^{t-1}\succeq^*\widehat{h}^{t-1}$ if $h^{t-1}_{\tau^*\left(h^{t-1},\widehat{h}^{t-1}\right)}\geq\widehat{h}^{t-1}_{\tau^*\left(h^{t-1},\widehat{h}^{t-1}\right)}$, i.e., if it is higher at the first point of divergence. It is easy to verify that the order $\succeq^*$ is complete, so without loss we can order the histories at time t from largest $(\overline{h}^{t-1})$ to smallest $(\underline{h}^{t-1})$, where the largest (smallest) history has all realizations equal to θ_0 (θ_N). Also, note that, $h^{t-1}\succeq^*\widehat{h}^{t-1}$ for all $\widehat{h}^{t-1}\in\Gamma\left(h^{t-1}\right)$.

Consider period t, and the smallest history of length $t-1$ (denoted, $\underline{h}^{t-1}$), in which all the realizations are θ_N. It is immediate to see that $q\left(\theta_N|\underline{h}^{t-1}\right)$ can not be distorted upward. To see this note that $q\left(\theta_N\left|\underline{h}^{t-1}\right.\right)$ is on the left hand side of no constraint.[46] If it were distorted upward, then a marginal decrease in $q\left(\theta_N|\underline{h}^{t-1}\right)$ would relax all constraints and increase surplus. Now, consider $q\left(\theta_{N-1}|\underline{h}^{t-1}\right)$: this quantity appears on the left hand side of only one constraint; $q\left(\theta_{N-1}|\underline{h}^{t-1}\right)\geq q\left(\theta_N|\underline{h}^{t-1}\right)$. If this constraint is not binding, then by the argument presented above, $q\left(\theta_{N-1}|\underline{h}^{t-1}\right)\leq\theta_{N-1}$. Assume it is binding. In this case $q\left(\theta_{N-1}|\underline{h}^{t-1}\right)=q\left(\theta_N|\underline{h}^{t-1}\right)\leq\theta_N\leq\theta_{N-1}$. Proceeding inductively with a similar argument, we can prove that $q\left(\theta_i\left|\underline{h}^{t-1}\right.\right)\leq\theta_i$ for all i.

Note that the case for first period quantities, when the history is just the empty set, is already covered by the above paragraph. Thus, now we consider $t\geq 2$. Assume, as an induction step, that there is a history $\widehat{h}^{t-1}$, where $\widehat{h}^{t-1}\succeq^*\underline{h}^{t-1}$, such that $\widehat{h}^{t-1}\succeq^* h^{t-1}\succeq^*\underline{h}^{t-1}$ implies $q\left(\theta_i|h^{t-1}\right)\leq\theta_i$ for all i. Let us also introduce a useful definition. For any h^{t-1} with $\overline{h}^{t-1}\succeq^* h^{t-1}, h^{t-1}\neq\overline{h}^{t-1}$ and $t\geq 2$, define $\left[h^{t-1}\right]^+$ to be the smallest t-period history larger than h^{t-1} according to the order $\succeq^*$in the following inductive way. If $t=2$, then $\left[h^{t-1}\right]^+=\left\{\kappa_{t-1}(h^{t-1}),h_{t-1}^{t-1}+\Delta\theta\right\}$; if $t>2$ then:

$$\left[h^{t-1}\right]^+=\left\{\begin{array}{ll}\left(\kappa_{t-1}(h^{t-1}),h_{t-1}^{t-1}+\Delta\theta\right), & \text{if } h_{t-1}^{t-1}<\theta_0 \\ \left(\left[\kappa_{t-1}(h^{t-1})\right]^+,\theta_N\right), & \text{if } h_{t-1}^{t-1}=\theta_0\end{array}\right.,$$

[46] We say that a quantity is on the left hand side of a given constraint if in that constraint it must be larger than some other quantity.

