Lecture Notes in Contract Theory

Holger M. Müller
Stockholm School of Economics

May 1997

preliminary and incomplete

Contents

Chapter 1

Introduction

Common sense suggests that contracts which are not enforceable are not worth the paper on which they are written. Enforcement, however, requires both the existence of a functioning legal system and the ability of courts to judge whether a contract has been broken. The latter criterion leads directly to the notion of verifiability. If the performance of a contractual duty is not verifiable vis-a-vis the court, then we can reasonably assume that this duty is not enforceable, which in turn implies that it should not be made part of a contract in the first place.

In many situations, parties would like to include a rule or instruction in a contract, but cannot do so by the above reasoning because either i) the performance of the rule is observable by all parties but not verifiable vis-a-vis the court, or ii) the performance is not even observable by all the parties involved. Throughout these notes, we refer to such rules as *social choice rules.* A social choice rule is a mapping $f : \Theta \rightarrow A$ from a set of states Θ into a set of outcomes or alternatives A. While we exclusively concentrate on social choice rules that are (Pareto-)efficient, we abstract from distributional consequences. For instance, in chapters 3 and 4, we deal with social choice rules that maximize the (expected) utility of one party subject to holding the utility of another party constant at the lowest possible level. This suggests that the term "social" must not always be taken literally.

Contract theory is concerned with the implementation of social choice rules in situations where these cannot be made part of a contract due to the presence of *incomplete information* (i.e. either non-observability and/or non-verifiability of performance). In such cases, we examine whether the social choice rule in question can be implemented indirectly or replicated through either a) an alternative rule that is enforceable by courts or b) some institutional arrangement. If the social choice rule can be fully replicated, then we speak of an *efficient* or *first-best solution.* Usually, however, this

is not possible due to the constraints imposed by incomplete information. We then typically search for an enforceable alternative rule or institutional arrangement that maximizes potential efficiency gains and thus comes as close as possible to the original social choice rule. Such rules or institutions are called *constrained efficient* or *second-best optimal*.

In chapters 2 to 4, we study contracts (i.e. rules or instructions) that are *comprehensive* in the sense that they optimally take into account all commonly observable information. There, it is implicitly assumed that this information is also verifiable in front of courts. Often, however, it is too costly to write a contract that is contingent on all jointly observable information. For instance, suppose that an employment contract had to specificy what a bricklayer should do next Monday when the weather is x, his health is y, and the price of concrete is z. If each of these three variables had only ten possible realizations, the contract would have to include 1,000 different contingencies. Alternatively, some commonly observable information may simply not be verifiable vis-a-vis the court. In both cases, contracts will be left *incomplete*, i.e. they will not optimally utilize all available information. Incomplete contracts are the subject of chapter 5. When the incompleteness is severe such that there is no hope of replicating the SCR through an alternative enforceable rule, institutional arrangements become important. In the case of the bricklayer, we would expect that either he or his supervisor has the *authority* do decide what to do in each situation. If (nonhuman) assets are involved, *ownership rights* usually determine how the asset is used in unforeseen contingencies.

To be completed

Chapter 2

Mechanism Design

2.1 The Implementation Problem

Mechanism design is based on two canonical examples, both of which can be addressed in the same theoretical framework. The first example is the *planner's problem,* where an uninformed agent (the planner) faces a group of informed agents. The agents' private information concerns the state θ, which determines their preferences over outcomes in A. Clearly, the planner cannot directly implement the social choice rule since she does not know the true state.

For instance, suppose that a government decides that it will provide a public good if and only if the aggregate valuation of all citizens exceeds the cost of the public good. Moreover, the citizens shall be taxed according to their valuations. If the citizens are asked to reveal their valuations, each individual citizen has an incentive to understate his valuation since the tax savings outweigh the potentially adverse effects on the collective decision. This is the well-known free-rider problem. Similarly, bidders in an auction where the object is sold at the price of the highest bid have an incentive to bid less than their true valuation. For each bidder, the reduction in price if he wins the object outweighs the reduction in the probability of winning. In both examples, the planner (i.e. the government or the auctioneer) has failed to elicit the agents' private information.

As we will see shortly, the planner can do better by designing a *mechanism* or *game form* that defines a game to be played by the agents in each state. Formally, a mechanism consists of a collection of strategy sets and an outcome function $g : \times S_i \to A$ that assigns an outcome to each strategy profile $s \in \times S_i$. The planner's objective is to select the function $g(s)$ such that in each state, the set of equilibrium outcomes "coincides" with the set of

outcomes determined by the social choice rule. Since the equilibrium strategy profile s is publicly observable, the outcome function $g(s)$ is verifiable vis-a-vis the court and can -unlike the social choice rule- be included in a (ficticious) contract between the agents and the planner.

The other canonical example is the *trade model*, which involves no planner, but a group of agents who would like to implement a social choice rule by means of a contract. The problem is that the true state is not verifiable so that the contract is not enforceable. For instance, consider a bilateral trading problem where agent 1 owns a good that agent 2 likes to buy. Here, a state is a profile of valuations, where each agent knows only his own valuation. A mechanism can then be interpreted as a bargaining rule which specifies an allocation of the good and a monetary transfer from agent 2 to agent 1 as a function of the agent's (verifiable) bid and ask prices s_1 and s_2.

The Model

We begin by assuming that agents possess complete information about each others' preferences. However, this restriction does not become relevant until section 2.3 when we study Nash implementation. In section 2.4, we then drop the assumption of complete information and turn to environments where preferences are no longer mutually observable.

Consider the following model:

1. There are n agents indexed by $i \in I = \{1, ..., n\}$.
2. There is a finite set A of feasible outcomes.
3. Each agent has a characteristic or *type* $\theta_i \in \Theta_i$.
4. A *state* is a profile of types $\theta = (\theta_1, ..., \theta_n) \in \Theta$ which defines a profile of preference orderings $\succsim(\theta) = (\succsim_1(\theta), ..., \succsim_n(\theta)) \in \Re$ on the set of feasible outcomes A. $\succsim_i(\theta) \in \Re_i$ denotes agent i's preference ordering on A in state θ.
5. Each agent (but no outside party) observes the entire vector $\theta = (\theta_1, ..., \theta_n)$, i.e. agents have *complete information*.
6. A *social choice rule* (SCR) is a correspondence $f : \Theta \twoheadrightarrow A$ which specifies a nonempty choice set $f(\theta) \subseteq A$ for every state θ.

An SCR is a selection rule that determines a set of socially desirable outcomes for each state $\theta \in \Theta$. Two examples of social choice rules which feature prominently in the literature are the *Paretian* SCR, which comprises

only Pareto optimal allocations, and the *dictatorial* SCR, where for all θ, the social choice set $f(\theta)$ is a subset of the most preferred outcomes of a particular agent. Note that implementation theory takes the existence of social choice rules as given - the underlying problem of constructing an SCR via aggregation of preferences is the subject of *social choice theory*.

In principle, the above formulation allows for any degree of correlation among the agents' preference orderings. For convenience, let us therefore narrow down the set of admissible preferences by requiring that the domains of the individual preference orderings be independent.

7. The set of possible preference orderings $\Re$ is the Cartesian product of the n sets of preference orderings, i.e. $\Re = \times\Re_i$.

Assumption 7 can be illustrated by means of a simple example: Let $I = \{1, 2\}$, $A = \{a, b\}$, $\Re_1 = \{a \succ_1 b, b \succ_1 a\}$, and $\Re_2 = \{a \succ_2 b, b \succ_2 a, aI_2b\}$. With independent domains, the set $\Re = \Re_1 \times \Re_2$ becomes

$a \succ_1 b$ $a \succ_2 b$	$a \succ_1 b$ $b \succ_2 a$	$a \succ_1 b$ $a \sim_2 b$
$b \succ_1 a$ $a \succ_2 b$	$b \succ_1 a$ $b \succ_2 a$	$b \succ_1 a$ $a \sim_2 b$

Finally, let us concentrate on a special, but much studied case known as *private values* which assumes that the mapping from preference orderings to types is one-to-one.

8. Agent i's preferences in state θ depend only on his type θ_i, i.e. $\succsim_i (\theta) = \succsim_i (\theta_i)$.

From assumption 8, it follows that we can label the rows and columns in the above example with θ_1^1 and θ_2^1, and θ_1^2, θ_2^2, and θ_3^2, respectively. Moreover, assumptions 7 and 8 together imply that $\Theta = \times\Theta_i$, i.e. that the state-space is spanned by the Cartesian product of the individual sets of possible types.

Example: Provision of a Public Good

The city council ("the social planner") considers the construction of a road for the n inhabitants of the city. In this example, θ_i represents agent i's valuation or willingness to pay for the road. An outcome is a profile $y = (x, t_1, ..., t_n)$, where x can be either 1 ("the road is built") or 0 ("the road is not built"), and where t_i denotes a transfer to agent i Note that transfers can be negative. Preferences are assumed to be quasilinear of the form $\theta_i x + t_i$. The city

council faces the restriction that it cannot provide additional funds, i.e. the cost $c \geq 0$ must be covered entirely by the inhabitants. This defines the set of feasible outcomes as $A = \{(x, t_1, ..., t_n) \mid x = \{0, 1\} \text{ and } \sum_i t_i \leq -cx\}$. A particularly desirable SCR is one where the road is built if and only if the the sum of the agent's valuations exceeds the construction cost and where the budget constraint is satisfied with equality, i.e.

$$x(\theta) = \begin{cases} 1 \text{ if } \sum_i \theta_i \geq cx \\ 0 \text{ otherwise,} \end{cases} \tag{2.1}$$

$$\sum_i t_i = -cx. \tag{2.2}$$

The set of outcomes defined by (2.1)-(2.2) coincides with the set of Pareto optimal allocations (with respect to both the public good and money). Unfortunately, it turns out that this SCR is not implementable in dominant strategies (cf. section 2.2).

Implementation and Full Implementation

At the beginning of this chapter, we distinguished between two canonical examples of mechanism design problems: the planner's problem and the trade model. In both examples, the social choice rule $f : \Theta \twoheadrightarrow A$ is not directly implementable as it depends on the true state θ, which is not verifiable vis-a-vis the court (in the planner's problem, θ is not even known to the planner herself). We also argued that when the agents are asked to reveal their preferences honestly, each individual agent has an incentive to misrepresent his information.

We then showed that the planner (or in the trade model, the agents themselves) can improve the situation by constructing a *mechanism* or *game form* that uses only publicly observable (and thus verifiable) information. Formally, a mechanism Γ consists of a collection of strategy sets $\Sigma = \{S_1, ..., S_n\}$ and an outcome function $g : \times S_i \to A$ which assigns an outcome $y \in A$ to each strategy profile $s = (s_1, ..., s_n) \in \times S_i$. Since the state θ is not verifiable, the outcomes themselves cannot be made contingent on the state. However, the agents' payoffs (utilities) from a particular outcome typically vary with θ since preferences $\succsim_i (\theta_i)$ are state-dependent. Thus, the mechanism Γ combined with the state-space Θ defines a game of complete information with a (possibly) different payoff structure in every state θ. The implementation problem is then to construct Γ such that in each state, the equilibrium outcomes of the resulting game coincide (in a way yet to be defined) with the elements in $f(\theta)$.

The idea which underlies mechanism design is information revelation through strategy choice. Since the collection of strategy sets $\Sigma = \{S_1, ..., S_n\}$ and the outcome function $g : \times S_i \to A$ are public information, outsiders such as courts (or the planner) can compute the agents' equilibrium strategy rules $s^*(\theta) = (s_1^*(\theta), ..., s_n^*(\theta))$. By observing a particular equilibrium strategy profile $(s_1^*, ..., s_n^*)$, these outsiders can then infer the true state θ.

Definition 1 (Mechanism) A mechanism or game form Γ is a collection of strategy sets $\Sigma = \{S_1, ..., S_n\}$ and a mapping $g : \times S_i \to A$.

Definition 2 (Implementation) Denote by $E_g(\theta)$ the set of equilibrium profiles $s^*(\theta)$ of Γ in state θ, and define the set of equilibrium outcomes of Γ in θ as $g(E_g(\theta)) \equiv \{g(s^*(\theta)) \mid s^*(\theta) \in E_g(\theta)\}$. The mechanism Γ implements the social choice rule $f(\theta)$ if $E_g(\theta)$ is nonempty, and if for every $\theta \in \Theta$, $g(E_g(\theta)) \subseteq f(\theta)$.

Definition 3 (Full Implementation) The mechanism Γ fully implements the social choice rule $f(\theta)$ if for every $\theta \in \Theta$, $g(E_g(\theta)) = f(\theta)$.

The distinction between implementation and full implementation is subtle. While implementation requires that all equilibrium outcomes are in the social choice set $f(\theta)$, it does not require that all elements in $f(\theta)$ correspond to some equilibrium outcome. Clearly, the notion of full implementation is stronger since it requires that the set of equilibrium outcomes exactly coincides with the choice set $f(\theta)$. Notice that there can be more equilibria than equilibrium outcomes if several equilibria give rise to the same outcome.

Truthful Implementation

The identification of all social choice rules that are implementable for a specific equilibrium concept requires knowledge of the entire set of possible mechanisms. Fortunately, a very useful result known as the *revelation principle* allows us to restrict attention to a particularly simple class of mechanisms called *direct mechanisms*. In a direct mechanism, an agent's strategy set consists of his possibly types Θ_i. In the underlying game, each agent announces a type $\hat{\theta}_i$, and the outcome function $g : \hat{\Theta} \to A$ subsequently selects an outcome $g(\hat{\theta}) = g(\hat{\theta}_1, ..., \hat{\theta}_n)$. If in each state, there exists an equilibrium in which all agents report their types truthfully (i.e. $\hat{\theta}_i = \theta_i$ for all i) and the equilibrium outcome $g(\theta)$ is an element in $f(\theta)$, then we say that the direct mechanism Γ_d *truthfully implements* the social choice rule $f(\theta)$.

Definition 4 (Direct Mechanism) A direct mechanism Γ_d is a mechanism in which $S_i = \Theta_i$.

Definition 5 (Truthful Implementation) The direct mechanism Γ_d truthfully implements the social choice rule $f(\theta)$ if for every $\theta \in \Theta$, $\theta \in E_g(\theta)$ and $g(\theta) \in f(\theta)$.

Observe that truthful implementation is a weaker concept than implementation or full implementation. Truthful implementation only requires that the profile $\theta = (\theta_1, ..., \theta_n)$ of truthful announcements is an equilibrium in each state θ and that the equilibrium outcome $g(\theta)$ is an element in $f(\theta)$. However, truthful implementation does not rule out the existence of further equilibria with outcomes $g(\hat{\theta}) \notin f(\theta)$ in which some agents lie (i.e. $\hat{\theta} \neq \theta$). Both implementation and full implementation rule out such "unwanted" equilibria as they require either that the set of equilibrium outcomes constitutes a subset of $f(\theta)$ (implementation) or that it coincides with $f(\theta)$ (full implementation).

Let us illustrate the notion of truthful implementation by means of an example. Suppose $I = \{1, 2\}$, $A = \{a, b, c, d\}$, $\Theta_1 = \{\theta_1^1, \theta_1^2\}$, and $\Theta_2 = \{\theta_2^1, \theta_2^2\}$, where

θ_1^1	θ_1^2	θ_2^1	θ_2^2
a	b		
b	a	$a \sim b$	$c \sim d$
c	d	$c \sim d$	$a \sim b$
d	c		

Consider the SCR

$$f(\theta_1^1, \theta_2^1) = \{a\}$$
$$f(\theta_1^1, \theta_2^2) = \{c\}$$
$$f(\theta_1^2, \theta_2^1) = \{b\}$$
$$f(\theta_1^2, \theta_2^2) = \{d\}$$

This SCR has the desirable property that in each state θ it contains only strong Pareto optimal allocations. The following direct mechanism Γ_d truthfully implements $f(\theta)$ in dominant strategies (agent 1 plays "row" and agent 2 plays "column"):

	θ_2^1	θ_2^2
θ_1^1	a	c
θ_1^2	b	d

Incidentally, Γ_d also fully implements $f(\theta)$ in dominant strategies since for each θ, the truthtelling outcome is the unique dominant strategy equilibrium outcome (this proves the "if"-part of a theorem stated in section 2.2 that an SCR is fully implementable in dominant strategies if and only if it is single-valued and truthfully implementable with a unique dominant strategy equilibrium outcome).

2.2 Dominant Strategy Implementation

The Revelation Principle

The great virtue of dominant strategy equilibrium is that agents need not forecast how other agents choose their strategies. In other words, agents do not have to know each others' preferences. This is the basis for an extremely convenient result known as the revelation principle (Gibbard (1973), Green and Laffont (1977), Dasgupta, Hammond, and Maskin (1979)), which says that we can restrict attention to direct mechanisms in which agents report only their own types. Thus, the assumption of complete information made at the beginning of this chapter is irrelevant and any result derived in this section continues to hold if this assumption is dropped. Due to this robustness property, SCRs that are truthfully implementable in dominant strategies are of particular interest.

Definition 6 (TIDS Social Choice Rule) The social choice rule $f(\theta)$ is truthfully implementable in dominant strategies (TIDS) or strategy-proof if there exists a direct mechanism Γ_d such that i) truthtelling is a dominant strategy equilibrium, i.e. if for all $i \in I$ and $\theta_i \in \Theta_i$,

$$g(\theta_i, \hat{\theta}_{-i}) \succsim_i (\theta_i)\, g(\hat{\theta}_i, \hat{\theta}_{-i}) \tag{2.3}$$

for all $\hat{\theta}_i \in \Theta_i$, $\hat{\theta}_{-i} \in \Theta_{-i}$, and ii) $g(\theta) \in f(\theta)$ for all $\theta \in \Theta$.

We now present the revelation principle, which asserts that for every mechanism Γ that implements $f(\theta)$ in dominant strategies, we can find a direct mechanism Γ_d that truthfully implements $f(\theta)$ in dominant strategies.

Theorem 1 (Revelation Principle) If an SCR is implementable in dominant strategies, then it is TIDS.

Proof (direct) Suppose Γ implements the social choice rule $f(\theta)$ in dominant strategies, and let $E_g(\theta)$ be non-empty for all θ. Define an *equilibrium selection* as a mapping $s^* : \Theta \to \times S_i$ which selects exactly one equilibrium profile $s^*(\theta) \in E_g(\theta)$ for each $\theta \in \Theta$. Since $s^*(\theta)$ is a dominant strategy profile, we have

$$g(s_i^*(\theta_i), s_{-i}) \succsim_i (\theta_i)\, g(s_i, s_{-i}) \tag{2.4}$$

for all $i \in I$, $\theta_i \in \Theta_i$, $s_i \in S_i$, and $s_{-i} \in S_{-i}$. In particular, it is true that

$$g(s_i^*(\theta_i), s_{-i}^*(\hat{\theta}_{-i})) \succsim_i (\theta_i)\, g(s_i^*(\hat{\theta}_i), s_{-i}^*(\hat{\theta}_{-i})) \tag{2.5}$$

for all $i \in I$, $\theta_i, \hat{\theta}_i \in \Theta_i$ and $\hat{\theta}_{-i} \in \Theta_{-i}$ since $s_i^*(\hat{\theta}_i) \in S_i$ and $s_{-i}^*(\hat{\theta}_{-i}) \in S_{-i}$ are merely specific strategies.

Next, define the composed mapping $h : \Theta \to A$ with $h(\theta) \equiv g(s^*(\theta))$. The function $h(\theta)$ together with the collection of possible types $\{\Theta_1, ..., \Theta_n\}$ defines a direct mechanism Γ_d. But Γ_d truthfully implements $f(\theta)$ in dominant strategies because $h(\theta) \equiv g(s^*(\theta)) \in f(\theta)$ and

$$h(\theta_i, \hat{\theta}_{-i}) \succsim_i (\theta_i)\, h(\hat{\theta}_i, \hat{\theta}_{-i}) \tag{2.6}$$

for all $i \in I$, $\theta_i, \hat{\theta}_i \in \Theta_i$ and $\hat{\theta}_{-i} \in \Theta_{-i}$. Thus, $f(\theta)$ is TIDS. ■

Remarks

1. The intuition that underlies the proof is straightforward. For each state θ, consider an equilibrium profile $s^*(\theta) = (s_1^*(\theta_1), ..., s_n^*(\theta_n))$ induced by the (indirect) mechanism Γ with outcome $g(s^*(\theta)) \in f(\theta)$. The planner can mimick Γ by asking each agent to announce a type $\hat{\theta}_i$ and playing on his behalf the strategy $s_i^*(\hat{\theta}_i)$. Since $s_i^*(\theta_i)$ is a dominant strategy in the game induced by Γ, reporting the true type $\hat{\theta}_i = \theta_i$ must also be a dominant strategy in the new game. Notice that we have assumed that the planner can commit to playing $s_i^*(\hat{\theta}_i)$ after the agents have revealed their types.

2. Since full implementation implies implementation, theorem 1 continues to hold if we substitute "implementable" with "fully implementable".

3. In the remainder of this section, we will characterize the set of SCRs that are implementable in dominant strategies. In theory, this implies that we have to consider all possible mechanisms. However, due to the revelation principle, we can restrict attention (subject to a caveat) to direct mechanisms and identify the set of SCRs that are TIDS.

4. Here is the caveat mentioned in 3. According to the revelation principle, TIDS is a necessary, but not a sufficient condition for dominant strategy implementation. Hence, if an SCR is not TIDS, we can be sure that it is not implementable in dominant strategies. However, the converse is not true, i.e. there may exist SCRs that are TIDS but not implementable in dominant strategies. For instance, if truthtelling is only a weakly dominant strategy and there exist other (untruthful) equilibria in weakly dominant strategies with outcomes $g(\hat{\theta}) \notin f(\theta)$, then the direct mechanism Γ_d does not implement $f(\theta)$, even though it truthfully implements $f(\theta)$. Conditions under which the revelation principle also holds in the other direction are presented in the following subsection.

Necessary and Sufficient Conditions for Implementation

Consider a direct mechanism Γ_d that truthfully implements $f(\theta)$ in dominant strategies with a unique outcome in each state. Per definition, Γ_d implements $f(\theta)$ in dominant strategies. Moreover, when $f(\theta)$ is single-valued, the concepts of implementation and full implementation coincide and Γ_d also fully implements $f(\theta)$ in dominant strategies. This suggests that the revelation principle holds in the other direction as well if we can ensure that the outcome associated with truthtelling is the unique dominant strategy equilibrium outcome. A sufficient condition for a game to have at most one dominant strategy outcome is that $\Re$ contains only strict preference orderings.

Theorem 2 Suppose that $\Re$ contains only strict preference orderings. If an SCR is TIDS, then it is implementable in dominant strategies.

Proof (direct) Assume that the direct mechanism Γ_d truthfully implements $f(\theta)$ in dominant strategies. Since $\Re$ contains only strict orderings, the set $g(E_g(\theta)) \equiv \{g(\hat{\theta}) \mid \hat{\theta} \in E_g(\theta)\}$ of dominant strategy equilibrium outcomes must be a singleton set for all $\theta \in \Theta$. Because the social choice rule $f(\theta)$ is TIDS, $\theta \in E_g(\theta)$ and $g(\theta) \in f(\theta)$. This implies $g(E_g(\theta)) \subseteq f(\theta)$ for all $\theta \in \Theta$, i.e. $f(\theta)$ is implementable. ■

Theorem 2 shows that strict preference orderings and truthful implementation imply implementation. The following theorem goes beyond theorem 2 by showing that strict preference orderings, truthful implementation and single-valuedness of $f(\theta)$ imply full implementation. In addition, it shows that the reverse also holds.

Theorem 3 Suppose that $\Re$ contains only strict preference orderings. An SCR is fully implementable in dominant strategies if and only if it is TIDS and single-valued.

Proof (direct) "if"-part: By theorem 2, strict preference orderings and TIDS imply implementability. When the social choice rule $f(\theta)$ is single-valued, the concepts of implementability and full implementability coincide. It follows that $f(\theta)$ is fully implementable.

"only if"-part: By Theorem 1, full implementability implies TIDS. Moreover, if $\Re$ contains only strict preference orderings, the set of dominant strategy equilibrium outcomes $g(E_g(\theta)) \equiv \{g(s^*(\theta)) \mid s^*(\theta) \in E_g(\theta)\}$ is a singleton set for all $\theta \in \Theta$. Hence every fully implementable SCR must be single-valued. ■

The Gibbard-Satterthwaite Theorem

In virtually all economic applications of interest, dictatorial SCRs are viewed as undesirable (recall that the social choice set $f(\theta)$ of a dictatorial SCR is a subset of the most preferred outcomes of a particular agent i in each state). For instance, democratic voting rules such as the majority rule are generically non-dictatorial. Also, the decision whether to provide a public good is typically not based on the valuation of a particular individual, but depends on the valuations of all the agents in the economy. Likewise, auctions do typically not assign the object in question to a single predetermined agent, but to the bidder with the highest valuation.

Given the prevalence of non-dictatorial SCRs, it is disturbing to learn that under some very general conditions, none of these SCRs is implementable in dominant strategies. In fact, the following theorem due to Gibbard (1973) and Satterthwaite (1975) tells us that when the domain of the agents' preference orderings is unrestricted, only dictatorial SCRs can be implemented in dominant strategies.

Definition 7 (Dictatorial Social Choice Rule) The social choice rule $f(\theta)$ is dictatorial on the set $A' \subseteq A$ if there exists an agent $i \in I$ such that for all $\theta \in \Theta$, the choice set $f(\theta)$ is a subset of agent i's most preferred outcomes in A', i.e. $f(\theta) \subseteq \{y \in A' \mid y \succsim_i (\theta_i)\, z \text{ for all } z \in A'\}$.

Theorem 4 (Gibbard-Satterthwaite Theorem) Let the social choice rule $f(\theta)$ be single-valued and let $A' \subseteq A$ denote the range of $f(\theta)$. Suppose that A is finite, that A' contains at least three elements, and that for each agent $i \in I$, the set of possible preference orderings $\Re_i$ is the set of strict preference orderings on A. Then $f(\theta)$ is TIDS if and only if it is dictatorial on A'.

Proof The "if"-part is obvious: Any dictatorial single-valued SCR is TIDS (assume that agent i is the dictator. In the direct mechanism Γ_d, we can then simply assign an element in the set $\{y \in A' \mid y \succsim_i (\theta_i)\, z \text{ for all } z \in A'\}$ to any profile of announcements $\hat{\theta}$ containing θ_i). The outline of the "only if"-part is as follows: From $f(\theta)$, we construct a *social welfare function* (SWF) $F(\theta)$ that is maximized by $f(\theta)$ and fulfills the conditions of Arrow's impossibility theorem. It follows that $F(\theta)$ is dictatorial, which in turn implies that $f(\theta)$ is dictatorial. ■

The full proof of the "only if"-part is lengthy and is omitted for the sake of brevity. For a complete version of the proof, see Green and Laffont (1979), theorem 2.2.

Remarks

1. Some intuition for the Gibbard-Satterthwaite theorem can be gained by looking at definition 6 which defines SCRs that are TIDS: In a simple setting with two agents and two possible types per agent, TIDS implies that an SCR must satisfy 16 incentive compatibility constraints. Given these restrictive requirements, it is somewhat less surprising that only dictatorial SCRs are TIDS.

2. In the face of this daunting result, we can choose between two possibilities: We can either relax the assumption of unrestricted preferences or we can abandon the concept of dominant strategy implementation altogether. In the remainder of this section, we pursue the first approach and assume that preferences are restricted to the quasilinear domain. The second approach is pursued in sections 2.3 and 2.4 where we study Nash and Bayesian implementation, respectively.

Groves Mechanisms

In the remainder of this section, we concentrate on the special, but much studied problem introduced at the beginning of this chapter whether to provide a public good. Consider the following additional assumptions:

9. A feasible outcome is a profile $y = (x, t_1, ..., t_n) \in A$ consisting of a decision $x \in \{0, 1\}$ and a vector of monetary transfers $t = (t_1, ..., t_n)$.

10. The agents' preferences are quasilinear of the form $\theta_i x + t_i$. Here, the type θ_i represents agent i's valuation or willingness to pay for the public good.

In the public good context, $x = 1$ means that the public good is provided and $x = 0$ means that it is not provided. However, the same framework can be used to represent an auction setting where an indivisible good is auctioned off to one of n agents. In this case, the decision x is a profile $x = (x_1, ..., x_n)$, where $x_i = 1$ means that agent i receives the good and $x_i = 0$ means that he does not receive the good (which imposes the additional constraint $\sum_i x_i = 1$). The agents' preferences then take the form $\theta_i x_i + t_i$. Without loss of generality, we can assume that the cost of the public good is zero (if the cost is $c \geq 0$, we simply let each agent pay an equal share $\frac{c}{n}$ and redefine the agents' valuations as $\overline{\theta}_i = \theta_i - \frac{c}{n}$, where θ_i is now a net valuation). Note that t_i can be negative.

When the agents' utilities are cardinally and interpersonally comparable, a reasonable social objective is the maximization of the utilitarian SWF

$F(\theta) = \sum_i (\theta_i x + t_i)$. The solution to this problem is $x = 1$ if and only if $\sum_i \theta_i \geq 0$, i.e. the public good is provided if and only if the sum of the agents' valuations exceeds the cost of the public good $c = 0$. Let us henceforth restrict attention to SCRs that meet this welfare criterion. Such SCRs are called *successful*.

Definition 8 (Successful Social Choice Rule) The social choice rule $f(\theta) = (x(\theta), t_1(\theta), ..., t_n(\theta))$ is successful if

$$x(\theta) = \begin{cases} 1 \text{ if } \sum_i \theta_i \geq 0 \\ 0 \text{ otherwise.} \end{cases} \tag{2.7}$$

Since successful SCRs have such desirable welfare properties, we would like to find out whether they are also TIDS. In what follows, we show that successful SCRs are indeed truthfully implementable in dominant strategies by a class of direct mechanisms known as *Groves mechanisms* due to Groves (1973).

Definition 9 (Groves Mechanism) A Groves mechanism Γ_G is a direct mechanism with

$$x(\hat{\theta}) = \begin{cases} 1 \text{ if } \sum_i \hat{\theta}_i \geq 0 \\ 0 \text{ otherwise,} \end{cases} \tag{2.8}$$

$$t_i(\hat{\theta}) = \begin{cases} \sum_{j \neq i} \hat{\theta}_j + h_i(\hat{\theta}_{-i}) \text{ if } \sum_i \hat{\theta}_i \geq 0 \\ h_i(\hat{\theta}_{-i}) \text{ otherwise,} \end{cases} \tag{2.9}$$

for all $i \in I$, where $\hat{\theta}$ is a profile of announcements, and where $h_i(\hat{\theta}_{-i})$ is an arbitrary function of $\hat{\theta}_{-i}$.

In a Groves mechanism, agent i's transfer $t_i(\hat{\theta})$ depends on his announcement $\hat{\theta}_i$ only insofar as this announcement affects the decision $x(\hat{\theta})$, given the announcements $\sum_{j \neq i} \hat{\theta}_j$ of the remaining $j \neq i$ agents. If $x(\hat{\theta})$ is changed, agent i's transfer is reduced by an amount equal to the sum of the other agents' valuations $\mid \sum_{j \neq i} \hat{\theta}_j \mid$, which corresponds exactly to the negative externality that agent i is imposing on these agents. Since externalities are now fully internalized, agent i no longer benefits from free riding at the expense of the other agents by misreporting his type and truthtelling becomes a dominant strategy.

Theorem 5 In a Groves mechanism Γ_G, truthtelling is a dominant strategy.

Proof (direct) Denote the true and the announced type of agent i by θ_i and $\hat{\theta}_i$, respectively, and denote the announcements of the other $j \neq i$ agents by $\hat{\theta}_{-i}$.

i) Suppose that $\theta_i + \sum_{j \neq i} \hat{\theta}_j \geq 0$. Any announcement $\hat{\theta}_i \neq \theta_i$ such that $\hat{\theta}_i + \sum_{j \neq i} \hat{\theta}_j \geq 0$ yields the same utility as truthtelling, viz. $\theta_i + \sum_{j \neq i} \hat{\theta}_j + h_i(\hat{\theta}_{-i})$. Any announcement $\hat{\theta}_i \neq \theta_i$ such that $\hat{\theta}_i + \sum_{j \neq i} \hat{\theta}_j < 0$ yields $h_i(\hat{\theta}_{-i})$, which is less than or equal to $\theta_i + \sum_{j \neq i} \hat{\theta}_j + h_i(\hat{\theta}_{-i})$ since $\theta_i + \sum_{j \neq i} \hat{\theta}_j \geq 0$. Thus, agent i is never worse off by telling the truth.

ii) Suppose now that $\theta_i + \sum_{j \neq i} \hat{\theta}_j < 0$. Any announcement $\hat{\theta}_i \neq \theta_i$ such that $\hat{\theta}_i + \sum_{j \neq i} \hat{\theta}_j < 0$ yields the same utility as truthtelling, viz. $h_i(\hat{\theta}_{-i})$. Any announcement $\hat{\theta}_i \neq \theta_i$ such that $\hat{\theta}_i + \sum_{j \neq i} \hat{\theta}_j \geq 0$ yields $\theta_i + \sum_{j \neq i} \hat{\theta}_j + h_i(\hat{\theta}_{-i})$, which is less than $h_i(\hat{\theta}_{-i})$ since $\theta_i + \sum_{j \neq i} \hat{\theta}_j < 0$. Again, agent i is never worse off by telling the truth. ■

Since truthtelling is a dominant strategy, (2.7) and (2.8) coincide, from which it follows that the Groves mechanism truthfully implements successful SCRs in dominant strategies. Perhaps even more intriguing is the following result, which states that the Groves mechanism Γ_G is the only mechanism that truthfully implements successful SCRs in dominant strategies.

Theorem 6 Any mechanism Γ_d that truthfully implements a successful SCR in dominant strategies coincides with the Groves mechanism.

Proof (indirect) We prove the logically equivalent statement that any mechanism Γ_d that does not coincide with the Groves mechanism cannot truthfully implement a successful SCR in dominant strategies. From definition 9, a mechanism Γ_d is a Groves mechanism if and only if it has the following properties:

i) $x(\hat{\theta}) = 1$ if and only if $\sum_i \hat{\theta}_i \geq 0$,

ii) Given $\hat{\theta}_{-i}$, $t_i(\hat{\theta})$ is constant for all i and $\hat{\theta}_i$ such that $\hat{\theta}_i + \sum_{j \neq i} \hat{\theta}_j \geq 0$,

iii) Given $\hat{\theta}_{-i}$, $t_i(\hat{\theta})$ is constant for all i and $\hat{\theta}_i$ such that $\hat{\theta}_i + \sum_{j \neq i} \hat{\theta}_j < 0$,

iv) $x(\hat{\theta}_i, \hat{\theta}_{-i}) = 1$ and $x(\hat{\theta}'_i, \hat{\theta}_{-i}) = 0$ imply $t_i(\hat{\theta}_i, \hat{\theta}_{-i}) - t_i(\hat{\theta}'_i, \hat{\theta}_{-i}) = \sum_{j \neq i} \hat{\theta}_j$.

We now show that if Γ_d lacks one or more of these properties, then it either violates success or TIDS or both. Since each property can either hold or fail, there are 2^4-1=15 possible events which contain at least one failure.

1) Suppose property i) does not hold (8 cases). Then Γ_d either violates success (if $\hat{\theta}$ is the true state) or TIDS (if $\hat{\theta}$ is not the true state).

2) Assume that property i) holds but that property ii) fails (4 cases). Then there exists an agent i and announcements θ_i (truthtelling), $\hat{\theta}_i$, and $\hat{\theta}_{-i}$ such that $\hat{\theta}_i + \sum_{j \neq i} \hat{\theta}_j \geq 0$, $\theta_i + \sum_{j \neq i} \hat{\theta}_j \geq 0$ (i.e. given $\hat{\theta}_{-i}$, both $\hat{\theta}_i$ and θ_i yield $x = 1$) and $t(\hat{\theta}_i, \hat{\theta}_{-i}) > t(\theta_i, \hat{\theta}_{-i})$. It follows that truthtelling is not a dominant strategy for agent i.

3) Suppose properties i) and ii) hold but property iii) does not hold (2 cases). Then there exists an agent i and announcements θ_i (truthtelling), $\hat{\theta}_i$, and $\hat{\theta}_{-i}$ such that $\hat{\theta}_i + \sum_{j \neq i} \hat{\theta}_j < 0$, $\theta_i + \sum_{j \neq i} \hat{\theta}_j < 0$ (i.e. given $\hat{\theta}_{-i}$, both $\hat{\theta}_i$ and θ_i yield $x = 0$) and $t(\hat{\theta}_i, \hat{\theta}_{-i}) > t(\theta_i, \hat{\theta}_{-i})$. Again, it follows that truthtelling is not a dominant strategy for agent i.

4) Finally, assume that properties i), ii), and iii) hold and that property iv) fails (1 case). Then $x(\hat{\theta}_i, \hat{\theta}_{-i}) = 1$ and $x(\hat{\theta}_i', \hat{\theta}_{-i}) = 0$ imply that either a) $t_i(\hat{\theta}_i, \hat{\theta}_{-i}) - t_i(\hat{\theta}_i', \hat{\theta}_{-i}) = \sum_{j \neq i} \hat{\theta}_j - \epsilon$ or b) $t_i(\hat{\theta}_i, \hat{\theta}_{-i}) - t_i(\hat{\theta}_i', \hat{\theta}_{-i}) = \sum_{j \neq i} \hat{\theta}_j + \epsilon$ is true for some $\epsilon > 0$. Let us only consider the former possibility (the proof of b) is along the same lines). If a) holds, then there exists an agent i and announcements $\hat{\theta}_i = \theta_i$ (truthtelling), $\hat{\theta}_i'$, and $\hat{\theta}_{-i}$ such that $\hat{\theta}_i' + \sum_{j \neq i} \hat{\theta}_j < 0$, $\theta_i = -\sum_{j \neq i} \hat{\theta}_j + \frac{\epsilon}{2}$ (hence $\theta_i + \sum_{j \neq i} \hat{\theta}_j > 0$) and $t_i(\theta_i, \hat{\theta}_{-i}) - t_i(\hat{\theta}_i', \hat{\theta}_{-i}) = \sum_{j \neq i} \hat{\theta}_j - \epsilon$. Agent i's utility from telling the truth is $\theta_i + t_i(\theta_i, \hat{\theta}_{-i}) = t_i(\hat{\theta}_i', \hat{\theta}_{-i}) - \frac{\epsilon}{2}$, whereas his utility from announcing $\hat{\theta}_i'$ is $t_i(\hat{\theta}_i', \hat{\theta}_{-i}) > t_i(\hat{\theta}_i', \hat{\theta}_{-i}) - \frac{\epsilon}{2}$. It follows that truthtelling is not a dominant strategy for agent i. ■

One implication of theorem 6 is that the only SCRs that are both successful and TIDS are those defined in (2.8)-(2.9). We can think of many examples where we may want to place even further restrictions on the transfer function $t_i(\theta)$. For instance, in some cases the planner may not be allowed to run a deficit when implementing $f(\theta)$, i.e. the sum $\sum_i t_i(\theta)$ must not exceed zero. An SCR with this property is called *feasible.*

Definition 10 (Feasible Social Choice Rule) An SCR is feasible if $\sum_i t_i(\theta) \leq 0$ for all $\theta \in \Theta$.

An even stronger requirement is that $f(\theta)$ be *budget-balanced*, i.e. that the sum of transfers $\sum_i t_i(\theta)$ be identically equal to zero.

Definition 11 (Budget-Balanced Social Choice Rule) An SCR is budget-balanced if $\sum_i t_i(\theta) = 0$ for all $\theta \in \Theta$.

Since the planner's preferences do not enter into our welfare considerations, any net surplus $|\sum_i t_i(\theta)| > 0$ collected from the agents is wasteful. It is therefore not surprising that the social choice rule $f^*(\theta)$ which maximizes the utilitarian SWF $F(\theta) = \sum_i (\theta_i x + t_i)$ subject to the feasibility constraint $\sum_i t_i(\theta) \leq 0$ is both successful and budget-balanced. Such an SCR is called *ex-post efficient.*

Definition 12 (Ex-Post Efficient Social Choice Rule) An SCR is ex-post efficient if it is both successful and budget-balanced.

Next, we will address the question whether ex-post efficient SCRs are TIDS. An important member of the class of Groves mechanisms is the *pivotal* or *Clarke mechanism* suggested by Clarke (1971). It has the convenient property that an agent's transfer $t_i(\hat{\theta})$ is zero unless he is pivotal, i.e. unless his announcement $\hat{\theta}_i$ affects the decision $x(\hat{\theta})$, in which case $t_i(\hat{\theta})$ is negative and equal to the externality $\mid \sum_{j \neq i} \hat{\theta}_j \mid$ imposed on the other agents. Thus, the Clarke mechanism embodies a fundamental principle of resource allocation in the presence of externalities according to which an individual should compensate the others for the harm he causes (albeit with the caveat that the transfers go to the planner and not to the other agents).

Definition 13 (Clarke Mechanism) A Clarke mechanism Γ_C is a Groves mechanism where for all $i \in I$

$$h_i(\hat{\theta}_{-i}) = \min\left(-\sum\nolimits_{j \neq i} \hat{\theta}_j, 0\right). \tag{2.10}$$

Since any Groves scheme induces truthtelling, we henceforth set $\hat{\theta} = \theta$ for ease of notation. The transfers $t_i(\hat{\theta})$ that ensue from (2.7) can be depicted in a 2×2 matrix as follows:

	$\sum_{j \neq i} \theta_j < 0$	$\sum_{j \neq i} \theta_j \geq 0$
$\int \sum_i \theta_i < 0$	0	$-\sum_{j \neq i} \theta_j$
$\int \sum_i \theta_i \geq 0$	$\sum_{j \neq i} \theta_j$	0

In what follows, we show that there exists no mechanism Γ_d that truthfully implements ex-post efficient SCRs in dominant strategies. From theorem 6, we know that if such a mechanism exists, it must belong to the class of Groves mechanisms. First, let us consider the Clarke mechanism Γ_C. A simple counterexample reveals that the Clarke mechanism entails a strictly positive surplus $|\sum_i t_i(\theta)| > 0$ in some states, which implies that it is not budget-balanced and therefore unable to implement budget-balanced SCRs.

Theorem 7 The Clarke mechanism Γ_C is not budget-balanced.

Proof (by contradiction) Suppose Γ_C is budget-balanced. Then $\sum_i t_i(\theta) = 0$ must hold for all $\theta \in \Theta$. Consider the example where $n = 3$ and $\theta = (-1, -1, 3)$. The Clarke transfers are $t = (0, 0, -2)$, which implies $\sum_i t_i(\theta) < 0$, a contradiction. ■

Next, we prove that there is no other feasible mechanism in the class of Groves mechanisms whose surplus dominates that of the Clarke mechanism.

Theorem 8 There exists no feasible Groves mechanism Γ_G with $|\sum_i t_i(\theta)| \leq |\sum_i \bar{t}_i(\theta)|$ for all $\theta \in \Theta$ and $|\sum_i t_i(\theta)| < |\sum_i \bar{t}_i(\theta)|$ for some $\theta \in \Theta$, where $\bar{t}_i(\theta)$ denotes the Clarke transfer to agent i.

The proof is lengthy and is omitted here. See Laffont and Maskin (1982), theorem 3.3. for a complete proof.

Theorems 7 and 8 together imply that no budget-balanced Groves mechanism exists. But by theorem 6, Groves mechanisms are the only candidates for truthful implementation of ex-post efficient SCRs. This leads to the following trivial, but important corollary:

Corollary 1 There exists no SCR that is both TIDS and ex-post efficient.

Remarks

1. While there is no SCR that is both ex-post efficient and TIDS, there do exist SCRs which are successful, feasible, and TIDS. In fact, any SCR that is implemented by the Clarke mechanism has this property.

2. More encouraging results can be obtained when the underlying equilibrium concept is weakened so that less incentive compatibility constraints must be satisfied. For instance, in sections 2.3 and 2.4 we will show that ex-post efficient SCRs are implementable in Nash and Bayesian equilibrium, respectively.

3. At the beginning of this subsection, we pointed out the analogy between the public good model and a setting where an indivisible good is auctioned off to one of n agents. In the auction setting, the analogue of the Clarke mechanism is known as *second-price sealed-bid* or *Vickrey auction*. There, agent i is pivotal if he is the bidder with the highest valuation, and his transfer $t_i(\theta)$ is equal to the second-highest valuation $\max\{\theta_j \mid j \neq i\}$, which is again the externality caused by agent i.

4. Despite the striking similarity, the public good setting and the auction setting differ in an important aspect: In the auction setting, the planner coincides with the seller, whose utility enters into our welfare considerations. An immediate consequence of this is that the SCR implemented by the Clarke-Vickrey mechanism is then ex-post efficient since the surplus $|\sum_i t_i(\theta)| > 0$ collected by the planner ("agent 0") is no longer wasteful. More generally, any profile of transfers is compatible with a Pareto optimal allocation.

Finally, we look at situations where the planner cannot force the agents to take part in the game. When participation is voluntary, an implementable SCR must satisfy an additional set of constraints known as *individual rationality* or *participation constraints*. These constraints ensure that each agent can guarantee himself a certain minimum utility (typically normalized to zero) by telling the truth.

Definition 14 (Ex-Post Individually Rational Social Choice Rule) An SCR is ex-post individually rational if $\theta_i x(\theta) + t(\theta)_i \geq 0$ for all $i \in I$ and $\theta \in \Theta$.

In the absence of voluntary participation, it was shown that successful and feasible SCRs can be truthfully implemented in dominant strategies. When individual rationality constraints are added, this turns out impossible.

Theorem 9 There exists no SCR that is TIDS, successful, feasible, and ex-post individually rational.

Proof (by contradiction) Suppose there exist SCRs that are TIDS, successful, feasible, and ex-post individually rational. By theorems 6 and 5, we can restrict attention to Groves mechanisms. Choose a profile $\theta = (\theta_1, ..., \theta_n)$ such that $\sum_i \theta_i \geq 0$ and such that for all $i \in I$, there exists a θ_i' with $\theta_i' + \sum_{j \neq i} \theta_j < 0$. Given that the other agents are of type θ_{-i}, the function $h_i(\theta_{-i})$ to agent i is the same for type θ_i or type θ_i'. If agent i is of type θ_i', success implies $x = 0$. From individual rationality, it then follows that $h_i(\theta_{-i}) \geq 0$. This is true for all $i \in I$. Consider now the case where $\theta = (\theta_1, ..., \theta_n)$. Success implies $x = 1$, and feasibility requires

$$\sum_i t_i = \sum_i \sum_{j \neq i} \theta_j + \sum_i h_i(\theta_{-i}) = (n-1) \sum_i \theta_i + \sum_i h_i(\theta_{-i}) \leq 0. \quad (2.11)$$

Since $\sum_i \theta_i \geq 0$, this implies $\sum_i h_i(\theta_{-i}) < 0$, a contradiction. ■

2.3 Nash Implementation

Direct vs. Indirect Mechanisms

The previous section made clear that very little is implementable in dominant strategies: For unrestricted preference domains, it was shown that only dictatorial SCRs are TIDS, and when preferences were restricted to the quasilinear domain, it was shown that no SCR is both ex-post efficient and TIDS. A less restrictive equilibrium concept than dominant strategy equilibrium is

Nash equilibrium. For a strategy s_i^* to be an equilibrium strategy, it need only be optimal with respect to the other players' equilibrium strategies s_{-i}^*, as opposed to dominant strategy equilibrium, where s_i^* must be optimal with respect to any profile s_{-i} in the other players' strategy set $\times S_{-i}$.

In section 2.2, it turned out to be convenient to restrict attention to direct mechanisms and concentrate on SCRs that are truthfully implementable.

Definition 15 (TINS Social Choice Rule) The social choice rule $f(\theta)$ is truthfully implementable in Nash strategies (TINS) if there exists a direct mechanism Γ_d such that i) truthtelling is a Nash equilibrium, i.e. if for all $i \in I$, $\theta_i \in \Theta_i$, and $\theta_{-i} \in \Theta_{-i}$,

$$g(\theta_i, \theta_{-i}) \succsim_i (\theta_i)\, g(\hat{\theta}_i, \theta_{-i}) \tag{2.12}$$

for all $\hat{\theta}_i \in \Theta_i$, and ii) $g(\theta) \in f(\theta)$ for all $\theta \in \Theta$.

We could now proceed by deriving a corresponding version of theorem 1 (revelation principle) for Nash implementation. Unfortunately, restricting attention to direct mechanisms does not get us any further since the set of SCRs that are TINS is no greater than the set of SCRs that are TIDS, which brings us back to dominant strategy implementation.

Theorem 10 An SCR is TINS if and only if it is TIDS.

Proof (direct) The "if"-part is obvious, since any dominant strategy equilibrium is a Nash equilibrium.

"Only if"-part: Assume that the direct mechanism Γ_d truthfully implements $f(\theta)$ in Nash strategies. Then for all i and θ_i, $\hat{\theta}_i$, and $\theta_{-i} \in \Theta_{-i}$, $g(\theta_i, \theta_{-i}) \succsim_i (\theta_i)\, g(\hat{\theta}_i, \theta_{-i})$, from which it follows that truthtelling is a dominant strategy. ■

The "only if"-part is nothing but a restatement of definition 15. There, we required that θ_i is optimal with respect to any possible profile θ_{-i} in the other players' strategy set Θ_{-i}, which coincides with the definition of a dominant strategy. Alternatively, by setting $\theta_{-i} = \hat{\theta}_{-i}$, we see that definitions 15 and 6 are equivalent. Intuitively, remember that the direct mechanism Γ_d induces a game of complete information in every state θ. In any such game, TINS requires only that announcing θ_i is optimal with respect to the profile of true announcements θ_{-i}. Thus, fix θ_i and consider the states (θ_i,θ_{-i}), (θ_i,θ'_{-i}), (θ_i,θ''_{-i}) etc. such that $\{\theta_{-i}, \theta'_{-i}, \theta''_{-i}, ...\} = \Theta_{-i}$. Clearly, requiring that θ_i is optimal with respect to any possible profile of truthful reports $\theta_{-i}, \theta'_{-i}, \theta''_{-i}, ... \in \Theta_{-i}$ amounts to the same as requiring that θ_i is a dominant strategy.

The reasoning which underlies theorem 10 does not apply if we define a strategy set in Γ_d as the entire state space $\Theta = \times\Theta_i$. That is, instead of having agents announce their types θ_i, we let each agent announce a complete profile $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_n)$. In this case, SCRs that are TINS are no longer automatically TIDS, and we can hope that by restricting attention to direct mechanisms, more SCRs are truthfully implementable in Nash strategies than in dominant strategies. Unfortunately, it turns out that with this enlarged definition of strategy sets, *any* SCR is TINS. To see this, let the true state be θ and consider the direct mechanism Γ_d which implements outcome $y \in f(\hat{\theta})$ if and only if all agents announce the same $\hat{\theta}$. If one or more agents disagree, a "bad" outcome is implemented (e.g. all agents are shot). Clearly, this mechanism has $\hat{\theta} = \theta$ as a Nash equilibrium and thus truthfully implements $f(\theta)$. But any common report $\hat{\theta} \neq \theta$ is also a Nash equilibrium, from which it follows that $f(\theta)$ is not implemented unless all equilibria have outcomes that are in $f(\theta)$.

Although there exists a version of the revelation principle for Nash implementation (cf. Repullo (1986), theorem 6.1), it is of little use here because restricting attention to direct mechanisms implies a great loss of generality due to the multiple equilibrium problem. In the remainder of this section, we therefore focus on indirect mechanisms.

Necessary and Sufficient Conditions for Implementation

Consider a social choice rule $f(\theta)$ that is implementable in Nash strategies and select a particular state θ'. This state has at least one Nash equilibrium whose outcome, say y, is in $f(\theta')$. Next, select a different state θ'' and assume that y (weakly) moves up everyone's ranking when switching from θ' to θ''. Clearly, the Nash equilibrium with outcome y continues to be a Nash equilibrium in state θ''. By the definition of implementability, y must therefore also be an element of $f(\theta'')$. An SCR with the property that an outcome y that is an element of $f(\theta')$ and does not fall in anyone's ranking when moving from θ' to θ'' is also an element of $f(\theta'')$ is called *monotonic*. By the above reasoning, any Nash implementable SCR must be monotonic.

Definition 16 (Monotonic Social Choice Rule) The social choice rule $f(\theta)$ is monotonic if for all $y \in A$ and $\theta', \theta'' \in \Theta$, the following holds: if i) $y \in f(\theta')$ and ii) for all $i \in I$ and $z \in A$, $y \succsim_i (\theta')\, z \Rightarrow y \succsim_i (\theta'')\, z$, then $y \in f(\theta'')$.

An immediate consequence of definition 16 is the following property of monotonic SCRs: If $y \in f(\theta')$ and $y \notin f(\theta'')$, then there exists at least one

agent i and outcome z such that $y \succsim_i (\theta_i')z$ and $z \succ_i (\theta_i'')y$. Let us henceforth call z a *test outcome* with respect to (y, θ', θ'') and the agent i for whom this preference reversal holds a *test agent* with respect to (y, θ', θ'').

The assertion that Nash implementability implies monotonicity is part of a theorem due to Maskin (1977), which is considered as the most important result in the theory of Nash implementation.

Theorem 11 (Maskin's Theorem I: Necessity) If an SCR is implementable in Nash strategies, then it is monotonic.

Proof (by contradiction) If $f(\theta)$ is implementable in Nash strategies, there exists a state $\theta' \in \Theta$ and a Nash equilibrium profile $s^*(\theta') = (s_i^*(\theta'), s_{-i}^*(\theta'))$ with $g(s^*(\theta')) \in f(\theta')$. By the definition of Nash equilibrium,

$$g(s_i^*(\theta'), s_{-i}^*(\theta')) \succsim_i (\theta_i') g(s_i, s_{-i}^*(\theta')) \tag{2.13}$$

for all $i \in I$ and $s_i \in S_i$. Suppose now that $f(\theta)$ is not monotonic. Then there exists a state $\theta'' \neq \theta'$ such that

$$g(s_i^*(\theta'), s_{-i}^*(\theta')) \succsim_i (\theta_i'') g(s_i, s_{-i}^*(\theta')) \tag{2.14}$$

for all $i \in I$ and $s_i \in S_i$ but $g(s^*(\theta')) \notin f(\theta'')$. However, from (2.14) it follows that the profile $s^*(\theta')$ with outcome $g(s^*(\theta'))$ continues to be a Nash equilibrium in state θ'', which contradicts the assumption that $f(\theta)$ is implementable. ■

Remarks

1. Since full implementation implies implementation, theorem 11 continues to hold if we substitute "implementable" with "fully implementable".