where κ_s projects the first s elements of a vector.[47] We intend to show that $q\left(\theta_i|\left[\widehat{h}^{t-1}\right]^+\right) \leq \theta_i$ for all i. Now, $q\left(\theta_N|\left[\widehat{h}^{t-1}\right]^+\right)$ appears on the left hand side in the following constraints: $q\left(\theta_N|\left[\widehat{h}^{t-1}\right]^+\right) \geq q\left(\theta_N|\widetilde{h}^{t-1}\right)$ for all $\widetilde{h}^{t-1} \in \Gamma\left(\left[\widehat{h}^{t-1}\right]^+\right)$. If none of these constraints bind, then as before, we have the desired inequality. Suppose at least one of them binds. Clearly, by the definition of $\left[\widehat{h}^{t-1}\right]^+$, we have $\widehat{h}^{t-1} \succeq^* \widetilde{h}^{t-1}$ for all $\widetilde{h}^{t-1} \in \Gamma\left(\left[\widehat{h}^{t-1}\right]^+\right)$. Thus, by the induction hypothesis $q\left(\theta_N|\widetilde{h}^{t-1}\right) \leq \theta_N$ for all $\widetilde{h}^{t-1} \in \Gamma\left(\left[\widehat{h}^{t-1}\right]^+\right)$. Since the inequality constraint binds for some $\widetilde{h}^{t-1}$, we have $q\left(\theta_N|\left[\widehat{h}^{t-1}\right]^+\right) = q\left(\theta_N|\widetilde{h}^{t-1}\right) \leq \theta_N$.

Next, consider $q\left(\theta_{N-1}|\left[\widehat{h}^{t-1}\right]^+\right)$. It appears **o**n the left hand side in the following constraints: $q\left(\theta_{N-1}|\left[\widehat{h}^{t-1}\right]^+\right) \geq q\left(\theta_N|\left[\widehat{h}^{t-1}\right]^+\right)$ and $q\left(\theta_{N-1}|\left[\widehat{h}^{t-1}\right]^+\right) \geq q\left(\theta_{N-1}|\widetilde{h}^{t-1}\right)$ for all $\widetilde{h}^{t-1} \in \Gamma\left(\left[\widehat{h}^{t-1}\right]^+\right)$. If none of these constraints bind, then as before, we have the desired inequality. If the first one binds then, $q\left(\theta_{N-1}|\left[\widehat{h}^{t-1}\right]^+\right) \leq \theta_N < \theta_{N-1}$. If any of the latter one binds, then invoking the induction hypothesis, as argued in the case above, we have the desired inequality. Proceeding inductively, we can show $q\left(\theta_i|h^{t-1}\right) \leq \theta_i$ for all i and h^{t-1}.

Step 2. We now prove that the allocation is asymptotically efficient. Consider problem (19). From this problem eliminate the constraint $q\left(\theta_0|h^0\right) \geq q\left(\theta_1|h^0\right)$ and all the monotonicity constraints that involve quantities following an history in which the agents reports to be a type θ_0. It is easy to see that in this problem the quantities offered after the agent reports (or has reported) to be θ_0 are efficient: $q\left(\theta_i|h^{t-1}\right) = \theta_i$ for $i = 0$ and/or $\forall h^{t-1} \in \overline{H}^{t-1}, t \geq 2$, where $\overline{H}^{t-1} = \left\{h^{t-1} \left| \exists \tau \leq t-1 \text{ s.t. } h_\tau^{t-1} = \theta_0 \right.\right\}$. Following the same approach as in Step 1, it can be shown that the solution of this relaxed problem is monotonic and so it coincides with the optimal monotonic contract. Since the probability of the event in which no type realization in t periods is equal to θ_0 converges to zero as $t \to \infty$, this solution is, is asymptotically efficient, and so the optimal monotonic contract. ■

10.7 Proof of Proposition 9

We prove that for any given T, the optimal monotonic contract converges in probability to the optimal contract. Let $\Pi^s(\boldsymbol{\alpha})$, $\Pi^m(\boldsymbol{\alpha})$ and $\Pi^{**}(\boldsymbol{\alpha})$ be the expected profits obtained by the seller from, respectively, the repetition of the optimal static contract, the optimal monotonic contract and the optimal contract when the Markov matrix is $\boldsymbol{\alpha}$. Because the repetition of the optimal static contract is a monotonic dynamic contract, we must have $\Pi^m(\boldsymbol{\alpha}) \in [\Pi^s(\boldsymbol{\alpha}), \Pi^{**}(\boldsymbol{\alpha})]$. Now note that when types are constant and $\boldsymbol{\alpha} = I$, it is well known that the repetition in every period