2. For the case of single-valued SCRs, it can be shown that an SCR is implementable in Nash strategies only if it is truthfully implementable in dominant strategies (given that the domain of preferences is monotonically closed - a property that we will not define here). Thus, nothing is gained from weakening the underlying equilibrium concept. Even worse, under some fairly innocuous assumptions, a result reminiscent of the Gibbard-Satterthwaite theorem can be proven which says that an SCR is implementable in Nash strategies if and only if it is dictatorial. For a proof, see Dasgupta, Hammond, and Maskin (1979), theorem 7.2.3. and corollary 7.2.5.

3. Monotonicity is satisfied by such common SCRs as the Paretian choice rule and the majority rule (if $\Re$ consists of strict orderings). Furthermore, monotonicity is closely related to Arrow's well-known "independence of irrelevant alternatives" condition. For instance, suppose that by switching from state θ' to state θ'', the set $L(y)$ of outcomes that all agents value less than y remains the same ($L(y)$ is called the *lower contour set* with respect to y), but that the relative rank order of some of the elements in $L(y)$ changes for some agents. Then, monotonicity requires that if y is in $f(\theta')$, it must also be in $f(\theta'')$, regardless of the changes in $L(y)$.

4. Monotonicity rules out interpersonal comparisons of the kind inherent in utilitarian or Rawlsian SCRs: The only thing that matters when switching from state θ' to θ'' is that no agent values y less than before. Whether some agents value y much higher while others value y only slightly higher is inconsequential.

The second part of Maskin's theorem shows that monotonicity together with the additional condition of *no veto power* implies full implementability.

Definition 17 (No Veto Power) The social choice rule $f(\theta)$ satisfies no veto power if for all $i \in I$ and $y \in A$, the following holds: if for all $j \neq i$ and $z \in A$, $y \succsim_j (\theta_j) z$, then $y \in f(\theta)$.

In words, no veto power says that whenever in some state θ an outcome y is top-ranked for $n-1$ agents, then y should be in the choice set $f(\theta)$, i.e. the remaining agent cannot veto it.

Theorem 12 (Maskin's Theorem II: Sufficiency) Suppose $n \geq 3$. If an SCR is monotonic and satisfies no veto power, then it is fully implementable in Nash strategies.

Proof (by contradiction) Consider the following mechanism: Each agent announces a state, an outcome, and a nonnegative integer.

i) If all agents agree on some state θ and outcome $y \in f(\theta)$, then y is implemented.

ii) If $n-1$ agents agree on some state θ and outcome $y \in f(\theta)$, then y is implemented unless the remaining agent i announces a state θ' and outcome z such that I) $y \notin f(\theta')$, II) i is a test agent for (y, θ, θ'), and III) z is a test outcome for (y, θ, θ'), in which case z is implemented.

iii) In all other cases, the outcome of the agent with the highest integer is implemented.

First, we show that if the true state is θ, there exists a Nash equilibrium for each $y \in f(\theta)$ where all agents announce (y, θ). Suppose that this is not the case. Then there must exist an agent i who strictly benefits from announcing a pair (z, θ') that satisfies conditions I)-III) (any other unilateral deviation leads to y being implemented and thus cannot be strictly profitable). By definition, i is a test agent and z is a test outcome for (y, θ, θ'), from which it follows that $y \succsim_i (\theta_i) z$. However, strict profitability implies that $z \succ_i (\theta_i) y$, a contradiction.

Next, we show that if the true state is θ, no Nash equilibrium with outcome $z \notin f(\theta)$ exists. Suppose that such an equilibrium exists. From i)-iii), we conclude that this Nash equilibrium must belong to one of the following two categories:

1) All agents agree on some state θ' and outcome z, where $z \in f(\theta')$ and $z \notin f(\theta)$. By monotonicity, there exists a test outcome y and a test agent i who strictly prefers to unilaterally deviate by announcing (y, θ) a contradiction.

2) $n-1$ agents agree on some state and outcome and the remaining agent disagrees. Denote the outcome from this equilibrium by $z \notin f(\theta)$. There are two possibilities: a) There exists an $x \in A$ such that one of the $n-1$ agents strictly prefers x to z. This agent is strictly better off by unilaterally deviating from the proposed equilibrium, a contradiction (he will announce a different state, together with x and some integer that exceeds the integers of the other agents). b) z is top-ranked for all $n-1$ agents. By no veto power, $z \in f(\theta)$, which contradicts the assumption that $z \notin f(\theta)$. ■

Remarks

1. The role of the integers is solely to rule out Nash equilibria based on step iii) of the mechanism with unwanted outcomes $z \notin f(\theta)$. Since the set of integers is unbounded, such equilibria cannot exist. For any profile of integers announced by the remaining $j \neq i$ agents, agent i would always want to deviate by announcing a higher integer.

2. The assumption of no veto power is trivially satisfied in any example with $n \geq 3$ agents in which there is a private good that yields positive utility (e.g. money). Because each agent wants to have all of the private good himself, there cannot exist situations in which $n-1$ agents agree on the same outcome. Then, monotonicity constitutes both a necessary and sufficient condition for full Nash implementability.

3. When $n = 2$, a result similar to theorem 12 can be proven if in addition to monotonicity, $f(\theta)$ satisfies a condition called *restricted no veto power*. See Moore and Repullo (1990), corollary 3, for details.

The Public Good Problem Revisited

In section 2.2, we saw that even if preferences are restricted to the quasi-linear domain, ex-post efficient SCRs cannot be implemented in dominant strategies. We now reconsider the public good problem studied earlier and show that with Nash implementation, life is much easier.

First, note that no veto power is trivially satisfied in the public good setting since preferences depend positively on the monetary transfers t_i. It then follows from Maskin's theorem that an SCR is fully implementable in Nash strategies if and only if it is monotonic. We can now apply definition 16 to the present context and conclude that an SCR is monotonic iff

$$(0, t_1, ..., t_n) \in f(\theta) \text{ and } \theta \geq \theta' \text{ implies } (0, t_1, ..., t_n) \in f(\theta'), \tag{2.15}$$

$$(1, t_1, ..., t_n) \in f(\theta) \text{ and } \theta \leq \theta' \text{ implies } (1, t_1, ..., t_n) \in f(\theta'). \tag{2.16}$$

In words: If an outcome y in the choice set implies $x = 0$ and the profile of valuations $\theta = (\theta_1, ..., \theta_n)$ decreases in the vector sense, then y should remain in the choice set. Conversely, if an element in the choice set implies $x = 1$ and the profile θ increases in the vector sense, then y should remain in the choice set. Notice that a decrease in θ implies an increase in each agent's valuation for outcomes that contain $x = 0$.

Conditions (2.15)-(2.16) reveal that monotonicity places only very little restriction on the vector of transfers $(t_1, ..., t_n)$. In particular, the following ex-post efficient SCR is monotonic and therefore Nash implementable:

$$x(\theta) = \begin{cases} 1 \text{ if } \sum_i \theta_i \geq 0 \\ 0 \text{ otherwise,} \end{cases} \tag{2.17}$$

$$t_i(\theta) = 0 \text{ for all } i \in I. \tag{2.18}$$

2.4 Bayesian Implementation

The Revelation Principle

We now turn to environments where agents cannot observe each others' preferences. As was remarked earlier, agents who possess a dominant strategy will use it even if they do not have complete information about the other agents' types. One implication of this is that the results derived in connection with dominant strategy implementation continue to hold in the presence of incomplete information. In this section, we employ the weaker concept of Bayesian (Nash) equilibrium. Consider the following assumptions which replace assumption 5 of the basic model:

5a. Each agent observes only his own type θ_i, i.e. agents have *incomplete information*.

5b. The profile of types $\theta = (\theta_1, ..., \theta_n)$ is drawn from the set $\Theta = \times\Theta_i$ according to the distribution function $\Pi(\theta)$ with density $\pi(\theta)$, which is common knowledge.

5c. The agents' types are statistically *independent*, i.e. $\pi(\theta) = \prod_{i \in I} \pi_i(\theta_i)$.

The assumption of common knowledge in 5b is important and known in game theory as *common prior assumption* or *Harsanyi doctrine*. It is crucial, because in a game of incomplete information a player's strategy not only depends on his beliefs about $\pi(\theta)$, but also on his beliefs about others' beliefs about $\pi(\theta)$, beliefs about beliefs about beliefs, etc.

If assumption 5c fails and types are correlated, a "shoot-them-all" mechanism along the lines of that presented in section 2.2 can be used to truthfully implement any social choice rule $f(\theta)$ as if information was complete. However, as was also shown in section 2.2, *any* common report is then an equilibrium and we are very far from actually implementing $f(\theta)$.

In environments with incomplete information, a mechanism Γ combined with the state-space Θ and density $\pi(\theta)$ defines a game of incomplete information with a (possibly) different payoff structure for every $\theta \in \Theta$. From now on, let $u_i(y, \theta_i)$ denote agent i's von Neumann-Morgenstern utility over outcomes when he is of type θ_i. As in the previous sections, we are especially interested in SCRs that are truthfully implementable.

Definition 18 (TIBS Social Choice Rule) The social choice rule $f(\theta)$ is truthfully implementable in Bayesian strategies (TIBS) if there exists a direct mechanism Γ_d such that i) truthtelling is a Bayesian equilibrium, i.e. if for all $i \in I$ and $\theta_i \in \Theta_i$,

$$\int_{\Theta_{-i}} u_i(g(\theta_i, \theta_{-i}), \theta_i)\, d\Pi_{-i}(\theta_{-i}) \geq \int_{\Theta_{-i}} u_i(g(\hat{\theta}_i, \theta_{-i}), \theta_i) d\Pi_{-i}(\theta_{-i}) \quad (2.19)$$

for all $\hat{\theta}_i \in \Theta_i$, and ii) $g(\theta) \in f(\theta)$ for all $\theta \in \Theta$.

According to (2.19), truthtelling need only be optimal in expected terms, which is a weaker requirement than both TIDS (where $\hat{\theta}_i = \theta_i$ is to be optimal for any profile $\hat{\theta}_{-i}$) and TINS (where $\hat{\theta}_i = \theta_i$ is to be optimal for any profile of truthful reports θ_{-i}).

If the density function $\pi(\theta)$ is degenerate (i.e. if it has point mass only on a single vector θ), Bayesian equilibrium reduces to ordinary Nash equilibrium. Therefore, if we require that an SCR is implementable for *all* densities

(including degenerate ones), we inevitably run into the problems associated with Nash implementation. In particular, there is no point in restricting attention to direct mechanisms as was shown in theorem 10.

Theorem 13 An SCR is TIBS for all possible densities $\pi(\theta)$ if and only if it is TIDS.

Proof (indirect) The "if"-part is obvious, since any dominant strategy equilibrium is a Bayesian equilibrium.

"Only if"-part: Suppose $f(\theta)$ is not truthfully implementable in dominant strategies. Then there exists an $i \in I$ and a $\theta' \in \Theta$ such that truthtelling is not a Nash equilibrium strategy for agent i, given the profile of truthful announcements θ'_{-i}. Now let $\pi(\theta') = 1$. Then truthtelling is not a Bayesian equilibrium strategy either. ■

In the remainder of this section, we confine ourselves to density functions $\pi(\theta)$ which are not degenerate. Analogous to theorem 1, we can now derive a version of the revelation principle for Bayesian equilibrium.

Theorem 14 (Revelation Principle) If an SCR is implementable in Bayesian strategies, then it is TIBS.

Proof (direct) Suppose that Γ implements the social choice rule $f(\theta)$ in Bayesian strategies, and let $E_g(\theta)$ be non-empty for all θ. As in the proof of theorem 1, $s^* : \Theta \to \times S_i$ is a mapping which selects exactly one equilibrium profile $s^*(\theta) \in E_g(\theta)$ for each $\theta \in \Theta$. Since $s^*(\theta)$ is a profile of Bayesian equilibrium strategies, we have

$$\begin{aligned} & \int_{\Theta_{-i}} u_i\left(g(s_i^*(\theta_i), s_{-i}^*(\theta_{-i})), \theta_i\right) d\Pi_{-i}(\theta_{-i}) \qquad (2.20) \\ \geq & \int_{\Theta_{-i}} u_i\left(g(s_i, s_{-i}^*(\theta_{-i})), \theta_i\right) d\Pi_{-i}(\theta_{-i}) \end{aligned}$$

for all $i \in I$, $\theta_i \in \Theta_i$, and $s_i \in S_i$. In particular, it is true that

$$\begin{aligned} & \int_{\Theta_{-i}} u_i\left(g(s_i^*(\theta_i), s_{-i}^*(\theta_{-i})), \theta_i\right) d\Pi_{-i}(\theta_{-i}) \qquad (2.21) \\ \geq & \int_{\Theta_{-i}} u_i(g(s_i^*(\hat{\theta}_i), s_{-i}^*(\theta_{-i})), \theta_i) d\Pi_{-i}(\theta_{-i}) \end{aligned}$$

for all $i \in I$ and $\theta_i, \hat{\theta}_i \in \Theta_i$, since $s_i^*(\hat{\theta}_i) \in S_i$ is merely a specific strategy rule.

Next, define the composed mapping $h : \Theta \to A$ with $h(\theta) \equiv g(s^*(\theta))$. The function $h(\theta)$ together with the collection of possible types $\{\Theta_1, ..., \Theta_n\}$ constitutes a direct mechanism Γ_d. But Γ_d truthfully implements $f(\theta)$ in Bayesian strategies because $h(\theta) \equiv g(s^*(\theta)) \in f(\theta)$ and

$$\int_{\Theta_{-i}} u_i\left(h\left(\theta_i, \theta_{-i}\right), \theta_i\right) d\Pi_{-i}\left(\theta_{-i}\right) \geq \int_{\Theta_{-i}} u_i(h(\hat{\theta}_i, \theta_{-i}), \theta_i) d\Pi_{-i}\left(\theta_{-i}\right) \quad (2.22)$$

for all $i \in I$ and $\theta_i, \hat{\theta}_i \in \Theta_i$. It follows that $f(\theta)$ is TIBS. ■

Remarks

1. The revelation principle for Bayesian equilibrium is based on the same intuition as the revelation principle for dominant strategy equilibrium. We therefore refer to the remarks made subsequent to theorem 1.

2. Contrary to an assertion by Laffont and Maskin (1982), p.44, the set of Bayesian equilibria in any indirect mechanism Γ is *not* isomorphic to that in a corresponding direct mechanism Γ_d (cf. Repullo (1986), p.185 for a counterexample). Hence, even if Γ gives rise to a unique Bayesian equilibrium, truthtelling may not be the unique equilibrium in Γ_d and restricting attention to direct mechanisms leads to a loss of generality.

3. If it can be ensured that the truthtelling outcome is the sole equilibrium outcome in the direct mechanism Γ_d, then the reverse of theorem 14 is also true. Sufficient conditions for uniqueness are given by Repullo (1986), section 5, and Palfrey (1992), theorem 1.

By definition 18, TIBS imposes fewer incentive constraints than TIDS, which suggests that a wider range of SCRs is implementable in Bayesian strategies than in dominant strategies. We will now show for the quasilinear framework that this is indeed true.

AGV Mechanisms

Let us return to the public good problem analyzed in section 2.2. There, we concluded that ex-post efficient SCRs are not implementable in dominant strategies. For environments with complete information, we then showed that this problem can be resolved by employing the weaker notion of Nash equilibrium. As it turns out, a similar result also holds in environments where information is incomplete, i.e. there exists a mechanism that (truthfully) implements ex-post efficient SCRs. The mechanism in question in known

as *AGV mechanism* and was independently discovered by d'Aspremont and Gérard-Varet (1979) and Arrow (1979).

Definition 19 (AGV Mechanism) An AGV mechanism Γ_{AGV} is a direct mechanism with

$$x(\hat{\theta}) = \begin{cases} 1 \text{ if } \sum_i \hat{\theta}_i \geq 0 \\ 0 \text{ otherwise,} \end{cases} \tag{2.23}$$

$$t_i(\hat{\theta}) = \int_{\Theta_{-i}} \left(\sum\nolimits_{j \neq i} \theta_j x \left(\hat{\theta}_i, \theta_{-i} \right) \right) d\Pi_{-i} \left(\theta_{-i} \right) + h_i(\hat{\theta}_{-i}), \tag{2.24}$$

where

$$h_i(\hat{\theta}_{-i}) = -\frac{1}{n-1} \sum\nolimits_{j \neq i} \int_{\Theta_{-j}} \left(\sum\nolimits_{k \neq j} \theta_k x \left(\hat{\theta}_j, \theta_{-j} \right) \right) d\Pi_{-i} \left(\theta_{-j} \right) \tag{2.25}$$

for all $i \in I$, and where $\hat{\theta}$ is a profile of announcements.

The logic which underlies the AGV mechanism is very similar to that of the Groves mechanism: Agent i's transfer $t_i(\hat{\theta})$ depends on his announcement $\hat{\theta}_i$ only insofar as this announcement changes the decision $x(\hat{\theta})$, given that all other agents tell the truth. For any given profile of truthful announcements θ_{-i}, such a change in $x(\hat{\theta})$ reduces agent i's transfer by an amount equal to the other agents' valuations $\mid \sum_{j \neq i} \theta_j \mid$, which represents the negative externality that he is imposing on these agents. Thus, the integral in (2.24) constitutes the expected (negative) externality from agent i's announcement. Since all externalities are now fully internalized, no agent has an incentive to misreport his type.

Theorem 15 In the AGV mechanism Γ_{AGV}, truthtelling is a Bayesian equilibrium.

Proof (direct) Given that the remaining $j \neq i$ agents tell the truth, agent i solves

$$\max_{\hat{\theta}_i} \int_{\Theta_{-i}} (\theta_i x(\hat{\theta}_i, \theta_{-i}) + t_i(\hat{\theta}_i, \theta_{-i})) d\Pi_{-i} \left(\theta_{-i} \right). \tag{2.26}$$

Because $h_i(\hat{\theta}_{-i})$ is independent of $\hat{\theta}_i$, this is equivalent to

$$\max_{\hat{\theta}_i} \int_{\Theta_{-i}} \left(\theta_i + \sum\nolimits_{j \neq i} \theta_j \right) x(\hat{\theta}_i, \theta_{-i}) d\Pi_{-i} \left(\theta_{-i} \right). \tag{2.27}$$

By (2.23), $\hat{\theta}_i = \theta_i$ maximizes the integrand in (2.27) for any profile θ_{-i}, which implies that $\hat{\theta}_i = \theta_i$ also maximizes the integral. ■

From theorem 15, it follows that an SCR with decision rule and transfer functions given by (2.23)-(2.24) is truthfully implementable in Bayesian strategies. It is now a straightforward exercise to show that any such SCR is ex-post efficient.

Theorem 16 There exist SCRs that are both TIBS and ex-post efficient.

Proof (direct) Consider the social choice rule $f(\theta)$ with decision rule $x(\theta)$ and transfer functions $t_i(\theta)$ given by (2.23)-(2.24). Success is obvious. Furthermore, by theorem 15, $f(\theta)$ is TIBS. In order to prove that it is also budget-balanced, let us define

$$\tau_i(\hat{\theta}_i) \equiv \int_{\Theta_{-i}} \left(\sum\nolimits_{j \neq i} \theta_j x\left(\hat{\theta}_i, \theta_{-i}\right) \right) d\Pi_{-i}(\theta_{-i}). \tag{2.28}$$

Since $f(\theta)$ is TIBS, we can set $\hat{\theta}_i = \theta_i$. Summing over all i yields

$$\begin{aligned} \sum\nolimits_i t_i(\theta) &= \sum\nolimits_i \tau_i(\theta_i) - \frac{1}{n-1} \sum\nolimits_i \sum\nolimits_{j \neq i} \tau_j(\theta_j) \\ &= \sum\nolimits_i \tau_i(\theta_i) - \frac{n-1}{n-1} \sum\nolimits_i \tau_i(\theta_i) \\ &= 0 \end{aligned} \tag{2.29}$$

for all $\theta \in \Theta$. ■

Hitherto, we have assumed that participation in the AGV mechanism is voluntary. If this assumption is relaxed, it may no longer be true that Γ_{AGV} truthfully implements ex-post efficient SCRs in Bayesian strategies. In fact, as we will show later in the context of a bilateral trade problem, there is *no* direct mechanism which truthfully implements ex-post efficient SCRs in Bayesian strategies and satisfies individual rationality at the same time. Prior to that, however, we characterize for a rather general class of problems the set of SCRs that are both TIBS and individually rational.

Necessary and Sufficient Conditions for Truthful Implementation with Individual Rationality Constraints

For convenience, we maintain the assumption that preferences are quasilinear, albeit we consider now the more general form $\theta_i v_i(x) + t_i$, where $x \in X \subseteq \mathrm{R}^k$. In addition, let us replace assumption 3 of the basic model with

3a. Each agent has a characteristic or type $\theta_i \in \Theta_i = \left[\underline{\theta}_i, \overline{\theta}_i\right]$ with $\underline{\theta}_i \neq \overline{\theta}_i$ and strictly positive density $\pi_i(\theta_i) > 0$ for all $\theta_i \in \Theta_i$.

For ease of exposition, let $\overline{v}_i(\hat{\theta}_i) \equiv \int_{\Theta_{-i}} v_i(x(\hat{\theta}_i, \theta_{-i}))d\Pi_{-i}(\theta_{-i})$ and $\bar{t}_i(\hat{\theta}_i) \equiv \int_{\Theta_{-i}} t_i(\hat{\theta}_i, \theta_{-i})d\Pi_{-i}(\theta_{-i})$ denote agent i's expected net benefit and expected transfer, respectively, when he announces $\hat{\theta}_i$ and the remaining $j \neq i$ agents tell the truth. This allows us to write agent i's expected utility from the profile $(\hat{\theta}_i, \theta_{-i})$ as $\theta_i \bar{v}_i(\hat{\theta}_i) + \bar{t}_i(\hat{\theta}_i)$. Finally, let $U_i(\theta_i) \equiv \theta_i \bar{v}_i(\theta_i) + \bar{t}_i(\theta_i)$ denote agent i's expected utility if everyone (including him) reveals his type truthfully.

The ex-post version of individual rationality given in definition 14 appears overly strong for environments where information is incomplete. If we require that $\theta_i v_i(x(\theta)) + t(\theta)_i \geq 0$ for all $i \in I$ and $\theta \in \Theta$, then we essentially allow agents to withdraw from the game *after* everybody (truthfully) announced his type. It seems therefore more appropriate to require that an SCR be *interim individually rational* in the sense that agents can only withdraw at a stage where they do not yet know each others' types.

Definition 20 (Interim Individually Rational Social Choice Rule) An SCR is interim individually rational (IIR) if $U_i(\theta_i) \geq 0$ for all $i \in I$ and $\theta_i \in \Theta_i$.

As in definition 14, the agents' reservation utilities are normalized to zero. We can now characterize the set of SCRs that are both TIBS and IIR.

Theorem 17 An SCR is both TIBS and IIR if and only if for all $i \in I$,

1) $\bar{v}_i(\theta_i)$ is nondecreasing,

2) $U_i(\theta_i) = U_i(\underline{\theta}_i) + \int_{\underline{\theta}_i}^{\theta_i} \bar{v}_i(\eta)d\eta$ for all θ_i, and

3) $U_i(\underline{\theta}_i) \geq 0$.

Proof (direct/by contradiction) "if"-part: First, we show that 1) and 2) imply TIBS. Take any two values $\theta_i, \theta_i' \in \left[\underline{\theta}_i, \overline{\theta}_i\right]$ with $\theta_i > \theta_i' > \underline{\theta}_i$. From 2), we have

$$U_i(\theta_i) - \int_{\underline{\theta}_i}^{\theta_i} \bar{v}_i(\eta)d\eta = U_i(\theta_i') - \int_{\underline{\theta}_i}^{\theta_i'} \bar{v}_i(\eta)d\eta. \tag{2.30}$$

Rearranging terms and using 1) gives

$$U_i(\theta_i) - U_i(\theta_i') = \int_{\theta_i'}^{\theta_i} \bar{v}_i(\eta)d\eta \geq \int_{\theta_i'}^{\theta_i} \bar{v}_i(\theta_i')d\eta = (\theta_i - \theta_i')\bar{v}_i(\theta_i'). \tag{2.31}$$

Note that $\bar{v}_i(\theta_i')$ is a constant. Hence

$$U_i(\theta_i) \geq U_i(\theta_i') + (\theta_i - \theta_i')\bar{v}_i(\theta_i') \equiv \theta_i \bar{v}_i(\theta_i') + \bar{t}_i(\theta_i'). \tag{2.32}$$

Similarly, suppose that $\theta_i' > \theta_i > \underline{\theta}_i$. By the same reasoning, we have

$$U_i(\theta_i') \geq U_i(\theta_i) + (\theta_i' - \theta_i)\bar{v}_i(\theta_i) \equiv \theta_i'\bar{v}_i(\theta_i) + \bar{t}_i(\theta_i). \tag{2.33}$$

Together, (2.32) and (2.33) imply TIBS. Next, we prove that 1), 2), and 3) imply IIR. Suppose not. Then there exists some $\theta_i > \underline{\theta}_i$ with $U_i(\theta_i) < 0$. We just established that 1) and 2) imply TIBS. However, TIBS in conjunction with 3) implies

$$U_i(\theta_i) \geq \theta_i\bar{v}_i(\underline{\theta}_i) + \bar{t}_i(\underline{\theta}_i) > \underline{\theta}_i\bar{v}_i(\underline{\theta}_i) + \bar{t}_i(\underline{\theta}_i) \equiv U_i(\underline{\theta}_i) \geq 0, \tag{2.34}$$

a contradiction.

"only if"-part: We now show that TIBS implies 1) and 2). For any $i \in I$ and any two types $\theta_i, \theta_i' \in \left[\underline{\theta}_i, \overline{\theta}_i\right]$, TIBS requires that

$$U_i(\theta_i) \geq \theta_i\bar{v}_i(\theta_i') + \bar{t}_i(\theta_i') \equiv U_i(\theta_i') + (\theta_i - \theta_i')\bar{v}_i(\theta_i'), \tag{2.35}$$

and

$$U_i(\theta_i') \geq \theta_i'\bar{v}_i(\theta_i) + \bar{t}_i(\theta_i) \equiv U_i(\theta_i) - (\theta_i - \theta_i')\bar{v}_i(\theta_i). \tag{2.36}$$

Suppose without loss of generality that $\theta_i < \theta_i'$. From (2.35) and (2.36), it follows that

$$\bar{v}_i(\theta_i') \geq \frac{U_i(\theta_i) - U_i(\theta_i')}{\theta_i - \theta_i'} \geq \bar{v}_i(\theta_i), \tag{2.37}$$

which shows that $\bar{v}_i(\cdot)$ is nondecreasing. Next, letting $\theta_i' \to \theta_i$, we obtain $\frac{dU_i(\theta_i)}{d\theta_i} = \bar{v}_i(\theta_i)$ for all θ_i. Integrating both sides over $[\underline{\theta}_i, \theta_i]$ gives

$$U_i(\theta_i) = U_i(\underline{\theta}_i) + \int_{\underline{\theta}_i}^{\theta_i} \bar{v}_i(\eta)d\eta. \tag{2.38}$$

for all θ_i Finally, note that IIR obviously implies 3) by definition 20. ■

Remarks

1. Theorem 17 is an extremely powerful tool in Bayesian implementation theory and will be used repeatedly in the remainder of this section. An analogous version for the one-agent case was developed by Mirrlees (1971) and plays a central role in our analysis of adverse selection in chapter 3.

2. The great merit of theorem 17 is that it allows us to replace the original TIBS and IIR constraints with the mathematically more tractable constraints 1)-3). Furthermore, direct inspection of conditions 1)-3) already yields many important insights as is shown in the following subsections in the context of a bilateral trading problem and auction design.

The Myerson-Satterthwaite Theorem

Consider a bilateral trading problem where agent 1 is the seller of an indivisible object that agent 2 likes to buy. Each agent has quasilinear utility $\theta_i x_i + t_i$, where θ_i and t_i denote agent i's valuation and transfer, respectively, and where x_i is the probability that agent i receives the object. This setting corresponds in the framework of the previous subsection to the case where $v_i(x) = x_i$, and where $X = \{(x_1, x_2) \mid x_i \in [0,1] \text{ for } i = 1,2 \text{ and } x_1 + x_2 \leq 1\}$. Similar to the previous subsection, let us define $\bar{x}_i(\theta_i) \equiv \int_{\Theta_{-i}} x_i(\theta_i, \theta_{-i}) d\Pi_{-i}(\theta_{-i})$, $\bar{t}_i(\theta_i) \equiv \int_{\Theta_{-i}} t_i(\theta_i, \theta_{-i}) d\Pi_{-i}(\theta_{-i})$, and $U_i(\theta_i) \equiv \theta_i \bar{x}_i(\theta_i) + \bar{t}_i(\theta_i)$.

Note that unlike in the public good problem where $x \in \{0,1\}$, we now also consider random decisions. Randomization convexifies the decision space X and thus allows us to prove our results (here: the Myerson-Satterthwaite theorem) for a wider class of SCRs. Besides, randomization also turns out to be convenient for technical reasons (cf. Laffont and Maskin (1982), p.44).

In the bilateral trading problem, an SCR is ex-post efficient if and only if i) there is no waste of either the object or money, and ii) whoever has a higher valuation receives the object with probability one.

Definition 21 (Ex-Post Efficient Social Choice Rule) An SCR is ex-post efficient if

1) $t_1(\theta) + t_2(\theta) = 0$ for all $\theta \in \Theta$,

2) $x_1(\theta) + x_2(\theta) = 1$ for all $\theta \in \Theta$, and

3) $x_1(\theta) = 1$ if $\theta_1 > \theta_2$ and $x_2(\theta) = 1$ if $\theta_1 < \theta_2$.

The Coase theorem predicts that in the presence of complete information, bargaining over the object in question leads to an ex-post efficient allocation. When information is incomplete, sellers typically overstate and buyers understate their valuations in order to maximize profits. The question is then whether there exists a trading mechanism (i.e. a bargaining or bidding procedure) that nonetheless attains an ex-post efficient outcome. By the revelation principle, it is not necessary to examine all possible bargaining games. Rather, we can restrict attention to direct mechanisms in which each agent simultaneously reports his valuation to a ficticious third party who then implements an outcome $(x(\theta), t(\theta)) \in f(\theta)$. Implicitly, this assumes that prior to announcing their valuations, both parties have signed an enforceable contract that specifies a social choice rule $f(\theta)$.

From our analysis of the public good problem, we already know that in the absence of individual rationality constraints, the AGV mechanism (truthfully) implements ex-post efficient SCRs in Bayesian strategies. However, in

the bilateral trading problem it is appropriate to require that $f(\theta)$ be (interim) individually rational, i.e. that both buyer and seller have nonnegative expected gains from trade if they are to participate. Since the seller can always consume the object, this implies that $U_1(\theta_1) \geq \theta_1$ for all $\theta_1 \in [\underline{\theta}_1, \overline{\theta}_1]$. By the same reasoning, the buyer's expected utility ought to be nonnegative, i.e. $U_2(\theta_2) \geq 0$ for all $\theta_2 \in [\underline{\theta}_2, \overline{\theta}_2]$. Unfortunately, the following result due to Myerson and Satterthwaite (1983) tells us that if gains from trade are possible but not certain, there is no SCR that is TIBS, IIR, and ex-post efficient at the same time.

Theorem 18 (Myerson-Satterthwaite Theorem) Suppose that $\underline{\theta}_1 < \overline{\theta}_2$ and $\overline{\theta}_1 > \underline{\theta}_2$. Then there exists no SCR that is TIBS, IIR and ex-post efficient.

Proof (by contradiction) Assume that there exists a social choice rule $f(\theta)$ that is TIBS, IIR and ex-post efficient. By theorem 17, we have

$$U_i(\theta_i) = U_i(\underline{\theta}_i) + \int_{\underline{\theta}_i}^{\theta_i} \bar{x}_i(\eta) d\eta \tag{2.39}$$

for all θ_i and $i = 1, 2$. Substituting $U_i(\theta_i) \equiv \theta_i \bar{x}_i(\theta_i) + \bar{t}_i(\theta_i)$ in (2.39) and solving for $\bar{t}_i(\theta_i)$ gives

$$\bar{t}_i(\theta_i) = U_i(\underline{\theta}_i) + \int_{\underline{\theta}_i}^{\theta_i} \bar{x}_i(\eta) d\eta - \theta_i \bar{x}_i(\theta_i) \tag{2.40}$$

for all θ_i and $i = 1, 2$. Taking expectations with respect to θ_i, we obtain

$$\begin{aligned} \int_{\underline{\theta}_i}^{\overline{\theta}_i} \bar{t}_i(\theta_i) d\Pi_i(\theta_i) &= U_i(\underline{\theta}_i) + \int_{\underline{\theta}_i}^{\overline{\theta}_i} \int_{\underline{\theta}_i}^{\theta_i} \bar{x}_i(\eta) d\eta d\Pi_i(\theta_i) \\ &\quad - \int_{\underline{\theta}_i}^{\overline{\theta}_i} \theta_i \bar{x}_i(\theta_i) d\Pi_i(\theta_i) \end{aligned} \tag{2.41}$$

for $i = 1, 2$. Consider the second term on the right-hand side of (2.41). Integration by parts yields

$$\begin{aligned} \int_{\underline{\theta}_i}^{\overline{\theta}_i} \int_{\underline{\theta}_i}^{\theta_i} \bar{x}_i(\eta) d\eta d\Pi_i(\theta_i) &= \int_{\underline{\theta}_i}^{\theta_i} \bar{x}_i(\eta) d\eta \Pi_i(\theta_i) \Big|_{\underline{\theta}_i}^{\overline{\theta}_i} \\ &\quad - \int_{\underline{\theta}_i}^{\overline{\theta}_i} \bar{x}_i(\theta_i) \Pi_i(\theta_i) d\theta_i \\ &= \int_{\underline{\theta}_i}^{\overline{\theta}_i} \bar{x}_i(\theta_i) (1 - \Pi_i(\theta_i)) d\theta_i. \end{aligned} \tag{2.42}$$

Inserting (2.42) back in (2.41), we have

$$\begin{aligned}\int_{\underline{\theta}_1}^{\overline{\theta}_1} \bar{t}_1(\theta_1) d\Pi_1(\theta_1) &= U_1(\underline{\theta}_1) - \int_{\underline{\theta}_1}^{\overline{\theta}_1} \bar{x}_1(\theta_1) \left(\theta_1 + \frac{\Pi_1(\theta_1)}{\pi_1(\theta_1)}\right) d\Pi_1(\theta_1) \\ &\quad + \int_{\underline{\theta}_1}^{\overline{\theta}_1} \bar{x}_1(\theta_1) d\theta_1, \end{aligned} \tag{2.43}$$

and

$$\begin{aligned}\int_{\underline{\theta}_2}^{\overline{\theta}_2} \bar{t}_2(\theta_2) d\Pi_2(\theta_2) &= U_2(\underline{\theta}_2) \\ &\quad - \int_{\underline{\theta}_2}^{\overline{\theta}_2} \bar{x}_2(\theta_2) \left(\theta_2 - \frac{1 - \Pi_2(\theta_2)}{\pi_2(\theta_2)}\right) d\Pi_2(\theta_2). \end{aligned} \tag{2.44}$$

By theorem 17, the third term on the right-hand side of (2.43) can be written as

$$\int_{\underline{\theta}_1}^{\overline{\theta}_1} \bar{x}_1(\theta_1) d\theta_1 = U_1\left(\overline{\theta}_1\right) - U_1(\underline{\theta}_1), \tag{2.45}$$

so that (2.43) is equal to

$$\int_{\underline{\theta}_1}^{\overline{\theta}_1} \bar{t}_1(\theta_1) d\Pi_1(\theta_1) = U_1\left(\overline{\theta}_1\right) - \int_{\underline{\theta}_1}^{\overline{\theta}_1} \bar{x}_1(\theta_1) \left(\theta_1 + \frac{\Pi_1(\theta_1)}{\pi_1(\theta_1)}\right) d\Pi_1(\theta_1). \tag{2.46}$$

For convenience, set $x(\theta) \equiv x_2(\theta)$ and $t(\theta) \equiv t_1(\theta)$. By definition 21, ex-post efficiency implies that $t_2(\theta) = -t(\theta)$ and $x_1(\theta) = 1 - x(\theta)$, where $x(\theta)$ is now simply the probability of trade. Next, define

$$\bar{x}_1(\theta_1) \equiv 1 - \int_{\underline{\theta}_2}^{\overline{\theta}_2} x(\theta_1, \theta_2) d\Pi_2(\theta_2), \tag{2.47}$$

$$\bar{x}_2(\theta_2) \equiv \int_{\underline{\theta}_1}^{\overline{\theta}_1} x(\theta_1, \theta_2) d\Pi_1(\theta_1), \tag{2.48}$$

$$\bar{t}_1(\theta_1) \equiv \int_{\underline{\theta}_2}^{\overline{\theta}_2} t(\theta_1, \theta_2) d\Pi_2(\theta_2), \tag{2.49}$$

and

$$\bar{t}_2(\theta_2) \equiv - \int_{\underline{\theta}_1}^{\overline{\theta}_1} t(\theta_1, \theta_2) d\Pi_1(\theta_1). \tag{2.50}$$

Inserting (2.47) and (2.49) in (2.46) and rearranging yields

$$\begin{aligned} U_1\left(\overline{\theta}_1\right) &= \int_{\underline{\theta}_1}^{\overline{\theta}_1}\int_{\underline{\theta}_2}^{\overline{\theta}_2} t(\theta_1,\theta_2) d\Pi_2\left(\theta_2\right) d\Pi_1\left(\theta_1\right) \\ &+ \int_{\underline{\theta}_1}^{\overline{\theta}_1}\int_{\underline{\theta}_2}^{\overline{\theta}_2} \left(\theta_1 + \frac{\Pi_1\left(\theta_1\right)}{\pi_1\left(\theta_1\right)}\right) d\Pi_2\left(\theta_2\right) d\Pi_1\left(\theta_1\right) \\ &- \int_{\underline{\theta}_1}^{\overline{\theta}_1}\int_{\underline{\theta}_2}^{\overline{\theta}_2} x(\theta_1,\theta_2)\left(\theta_1 + \frac{\Pi_1\left(\theta_1\right)}{\pi_1\left(\theta_1\right)}\right) d\Pi_2\left(\theta_2\right) d\Pi_1\left(\theta_1\right). \end{aligned} \tag{2.51}$$

Using integration by parts, the second term on the right-hand side of (2.51) can be written as (note that $\theta_1 + \frac{\Pi_1(\theta_1)}{\pi_1(\theta_1)}$ is a constant with respect to θ_2)

$$\begin{aligned} \int_{\underline{\theta}_1}^{\overline{\theta}_1}\int_{\underline{\theta}_2}^{\overline{\theta}_2} \left(\theta_1 + \frac{\Pi_1\left(\theta_1\right)}{\pi_1\left(\theta_1\right)}\right) d\Pi_2\left(\theta_2\right) d\Pi_1\left(\theta_1\right) &= \theta_1\Pi_1\left(\theta_1\right)\Big|_{\underline{\theta}_1}^{\overline{\theta}_1} \\ &= \overline{\theta}_1. \end{aligned} \tag{2.52}$$

Inserting (2.52) back in (2.51), we obtain

$$\begin{aligned} U_1\left(\overline{\theta}_1\right) - \overline{\theta}_1 &= \int_{\underline{\theta}_1}^{\overline{\theta}_1}\int_{\underline{\theta}_2}^{\overline{\theta}_2} t(\theta_1,\theta_2) d\Pi_2\left(\theta_2\right) d\Pi_1\left(\theta_1\right) \\ &- \int_{\underline{\theta}_1}^{\overline{\theta}_1}\int_{\underline{\theta}_2}^{\overline{\theta}_2} x(\theta_1,\theta_2)\left(\theta_1 + \frac{\Pi_1\left(\theta_1\right)}{\pi_1\left(\theta_1\right)}\right) d\Pi_2\left(\theta_2\right) d\Pi_1\left(\theta_1\right). \end{aligned} \tag{2.53}$$

Similarly, inserting (2.48) and (2.50) in (2.44) and rearranging, we have

$$\begin{aligned} U_2\left(\underline{\theta}_2\right) &= -\int_{\underline{\theta}_1}^{\overline{\theta}_1}\int_{\underline{\theta}_2}^{\overline{\theta}_2} t(\theta_1,\theta_2) d\Pi_2\left(\theta_2\right) d\Pi_1\left(\theta_1\right) \\ &+ \int_{\underline{\theta}_2}^{\overline{\theta}_2}\int_{\underline{\theta}_1}^{\overline{\theta}_1} x(\theta_1,\theta_2)\left(\theta_2 - \frac{1-\Pi_2\left(\theta_2\right)}{\pi_2\left(\theta_2\right)}\right) d\Pi_1\left(\theta_1\right) d\Pi_2\left(\theta_2\right). \end{aligned} \tag{2.54}$$

Adding up (2.53) and (2.54) gives

$$\begin{aligned} & U_1\left(\overline{\theta}_1\right) - \overline{\theta}_1 + U_2\left(\underline{\theta}_2\right) \\ = & \int_{\underline{\theta}_2}^{\overline{\theta}_2}\int_{\underline{\theta}_1}^{\overline{\theta}_1} x(\theta_1,\theta_2)\left(\theta_2 - \frac{1-\Pi_2\left(\theta_2\right)}{\pi_2\left(\theta_2\right)}\right) d\Pi_1\left(\theta_1\right) d\Pi_2\left(\theta_2\right) \\ & - \int_{\underline{\theta}_2}^{\overline{\theta}_2}\int_{\underline{\theta}_1}^{\overline{\theta}_1} x(\theta_1,\theta_2)\left(\theta_1 + \frac{\Pi_1\left(\theta_1\right)}{\pi_1\left(\theta_1\right)}\right) d\Pi_1\left(\theta_1\right) d\Pi_2\left(\theta_2\right). \end{aligned} \tag{2.55}$$

But IIR implies that $U_1\left(\overline{\theta}_1\right) \geq \overline{\theta}_1$ and $U_2\left(\underline{\theta}_2\right) \geq 0$, from which it follows that the right-hand side of (2.55) must be nonnegative.

The remainder of the proof establishes that if $f(\theta)$ is to be ex-post efficient, the right-hand side of (2.55) cannot be nonnegative, thus leading to a contradiction. Suppose that $f(\theta)$ is ex-post efficient. By definition 21, we then have $x(\theta) = 0$ if $\theta_1 > \theta_2$ and $x(\theta) = 1$ if $\theta_1 < \theta_2$, which implies that both integrands in (2.55) become zero whenever $\theta_1 > \theta_2$. Consequently, the right-hand side of (2.55) can be written as

$$\begin{aligned} &\int_{\underline{\theta}_2}^{\overline{\theta}_2} \int_{\underline{\theta}_1}^{\min[\theta_2,\overline{\theta}_1]} \left(\theta_2 - \frac{1-\Pi_2(\theta_2)}{\pi_2(\theta_2)}\right) \pi_1(\theta_1)\,\pi_2(\theta_2)\,d\theta_1 d\theta_2 \qquad (2.56)\\ &- \int_{\underline{\theta}_2}^{\overline{\theta}_2} \int_{\underline{\theta}_1}^{\min[\theta_2,\overline{\theta}_1]} \left(\theta_1 + \frac{\Pi_1(\theta_1)}{\pi_1(\theta_1)}\right) \pi_1(\theta_1)\,\pi_2(\theta_2)\,d\theta_1 d\theta_2. \end{aligned}$$

Integrating with respect to θ_1, the first term in (2.56) is equal to

$$\begin{aligned} &\int_{\underline{\theta}_2}^{\overline{\theta}_2} \Pi_1(\theta_1)|_{\underline{\theta}_1}^{\min[\theta_2,\overline{\theta}_1]} \left(\theta_2\pi_2(\theta_2) - 1 + \Pi_2(\theta_2)\right) d\theta_2 \qquad (2.57)\\ = &\int_{\underline{\theta}_2}^{\overline{\theta}_1} \Pi_1(\theta_2)\left(\theta_2\pi_2(\theta_2) - 1 + \Pi_2(\theta_2)\right) d\theta_2\\ &+ \int_{\overline{\theta}_1}^{\overline{\theta}_2} \left(\theta_2\pi_2(\theta_2) - 1 + \Pi_2(\theta_2)\right) d\theta_2. \end{aligned}$$

Using integration by parts, the second term in (2.56) can be written as

$$\begin{aligned} &-\int_{\underline{\theta}_2}^{\overline{\theta}_2} \theta_1\Pi_1(\theta_1)|_{\underline{\theta}_1}^{\min[\theta_2,\overline{\theta}_1]}\,\pi_2(\theta_2)\,d\theta_2 \qquad (2.58)\\ = &-\int_{\underline{\theta}_2}^{\overline{\theta}_1} \theta_2\Pi_1(\theta_2)\,\pi_2(\theta_2)\,d\theta_2 - \int_{\overline{\theta}_1}^{\overline{\theta}_2} \overline{\theta}_1\pi_2(\theta_2)\,d\theta_2. \end{aligned}$$

Adding up (2.57) and (2.58) yields

$$\begin{aligned} &-\int_{\underline{\theta}_2}^{\overline{\theta}_1} \Pi_1(\theta_2)\left(1-\Pi_2(\theta_2)\right) d\theta_2 \qquad (2.59)\\ &+\int_{\overline{\theta}_1}^{\overline{\theta}_2} \left(\left(\theta_2 - \overline{\theta}_1\right)\pi_2(\theta_2) - 1 + \Pi_2(\theta_2)\right) d\theta_2. \end{aligned}$$

Integrating by parts shows that the second integral in (2.59) is equal to zero:

$$\int_{\overline{\theta}_1}^{\overline{\theta}_2} \left(\left(\theta_2 - \overline{\theta}_1\right)\pi_2(\theta_2) - 1 + \Pi_2(\theta_2)\right) d\theta_2 \qquad (2.60)$$

$$\begin{aligned} &= \left(\left(\theta_2 - \overline{\theta}_1\right)\Pi_2\left(\theta_2\right) - \theta_2\right)\Big|_{\overline{\theta}_1}^{\overline{\theta}_2} \\ &= 0. \end{aligned}$$

Thus, if $f(\theta)$ is to be ex-post efficient, the right-hand side of (2.55) can be written as

$$-\int_{\underline{\theta}_2}^{\overline{\theta}_1} \Pi_1\left(\theta_2\right)\left(1 - \Pi_2\left(\theta_2\right)\right) d\theta_2, \tag{2.61}$$

which is strictly negative since $\overline{\theta}_1 > \underline{\theta}_2$ and $\underline{\theta}_1 < \overline{\theta}_2$, a contradiction. ■

Remarks

1. If $\overline{\theta}_1 > \underline{\theta}_2$ but $\underline{\theta}_1 > \overline{\theta}_2$ so that there are no gains from trade, (2.61) is equal to zero and IIR holds. Likewise, if $\underline{\theta}_1 < \overline{\theta}_2$ but $\overline{\theta}_1 < \underline{\theta}_2$ so that the gains from trade are certain, (2.61) is equal to $\underline{\theta}_2 - \overline{\theta}_1 > 0$ and IIR is satisfied. In addition, if the density $\pi_i(\theta_i)$ is not strictly positive everywhere on $\left[\underline{\theta}_i, \overline{\theta}_i\right]$, then SCRs that are TIBS, IIR, and ex-post efficient exist. For an example where $\pi_i(\theta_i)$ does not have strictly positive mass everywhere, see Myerson and Satterthwaite, p.273.

2. A weaker requirement than IIR is *ex ante individual rationality*, which ensures that the agents do not opt out of the mechanism at the ex ante stage prior to learning their types. This implicitly assumes that agents can commit to be bound by the mechanism at both the interim stage (when they get to know their own types) and the ex-post stage (when the state θ is truthfully revealed in equilibrium). With ex-ante individual rationality, the Myerson-Satterthwaite theorem no longer holds and it can be shown that the AGV mechanism achieves ex-post efficiency.

3. It is easy to check that the Myerson-Satterthwaite theorem continues to hold if we replace the ex-post constraint "$t_1(\theta) + t_2(\theta) = 0$ for all $\theta \in \Theta$" in definition 21 with the *ex-ante budget-balancing condition* $\int_{\underline{\theta}_1}^{\overline{\theta}_1}\int_{\underline{\theta}_2}^{\overline{\theta}_2} t_1(\theta_1,\theta_2) d\Pi_2\left(\theta_2\right) d\Pi_1\left(\theta_1\right) + \int_{\underline{\theta}_1}^{\overline{\theta}_1}\int_{\underline{\theta}_2}^{\overline{\theta}_2} t_2(\theta_1,\theta_2) d\Pi_2\left(\theta_2\right) d\Pi_1\left(\theta_1\right) = 0$. After all, in the first part of the proof we essentially add up (2.44) and (2.46) and set the left-hand side equal to zero, which is equivalent to the requirement that $f(\theta)$ be budget-balancing ex ante.

4. Given that TIBS, IIR and ex-post efficiency cannot be satisfied simultaneously, Myerson and Satterthwaite go on to solve for the *optimal trading mechanism*, i.e. the mechanism that maximizes expected gains from trade subject to TIBS and IIR. We will derive such an optimal mechanism in the following subsection in the context of auction design.

Optimal Auctions

We have already encountered auction mechanisms in section 2.2 in connection with the Clarke mechanism. There, we pointed out that the Clarke mechanism is the public good analogue of what is known in a private good setting as *second-price sealed-bid* or *Vickrey auction*. In addition, we remarked that the Vickrey auction is ex-post efficient since the seller receives all transfers and budget-balancing is consequently not a problem.

In this subsection, we will discuss auctions in the context of Bayesian implementation. Consider the problem of a seller (agent 0) who wants to auction off an indivisible object to n bidders. We adopt the notation and assumptions from the previous subsection, except for the fact that the seller's valuation θ_0 is common knowledge. In particular, we require that $U_i(\theta_i) \geq 0$ for all $i \neq 0$ (IIR) and allow for random assignments of the object $x_i \in X = \{x_i \mid x_i \in [0,1] \text{ and } \sum_i x_i \leq 1 \text{ for } i \in \{0,...,n\}\}$. Finally, let us assume that for all $i \neq 0$, the distribution functions $\Pi_i(\theta_i)$ satisfy the *monotone hazard rate property*.

Definition 22 (Monotone Hazard Rate Property) The distribution function $\Pi(\theta)$ satisfies the monotone hazard rate property (MHRP) if $\frac{\pi(\theta)}{1-\Pi(\theta)}$ is nondecreasing.

MHRP holds for many distributions such as the normal, uniform, logistic, and exponential distribution. It has a very natural interpretation if θ is interpreted as, say, the lifetime of a machine. Then, MHRP states that the conditional probability that the machine fails in the interval $[\theta, \theta + d\theta]$ given that it lasts until time θ is nondecreasing.

Next, we pursue the question of optimal auction design from the seller's point of view. Without loss of generality, we can assume that $x_0(\theta) \equiv 1 - \sum_{i\neq 0} x_i(\theta)$ and $t_0(\theta) \equiv -\sum_{i\neq 0} t_i(\theta)$, since the seller is always better off by choosing an auction that entails no waste of either the object or money. At first glance, the task of finding an optimal auction mechanism appears quite formidable in view of the unlimited possibilities that are available. Fortunately, the revelation principle states that we can restrict attention to direct mechanisms in which each bidder announces only his valuation. The problem is then to find a social choice rule that maximizes the seller's expected utility subject to the constraint that it satisfies both TIBS and IIR. A remarkable result discovered by Vickrey (1961) and extended by Myerson (1981) and Riley and Samuelson (1981) known as the *revenue equivalence theorem* states that the sellers' expected revenue is completely determined by the decision $x(\theta)$ and the profile of rents earned by the lowest possible type of each bidder, and therefore independent of the transfer function $t(\theta)$.

Theorem 19 (Revenue Equivalence Theorem) Any two SCRs that are TIBS and have the same decision $x(\theta) = (x_1(\theta), ..., x_n(\theta))$ and profile of rents $U(\underline{\theta}) = (U_1(\underline{\theta}_1), ..., U_n(\underline{\theta}_n))$ generate the same expected revenue for the seller.

Proof (direct) By equation (2.44) (setting $2 = i$), TIBS implies that

$$\int_{\Theta_i} \bar{t}_i(\theta_i) d\Pi_i(\theta_i) = U_i(\underline{\theta}_i) - \int_{\Theta_i} \bar{x}_i(\theta_i) \left(\theta_i - \frac{1 - \Pi_i(\theta_i)}{\pi_i(\theta_i)} \right) d\Pi_i(\theta_i). \quad (2.62)$$

Substituting in (2.62) for $\bar{t}_i(\theta_i) \equiv \int_{\Theta_{-i}} t_i(\theta_i, \theta_{-i}) d\Pi_{-i}(\theta_{-i})$ and $\bar{x}_i(\theta_i) \equiv \int_{\Theta_{-i}} x_i(\theta_i, \theta_{-i}) d\Pi_{-i}(\theta_{-i})$, and using the fact that the θ_i's are statistically independent, we have

$$\int_{\Theta} t_i(\theta) d\Pi(\theta) = U_i(\underline{\theta}_i) - \int_{\Theta} x_i(\theta) \left(\theta_i - \frac{1 - \Pi_i(\theta_i)}{\pi_i(\theta_i)} \right) d\Pi(\theta). \quad (2.63)$$

The seller's expected revenue if all bidders tell the truth is

$$\int_{\Theta} \left(-\sum\nolimits_{i \neq 0} t_i(\theta) \right) d\Pi(\theta) = -\sum\nolimits_{i \neq 0} \int_{\Theta} t_i(\theta) d\Pi(\theta). \quad (2.64)$$

Inserting (2.63) in (2.64), the seller's expected revenue can be written as

$$\int_{\Theta} \sum\nolimits_{i \neq 0} x_i(\theta) \left(\theta_i - \frac{1 - \Pi_i(\theta_i)}{\pi_i(\theta_i)} \right) d\Pi(\theta) - \sum\nolimits_{i \neq 0} U_i(\underline{\theta}_i), \quad (2.65)$$

which is completely determined by $x(\theta) = (x_1(\theta), ..., x_n(\theta))$ and $U(\underline{\theta}) = (U_1(\underline{\theta}_1), ..., U_n(\underline{\theta}_n))$. ■

Remarks

1. One implication of theorem 19 is that any two auctions which award the object to the bidder with the highest valuation and leave zero rents to bidders of type $\underline{\theta}_i$ must yield the same revenue to the seller. In the symmetric case where the types of all bidders are drawn from the same probability distribution, the following standard auctions all have these properties and therefore generate the same revenue: i) the *Dutch auction*, where prices are called in descending order, ii) the *English auction*, where prices are announced in ascending order, iii) the *first-price sealed bid auction*, where the bidder with the highest bid acquires the object at that price, and iv) the *second-price sealed-bid* or *Vickrey auction*, where the bidder with the highest bid pays a price equal to the second highest bid.