[47] Recollect that h^{t-1} is a vector of length t: $h^{t-1} = (h_0^{t-1}, h_1^{t-1}, ..., h_{t-1}^{t-1})$, where $h_0^{t-1} = \emptyset$. So, $\kappa_{t-1}(h^{t-1}) = (h_0^{t-1}, ..., h_{t-2}^{t-1})$.

of the optimal static contract is optimal.[48] Since $\Pi^m(\boldsymbol{\alpha}), \Pi^s(\boldsymbol{\alpha})$ and $\Pi^{**}(\boldsymbol{\alpha})$ are continuous in α by the theorem of the maximum[49] , we must have that for any sequence $\boldsymbol{\alpha}_n \to I$ and $\varepsilon > 0$ there must be a n' such that for $n > n'$, we have $|\Pi^m(\boldsymbol{\alpha}_n) - \Pi^{**}(\boldsymbol{\alpha}_n)| \leq |\Pi^s(\boldsymbol{\alpha}_n) - \Pi^{**}(\boldsymbol{\alpha}_n)| < \varepsilon$. It is immediate to see that the fact that $\Pi^m(\boldsymbol{\alpha}_n)$ converges to $\Pi^{**}(\boldsymbol{\alpha}_n)$ and that by Proposition 8 quantities are bounded imply that the optimal monotonic contract must converge to a contract that maximizes profit in probability. ■

10.8 Proof of Proposition 10

The fact that $\lim_{\delta \to 1} \lim_{\boldsymbol{\alpha} \to I} \pi_m(\boldsymbol{\alpha}, \delta) = \lim_{\delta \to 1} \lim_{\boldsymbol{\alpha} \to I} \pi^*(\boldsymbol{\alpha}, \delta)$ follows immediately from the fact that for any δ, $\lim_{\boldsymbol{\alpha} \to I} \pi_m(\boldsymbol{\alpha}, \delta) = \lim_{\boldsymbol{\alpha} \to I} \pi^*(\boldsymbol{\alpha}, \delta)$. We now prove the remaining equality. Let $S(\boldsymbol{\alpha}, \delta, T)$ be the total expected surplus generated in the efficient contract, $U^*(\boldsymbol{\alpha}, \delta, T)$ be the agent's expected surplus obtained with the efficient contract and $U_i^*(\boldsymbol{\alpha}, \delta, T)$ be the agent's surplus obtained with the efficient contract conditional on being type i at $t = 1$. Define also $s(\boldsymbol{\alpha}, \delta, T) = (1-\delta) S(\boldsymbol{\alpha}, \delta, T)$, $u^*(\boldsymbol{\alpha}, \delta, T) = (1-\delta) U^*(\boldsymbol{\alpha}, \delta, T)$ and $u_i^*(\boldsymbol{\alpha}, \delta, T) = (1-\delta) U_i^*(\boldsymbol{\alpha}, \delta, T)$; and $s(\boldsymbol{\alpha}, \delta) = \lim_{T \to \infty} s(\boldsymbol{\alpha}, \delta, T)$, $u^*(\boldsymbol{\alpha}, \delta) = \lim_{T \to \infty} u^*(\boldsymbol{\alpha}, \delta, T)$ and $u_i^*(\boldsymbol{\alpha}, \delta) = \lim_{T \to \infty} u_i^*(\boldsymbol{\alpha}, \delta, T)$. Profits $\pi_m(\boldsymbol{\alpha}, \delta)$ must be larger or equal to the profits obtained by offering the efficient quantity and charging a fixed per period price equal to $u_N^*(\boldsymbol{\alpha}, \delta)$, since this is an incentive compatible monotonic contract. Note that since types follow an irreducible Markov process, their distribution converges to a stationary distribution that is independent from the realization at $t = 1$. It follows that, for all α, $\lim_{\delta \to 1} u_i^*(\boldsymbol{\alpha}, \delta) = \lim_{\delta \to 1} u^*(\boldsymbol{\alpha}, \delta)$ and so the per period profits in this contract converge to $s(\boldsymbol{\alpha}, \delta)$, implying that, for all α**,** $\lim_{\delta \to 1} \pi_m(\boldsymbol{\alpha}, \delta) = \lim_{\delta \to 1} s(\boldsymbol{\alpha}, \delta)$. Similarly we can show that for all α**,** $\lim_{\delta \to 1} \pi^*(\boldsymbol{\alpha}, \delta) = \lim_{\delta \to 1} s(\boldsymbol{\alpha}, \delta)$. It follows that: $\lim_{\boldsymbol{\alpha} \to I} \lim_{\delta \to 1} \pi_m(\boldsymbol{\alpha}, \delta) = \lim_{\boldsymbol{\alpha} \to I} \lim_{\delta \to 1} s(\boldsymbol{\alpha}, \delta) = \lim_{\boldsymbol{\alpha} \to I} \lim_{\delta \to 1} \pi^*(\boldsymbol{\alpha}, \delta)$. This proves the result. ■