2. The seller's expected utility is $\int_\Theta (\theta_0 x_0(\theta) + t_0(\theta))\, d\Pi(\theta)$, where the probability of keeping the object $x_0(\theta) = 1 - \sum_{i\neq 0} x_i(\theta)$ is completely determined by the decision $x(\theta)$. Thus, theorem 19 continues to hold if we replace "expected revenue" by "expected utility".

3. Assumptions 5c (independence) and 10 (risk-neutrality) are crucial for the revenue equivalence theorem to hold. If one of these assumptions fails, the seller is typically no longer indifferent between the four standard auctions listed in 1.

While the revenue equivalence theorem permits a comparison among different auction mechanisms, it says nothing about the maximum revenue that can be attained. Let us now finally derive the social choice rule that maximizes the seller's expected utility subject to TIBS and IIR.

Theorem 20 The social choice rule $f^*(\theta) = (x^*(\theta), t^*(\theta))$ maximizes the seller's expected utility subject to TIBS and IIR if and only if

$$x_i^*(\theta) = \begin{cases} 1 \text{ if } J_i(\theta_i) = \max\limits_{j\in\{0,\ldots,n\}} J_j(\theta_j) \\ 0 \text{ otherwise,} \end{cases} \tag{2.66}$$

and

$$\bar{t}_i^*(\theta_i) = -\theta_i \bar{x}_i^*(\theta_i) + \int_{\underline{\theta}_i}^{\theta_i} \bar{x}_i^*(\eta) d\eta, \tag{2.67}$$

for all $i \neq 0$ and $\theta \in \Theta$, where $J_i(\theta_i) \equiv \theta_i - \frac{1-\Pi_i(\theta_i)}{\pi_i(\theta_i)}$ and $J_0(\theta_0) \equiv \theta_0$.

Proof (direct) Replacing TIBS and IIR with conditions 1)-3) in theorem 17, the seller's relaxed problem can be written as

$$\max_{\{x_i(\cdot),t_i(\cdot)\}_{i\neq 0}} \int_\Theta \left(\theta_0 \left(1 - \sum\nolimits_{i\neq 0} x_i(\theta)\right) - \sum\nolimits_{i\neq 0} t_i(\theta)\right) d\Pi(\theta) \tag{2.68}$$

$$\begin{aligned} \text{s.t.} \quad & 1)\ \bar{x}_i(\theta_i) \text{ is nondecreasing for all } i \neq 0, \\ & 2)\ U_i(\theta_i) = U_i(\underline{\theta}_i) + \textstyle\int_{\underline{\theta}_i}^{\theta_i} \bar{x}_i(\eta) d\eta \text{ for all } \theta_i \text{ and } i \neq 0, \\ & 3)\ U_i(\underline{\theta}_i) \geq 0 \text{ for all } i \neq 0, \\ & 4) \text{ for all } \theta\text{: } x_i(\theta) \in [0,1] \text{ for all } i \neq 0 \text{ and } \textstyle\sum_i x_i(\theta) \leq 1. \end{aligned}$$

By (2.63), constraint 2) is equivalent to

$$\int_\Theta t_i(\theta) d\Pi(\theta) = U_i(\underline{\theta}_i) - \int_\Theta x_i(\theta) \left(\theta_i - \frac{1-\Pi_i(\theta_i)}{\pi_i(\theta_i)}\right) d\Pi(\theta). \tag{2.69}$$

Inserting (2.69) in the objective function (2.68), the seller's relaxed problem can be written as

$$\max_{\{x_i(\cdot),U_i(\underline{\theta}_i)\}_{i\neq 0}} \int_{\Theta} \sum\nolimits_{i\neq 0} x_i(\theta) \left(\theta_i - \frac{1-\Pi_i(\theta_i)}{\pi_i(\theta_i)} - \theta_0\right) d\Pi(\theta) \quad (2.70)$$
$$+ \theta_0 - \sum\nolimits_{i\neq 0} U_i(\underline{\theta}_i)$$
$$\text{s.t.} \quad 1), 3), \text{ and } 4).$$

Let us ignore constraint 1) for the moment. Inspection of (2.70) immediately reveals that the solution to the relaxed problem is

$$x_i^*(\theta) = \begin{cases} 1 \text{ if } J_i(\theta_i) = \max\limits_{j\in\{0,\ldots,n\}} J_j(\theta_j) \\ 0 \text{ otherwise,} \end{cases} \quad (2.71)$$

and

$$U_i^*(\underline{\theta}_i) = 0 \quad (2.72)$$

for all $i \neq 0$ and $\theta \in \Theta$ (we disregard ties as they occur with probability zero). By theorem 17, (2.72) can be rewritten as

$$\bar{t}_i^*(\theta_i) = -\theta_i \bar{x}_i^*(\theta_i) + \int_{\underline{\theta}_i}^{\theta_i} \bar{x}_i^*(\eta) d\eta, \quad (2.73)$$

where we used the fact that $U_i(\theta_i) \equiv \theta_i \bar{x}_i(\theta_i) + \bar{t}_i(\theta_i)$.

Finally, let us verify that 1) holds. MHRP implies that $J_i(\theta_i)$ is nondecreasing, which in turn implies that $x_i^*(\theta) \equiv x_i^*(\theta_i, \theta_{-i})$ is nondecreasing in θ_i. From $\bar{x}_i(\theta_i) \equiv \int_{\Theta_{-i}} x_i(\theta_i, \theta_{-i}) d\Pi_{-i}(\theta_{-i})$, it then follows that $\bar{x}_i^*(\theta_i)$ is nondecreasing and that constraint 1) is satisfied. ■

Remarks

1. The expression $J_i(\theta_i) \equiv \theta_i - \frac{1-\Pi_i(\theta_i)}{\pi_i(\theta_i)}$ is known as a bidder's *virtual valuation*. Hence, the optimal solution (2.66) states that the object should be awarded to the agent (bidder or seller) with the highest virtual valuation. Since this may not always be the agent with the highest actual valuation, the optimal solution is not necessarily ex-post efficient. One particular case where the optimal solution is ex-post efficient is when i) the types of all bidders are drawn from the same probability distribution (i.e. $J_i(\theta_i) = J(\theta_i)$ for all $i \neq 0$ so that the bidder with the highest actual valuation coincides with the bidder with the highest virtual valuation), and ii) $J(\underline{\theta}) > \theta_0$ (i.e. the object is always transferred to one of the bidders).

2. In the symmetric case, the optimal solution (2.66) translates into

$$x_i^*(\theta) = \begin{cases} 1 \text{ if } \theta_i = \max\limits_{j \in \{1,\ldots,n\}} \left[\theta_j, \tilde{\theta}\right] \\ 0 \text{ otherwise,} \end{cases} \tag{2.74}$$

where $\tilde{\theta}$ is defined by $\tilde{\theta} - \frac{1-\Pi(\tilde{\theta})}{\pi(\tilde{\theta})} \equiv \theta_0$. It follows that with a reservation or minimum price of $\tilde{\theta}$, the four standard auctions introduced earlier are all optimal.

3. Notice that the optimal solution (2.67) is only defined in terms of expected transfers $\bar{t}_i^*(\theta_i)$. This suggests that there is a great deal of flexibility in choosing the ex-post transfers $t_i^*(\theta_i)$, which results in a multiplicity of optimal SCRs. For instance, the SCRs implemented by the first- and second-price auctions are under certain conditions both optimal (cf. point 3), even though they differ in their transfer functions $t_i(\theta_i)$.

2.5 Bibliographic Notes

Our discussion of the revelation principle and the necessary and sufficient conditions for dominant strategy implementation in section 2.2 borrows from Dasgupta, Hammond, and Maskin (1979). The section on Groves mechanisms is based on Green and Laffont (1977, 1979) and Laffont and Maskin (1982). Laffont and Maskin (1980) obtain similar results by using an alternative approach which rests on the differentiability of the agent's valuation function.

The section on Nash implementation draws on the works of Dasgupta, Hammond, and Maskin (1979) and Maskin (1985). The proof of the sufficiency part of Maskin's theorem is adopted from Repullo (1987), and the application of Maskin's theorem to the public good problem is due to Laffont and Maskin (1982). Let us also mention the comprehensive survey by Moore (1992), who deals with many issues related to implementation in complete information environments that have been neglected here.

The treatment of the revelation principle for Bayesian implementation is, once again, based on Dasgupta, Hammond, and Maskin (1979). Much of the remaining discussion of Bayesian implementation follows Mas-Colell, Whinston, and Green (1995), chapter 23, and Fudenberg and Tirole (1991), chapter 7. In addition, our analysis of the Myerson-Satterthwaite theorem borrows from Myerson and Satterthwaite (1983), and that of optimal auctions draws on Myerson (1981).

2.6 References

Arrow, K. (1979), "The Property Rights Doctrine and Demand Revelation under Incomplete Information," in M. Boskin (ed.), *Economics and Human Welfare*. New York: Academic Press

Clarke, E. (1971), "Multipart Pricing of Public Goods," *Public Choice* **8**, 19-33.

Dasgupta, P., Hammond, P., and Maskin, E. (1979), "The Implementation of Social Choice Rules: Some General Results on Incentive Compatibility," *Review of Economic Studies* **46**, 185-216.

D'Aspremont, C., and Gérard-Varet, L. (1979), "Incentives and Incomplete Information," *Journal of Public Economics* **11**, 25-45.

Fudenberg, D., and Tirole, J. (1991), *Game Theory*. Cambridge (MA): MIT Press.

Gibbard, A. (1973), "Manipulation of Voting Schemes: A General Result," *Econometrica* **41**, 587-601.

Green, J., and Laffont, J.-J. (1977), "Characterization of Satisfactory Mechanisms for the Revelation of Preferences for Public Goods," *Econometrica* **45**, 427-438.

Green, J., and Laffont, J.-J. (1979), *Incentives in Public Decision Making*. Amsterdam: North-Holland.

Groves, T. (1973), "Incentives in Teams," *Econometrica* **41**, 617-631.

Laffont, J.-J., and Maskin, E. (1980), "A Differentiable Approach to Dominant Strategy Mechanisms," *Econometrica* **48**, 1507-1520.

Laffont, J.-J., and Maskin, E. (1982), "The Theory of Incentives: An Overview," in W. Hildenbrand (ed.), *Advances in Economic Theory - Fourth World Congress*. Cambridge (UK): Cambridge University Press.

Mas-Colell, A., Whinston, M., and Green, J. (1995), *Microeconomic Theory*. New York: Oxford University Press.

Maskin, E. (1977), "Nash Equilibrium and Welfare Optimality," mimeo, Massachusetts Institute of Technology.

Maskin, E. (1985), "The Theory of Implementation in Nash Equilibrium: A Survey," in L. Hurwicz, D. Schmeidler, and H. Sonnenschein (eds.), *Social Goals and Social Organization: Essays in Honor of Elisha Pazner*. Cambridge (UK): Cambridge University Press.

Mirrlees, J. (1971), "An Exploration in the Theory of Optimum Income Taxation," *Review of Economic Studies* **38**, 175-208.

Moore, J. (1992), "Implementation, Contracts, and Renegotiation in Environments with Complete Information," in J.-J. Laffont (ed.), *Advances in Economic Theory - Sixth World Congress*. Cambridge (UK): Cambridge University Press.

Moore, J., and Repullo, R. (1990), "Nash Implementation: A Full Characterization," *Econometrica* **58**, 1083-1099.

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

Myerson, R., and Satterthwaite, M. (1983), "Efficient Mechanisms for Bilateral Trading," *Jounal of Economic Theory* **29**, 265-281.

Palfrey, T. (1992), "Implementation in Bayesian Equilibrium: The Multiple Equilibrium Problem in Mechanism Design," in J.-J. Laffont (ed.), *Advances in Economic Theory - Sixth World Congress*. Cambridge (UK): Cambridge University Press.

Repullo, R. (1986), "On the Revelation Principle under Complete and Incomplete Information," in K. Binmore and P. Dasgupta (eds.), *Economic Organizations as Games*. Oxford: Basil Blackwell.

Repullo, R. (1987), "A Simple Proof of Maskin's Theorem on Nash Implementation," *Social Choice and Welfare* **4**, 39-41.

Riley, J., and Samuelson, W. (1981), "Optimal Auctions," *American Economic Review* **71**, 381-392.

Satterthwaite, M. (1975), "Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions," *Journal of Economic Theory* **10**, 187-217.

Vickrey, W. (1961), "Counterspeculation, Auctions, and Competitive Sealed Tenders," *Journal of Finance* **16**, 8-37.

Chapter 3

Adverse Selection

3.1 Static Adverse Selection

The previous chapter ended with a discussion of optimal mechanisms that maximize the planner's utility subject to the constraints that the agents reveal their preferences truthfully and receive at least their reservation utility in equilibrium. In this chapter, we examine the special case in which there is only one agent. This case is commonly known as *adverse selection*, which constitutes a somewhat inappropriate use of language as it refers to the consequences of asymmetric information rather than to the problem as such. In order to distinguish the one agent-case from the more general implementation problem, we will call the planner from now on *principal* - an expression that carries over to the next chapter where we deal with moral hazard. With only one agent, the principal need no longer be concerned with the intricate problems of equilibrium play, and the concepts of dominant strategy, Nash, and Bayesian equilibrium all coincide. In particular, all versions of the revelation principle are now equivalent, and TIDS, TINS, and TIBS boil down to the same requirement that the social choice rule be implementable.

In many applications of adverse selection models, the notion of a mechanism has received a very natural interpretation as a menu of outcomes from which the agent can choose. For instance, nonlinear pricing models typically assume that firms offer different price/quantity-pairs in order to separate consumers with high valuations from those with low valuations. Similarly, models of insurance markets assume that firms offer a menu of insurance policies in order to separate individuals with different risk attributes. The act of offering a menu of outcomes in order to separate agents with different types is known as *screening*, and the subsequent revelation of preferences through the choice of an outcome is called *self-selection*. However, by the

revelation principle, we can ignore mechanisms in which the agent chooses from a menu of outcomes and concentrate on direct mechanisms in which the agent reports his type to the principal who then makes a choice from the menu on his behalf.

The Model

Consider the following model:

1. There is a finite set A of feasible outcomes. A feasible outcome $y = (x, t)$ consists of a decision $x \in X$ and a monetary transfer $t \in \mathrm{R}$ from the principal to the agent.

2. The agent has a type $\theta \in \Theta = \left[\underline{\theta}, \overline{\theta}\right]$ with $\underline{\theta} \neq \overline{\theta}$ and density $\pi(\theta) > 0$ for all $\theta \in \Theta$. The principal cannot observe θ.

3. The agent has quasilinear preferences of the form $U(x, t, \theta) = u(x, \theta) + t$, where $u(\cdot)$ is thrice continuously differentiable and strictly concave in x.

4. The principal has quasilinear preferences of the form $V(x, t) = v(x) - t$, where $v(\cdot)$ is twice continuously differentiable and concave in x.

5. A social choice rule (SCR) is a correspondence $f : \Theta \twoheadrightarrow A$ which specifies a nonempty choice set $f(\theta) \subseteq A$ for every type θ.

Before we proceed, let us compare these assumptions with those used in chapter 2. Assumption 1 is equivalent to assumption 9, except that x is no longer binary, and assumption 2 is identical to assumption 3a. Assumptions 3 and 4 are similar to assumption 10, albeit here, the valuation functions take a more general form. However, novel are the concavity and differentiability assumptions. Incidentally, assumption 8 (private values) is also satisfied. Finally, assumption 5 coincides with assumption 6 of chapter 2. Since we concentrate on SCRs that maximize the principal's expected utility, the term "social" is rather unsuitable. We have nonetheless kept this terminology in order to emphasize the resemblance to the model of chapter 2.

6. a) $\frac{\partial u^3(x,\theta)}{\partial x \partial \theta^2} \leq 0$, b) $\frac{\partial u^3(x,\theta)}{\partial x^2 \partial \theta} \leq 0$, c) $\frac{\partial^2}{\partial x^2}\left(u(x,\theta) - \frac{\partial u(x,\theta)}{\partial \theta}\frac{1-\Pi(\theta)}{\pi(\theta)}\right) < 0$, and d) $\frac{\partial u(x,\theta)}{\partial \theta} > 0$.

7. $X \subseteq [0, \overline{x}]$, where $\overline{x} > \arg\max_x [v(x) + u(x, \theta)] > 0$ for all θ.

8. $\frac{d}{d\theta}\frac{\pi(\theta)}{1-\Pi(\theta)} \geq 0$.

Let us once more comment on the assumptions. Assumptions 6a, b, and c involve third derivatives and are therefore controversial. Observe that assumption 6c requires that either $u(x,\theta)$ is sufficiently concave or that $-\frac{\partial u(x,\theta)}{\partial \theta}$ is not too convex in x. Assumption 6d states that agents with a higher type have a higher valuation. Incidentally, this assumption is met by the quasilinear utility function $\theta x + t$, which was used repeatedly in chapter 2. In 7, the values of x that solve the right-hand side of the inequality represent the first-best solution. Thus, assumption 7 requires that the first-best problem has an interior solution. Assumption 8 is the by now familiar monotone hazard rate property (MHRP) encountered in our discussion of optimal auctions. Finally, let us introduce an assumption known as *single-crossing property* or *Spence-Mirrlees condition* on which the entire analysis in this chapter (and that of optimal mechanisms in general) is based.

Definition 1 (Single-Crossing Property) The agent's utility function $U(x,t,\theta)$ satisfies the single-crossing property if $\frac{\partial U(x,t,\theta)}{\partial x}/\frac{\partial U(x,t,\theta)}{\partial t}$ is either strictly increasing (SC^+) or decreasing (SC^-) in θ.

Here, we assume that SC^+ holds, so that the single-crossing property implies $\frac{\partial^2 u(x,\theta)}{\partial x \partial \theta} > 0$. Also, note that the (quasi-)linear utility function $\theta x + t$ used in the previous chapter trivially satisfies the single-crossing property. In the (t,x) space, assumption 3 together with SC^+ implies that agents with higher values of θ have steeper indifference curves, which in turn implies that the indifference curves of two agents with different types cross only once.

Necessary and Sufficient Conditions for Truthful Implementation with Individual Rationality Constraints

As in the case of optimal auctions, our objective is to identify the set of SCRs that maximize the principal's expected utility subject to being implementable and individually rational. By the revelation principle, we can again restrict attention to direct mechanisms and concentrate on SCRs that are truthfully implementable or *incentive compatible*.

Definition 2 (Incentive Compatible Social Choice Rule) The social choice rule $f(\theta)$ is truthfully implementable or incentive compatible (IC) if there exists a direct mechanism Γ_d such that i) truthtelling is optimal for the agent, i.e. if for all $\theta \in \Theta$,

$$U(x(\theta), t(\theta), \theta) \geq U(x(\hat{\theta}), t(\hat{\theta}), \theta) \tag{3.1}$$

for all $\hat{\theta} \in \Theta$, and ii) $(x(\theta), t(\theta)) \in f(\theta)$.

In the applied literature, the incentive compatibility constraints (3.1) are sometimes called *self-selection constraints*. In adverse selection models, there is a loss of generality from restricting attention to direct mechanisms only if the agent's problem has several global maxima, some of which have outcomes $(x(\hat{\theta}), t(\hat{\theta})) \notin f(\theta)$. In this case, it is usually assumed that the agent chooses the outcome preferred by the principal.

For convenience, let us define $U(\hat{\theta}, \theta) \equiv (x(\hat{\theta}), t(\hat{\theta}), \theta)$ and $U(\theta, \theta) \equiv U(\theta)$. Since there is no uncertainty with respect to other agents, ex-post individual rationality and IIR coincide and reduce to the concept of *individual rationality*.

Definition 3 (Individually Rational Social Choice Rule) An SCR is individually rational (IR) if $U(\theta) \geq 0$ for all $\theta \in \Theta$.

Again, we have normalized the reservation utilities to zero. Note that this is only possible if we assume that the agent's reservation utility is independent of his type. We can now characterize the set of SCRs that satisfy both IC and IR. For technical reasons, we need to limit the analysis to functions $x(\theta)$ and $t(\theta)$ that are continuously differentiable.

Theorem 1 An SCR satisfies both IC and IR if and only if

1) $x(\theta)$ is nondecreasing,

2) $U(\theta) = U(\underline{\theta}) + \int_{\underline{\theta}}^{\theta} \frac{\partial u(x(\eta),\eta)}{\partial \eta} d\eta$ for all θ, and

3) $U(\underline{\theta}) \geq 0$.

Proof (direct/by contradiction) "if"-part: First, we show that 1) and 2) imply IC. Suppose not. Then there exists a value $\hat{\theta} \neq \theta$ such that

$$U(\theta) < U(\hat{\theta}, \theta) \tag{3.2}$$

or equivalently,

$$\begin{aligned} U(\theta) - U(\hat{\theta}, \theta) &= \int_{\hat{\theta}}^{\theta} \frac{\partial U(\eta, \theta)}{\partial \eta} d\eta \\ &= \int_{\hat{\theta}}^{\theta} \left(\frac{\partial u(x(\eta), \theta)}{\partial x} \frac{dx(\eta)}{d\eta} + \frac{dt(\eta)}{d\eta} \right) d\eta \\ &< 0. \end{aligned} \tag{3.3}$$

Differentiating 2) with respect to θ yields

$$-\frac{\partial u(x(\theta), \theta)}{\partial x} \frac{dx(\theta)}{d\theta} = \frac{dt(\theta)}{d\theta} \tag{3.4}$$

for all $\theta \in \Theta$. Inserting (3.4) in (3.3), we have

$$\int_{\hat{\theta}}^{\theta} \left(\frac{\partial u\left(x\left(\eta\right),\theta\right)}{\partial x} - \frac{\partial u\left(x\left(\eta\right),\eta\right)}{\partial x} \right) \frac{dx\left(\eta\right)}{d\eta} d\eta < 0. \tag{3.5}$$

Suppose first that $\theta > \hat{\theta}$. The single-crossing property then implies that $\frac{\partial u(x(\eta),\theta)}{\partial x} > \frac{\partial u(x(\eta),\eta)}{\partial x}$ for all $\eta \in [\hat{\theta}, \theta)$. Together with 1), this implies that the integral in (3.5) is nonnegative, a contradiction. Next, suppose that $\theta < \hat{\theta}$. Changing the order of integration, we can rewrite (3.5) as

$$-\int_{\theta}^{\hat{\theta}} \left(\frac{\partial u\left(x\left(\eta\right),\theta\right)}{\partial x} - \frac{\partial u\left(x\left(\eta\right),\eta\right)}{\partial x} \right) \frac{dx\left(\eta\right)}{d\eta} d\eta < 0. \tag{3.6}$$

Due to the single-crossing property, we have $\frac{\partial u(x(\eta),\theta)}{\partial x} < \frac{\partial u(x(\eta),\eta)}{\partial x}$ for all $\eta \in (\theta, \hat{\theta}]$. In conjunction with 1), this implies that the left-hand side in (3.6) is nonnegative, a contradiction. Finally, let us prove that 2) and 3) imply IR. Suppose not. Then there exists a type $\theta > \underline{\theta}$ such that $U\left(\theta\right) < 0$. In this case, 2) implies that

$$U\left(\theta\right) = U\left(\underline{\theta}\right) + \int_{\underline{\theta}}^{\theta} \frac{\partial u\left(x\left(\eta\right),\eta\right)}{\partial \eta} d\eta < 0, \tag{3.7}$$

where $U\left(\underline{\theta}\right) \geq 0$ due to 3). But from assumption 6d, it follows that the integral in (3.7) is strictly positive, a contradiction.

"only if"-part: We now show that IC implies 1) and 2). For any two values $\theta, \hat{\theta} \in \Theta$, IC requires that

$$U\left(\theta\right) \geq u(x(\hat{\theta}), \theta) + t(\hat{\theta}) \equiv U(\hat{\theta}) + u(x(\hat{\theta}), \theta) - u(x(\hat{\theta}), \hat{\theta}), \tag{3.8}$$

and

$$U(\hat{\theta}) \geq u(x(\theta), \hat{\theta}) + t(\theta) \equiv U(\theta) + u(x(\theta), \hat{\theta}) - u\left(x\left(\theta\right), \theta\right), \tag{3.9}$$

which implies

$$u\left(x\left(\theta\right), \theta\right) - u(x(\theta), \hat{\theta}) \geq U\left(\theta\right) - U(\hat{\theta}) \geq u(x(\hat{\theta}), \theta) - u(x(\hat{\theta}), \hat{\theta}). \tag{3.10}$$

Suppose without loss of generality that $\theta > \hat{\theta}$. Rearranging, we obtain

$$u\left(x\left(\theta\right), \theta\right) - u(x(\hat{\theta}), \theta) \geq u(x(\theta), \hat{\theta}) - u(x(\hat{\theta}), \hat{\theta}). \tag{3.11}$$

Together with the single-crossing property, (3.11) implies that $x\left(\theta\right)$ is nondecreasing (note that this is true regardless of the sign of $\frac{\partial u(\cdot)}{\partial x}$). Dividing (3.10)

by $(\theta - \hat{\theta})$ and letting $\theta \to \hat{\theta}$, we have $\frac{dU(\theta)}{d\theta} = \frac{\partial u(x(\theta),\theta)}{\partial \theta}$ for all θ. Integrating both sides over $[\underline{\theta}, \theta]$ yields

$$U(\theta) = U(\underline{\theta}) + \int_{\underline{\theta}}^{\theta} \frac{\partial u(x(\eta), \eta)}{\partial \eta} d\eta \tag{3.12}$$

for all θ. Finally, note that IR obviously implies 3) by definition 3. ■

Remarks

1. Theorem 1 is the analogue of theorem 17 in chapter 2. Like theorem 17, it allows us to replace the original IC and IR constraints with the mathematically more tractable conditions 1)-3). This approach was first used by Mirrlees (1971) in his study of optimal income taxation.

2. Differentiating 2) with respect to θ, we obtain the agent's first-order condition at $\hat{\theta} = \theta$. Thus, theorem 1 states that if $x(\theta)$ is nondecreasing, the (global) incentive compatibility constraints can be replaced by the agent's (local) first-order condition. A similar technique also exists in moral hazard models and is known as *first-order approach* (cf. chapter 4), albeit there, the conditions under which the first-order condition is sufficient are less appealing.

Optimal Mechanisms

The design of optimal mechanisms can be viewed as a two step-procedure: The first step characterizes the set of SCRs that satisfy both IC and IR, and the second step selects from this set those SCRs that maximize the principal's expected utility. We now proceed with the second step. As a benchmark, let us first determine the optimal solution for the complete information case.

Theorem 2 Suppose that the agent's type θ is observable. The social choice rule $f^*(\theta) = (x^*(\theta), t^*(\theta))$ maximizes the principal's utility subject to IR if and only if

$$\frac{dv(x^*(\theta))}{dx} + \frac{\partial u(x^*(\theta), \theta)}{\partial x} = 0 \tag{3.13}$$

and

$$t^*(\theta) = -u(x^*(\theta), \theta). \tag{3.14}$$

Proof (direct) The principal's first-best problem is

$$\max_{x(\theta), t(\theta)} v(x(\theta)) - t(\theta) \tag{3.15}$$

$$\text{s.t.} \quad u\left(x\left(\theta\right),\theta\right)+t\left(\theta\right)\geq 0.$$

Inspection of (3.15) immediately reveals that IR must be binding. Thus, the principal's relaxed problem can be written as

$$\max_{x(\theta)} v\left(x\left(\theta\right)\right)+u\left(x\left(\theta\right),\theta\right), \tag{3.16}$$

which corresponds to the maximization of total welfare. Differentiating with respect to x gives

$$\frac{dv\left(x^{*}\left(\theta\right)\right)}{dx}+\frac{\partial u\left(x^{*}\left(\theta\right),\theta\right)}{\partial x}=0, \tag{3.17}$$

and inserting $x^{*}\left(\theta\right)$ in IR yields

$$t^{*}\left(\theta\right)=-u\left(x^{*}\left(\theta\right),\theta\right). \tag{3.18}$$

Since $v\left(\cdot\right)$ is concave and $u\left(\cdot\right)$ is strictly concave , the first-order condition (3.17) is sufficient for a (unique) global maximum. ■

Back to the incomplete information case, let us now determine the optimal SCR from the set of SCRs that satisfy both IC and IR.

Theorem 3 The social choice rule $f^{*}\left(\theta\right)=\left(x^{*}\left(\theta\right),t^{*}\left(\theta\right)\right)$ maximizes the principal's expected utility subject to IC and IR if and only if

$$\frac{dv\left(x^{*}\left(\theta\right)\right)}{dx}+\frac{\partial u\left(x^{*}\left(\theta\right),\theta\right)}{\partial x}-\frac{\partial^{2}u\left(x^{*}\left(\theta\right),\theta\right)}{\partial x\partial\theta}\frac{1-\Pi\left(\theta\right)}{\pi\left(\theta\right)}=0 \tag{3.19}$$

and

$$t^{*}\left(\theta\right)=-u\left(x^{*}\left(\theta\right),\theta\right)+\int_{\underline{\theta}}^{\theta}\frac{\partial u\left(x^{*}\left(\eta\right),\eta\right)}{\partial\eta}. \tag{3.20}$$

Proof (direct) By theorem 1, we can replace IC and IR with conditions 1)-3). The principal's relaxed problem can then be written as

$$\max_{x(\theta),t(\theta)}\int_{\Theta}\left(v\left(x\left(\theta\right)\right)-t\left(\theta\right)\right)d\Pi\left(\theta\right) \tag{3.21}$$

$$\begin{aligned}\text{s.t.}\quad & 1)\ x\left(\theta\right)\ \text{is nondecreasing,}\\ & 2)\ U\left(\theta\right)=U\left(\underline{\theta}\right)+\textstyle\int_{\underline{\theta}}^{\theta}\frac{\partial u(x(\eta),\eta)}{\partial\eta}d\eta\ \text{for all}\ \theta,\\ & 3)\ U\left(\underline{\theta}\right)\geq 0.\end{aligned}$$

Clearly, constraint 3) must be binding since transfers are costly to the principal. Consequently, 2) reduces to

$$U(\theta) = \int_{\underline{\theta}}^{\theta} \frac{\partial u(x(\eta),\eta)}{\partial \eta} d\eta, \tag{3.22}$$

or

$$t(\theta) = -u(x(\theta),\theta) + \int_{\underline{\theta}}^{\theta} \frac{\partial u(x(\eta),\eta)}{\partial \eta} d\eta \tag{3.23}$$

for all θ, where we used the fact that $U(\theta) \equiv u(x(\theta),\theta) + t(\theta)$. Inserting (3.23) in the objective function (3.21), the principal's relaxed problem is equivalent to

$$\max_{x(\theta)} \int_{\Theta} \left(v(x(\theta)) + u(x(\theta),\theta)\right) d\Pi(\theta) - \int_{\Theta} \int_{\underline{\theta}}^{\theta} \frac{\partial u(x(\eta),\eta)}{\partial \eta} d\eta d\Pi(\theta) \tag{3.24}$$

s.t. 1).

Using integration by parts, the second term in (3.24) can be expressed as

$$\begin{aligned} \int_{\Theta} \int_{\underline{\theta}}^{\theta} \frac{\partial u(x(\eta),\eta)}{\partial \eta} d\eta d\Pi(\theta) &= \int_{\underline{\theta}}^{\theta} \frac{\partial u(x(\eta),\eta)}{\partial \eta} d\eta \Pi(\theta) \Big|_{\underline{\theta}}^{\overline{\theta}} \\ &\quad - \int_{\Theta} \frac{\partial u(x(\theta),\theta)}{\partial \theta} \Pi(\theta)\, d\theta \\ &= \int_{\Theta} \frac{\partial u(x(\theta),\theta)}{\partial \theta} \frac{1-\Pi(\theta)}{\pi(\theta)} d\Pi(\theta). \end{aligned} \tag{3.25}$$

Inserting (3.25) in (3.24), the principal's relaxed problem reduces to

$$\max_{x(\theta)} \int_{\Theta} \left(v(x(\theta)) + u(x(\theta),\theta) - \frac{\partial u(x(\theta),\theta)}{\partial \theta} \frac{1-\Pi(\theta)}{\pi(\theta)} \right) d\Pi(\theta) \tag{3.26}$$

s.t. 1).

Let us ignore constraint 1) for the moment. Pointwise maximizing (3.26) with respect to x, we obtain (after dividing through by $\pi(\theta)$)

$$\frac{dv(x^*(\theta))}{dx} + \frac{\partial u(x^*(\theta),\theta)}{\partial x} - \frac{\partial^2 u(x^*(\theta),\theta)}{\partial x \partial \theta} \frac{1-\Pi(\theta)}{\pi(\theta)} = 0 \tag{3.27}$$

for all θ. Inserting $x^*(\theta)$ in (3.23) yields

$$t^*(\theta) = -u(x^*(\theta),\theta) + \int_{\underline{\theta}}^{\theta} \frac{\partial u(x^*(\eta),\eta)}{\partial \eta} d\eta. \tag{3.28}$$

By assumption 6c, the objective function (3.26) is strictly concave in x, which implies that the first-order condition (3.27) is sufficient for a (unique) global maximum.

Finally, let us verify that the monotonicity constraint 1) holds. Totally differentiating (3.27), we have

$$\begin{aligned} & \frac{d\left(x^*\left(\theta\right)\right)}{d\theta}\left(\frac{d^2v\left(x^*\left(\theta\right)\right)}{dx^2}+\frac{\partial^2u\left(x^*\left(\theta\right),\theta\right)}{\partial x^2}-\frac{\partial^3u\left(x^*\left(\theta\right),\theta\right)}{\partial x^2\partial\theta}\frac{1-\Pi\left(\theta\right)}{\pi\left(\theta\right)}\right) \\ = & \frac{\partial^2u\left(x^*\left(\theta\right),\theta\right)}{\partial x\partial\theta}\left(\frac{d}{d\theta}\frac{1-\Pi\left(\theta\right)}{\pi\left(\theta\right)}-1\right)+\frac{\partial^3u\left(x^*\left(\theta\right),\theta\right)}{\partial x\partial\theta^2}\frac{1-\Pi\left(\theta\right)}{\pi\left(\theta\right)}. \quad (3.29) \end{aligned}$$

for all θ. Concavity of $v\left(\cdot\right)$ together with assumption 6c implies that the expression in the parenthesis on the left-hand side of (3.29) is negative). Furthermore, the single-crossing property, in conjunction with assumption 8 (MHRP) and assumption 6a implies that the right-hand side is of (3.29) is also negative. Together, this establishes that $x^*\left(\theta\right)$ is (strictly) increasing and that constraint 1) is satisfied. ■

Remarks

1. The optimal solution (3.20) reveals that $U\left(\theta\right)>U\left(\underline{\theta}\right)$. Since $U\left(\underline{\theta}\right)=0$, this implies that IR is only binding at $\theta=\underline{\theta}$. If the agent's reservation utility is type-dependent, this need no longer be the case and we may observe that IR is also binding at other values.

2. The surplus $U\left(\theta\right)=\int_{\underline{\theta}}^{\theta}\frac{\partial u(x^*(\eta),\eta)}{\partial\eta}d\eta$ is known as type θ's *information rent* and is strictly increasing in θ. Intuitively, the information rent reflects the ability of agents with higher types to mimick those with lower types. To see this, suppose for a moment that contrary to (3.20), it is true that $U\left(x^*\left(\theta\right),\theta\right)=U\left(x^*\left(\theta'\right),\theta'\right)$ for $\theta>\theta'$. By the single-crossing property, we then have $U\left(x^*\left(\theta'\right),\theta\right)>U\left(x^*\left(\theta'\right),\theta'\right)$, i.e. type θ is better off by announcing θ' instead of his own type θ, which violates IC. In order to make type θ indifferent between lying and telling the truth, he must be offered some additional rent, i.e. $U\left(x^*\left(\theta\right),\theta\right)>U\left(x^*\left(\theta'\right),\theta'\right)$.

3. Inspection of the optimal solution shows that (3.19)-(3.20) coincides with the first-best solution (3.13)-(3.14) at $\theta=\overline{\theta}$ (recall that $\Pi\left(\overline{\theta}\right)=1$). This result is commonly referred to as "*no distortion at the top*". Moreover, a simple revealed preference argument shows that for all $\theta<\overline{\theta}$, the second-best solution entails an underprovision of x. To see this, recall that the first-best solution $x_{FB}^*\left(\theta\right)$ maximizes (3.16) for all

θ, while the second-best solution $x^*_{SB}(\theta)$ maximizes (3.26) for all θ. But this implies that

$$v\left(x^*_{FB}(\theta)\right) + u\left(x^*_{FB}(\theta), \theta\right) \geq v\left(x^*_{SB}(\theta)\right) + u\left(x^*_{SB}(\theta), \theta\right)$$

and

$$\begin{aligned} & v\left(x^*_{SB}(\theta)\right) + u\left(x^*_{SB}(\theta), \theta\right) - \frac{\partial u\left(x^*_{SB}(\theta), \theta\right)}{\partial \theta} \frac{1 - \Pi(\theta)}{\pi(\theta)} \\ \geq \; & v\left(x^*_{FB}(\theta)\right) + u\left(x^*_{FB}(\theta), \theta\right) - \frac{\partial u\left(x^*_{FB}(\theta), \theta\right)}{\partial \theta} \frac{1 - \Pi(\theta)}{\pi(\theta)}. \end{aligned}$$

Adding up both inequalities yields $\frac{\partial u\left(x^*_{SB}(\theta), \theta\right)}{\partial \theta} \leq \frac{\partial u\left(x^*_{FB}(\theta), \theta\right)}{\partial \theta}$ for all θ. By the single-crossing property, $\frac{\partial u(x,\theta)}{\partial \theta}$ is strictly increasing in x, which implies $x^*_{SB}(\theta) \leq x^*_{FB}(\theta)$ for all θ. Finally, a comparison of the first-order conditions (3.13) and (3.19) shows that $x^*_{SB}(\theta) \neq x^*_{FB}(\theta)$ for all $\theta < \overline{\theta}$ because $\frac{\partial u(x(\theta),\theta)}{\partial \theta} \frac{1-\Pi(\theta)}{\pi(\theta)}$ is strictly positive for all $\theta \neq \overline{\theta}$. But since we established that $x^*_{SB}(\theta) \leq x^*_{FB}(\theta)$, this implies that $x^*_{SB}(\theta) < x^*_{FB}(\theta)$ for all $\theta < \overline{\theta}$, i.e. the second-best allocation is strictly inefficient.

4. In (3.26), the principal's relaxed problem reflects a tradeoff between efficiency (represented by the total surplus $v(x(\theta)) + u(x(\theta), \theta)$) and rent extraction (represented by the term $\frac{\partial u(x(\theta),\theta)}{\partial \theta} \frac{1-\Pi(\theta)}{\pi(\theta)} = \int_{\underline{\theta}}^{\theta} \frac{\partial u(x(\eta),\eta)}{\partial \eta} d\eta$ (cf. (3.25)). The first-order condition (3.27) then says that at the optimum, the increase in total surplus from a marginal increase in $x(\cdot)$ must be offset by the increase in the agent's rent.

5. In the proof, we have implicitly assumed that the principal's relaxed problem (3.26) has an interior solution, i.e. we have ignored the constraint $x \in X \subseteq [0, \overline{x}]$. Since $x^*(\theta)$ is nondecreasing (in fact, it is strictly increasing), it is true that $x^*(\theta) \leq x^*\left(\overline{\theta}\right)$. But $x^*\left(\overline{\theta}\right) = \arg\max_x \left[v(x) + u\left(x, \overline{\theta}\right)\right]$ (this can be seen by comparing (3.19) with the welfare optimum (3.14) at $\theta = \overline{\theta}$), which is strictly smaller than $\overline{x}$ by assumption 7. Hence, $x \leq \overline{x}$ is never binding. However, $x \geq 0$ may be binding, in which case the optimal solution is $x^*(\theta) = 0$.

6. If assumption 8 (MHRP) is not satisfied, the optimal solution $x^*(\theta)$ may (!) not be nondecreasing everywhere. If this is the case, the monotonicity constraint 1) is binding over some range and the optimal solution involves *bunching* or *pooling*, i.e. the principal chooses the same value of $x(\theta)$ for a subset of $\left[\underline{\theta}, \overline{\theta}\right]$ (see Guesnerie and Laffont (1984) for a more detailed analysis).

Application: Nonlinear Pricing

Consider a monopolist who produces a good at a constant marginal cost of c. The monopolist has utility $V(x,t) = t - xc$, where t denotes the monopolist's revenue and x is the quantity of the good sold. Consumers have utility functions $U(x,t,\theta) = u(x,\theta) - t$ indexed by their valuation $\theta \in \left[\underline{\theta}, \overline{\theta}\right]$. Unlike in the previous setting, we now assume that SC^- holds. Since transfers are negative, this implies again that $\frac{\partial^2 u(x,\theta)}{\partial x \partial \theta} > 0$, i.e. consumers with a high valuation have a higher marginal utility from the good than consumers with a low valuation.

As a benchmark, let us first derive the first-best solution. From (3.13)-(3.14), it follows that

$$\frac{\partial u\left(x^*(\theta),\theta\right)}{\partial x} = c \tag{3.30}$$

and

$$t^*(\theta) = u\left(x^*(\theta),\theta\right) \tag{3.31}$$

for all θ. Equation (3.30) requires that marginal utility is equal to marginal cost, and (3.31) implies that after purchasing $x^*(\theta)$ units of the good, a consumer with valuation θ should be left with zero utility.

In the presence of incomplete information, the welfare optimum cannot generally be attained. By theorem 3, we have

$$\frac{\partial u\left(x^*(\theta),\theta\right)}{\partial x} - \frac{\partial^2 u\left(x^*(\theta),\theta\right)}{\partial x \partial \theta} \frac{1 - \Pi(\theta)}{\pi(\theta)} = c \tag{3.32}$$

and

$$t^*(\theta) = u\left(x^*(\theta),\theta\right) - \int_{\underline{\theta}}^{\theta} \frac{\partial u\left(x^*(\eta),\eta\right)}{\partial \eta} \tag{3.33}$$

for all θ. Setting $\theta = \overline{\theta}$ in (3.32) reveals that the consumer with the highest valuation consumes the (first-best) efficient quantity. Moreover, using the revealed preference argument from the previous subsection, it can be shown that consumers with valuations $\theta < \overline{\theta}$ consume an inefficiently small amount. This underconsumption reflects the fundamental tradeoff between efficiency and rent extraction. Suppose that the monopolist wanted to raise the quantity $x^*(\theta)$ for all types $\theta \in [\theta_1, \theta_2]$. By the single-crossing property, this increases the integrand in (3.33) for all $\theta \in [\theta_1, \theta_2]$, which in turn implies less rent extraction from consumers with valuations $\theta > \theta_2$. The effect on rent extraction from types $\theta \in [\theta_1, \theta_2]$ is ambiguous since both components of $t^*(\theta)$ are now larger. In addition, production costs rise proportionally at a rate of c. Balancing these effects, the monopolist sacrifices some efficiency in order to limit the rent earned by consumers with high valuations.

In this framework, it can also be shown that the optimal nonlinear pricing rule entails quantity discounts. Since $x^*(\theta)$ is monotonic, it has a monotonic inverse $x^{*-1}(x)$ such that $x^{*-1}(x) = \theta \Leftrightarrow x = x^*(\theta)$. Thus, $x^{*-1}(x)$ denotes the type that consumes exactly x units according to the optimal decision $x^*(\theta)$. We can then rewrite the optimal transfer function $t^*(\theta)$ as $t(x^{*-1}(x))$, which is a function of x only. Our objective is to show that $t(x^{*-1}(x))$ is strictly concave in x.

Theorem 4 The optimal nonlinear pricing rule involves quantity discounts.

Proof (direct) Differentiating $t(x^{*-1}(x))$ with respect to x yields

$$\frac{dt(x^{*-1}(x))}{dx} = \frac{dt(x^{*-1}(x))}{d\theta}\frac{dx^{*-1}(x)}{dx}. \tag{3.34}$$

By the inverse function theorem, it is true that $\frac{dx^{*-1}(x)}{dx} = 1 \Big/ \frac{dx^*(\theta)}{d\theta}$. Equation (3.34) can then be rewritten as

$$\frac{dt(x^{*-1}(x))}{d\theta} = \frac{dt(x^{*-1}(x))}{dx}\frac{dx^*(\theta)}{d\theta}. \tag{3.35}$$

Totally differentiating (3.33), we have

$$\frac{dt(x^{*-1}(x))}{d\theta} = \frac{\partial u(x, x^{*-1}(x))}{\partial x}\frac{dx^*(\theta)}{d\theta}. \tag{3.36}$$

Subtracting (3.36) from (3.35) gives

$$\frac{dt(x^{*-1}(x))}{dx} = \frac{\partial u(x, x^{*-1}(x))}{\partial x}, \tag{3.37}$$

and totally differentiating (3.37), we obtain

$$\frac{d^2t(x^{*-1}(x))}{dx^2} = \frac{\partial^2 u(x, x^{*-1}(x))}{\partial x^2} + \frac{\partial^2 u(x, x^{*-1}(x))}{\partial x \partial \theta}\frac{1}{\frac{dx^*(\theta)}{d\theta}}, \tag{3.38}$$

where we again used the inverse function theorem. For convenience, let us now use subscripts in order to denote partial derivatives. Equation (3.38) implies that $\frac{d^2t\left(x^{*-1}(x)\right)}{dx^2}$ is strictly negative if and only if

$$-\frac{u_{x\theta}}{u_{xx}} < \frac{dx^*(\theta)}{d\theta}. \tag{3.39}$$

From (3.29) we know that (note that $\frac{d^2v(x^*(\theta))}{dx^2}=0$)

$$\frac{d\left(x^*\left(\theta\right)\right)}{d\theta}\left(u_{xx}-u_{xx\theta}\frac{1-\Pi\left(\theta\right)}{\pi\left(\theta\right)}\right) = u_{x\theta}\left(\frac{d}{d\theta}\frac{1-\Pi\left(\theta\right)}{\pi\left(\theta\right)}-1\right) + u_{x\theta\theta}\frac{1-\Pi\left(\theta\right)}{\pi\left(\theta\right)}. \quad (3.40)$$

Solving for $\frac{dx^*(\theta)}{d\theta}$ and inserting the result in (3.39) yields

$$-\frac{u_{x\theta}}{u_{xx}}<\frac{u_{x\theta}\left(\frac{d}{d\theta}\frac{1-\Pi(\theta)}{\pi(\theta)}-1\right)+u_{x\theta\theta}\frac{1-\Pi(\theta)}{\pi(\theta)}}{u_{xx}-u_{xx\theta}\frac{1-\Pi(\theta)}{\pi(\theta)}}. \quad (3.41)$$

which can be rearranged as

$$0<\left(u_{xx}u_{x\theta\theta}-u_{x\theta}u_{xx\theta}\right)\frac{1-\Pi\left(\theta\right)}{\pi\left(\theta\right)}+u_{xx}u_{x\theta}\frac{d}{d\theta}\frac{1-\Pi\left(\theta\right)}{\pi\left(\theta\right)}. \quad (3.42)$$

Given that assumptions 3, 6a, 6b, 8, and the single-crossing property hold, the inequality (3.42) is indeed satisfied, which implies that $t\left(x^{*-1}\left(x\right)\right)$ is strictly concave. ■

3.2 Repeated Adverse Selection

From the principal's point of view, the optimal static solution possesses two unfavorable properties. Firstly, it involves an allocative inefficiency for agents with types $\theta<\overline{\theta}$. Secondly, the principal must grant agents with types $\theta>\overline{\theta}$ an information rent $U\left(\theta\right)-U\left(\underline{\theta}\right)$. In a repeated version of the static problem, information about the agent's type is revealed over time. Naturally, the principal would then like to use this information in order to reduce both the productive inefficiency and the agent's rent. This suggests that the analysis of the repeated problem depends on whether the principal can credibly commit to ignore this information. In line with the majority of the adverse selection literature, let us from now on call an outcome $y=(t,x)$ a *contract*. For instance, in the nonlinear pricing example, the pair (x,t) can be interpreted as a sales contract, and in the insurance example mentioned at the beginning of this chapter, (x,t) constitutes an insurance contract.

Consider a T-period version of the static problem. By the above reasoning, we can distinguish between three possibilities:

1. *Full commitment*: Principal and agent can commit ex ante (at date 0) to a T-period contract $\left(x_1,t_1\right),...,\left(x_T,t_T\right)$.

2. *Commitment with renegotiation:* Principal and agent can commit ex ante to a T-period contract $(x_1, t_1), ..., (x_T, t_T)$. However, the parties cannot commit *not* to renegotiate if a Pareto-improvement arises.

3. *No commitment*: Principal and agent cannot commit to a long-term contract.

The full commitment case rests on the strong assumption that even if there is a situation in which both parties can be made better off by replacing the original contract with a new contract, they must nonetheless adhere to the former. In real-world situations, the assumption of full commitment is hopelessly unrealistic as courts do typically not intervene if all parties agree to break an agreement. The second case allows for renegotiation, but does not permit that either party unilaterally deviates from the initial T-period contract. Finally, the case where no commitment exists is essentially equivalent to a sequence of spot contracts where the principal offers a new contract (x_τ, t_τ) at the beginning of each period.

For convenience, assume that the agent's utility function is of the form $\sum_{\tau=1}^{T} \delta^{\tau-1} U(x_\tau, t_\tau, \theta) = \sum_{\tau=1}^{T} \delta^{\tau-1} (\theta x_\tau + t_\tau)$, where $\delta \in (0,1)$ denotes the discount factor. The principal's utility is $\sum_{\tau=1}^{T} \delta^{\tau-1} V(x_\tau, t_\tau)$. Let us now determine the optimal solution for $T = 2$ for all three environments.

Full Commitment

In the full commitment case, the principal is bound to ignore any information in period 2 that is revealed about the agent's type in period 1. Thus, there is no genuine dynamic interaction between periods and the principal's two-period problem reduces to a static problem. We can therefore apply the revelation principle and restrict attention to direct two-period mechanisms in which the agent announces a type $\hat{\theta}$ at date 0 and receives in return a two-period contract $g_{1,2}(\hat{\theta}) = (x_1(\hat{\theta}), t_1(\hat{\theta})), (x_2(\hat{\theta}), t_2(\hat{\theta}))$. Since both $(x_1(\hat{\theta}), t_1(\hat{\theta}))$ and $(x_2(\hat{\theta}), t_2(\hat{\theta}))$ are based on the same information, the optimal two-period contract replicates the optimal static contract in every period.

Theorem 5 Let $(x^*(\theta), t^*(\theta))$ denote the optimal static solution. Under full commitment, the social choice rule $f^*_{1,2}(\theta) = ((x^*_1(\theta), t^*_1(\theta)), (x^*_2(\theta), t^*_2(\theta)))$ maximizes the principal's expected utility subject to the agent's two-period IC and IR constraints if and only if

$$\left(x^*_\tau(\theta), t^*_\tau(\theta)\right) = (x^*(\theta), t^*(\theta)) \tag{3.43}$$

for $\tau = 1, 2$.

Proof (by contradiction) First, note that since $f^*_{1,2}(\theta)$ satisfies IC and IR, it must also satisfy the two-period incentive compatibility constraint $\text{IC}_{1,2}$

$$\sum_{\tau=1}^{2} \delta^{\tau-1} \left(\theta x_\tau(\theta) + t_\tau(\theta)\right) \geq \sum_{\tau=1}^{2} \delta^{\tau-1} (\theta x_\tau(\hat{\theta}) + t_\tau(\hat{\theta})) \tag{3.44}$$

for all $\theta, \hat{\theta} \in \Theta$, and the two-period individual rationality constraint $\text{IR}_{1,2}$

$$\sum_{\tau=1}^{2} \delta^{\tau-1} \left(\theta x_\tau(\theta) + t_\tau(\theta)\right) \geq 0, \tag{3.45}$$

where the agent's reservation utility is time-invariant and normalized to zero as usual. Finally, suppose that there exists a social choice rule $f_{1,2}(\theta) = ((x_1(\theta), t_1(\theta)), (x_2(\theta), t_2(\theta)))$ which satisfies both $(\text{IC}_{1,2})$ and $(\text{IR}_{1,2})$ and generates a higher expected utility for the principal than $f^*_{1,2}(\theta)$, i.e.

$$\begin{aligned} &\int_\Theta \sum_{\tau=1}^{2} \delta^{\tau-1} V\left(x_\tau(\theta), t_\tau(\theta)\right) d\Pi(\theta) \\ > \;& (1+\delta) \int_\Theta V\left(x^*(\theta), t^*(\theta)\right) d\Pi(\theta). \end{aligned} \tag{3.46}$$

Consider now the static problem examined in section 3.1. In particular, consider a lottery which yields the social choice rule $f_1(\theta) = (x_1(\theta), t_1(\theta))$ with probability $1/(1+\delta)$ and the social choice rule $f_2(\theta) = (x_2(\theta), t_2(\theta))$ with probability $\delta/(1+\delta)$. Dividing (3.44)-(3.45) through by $(1+\delta)$ gives

$$\sum_{\tau=1}^{2} \frac{\delta^{\tau-1}}{(1+\delta)} \left(\theta x_\tau(\theta) + t_\tau(\theta)\right) \geq \sum_{\tau=1}^{2} \frac{\delta^{\tau-1}}{(1+\delta)} (\theta x_\tau(\hat{\theta}) + t_\tau(\hat{\theta})) \tag{3.47}$$

and

$$\sum_{\tau=1}^{2} \frac{\delta^{\tau-1}}{(1+\delta)} \left(\theta x_\tau(\theta) + t_\tau(\theta)\right) \geq 0, \tag{3.48}$$

which shows that the lottery is both incentive compatible and individually rational (note that (3.47)-(3.48) are the static IC and IR constraints for the lottery). However, dividing (3.46) through by $(1+\delta)$, we have

$$\begin{aligned} &\frac{1}{(1+\delta)} \int_\Theta V\left(x_1(\theta), t_1(\theta)\right) d\Pi(\theta) + \frac{\delta}{(1+\delta)} \int_\Theta V\left(x_2(\theta), t_2(\theta)\right) d\Pi(\theta) \\ > \;& \int_\Theta V\left(x^*(\theta), t^*(\theta)\right) d\Pi(\theta), \end{aligned} \tag{3.49}$$

which implies that the principal's expected utility from the lottery is strictly greater than from the optimal (!) social choice rule $f^*(\theta) = (x^*(\theta), t^*(\theta))$, a contradiction. ■

No Commitment

Let us now study the other polar case where either party can terminate the two-period contract at the beginning of the second period. One implication of this is that we can no longer restrict attention to direct two-period mechanisms in which the principal commits to an outcome function $g_{1,2}(\hat{\theta}) = ((x_1(\hat{\theta}), t_1(\hat{\theta})), (x_2(\hat{\theta}), t_2(\hat{\theta})))$. The fact that we cannot employ the revelation principle means that we must resort to a game-theoretic analysis. Consider the following two-stage game of incomplete information.