[48] This result can be easily deduced studying problem (19). To see it, note that the repetition of the static contract is incentive compatible and individually rational. Then note that when $\Lambda = I$, the first order optimal contract coincides with the static optimal contract along the histories in which the agent reports always the same type. Since the other histories have probability zero, the profit from the repetition of the static contract is the same of the profit from the FO-optimal contract. Since the FO-optimal contract yields a profit not inferior to the optimal contract, the result is proven.

[49] In order to apply the theorem of the maximum the space of quantities must be compact. It is clearly bounded below by zero. Also, Proposition 8 shows that it is bounded above by the efficient quantities. Hence, there is no loss of generality in assuming that set of quantities is contained in the interval $[0, \theta_0\ ]$.

	q_H	q_M	q_L	$q_H(\theta)$	$q_i(H)$	$q_M(M)$	$q_L(M)$	$q_M(L)$	$q_L(L)$
A1	θ_H	$\theta_M - \frac{\mu_H}{\mu_M}\Delta\theta$	$\theta_L - \frac{\mu_H+\mu_M}{\mu_M}\Delta\theta$	θ_H	θ_i	$\theta_M - \frac{\mu_H}{\mu_M}\frac{3\alpha-1}{2\alpha}\Delta\theta$	θ_L	θ_M	$\theta_L - \frac{\mu_H+\mu_M}{\mu_M}\frac{3\alpha-1}{2\alpha}\Delta\theta$
A2						$\frac{2\alpha}{1+\alpha}\theta_M + \frac{1-\alpha}{1+\alpha}\theta_L - \frac{\mu_H}{\mu_M}\frac{3\alpha-1}{1+\alpha}\Delta\theta$			
B1		$\theta_M - \frac{\mu_H-\lambda_1}{\mu_M}\Delta\theta$	$\theta_L - \frac{\mu_H+\mu_M+\lambda_1}{\mu_M}\Delta\theta$			$\theta_M - \frac{\mu_H-\lambda_1}{\mu_M}\frac{3\alpha-1}{2\alpha}\Delta\theta$	θ_L	$\theta_M - \frac{\lambda_1}{\mu_L}\frac{3\alpha-1}{1-\alpha}\Delta\theta$	
B2		$\theta_M - \frac{\mu_H-\lambda_2}{\mu_M}\Delta\theta$	$\theta_L - \frac{\mu_H+\mu_M+\lambda_2}{\mu_M}\Delta\theta$			$\frac{2\alpha}{1+\alpha}\theta_M + \frac{1-\alpha}{1+\alpha}\theta_L - \frac{\mu_H-\lambda_2}{\mu_M}\frac{3\alpha-1}{1+\alpha}\Delta\theta$		$\theta_M - \frac{\lambda_2}{\mu_L}\frac{3\alpha-1}{1-\alpha}\Delta\theta$	
B3		$\theta_M - \frac{\mu_H-\lambda}{\mu_M}\Delta\theta = \theta_L - \frac{\mu_H+\mu_M+\lambda}{\mu_M}\Delta\theta$				$\frac{2\alpha}{1+\alpha}\theta_M + \frac{1-\alpha}{1+\alpha}\theta_L - \frac{\mu_H-\lambda}{\mu_M}\frac{3\alpha-1}{2\alpha}\Delta\theta = \frac{1-\alpha}{1+\alpha}\theta_M + \frac{2\alpha}{1+\alpha}\theta_L - \frac{\mu_H+\mu_M+\lambda}{\mu_L}\frac{3\alpha-1}{1+\alpha}\Delta\theta$			

Table 1: The optimal contract when $|\Theta|=3$ and $T=2$.