At the beginning of period 1, the principal offers a menu of first-period contracts $(x_1(\hat{\theta}), t_1(\hat{\theta}))$ indexed by $\hat{\theta}$. The agent can then either accept or reject. If he accepts, he selects a contract from the menu by announcing a type $\hat{\theta}$. In case he rejects, the game ends and the agent receives his reservation utility of 0 in both periods. At the start of period 2, the principal offers a menu of second-period contracts $(x_2(\hat{\theta}), t_2(\hat{\theta}))$, which the agent can again either accept or reject. If the agent accepts, he chooses a contract from the menu by announcing a type $\hat{\theta}$, whereas if he rejects, he receives his second-period reservation utility of 0. Games like the one described in which the uninformed party (i.e. the principal) moves first are known as *screening games*, as opposed to *signaling games*, where the informed party moves first.

Suppose there exists a perfect Bayesian equilibrium (PBE) of the screening game that is fully separating (i.e. the function $\hat{\theta} : \Theta \to \Theta$ that maps types into first-period announcements is one-to-one). Without loss of generality, we can restrict attention to equilibria in which each agent announces his type truthfully. Since the principal knows the agent's type at the beginning of period 2, sequential rationality requires that she offers only a single second-period contract, viz. the first-best contract $\left(x^*_{FB}(\theta), t^*_{FB}(\theta)\right)$, in order to fully extract the agent's second-period rent. Given the principal's equilibrium strategy, the agent's (second-period) continuation payoff depends on his first-period announcement as follows:

1. If the agent makes a truthful report in the first period, he receives a continuation payoff $\theta x^*_{FB}(\theta) + t^*_{FB}(\theta) = 0$.

2. If the agent unilaterally deviates and announces $\hat{\theta} < \theta$, he receives a strictly positive continuation payoff $\theta x^*_{FB}(\hat{\theta}) + t^*_{FB}(\hat{\theta}) > \hat{\theta} x^*_{FB}(\hat{\theta}) + t^*_{FB}(\hat{\theta}) = 0$.

3. Finally, if the agent deviates in the other direction by announcing $\hat{\theta} > \theta$, he receives a continuation payoff $\max[\theta x^*_{FB}(\hat{\theta}) + t^*_{FB}(\hat{\theta}), 0] = 0$, which implies that he quits after the first period in order to avoid the negative second-period utility $\theta x^*_{FB}(\hat{\theta}) + t^*_{FB}(\hat{\theta})$.

Consider now a small deviation $\theta - d\theta < \theta$. Since truthtelling is optimal (recall that we are assuming a fully separating PBE), the envelope theorem implies that if type θ announces $\theta - d\theta$, he incurs only a second-order loss in the first-period. On the other hand, he enjoys a first-order profit $\theta x^*_{FB}(\theta - d\theta) + t^*_{FB}(\theta - d\theta) > 0$ in the second period, which suggests that he would like to pool with type $\theta - d\theta$. However, this contradicts our initial assumption that the equilibrium is fully separating.

Theorem 6 There exists no fully separating PBE.

Proof (by contradiction) Suppose there exists a PBE in which each type makes a truthful announcement in the first period. For technical reasons, we first need to establish that the functions $x_1(\cdot)$ and $t_1(\cdot)$ are differentiable almost everywhere.

Consider two types $\theta > \hat{\theta}$. Truthtelling requires that

$$\theta x_1(\theta) + t_1(\theta) \geq \theta x_1(\hat{\theta}) + t_1(\hat{\theta}) + \delta(\theta x^*_{FB}(\hat{\theta}) + t^*_{FB}(\hat{\theta})) \tag{3.50}$$

and

$$\hat{\theta} x_1(\hat{\theta}) + t_1(\hat{\theta}) \geq \hat{\theta} x_1(\theta) + t_1(\theta). \tag{3.51}$$

Recall that the agent receives a positive continuation payoff from announcing a lower type (cf. point 2), but a zero continuation payoff from announcing either the truth or a higher type (cf. points 1 and 3, respectively). Adding up (3.50)-(3.51) and rearranging, we obtain

$$(\theta - \hat{\theta})(x_1(\theta) - x_1(\hat{\theta})) \geq \delta(\theta x^*_{FB}(\hat{\theta}) + t^*_{FB}(\hat{\theta})) > 0, \tag{3.52}$$

which implies that $x_1(\cdot)$ is strictly increasing. Using this result in (3.51) immediately shows that $t_1(\cdot)$ is strictly decreasing. Thus, $x_1(\cdot)$ and $t_1(\cdot)$ are both differentiable almost everywhere.

Consider now a point of differentiability θ. Truthtelling implies that type θ has no incentive to claim that he is type $\theta + d\theta > \theta$, i.e.

$$\theta x(\theta) + t(\theta) \geq \theta x(\theta + d\theta) + t(\theta + d\theta), \tag{3.53}$$

which can be rearranged as

$$t(\theta + d\theta) - t(\theta) \leq \theta [x(\theta) - x(\theta + d\theta)]. \tag{3.54}$$

Dividing (3.54) through by $d\theta$ and letting $d\theta \to 0$, we have

$$\frac{dt(\theta)}{d\theta} \leq \theta \frac{dx(\theta)}{d\theta}. \tag{3.55}$$

Similarly, truthtelling also implies that type θ has no incentive to assert that he is type $\theta - d\theta < \theta$, i.e.

$$\begin{aligned} \theta x(\theta) + t(\theta) &\geq \theta x(\theta - d\theta) + t(\theta - d\theta) \\ &\quad + \delta\left(\theta x_{FB}^{*}(\theta - d\theta) + t_{FB}^{*}(\theta - d\theta)\right), \end{aligned} \tag{3.56}$$

which can be rearranged as

$$\begin{aligned} t(\theta) - t(\theta - d\theta) &\geq \theta\left[x(\theta - d\theta) - x(\theta)\right] \\ &\quad + \delta\left(\theta x_{FB}^{*}(\theta - d\theta) + t_{FB}^{*}(\theta - d\theta)\right). \end{aligned} \tag{3.57}$$

Note that

$$t_{FB}^{*}(\theta - d\theta) = -(\theta - d\theta)\, x_{FB}^{*}(\theta - d\theta) \tag{3.58}$$

by (3.14). Inserting (3.58) in (3.57), dividing the result by $d\theta$, and taking the limit as $d\theta \to 0$, we have

$$\frac{dt(\theta)}{d\theta} \geq \theta \frac{dx(\theta)}{d\theta} + \delta x_{FB}^{*}(\theta). \tag{3.59}$$

Together (3.55) and (3.59) imply

$$\theta \frac{dx(\theta)}{d\theta} \geq \frac{dt(\theta)}{d\theta} \geq \theta \frac{dx(\theta)}{d\theta} + \delta x_{FB}^{*}(\theta), \tag{3.60}$$

a contradiction, since $x_{FB}^{*}(\theta) > 0$ for all θ by assumption 7. ■

Remarks

1. Unlike in the static problem, the proof that $x_1(\cdot)$ is monotonic does not depend on the monotone hazard rate property.

2. Since the proof of theorem 6 is based on arbitrarily small deviations $\theta - d\theta$ and $\theta + d\theta$, we can replace theorem 6 with the stronger assertion that there exists no subinterval in $\left[\underline{\theta}, \overline{\theta}\right]$ with positive measure in which full separation occurs.

3. Under full commitment, the principal can replicate any outcome that is attainable under no commitment. As was shown, the full commitment outcome satisfies the two-period $\mathrm{IC}_{1,2}$ constraint (3.44) and is therefore separating, whereas the no commitment outcome involves pooling. Thus, by revealed preference, the inability to commit makes the principal strictly worse off.

4. If the agent reveals his type in the first period, the principal will exploit this information and fully extract the agent's second-period rent. This phenomenon is known as the *ratchet effect*. Hence, in order to maximize future rents, high type agents prefer to pool with low type agents.

5. While the gains or losses from infinitesimally small deviations $d\theta$ are of second order and therefore negligible, this is not the case for large deviations. Consider the $\text{IC}_{1,2}$ constraint (3.50). If type θ tells the truth, he receives no second period rent due to the ratchet effect, whereas if he mimicks type $\hat{\theta} < \theta$, he gets $\delta(\theta x^*_{FB}(\hat{\theta}) + t^*_{FB}(\hat{\theta}))$. Thus, in order to induce type θ to announce his type truthfully, the first-period contract $(x_1(\theta), t_1(\theta))$ must be sufficiently attractive. In particular, it must include the foregone rent $\delta(\theta x^*_{FB}(\hat{\theta}) + t^*_{FB}(\hat{\theta}))$. But this creates a new problem known as *take-the-money-and-run strategy*: Type $\hat{\theta}$ has now an incentive to mimick type θ, cash in the first-period profit, and quit at the end of the first period (recall that his continuation payoff from this strategy is $\max[\hat{\theta} x^*_{FB}(\theta) + t^*_{FB}(\theta), 0] = 0$ since he cannot be forced to accept the unfavorable second-period contract $(x_2(\theta), t_2(\theta))$). This suggests that the upward IC constraint (3.51) is also binding.

6. With only two types, separation is possible, although it is typically not optimal (except for small discount factors). See Laffont and Tirole (1987) for a detailed analysis.

Commitment with Renegotiation

Let us finally look at the intermediate case where renegotiation is possible. That is, neither party can unilaterally deviate from the initial contract, but nothing prevents the parties from altering the initial contract to their mutual advantage during the renegotiation stage at the beginning of period 2. A multiperiod contract that is never renegotiated is called *renegotiation-proof*.

Definition 4 (Renegotiation-Proof Contract) A long-term contract is renegotiation-proof if it is never renegotiated on the equilibrium path.

Without loss of generality, we can restrict attention to two-period contracts that are renegotiation-proof. This is because any second-period contract that results from the renegotiation process can be rationally anticipated and made part of the initial two-period contract. Obviously, the full commitment solution is not renegotiation-proof: At the renegotiation stage, the agent's type is known and it is strictly Pareto-improving to renegotiate away from the inefficient decision $x^*_2(\theta)$ toward a more efficient level.

In the no commitment case, it was shown that a fully separating menu of first-period contracts is not feasible. In the present case, the principal can credibly commit to ignore information that is revealed in the first period. As a consequence, complete separation is now possible (but not necessarily optimal). For convenience, let us be somewhat imprecise and call a menu of two-period contracts $((x_1(\hat{\theta}), t_1(\hat{\theta})), (x_2(\hat{\theta}), t_2(\hat{\theta})))$ simply a two-period contract (whose contents vary with the agent's announcement $\hat{\theta}$).

Theorem 7 There exist fully separating renegotiation-proof two-period contracts.

Proof (direct) Consider the following two-period contract: The second-period contract implements the first-best decision $x^*_{FB}(\hat{\theta})$ and leaves the principal (!) with no second-period rent (Laffont and Tirole (1990) call this contract *sell-out contract*). Since the second-period allocation is Pareto-efficient, this two-period contract is renegotiation-proof. In the first period, the agent receives the (fully separating) optimal static contract $(x^*(\hat{\theta}), t^*(\hat{\theta}))$. If there was no second period, the optimal static contract would be incentive compatible by definition. With two periods, the optimal static contract remains incentive compatible if truthtelling also maximizes the agent's second-period utility. But this is indeed the case since i) the agent is entitled to all welfare gains in the second period, and ii) $x^*_{FB}(\hat{\theta})$ maximizes second-period welfare at $\hat{\theta} = \theta$. ■

In the parlance of game theory, theorem 7 shows that if the principal offers the optimal static contract in the first period and a sell-out contract in the second period, the agent's best response is to tell the truth. However, theorem 7 does not assert that choosing such a contract (or any other fully separating two-period contract) is optimal from the principal's point of view. That is, theorem 7 does not assert the existence of a fully separating PBE. In fact, the following theorem claims that the opposite is true, i.e. that a fully separating equilibrium does not exist.

Theorem 8 A fully separating contract is never optimal for the principal.

The proof is cumbersome and omitted for the sake of brevity. See Laffont and Tirole (1990), appendix 3 for a complete proof.

Finally, note that the principal's utility under commitment with renegotiation is typically higher than under no commitment, but lower than under full commitment. Intuitively, this is obvious since under full commitment, the principal can replicate any contract that is possible in the case of commitment with renegotiation, and under commitment with renegotiation, she can always replicate the optimal sequence of spot contracts.

3.3 Bibliographic Notes

The section on static adverse selection borrows from Fudenberg and Tirole (1991), chapter 7, albeit some assumptions and proofs have been changed. Our discussion of nonlinear pricing was inspired by Maskin and Riley (1984). We have altered the Maskin-Riley model in order to make their results compatible with the more general framework used by Fudenberg and Tirole.

The treatment of repeated adverse selection is based on Laffont and Tirole (1988, 1990), as well as Laffont and Tirole (1993), chapters 1, 9, and 10. Laffont and Tirole perform their analysis in the context of a regulatory setting. We have adjusted their results so that they fit in the standard Fudenberg-Tirole framework. Results similar to those by Laffont and Tirole have also been derived by Hart and Tirole (1988) in the context of intertemporal price discrimination.

3.4 References

Guesnerie, R., and Laffont, J.-J., (1984), "A Complete Solution to a Class of Principal-Agent Problems with an Application to the Control of a Self-Managed Firm," *Journal of Public Economics* **25**, 329-369.

Hart, O. and Tirole, J. (1988), "Contract Renegotiation and Coasian Dynamics," *Review of Economic Studies* **55**, 509-540.

Laffont, J.-J., and Tirole, J. (1987), "Comparative Statics of the Optimal Dynamic Incentive Contract," *European Economic Review* **31**, 901-926.

Laffont, J.-J., and Tirole, J. (1988), "The Dynamics of Incentive Contracts," *Econometrica* **56**, 1153-1175.

Laffont, J.-J., and Tirole, J. (1990), "Adverse Selection and Renegotiation in Procurement," *Review of Economic Studies* **57**, 597-626.

Laffont, J.-J., and Tirole, J. (1993), *A Theory of Incentives in Procurement and Regulation*. Cambridge (MA): MIT Press.

Maskin, E., and Riley, J. (1984), "Monopoly with Incomplete Information," *Rand Journal of Economics* **15**, 171-96.

Mirrlees, J. (1971), "An Exploration in the Theory of Optimum Income Taxation," *Review of Economic Studies* **38**, 175-208.

Chapter 4

Moral Hazard

4.1 Static Moral Hazard

In our discussion of adverse selection, we followed standard conventions and interpreted an outcome y as a contract. It is therefore quite common to use the term *precontractual asymmetric information* (or simply asymmetric information) instead of adverse selection in order to emphasize the existence of incomplete information prior to the date when the contract is signed.

In problems of moral hazard, there is no asymmetric information between principal and agent at the precontractual stage. But there is *postcontractual asymmetric information* in the sense that after the contract is signed, the agent takes an unobservable action $a \in \Lambda$. Think of a as the parameter of a probability distribution with random outcome $x \in X$. In our earlier notation, a pair (a, x) corresponds to the state θ. What makes the moral hazard problem intricate is an underlying *risk-sharing problem*. Ideally, the principal would like to implement a social choice rule $f : \Lambda \times X \to R$ which makes the agent's pay dependent on both a (in order to compensate him for his action) and x (for risk-sharing purposes). However, since a is unobservable, the social choice rule cannot be implemented directly. Note that here, an outcome $y \in A$ is not a contract like in the previous chapter, but a monetary payment represented by a point on the real line.

Due to the unobservability of a, the principal must settle for an *incentive contract* or *sharing rule* $s : X \to R$ which is based only on observable (and thus verifiable) data. As before, we require that the sharing rule be both incentive compatible and individually rational. In the case of moral hazard, incentive compatibility means that given the sharing rule $s(x)$, the agent prefers a particular action a to any other action $\hat{a} \in \Lambda$. Hence, by using an incentive compatible sharing rule, the principal knows (!) which action

the agent will subsequently take (in equilibrium). This makes it unnecessary to construct a more complicated mechanism that, once the action is taken, also asks the agent in addition to truthfully reveal the selected action to the principal. The principal's problem is then to determine the action and sharing rule that maximize her expected utility subject to IC and IR.

Prominent examples of moral hazard are employment relationships and insurance markets. Since input (e.g. effort) is hard to measure, an employee's pay is typically based on output or performance (note that a fixed wage is merely a special kind of performance pay). Well-known examples are piece rates or bonus schemes. Likewise, insurance companies typically cannot monitor precautionary measures taken by individuals in order to reduce the likelihood of an accident. As a consequence, insurance payments are based on observable variables such as the occurence and/or size of an accident.

The Model

The first assumptions describe the chronological order in which events occur.

1. The principal offers the agent a sharing rule $s(x)$ which he can either accept or reject.

2. In case he accepts, the agent takes an unobservable action $a \in \Lambda$ which affects the distribution function $\Pi(x, a)$ of a (random) monetary outcome with support $X \subseteq R$. The agent incurs a cost $c(a)$.

3. An outcome $x \in X$ is realized and verifiable vis-a-vis the courts. The agent receives $s(x)$ and the principal keeps the remainder $x - s(x)$.

Before we proceed with some some technical assumptions, let us briefly comment on assumptions 1-3. In assumption 1, the principal can make a take-it-or-leave-it offer. This has no implication for the relative bargaining power between the two parties. The agent will accept the principal's offer only if he expects to receive at least his reservation utility (in equilibrium). By raising the agent's reservation utility, we can grant him more bargaining power and thus trace out the entire (constrained) Pareto frontier. In addition, we implicitly assumed that the parties can commit to $s(x)$ for the full duration of the contract. Later in this chapter, we will briefly address the question of what happens if renegotiation is possible. The above formulation is known as *parameterized distribution approach* and was first used by Mirrlees (1974, 1975, 1975) and Holmström (1979). The name is derived from the fact that the agent's action enters as a parameter in the distribution function $\Pi(x, a)$. The very first moral hazard models used a different representation known as *state-space approach*.

4. The support X is invariant with respect to a.

5. The agent has additively separable preferences of the form $U(s(x), a) = u(s(x)) - c(a)$, where $u'(\cdot) > 0$ and $u''(\cdot) \leq 0$, and where $c'(\cdot) > 0$ and $c''(\cdot) \geq 0$.

6. The principal has preferences of the form $V(x - s(x))$, where $V'(\cdot) > 0$ and $V''(\cdot) \leq 0$.

Throughout the chapter, we use primes to denote derivatives and subscripts to denote partial derivatives. By assumption 4, any x that occurs with positive probability under action a must also occur with positive probability under action $\hat{a}$ for all $a, \hat{a} \in \Lambda$. We will show later that if this assumption is not satisfied, trivial solutions to the moral hazard problem may be possible. The assumption of additive separability implies that the agent's preferences over (random) income are independent of his action. Notice that we have been deliberately vague regarding the sets Λ and X. In particular, we have not provided sufficient conditions that ensure the existence of a solution to the principal's problem. In fact, later we will show that for a rather broad class of problems, a solution does not exist. Finally, let us introduce two assumptions that characterize the effect of the agent's action on the distribution function $\Pi(x, a)$.

Definition 1 (Monotone Likelihood Ratio Property) The distribution function $\Pi(x, a)$ with density $\pi(x, a)$ satisfies the monotone likelihood ratio property (MLRP) if $\frac{\pi_a(x,a)}{\pi(x,a)}$ is nondecreasing in x for all $a \in \Lambda$.

MLRP is equivalent to the condition that for any two actions $a > \hat{a}$, the ratio $\frac{\pi(x,a)}{\pi(x,\hat{a})}$ is nondecreasing in x. Thus, MLRP implies that higher values of x are more likely to be generated by the distribution $\pi(x, a)$ than by the distribution $\pi(x, \hat{a})$. This condition is satisfied by many commonly used probability distributions such as the normal and uniform distribution. Incidentally, MLRP implies (but is not implied by) a weaker condition known as *first-order stochastic dominance* (*FOSD*) which states that $\Pi_a(x, a) < 0$ for all x in the interior of X, i.e. an increase in the agent's action induces a rightwards (and therefore favorable) shift in the distribution function. A far more controversial assumption than MLRP is the *convexity of the distribution function condition* which requires that $\Pi_{aa}(x, a) \geq 0$. Unlike MLRP, this condition is satisfied by virtually no standard probability distribution.

Definition 2 (Convexity of the Distribution Function Condition) The distribution function $\Pi(x, a)$ satisfies the convexity of the distribution function condition (CDFC) if $\Pi(x, a)$ is convex in a for all $x \in X$.

The First-Order Approach

As a benchmark, let us begin with the case where the agent's action is observable and verifiable vis-a-vis the court. The principal's problem is then 1) to determine the optimal risk-sharing rule $s(x)$ for any given distribution $\Pi(x,a)$, and 2) given step 1, to determine the "optimal distribution" $\Pi(x,a)$ by weighing up the costs and benefits from the agent's action. Implicitly, we assume that if the agent chooses $\hat{a} \neq a$, he receives a large penalty.

Under complete information, the principal offers the agent a contract which specifies a pair $(s(x), a)$. Under incomplete information, a is not observable and the contract can only specify a sharing rule $s(x)$. However, $s(x)$ will be designed such that in equilibrium, the agent finds it in his best interest to choose the action preferred by the principal. Hence, in both cases the principal (either explicitly or implicitly) offers a pair $(s(x), a)$. Let us call $(s(x), a)$ simply an *allocation*. Since the agent is free to reject the principal's offer, the allocation $(s(x), a)$ must satisfy the agent's *individual rationality constraint* (*IR*).

Definition 3 (Individually Rational Allocation) The allocation $(s(x), a)$ is individually rational (IR) if

$$\int_X u(s(x))\,\pi(x,a)\,dx - c(a) \geq 0. \tag{4.1}$$

Thus, an allocation is individually rational if by choosing a, the agent can guarantee himself an expected utility of at least zero (note that the normalization of the agent's reservation utility to zero is without any loss of generality).

Theorem 1 Suppose that the agent's action a is observable. The allocation $(s^*_{FB}(x), a^*_{FB})$ maximizes the principal's expected utility subject to IR only if

$$\frac{V'(x - s^*_{FB}(x))}{u'(s^*_{FB}(x))} = \lambda \tag{4.2}$$

for all $x \in X$.

Proof (direct) The principal's first-best problem is

$$\max_{a,s(x)} \int_X V(x - s(x))\,\pi(x,a)\,dx \tag{4.3}$$

s.t.

$$\int_X u(s(x))\,\pi(x,a)\,dx - c(a) \geq 0. \tag{4.4}$$

Clearly, IR must be binding at the optimal solution. Denote the Lagrange multiplier by λ. Assuming an interior solution, pointwise maximization of the Lagrangean gives

$$-V'\left(x-s_{FB}^{*}\left(x\right)\right)\pi\left(x,a_{FB}^{*}\right)+\lambda u'\left(s_{FB}^{*}\left(x\right)\right)\pi\left(x,a_{FB}^{*}\right)=0, \tag{4.5}$$

which can be rearranged as

$$\frac{V'\left(x-s_{FB}^{*}\left(x\right)\right)}{u'\left(s_{FB}^{*}\left(x\right)\right)}=\lambda \tag{4.6}$$

for all $x \in X$. ■

Remarks

1. Assuming an interior solution, the first-order condition (4.2) is necessary, but not sufficient for an optimal solution: Firstly, it constitutes only one of two first-order conditions (we have skipped the first-order condition with respect to a as it yields only meager insights). And secondly, our assumptions are not sufficient to ensure that the second-order conditions are satisfied globally.

2. The first-best solution (4.2) states that the ratio of the marginal utilities must be set equal for all values of x. This is an example of Borch's (1962) condition for optimal risk-sharing.

3. If the principal is risk neutral and the agent is risk averse, (4.2) requires that the agent's income is constant for all values of x. Borrowing an expression from the insurance literature, we then say that the agent is fully insured.

Let us now consider the case where the agent's action is unobservable. If the principal wishes to implement a certain action a, she can no longer penalize the agent for choosing $\hat{a} \neq a$, but must provide him with incentives via the sharing rule $s\left(x\right)$. We call an allocation $\left(s\left(x\right),a\right)$ *incentive compatible* if given the sharing rule $s\left(x\right)$, the agent finds it optimal to select a.

Definition 4 (Incentive Compatible Allocation) The allocation $\left(s\left(x\right),a\right)$ is incentive compatible (IC) if

$$\int_{X} u\left(s\left(x\right)\right)\pi\left(x,a\right)dx-c\left(a\right)\geq\int_{X} u\left(s\left(x\right)\right)\pi(x,\hat{a})dx-c(\hat{a}) \tag{4.7}$$

for all $\hat{a} \in \Lambda$.

The principal's problem is to determine the allocation $(s(x), a)$ that maximizes her expected utility subject to IR and IC. If Λ contains infinitely many elements (as is the case when $\Lambda \subseteq R$), (4.7) represents a continuum of constraints and standard optimization techniques are not applicable. Instead of IC, we will therefore work with the agent's first-order condition

$$\int_X u(s(x))\,\pi_a(x,a)\,dx - c'(a) = 0. \tag{4.8}$$

Assuming an interior solution, (4.8) constitutes a necessary condition for optimality in the agent's problem. The substitution of IC with (4.8) is known as *first-order approach* (*FOA*), and (4.3), (4.4) and (4.8) are called the principal's *relaxed problem*.

Let $s^*(x)$ be a solution to the principal's relaxed problem. Clearly, the FOA is valid if at $s^*(x)$, the agent's problem is globally concave. Mirrlees (1979) shows that this is indeed the case if the Lagrange multiplier associated with (4.8) is nonnegative and MLRP and CDFC hold. Unfortunately, the only known proof that the multiplier for (4.8) is nonnegative rests on the assumption (!) that the agent's problem is globally concave. In order to avoid this circularity, Rogerson (1985) introduces a further relaxation of the principal's relaxed problem. Let us replace (4.8) with

$$\int_X u(s(x))\,\pi_a(x,a)\,dx - c'(a) \geq 0 \tag{4.9}$$

and call (4.3), (4.4) and (4.9) the principal's *doubly relaxed problem*. Below we prove that this further relaxation is without consequence, i.e. that at the solution to the doubly relaxed problem, (4.9) is satisfied with equality (or equivalently, that any allocation that solves the doubly relaxed problem also solves the relaxed problem and vice versa). Moreover, since (4.9) is an inequality constraint, the affiliated Lagrange multiplier must be nonnegative. We can then apply Mirrlees' argument and conclude that at the optimal solution, the agent's problem is globally concave.

Theorem 2 Let $(s^*(x), a^*)$ solve the principal's relaxed problem. If MLRP and CDFC are satisfied, the agent's problem is globally concave at $s^*(x)$.

Proof (direct) Consider the doubly relaxed problem

$$\max_{a,s(x)} \int_X V(x - s(x))\,\pi(x,a)\,dx \tag{4.10}$$

s.t.

$$\int_X u(s(x))\,\pi(x,a)\,dx - c(a) \geq 0, \tag{4.11}$$

$$\int_X u\left(s\left(x\right)\right)\pi_a\left(x,a\right)dx - c'\left(a\right) \geq 0. \tag{4.12}$$

Denote by λ and δ the multipliers for (4.11) and (4.12), respectively. The Kuhn-Tucker necessary condition with respect to $s\left(\cdot\right)$ is

$$\frac{V'\left(x-s^*\left(x\right)\right)}{u'\left(s^*\left(x\right)\right)} = \lambda + \delta\frac{\pi_a\left(x,a^*\right)}{\pi\left(x,a^*\right)} \tag{4.13}$$

for all $x \in X$, and the necessary condition with respect to a is

$$\begin{aligned} 0 &= \int_X V\left(x-s^*\left(x\right)\right)\pi_a\left(x,a^*\right)dx \\ &\quad +\lambda\left[\int_X u\left(s^*\left(x\right)\right)\pi_a\left(x,a^*\right)dx - c'\left(a^*\right)\right] \\ &\quad +\delta\left[\int_X u\left(s^*\left(x\right)\right)\pi_{aa}\left(x,a^*\right)dx - c''\left(a^*\right)\right]. \end{aligned} \tag{4.14}$$

We will now show that if $\left(s^*\left(x\right),a^*\right)$ solves the doubly relaxed problem, then it also solves the relaxed problem (4.3), (4.4) and (4.8). In other words, we will show that (4.12) is binding at $\left(s^*\left(x\right),a^*\right)$. Complementary slackness implies that if $\delta > 0$, (4.12) must be binding, whereas if $\delta = 0$, it can be either binding or slack. It therefore remains to be proven that if $\delta = 0$, (4.12) is binding. Given that $\delta = 0$, (4.13) in conjunction with assumptions 5 and 6 implies $\lambda > 0$ and $s^{*\prime}\left(x\right) \in [0,1]$. Set $X = [\underline{x},\overline{x}]$, where $\underline{x}$ and $\overline{x}$ can be equal to $-\infty$ and $+\infty$, respectively. Using integration by parts, the first term on the right-hand side of (4.14) can be expressed as

$$\begin{aligned} &\frac{\partial}{\partial a}\int_X V\left(x-\left(s^*\left(x\right)\right)\right)\pi\left(x,a^*\right)dx \\ =\ &\frac{\partial}{\partial a}\left[V\left(\overline{x}-s^*\left(\overline{x}\right)\right) - \int_X V'\left(x-\left(s^*\left(x\right)\right)\right)\left(1-s^{*\prime}\left(x\right)\right)\Pi\left(x,a^*\right)dx\right] \\ =\ &-\int_X V'\left(x-\left(s^*\left(x\right)\right)\right)\left(1-s^{*\prime}\left(x\right)\right)\Pi_a\left(x,a^*\right)dx \\ \geq\ &0, \end{aligned} \tag{4.15}$$

where the inequality follows from assumption 6, $s^{*\prime}\left(x\right) \in [0,1]$ and MLRP (which implies FOSD, i.e. $\Pi_a\left(x,a\right) < 0$). Since $\lambda > 0$ and $\delta = 0$, it follows from (4.14) that

$$\int_X u\left(s^{*\prime}\left(x\right)\right)\pi_a\left(x,a^*\right)dx - c'\left(a^*\right) \leq 0. \tag{4.16}$$

But this is consistent with (4.12) if and only if (4.12) is binding. Hence, any allocation $(s^*(x), a^*)$ that solves the doubly relaxed problem also solves the relaxed problem.

Finally, we show that at the solution to the (doubly) relaxed problem $(s^*(x), a^*)$, the agent's expected utility is globally concave, which implies that $(s^*(x), a^*)$ also solves the unrelaxed problem (4.3), (4.4) and (4.7). First, let us show that $s^*(x)$ must be nondecreasing (above we proved that $s^{*\prime}(x) \in [0,1]$ only for the case $\delta = 0$): From $\delta \geq 0$ and MLRP, it follows that the right-hand side (and consequently also the left-hand side) of (4.13) is nondecreasing in x. By assumptions 5 and 6, this implies that $s^{*\prime}(x) \geq 0$. Consider now the agent's objective function at $s^*(x)$

$$\int_X u(s^*(x))\,\pi(x,a)\,dx - c(a). \tag{4.17}$$

Integrating by parts (again, we set $X = [\underline{x}, \overline{x}]$ without loss of generality), we have

$$u(s^*(\overline{x})) - \int_X u'(s^*(x))\,s^{*\prime}(x)\,\Pi(x,a)\,dx - c(a). \tag{4.18}$$

The first term in (4.18) is a constant. Furthermore, $u'(\cdot) > 0$ and $c''(a) \geq 0$ (by assumption 5), and $\Pi_{aa}(x,a) \geq 0$ (by CDFC). Since $s^{*\prime}(x) \geq 0$, the agent's problem is globally concave. ■

Remarks

1. For didactic reasons, let us once again repeat the main line of argument used in the proof. First, we established that any allocation $(s^*(x), a^*)$ that solves the doubly relaxed problem also solves the relaxed problem. Setting $\delta \geq 0$ in the principal's first-order condition (4.13) and using MLRP, we then showed that $s^*(x)$ is nondecreasing. Finally, using $s^{*\prime}(x) \geq 0$, CDFC and FOSD (which is implied by MLRP), we showed that at $s^*(x)$, the agent's problem is globally concave, which in turn implies that $(s^*(x), a^*)$ is also a solution to the unrelaxed problem.

2. The circular proof mentioned earlier is based on the relaxed problem (4.3), (4.4) and (4.8). The first-order condition for this problem with respect to $s(\cdot)$ is

$$\frac{V'(x - s^*(x))}{u'(s^*(x))} = \lambda + \mu\frac{\pi_a(x,a^*)}{\pi(x,a^*)}, \tag{4.19}$$

where μ is the Lagrange multiplier associated with (4.8). Unlike in the doubly relaxed problem, the Lagrange multiplier can now also take

negative values since (4.8) is an equality constraint. However, Holmström (1979) shows that $\mu > 0$ under the assumption that the FOA holds (and additionally, that $u''(\cdot) < 0$). Using $\mu > 0$ instead of $\delta \geq 0$, one can now repeat the last step of the above proof and come to the (erroneous) conclusion that the FOA is valid. Rogerson's (1985) ingenious proof avoids this circularity by using the inequality constraint (4.9), which naturally entails a nonnegative multiplier.

3. The drawback of theorem 2 is that CDFC is only rarely satisfied (incidentally, CDFC *is* satisfied if $\pi(x,a) = a\overline{\pi}(x) + (1-a)\underline{\pi}(x)$, where $\overline{\pi}$ dominates $\underline{\pi}$ in the sense of FOSD). An alternative proof of the validity of the FOA that does not rely on CDFC is provided by Jewitt (1988). For instance, Jewitt shows that the FOA holds if the agent has either square root, log, or exponential utility, and if x is drawn from either a Poisson, gamma, or chi-squared distribution.

We proceed with an economic analysis of the second-best solution. The results follow more or less immediately from the proof of theorem 2.

Corollary 1 Suppose MLRP and CDFC hold. The allocation $(s^*(x), a^*)$ maximizes the principal's expected utility subject to IR and IC only if

$$\frac{V'(x - s^*(x))}{u'(s^*(x))} = \lambda + \delta \frac{\pi_a(x, a^*)}{\pi(x, a^*)} \tag{4.20}$$

for all $x \in X$.

Proof (direct) Equation (4.20) is the first-order necessary condition (4.13) for the doubly relaxed problem. By theorem 2, $(s^*(x), a^*)$ solves the doubly relaxed problem if and only if it solves the unrelaxed problem. Hence, (4.20) is also the first-order necessary condition for the unrelaxed problem. ■

The following result was derived in the proof of theorem 2.

Corollary 2 If MLRP and CDFC hold, the optimal sharing rule $s^*(x)$ is nondecreasing.

Comparing (4.20) with (4.2), we see that $(s^*(x), a^*) = (s^*_{FB}(x), a^*_{FB})$ if and only if $\delta = 0$. We now show that if the agent is risk averse, it must be necessarily true that $\delta > 0$, which implies that the second-best solution is strictly inefficient.

Corollary 3 Suppose MLRP and CDFC are satisfied and $u''(\cdot) < 0$. At the solution to the principal's problem, $\delta > 0$.

Proof (by contradiction) Since (4.12) is an inequality constraint, it is true that $\delta \geq 0$. It therefore remains to be shown that $\delta \neq 0$. Suppose that $\delta = 0$. Since $u''(\cdot) < 0$, (4.13) together with assumptions 5 and 6 imply that $\lambda > 0$ and $s^{*\prime}(x) \in [0, 1)$. Repeating the argument in the proof of theorem 2, (4.15) is now satisfied with strict inequality, and the left-hand side of (4.16) is strictly negative, which contradicts (4.12). ■

Remarks

1. If the agent is risk averse, the principal faces a fundamental tradeoff between risk-sharing and incentives. This can be best illustrated if we assume a risk neutral principal. From Borch's rule (4.2), it then follows that the principal should bear all the risk, i.e. $s^{*\prime}(x) = 0$. However, full insurance implies that the agent's action has no effect on his income, which dulls his incentives to take a costly action. Therefore, if the principal wishes to implement an action other than the least costly action, she must deviate from first-best risk-sharing and make the agent's pay dependent on output, i.e. $s^{*\prime}(x) > 0$. Assuming that $\pi_a(x, a)/\pi(x, a) > 0$, this follows immediately from the first-order condition (4.13) and corollary 3.

2. If the agent is risk averse and output is informative about the agent's action (i.e. $\pi_a(x, a)/\pi(x, a) > 0$), $s^*(x)$ must be strictly increasing. At first glance, this looks like a statistical inference problem: The optimal sharing rule pays more for outcomes that signal a higher choice of a (in probabilistic terms) than for outcomes that signal a lower choice of a. Note, however, that this is only seemingly the case. Due to the incentive compatibility constraint, the principal knows exactly which action was taken in equilibrium. Nonetheless, she must commit to an incentive scheme that is based on the informativeness of x.

3. In the proof of corollary 3, it was claimed that $\delta = 0$ and $u''(\cdot) < 0$ imply that $\lambda > 0$ and $s^{*\prime}(x) \in [0, 1)$. From (4.13), it is obvious that $\delta = 0$ implies $\lambda > 0$ due to $V'(\cdot) > 0$. Consequently, the left-hand side (lhs) of (4.13) must be positive and constant for all x. Next, note that $s^*(x)$ cannot be strictly decreasing since this would imply that the lhs of (4.13) is strictly decreasing. Likewise, $s^*(x)$ cannot be increasing at a rate greater than 1 since this would imply that the lhs of (4.13) is strictly increasing. Finally, note that $s^{*\prime}(x) = 1$ is not feasible when the agent is risk averse since then, the lhs of (4.13) would be strictly increasing. This reveals that in the proof of theorem 2, $s^{*\prime}(x) = 1$ was only possible because we permitted that $u'(\cdot)$ is constant.

4. When the agent is risk neutral, corollary 3 no longer holds and the first-best outcome can be attained by "*selling the firm to the agent*". That is, $s^*(x) = x - K$, where K is a lump-sum transfer from the agent to the principal (K is chosen such as to hold the agent down at his reservation utility). Recall that due to risk neutrality, imposing risk on the agent involves no welfare loss. Being the residual claimant, the agent then faces proper incentives and selects the first-best action a^*_{FB}. The fact that $s^*(x)$ implies efficient risk-sharing can be verified by setting $u'(\cdot)$ equal to a constant in the first-order condition for the first-best problem (4.3). To see that the agent chooses a^*_{FB}, note that the first-order condition with respect to a in the principal's first-best problem (4.3)-(4.4) is

$$\int_X \left[V\left(x - s^*_{FB}(x)\right) + \lambda u\left(s^*_{FB}(x)\right)\right] \pi_a\left(x, a^*_{FB}\right) dx = \lambda c'\left(a^*_{FB}\right). \quad (4.21)$$

Integrating by parts and setting $X = [\underline{x}, \overline{x}]$, (4.21) can be written as

$$\begin{aligned} & \left[V\left(x - s^*_{FB}(x)\right) + \lambda u\left(s^*_{FB}(x)\right)\right] \Pi_a\left(x, a^*_{FB}\right)\Big|^{\overline{x}}_{\underline{x}} \quad (4.22)\\ & - \int_X V'\left(x - s^*_{FB}(x)\right)\left(1 - s^{*\prime}_{FB}(x)\right) \Pi_a\left(x, a^*_{FB}\right) dx \\ & - \lambda \int_X u'\left(s^*_{FB}(x)\right) s^{*\prime}_{FB}(x)\, \Pi_a\left(x, a^*_{FB}\right) dx \\ = \;& \lambda c'\left(a^*_{FB}\right). \end{aligned}$$

Dividing through by λ and inserting (4.2), (4.22) reduces to

$$-\int_X u'\left(s^*_{FB}(x)\right) \Pi_a\left(x, a^*_{FB}\right) dx = c'\left(a^*_{FB}\right). \quad (4.23)$$

Finally, setting $u'(\cdot) = 1$ and integrating by parts once more, we get

$$\int_X x f_a\left(x, a^*_{FB}\right) dx = c'\left(a^*_{FB}\right), \quad (4.24)$$

which implies that a^*_{FB} maximizes $\int_X x f(x,a)\, dx - c(a)$. But this is equivalent to a risk neutral agent's maximization problem who faces an incentive scheme $s(x) = x - K$.

5. A second case where the first-best can be attained is when assumption 4 is violated. In environments with a *shifting support*, there exist outcomes that indicate with probability 1 that the agent has taken an action that differs from the one preferred by the principal. A *forcing contract* that punishes the agent very hard whenever such an outcome is observed can then restore the first-best. See Harris and Raviv (1979) for a formal analysis.

Renegotiation

Consider the case of a risk-neutral principal and a risk-averse agent. The previous analysis has shown that in order to implement an action other than the least costly action, the principal must deviate from the first-best risk-sharing solution (here: full insurance) and make the agent's pay dependent on output. In this regard, we have implicitly assumed that both parties can commit to the optimal sharing rule for the full duration of the contract. Suppose now that after the action was taken, principal and agent can meet and renegotiate the initial contract. Since the action choice is irreversible, there is no reason why the agent should be still exposed to risk. The principal can then realize efficiency gains by offering the agent a new contract that provides him with full insurance (note that there is no adverse selection problem since the principal knows which action was taken in equilibrium). The problem with renegotiation is that while it is Pareto-improving ex post, it is detrimental ex ante: Foreseeing that the outcome from renegotiation will be full insurance, the agent supplies only the least costly action.

We know from our discussion of repeated adverse selection that if renegotiation is possible, we can without loss of generality restrict attention to renegotiation-proof contracts. The above reasoning suggests that in this case, no action other than the least costly action can be implemented.

Theorem 3 Suppose $V''(\cdot) = 0$ and $u''(\cdot) < 0$. The allocation $(s(x), a)$ is renegotiation-proof and incentive compatible only if a is the least costly action.

Proof (direct) If $(s(x), a)$ is incentive compatible, the principal "knows" a at the interim stage. Given that information is symmetric, renegotiation-proofness requires that $s(x)$ is the first-best sharing rule. By theorem 1, this implies that $s(x)$ must be a constant (say, $s(x) = K$). The agent's problem is then $\max_a \int u(K) f(x,a)\,dx - c(a)$ which is equivalent to $\min_a c(a)$. ■

Remarks

1. It is obvious that the reverse of theorem 3 also holds.

2. Theorem 3 no longer holds if we enlarge the agent's strategy set to include mixed strategies. Suppose the agent randomizes with probability distribution $\phi(a)$. At the interim stage (that is, after the action was taken but before output is realized), the principal no longer "knows" a and the situation corresponds to an adverse selection setting in which a represents the agent's type. Consequently, renegotiation will not lead to full insurance.

3. If mixed strategies are allowed, the optimal renegotiation-proof contract can be derived using backwards induction. Consider first the interim stage where the distribution $\phi(a)$ is given. Renegotiation-proofness requires that at the interim stage, the initial contract must be (constrained) efficient. It then follows from chapter 3 that the initial contract must specify an optimal menu of sharing rules $s_a(x)$, where each sharing rule is "designed" for a particular type a. Since low types (i.e. types with a higher risk of low outcomes) value insurance more than high types, the single-crossing property is satisfied, which implies that under the optimal menu, low types will receive full insurance while high types will be exposed to some risk. In a second step, we consider the ex ante stage and determine the distribution $\phi(a)$ and the interim menu $s_a(x)$ that maximize the principal's expected utility subject to the constraints that

   (a) the interim menu $s_a(x)$ is renegotiation-proof given the distribution $\phi(a)$ (this imposes optimality at the interim stage),

   (b) given $s_a(x)$, the agent finds it indeed optimal to randomize according to $\phi(a)$ (this implies that ex ante, the agent must be indifferent between any action in the support of $\phi(a)$), and

   (c) the agent receives at least his reservation utility in equilibrium.

   Due to the added renegotiation-proofness constraint, welfare under renegotiation can never be greater than under commitment. For a full characterization of the mixed strategy equilibrium, see Fudenberg and Tirole (1990).

4. Ma (1994) analyzes a signalling version of the renegotiation game in which the informed party (i.e. the agent) makes a take-it-or-leave-it offer. He shows that in this case, renegotiation does not lead to a welfare loss, i.e. the principal can implement the same actions at the same cost as under full commitment.

5. When the agent's action is observable but not verifiable (so that a contract contingent on a is not possible), renegotiation can even be welfare-enhancing. Consider the following construction due to Hermalin and Katz (1991): At the ex ante stage, the principal "sells the firm to the agent", i.e. $s(x) = x - K$. As we have shown earlier, this induces the agent to exert the first-best action. At the interim stage, the principal offers the agent to "buy back" the lottery $s(x)$ at a value equal to its certainty equivalent $\omega(a)$, where $u(\omega(a)) \equiv \int_X u(s(x))\,\pi(x,a)\,dx$. It follows that

(a) the agent will accept the new (full insurance) contract as it gives him the same expected utility as $s(x)$,

(b) the agent's incentives to exert the first-best action are not distorted by renegotiation since for any action a, he receives the same expected utility under the old and new contract, which implies that

(c) the first-best allocation is implementable.

Near First-Best Efficiency with Step Functions (technical)

An important insight from our discussion of the first-order approach is that the optimal sharing rule is likely to look very complicated. The reason for this is that $s^*(x)$ is a function of x only indirectly via the likelihood ratio $\pi_a(x,a)/\pi(x,a)$. Thus, unless the likelihood ratio depends on x in a simplistic fashion, there is little hope of getting further-reaching results than monotonicity. And yet, real-world incentive schemes appear to be simple. For instance, salespeople are often rewarded according to bonus schemes, and assembly line workers are typically paid by the piece. One way to deal with this puzzle is to turn to richer and more structured environments. This is done in section 4.3 where we discuss dynamic moral hazard. Alternatively, one could simply try out various functional forms of $s(x)$ and compare their performance relative to some benchmark (the first-best). In general, this procedure is pointless (given that there is an infinite number of possible sharing rules) unless one finds a sharing rule that actually *is* (at least nearly) first-best efficient. Surprisingly, such a sharing rule exists. As was shown by Mirrlees (1974) in the context of a concrete example, *step functions* can under certain conditions approach the first-best asymptotically.

Definition 5 (Step Function) A step function is a sharing rule $s(x)$ where

$$s(x) = \begin{cases} \overline{s} \text{ if } x \geq \hat{x} \\ \underline{s} \text{ otherwise.} \end{cases} \tag{4.25}$$

Thus, a step function is defined by three parameters: a high payment $\overline{s}$, a low payment $\underline{s}$, and a cutoff value $\hat{x}$. Consider now the following assumptions which complement our earlier assumptions 1-6.

1. The principal is risk neutral, i.e. $V(x - s(x)) = x - s(x)$.

2. The agent's utility for wealth is defined on the set $W = (\underline{w}, \infty)$, where $u''(w) < 0$ and $\lim_{w \to \underline{w}} u(w) = -\infty$.

3. The distribution function satisfies FOSD, i.e. $\Pi_a(x,a) < 0$ for all x.

4. There exists a value $\tilde{x}$ such that for all $x \leq \tilde{x}$, $\Pi(x,a)$ is concave in a.

5. The likelihood ratio satisfies $\lim_{x \to \underline{x}} \frac{\pi_a(x,a)}{\pi(x,a)} = -\infty$.

Let us briefly comment on the assumptions. Assumptions 1 and 2 imply that the first-best sharing rule is a constant. Unboundedness of $u(\cdot)$ is crucial and signifies that the agent's utility becomes infinitely negative as we approach the infimum of the domain. Moreover, assumption 2 restricts the set of feasible payments to W. Assumption 4 is rather innocuous and is satisfied, for example, by the normal distribution. Finally, assumption 5 implies that very low values of x are extremely informative with respect to the agent's action. We will discuss the role of this assumption in detail later in the text. Note that we have not assumed that MLRP or CDFC hold.

We will now show that by using a step function, the principal can implement the first-best action at a cost that is arbitrarily close to the first-best cost. Put it differently, we will show that the principal's expected utility from implementing a^*_{FB} converges to the value that obtains under complete information. Recall that in the case of a risk neutral principal and a risk averse agent, the first-best sharing rule is $s^*_{FB}(x) = K$, where the constant K is implicitly defined by the agent's binding IR constraint $u(K) = c(a^*_{FB})$. Since $u(w)$ is monotonic, it has a monotonic inverse $u^{-1}(u)$ such that $u^{-1}(u) = w \Leftrightarrow u(w) = u$. Thus, $u^{-1}(u)$ denotes the level of wealth that yields utility u. Using this notation, we can express the principal's expected utility from the full information optimum as

$$\int_X x\pi(x, a^*_{FB})\, dx - u^{-1}(c(a^*_{FB})), \tag{4.26}$$

which from now on shall be our benchmark.

Back to the incomplete information case, implementation of a^*_{FB} through a step function implies that a^*_{FB} must satisfy the agent's IC constraint

$$\begin{aligned} & u(\underline{s})\,\Pi(\hat{x}, a^*_{FB}) + u(\overline{s})\,(1 - \Pi(\hat{x}, a^*_{FB})) - c(a^*_{FB}) \qquad (4.27) \\ \geq\ & u(\underline{s})\,\Pi(\hat{x}, \hat{a}) + u(\overline{s})\,(1 - \Pi(\hat{x}, \hat{a})) - c(\hat{a}) \end{aligned}$$

for all $\hat{a} \in \Lambda$, where $\underline{s}$, $\overline{s}$ and $\hat{x}$ are given parameters. Instead of working with (4.27), we will work with the agent's first-order condition at a^*_{FB}

$$(u(\underline{s}) - u(\overline{s}))\,\Pi_a(\hat{x}, a^*_{FB}) - c'(a^*_{FB}) = 0. \tag{4.28}$$

From our discussion of the FOA, we know that the substitution of (4.27) with (4.28) is valid if under the step function $(\underline{s}, \overline{s}, \hat{x})$, the agent's objective

function is concave in a. We now proceed as follows: First, we determine $\underline{s}$ and $\overline{s}$ as a function of $\overline{x}$ by solving explicitly the agent's first-order condition and IR (note that at the optimum, IR must be binding since transfers are costly to the principal). Given $\underline{s}(\hat{x})$ and $\overline{s}(\hat{x})$, it follows trivially that for any $\hat{x} \leq \tilde{x}$, the agent's problem is globally concave. We then insert $\underline{s}(\hat{x})$ and $\overline{s}(\hat{x})$ in the principal's objective function and obtain a maximization problem with respect to $\hat{x}$, subject to the constraint that $\hat{x} \leq \tilde{x}$.

Theorem 4 For any step function $(\underline{s}(\hat{x}), \overline{s}(\hat{x}), \hat{x})$ where $\underline{s}(\hat{x})$ and $\overline{s}(\hat{x})$ satisfy the agent's first-order condition and IR with equality at some $\tilde{a} \in \Lambda$ and where $\hat{x}$ satisfies $\hat{x} \leq \tilde{x}$, the agent's problem is globally concave.

Proof (direct) The agent's first-order condition and indvidual rationality constraint are

$$(u(\underline{s}) - u(\overline{s}))\,\Pi_a(\hat{x}, \tilde{a}) - c'(\tilde{a}) = 0 \tag{4.29}$$

and

$$u(\underline{s})\,\Pi(\hat{x}, \tilde{a}) + u(\overline{s})\,(1 - \Pi(\hat{x}, \tilde{a})) - c(\tilde{a}) = 0, \tag{4.30}$$

respectively. Solving (4.29)-(4.30) for $\underline{s}$ and $\overline{s}$ as a function of $\hat{x}$, we have

$$\underline{s}(\hat{x}) = u^{-1}\left(c(\tilde{a}) + \frac{c'(\tilde{a})\,(1 - \Pi(\hat{x}, \tilde{a}))}{\Pi_a(\hat{x}, \tilde{a})}\right) \tag{4.31}$$

and

$$\overline{s}(\hat{x}) = u^{-1}\left(c(\tilde{a}) - \frac{c'(\tilde{a})\,\Pi(\hat{x}, \tilde{a})}{\Pi_a(\hat{x}, \tilde{a})}\right). \tag{4.32}$$

From FOSD and the fact that $u^{-1}(\cdot)$ is strictly increasing, it follows that $\overline{s}(\hat{x}) > \underline{s}(\hat{x})$ for all $\hat{x}$ in the interior of X.

Given a step function with cutoff value $\hat{x} \leq \tilde{x}$ and payments $\underline{s}(\hat{x})$ and $\overline{s}(\hat{x})$ as defined above, the agent's problem is

$$\max_a u(\underline{s}(\hat{x}))\,\Pi(\hat{x}, a) + u(\overline{s}(\hat{x}))\,(1 - \Pi(\hat{x}, a)) - c(a), \tag{4.33}$$

which is concave since $\overline{s}(\hat{x}) > \underline{s}(\hat{x})$ and $\Pi_{aa}(\hat{x}, a) \geq 0$ by assumption 4. ■

Next, we let the principal choose $\hat{x}$ in order to maximize her expected utility over the set of step functions that implement a^*_{FB} and satisfy IR with equality at a^*_{FB}. By theorem 4, this is equivalent to letting the principal maximize over the set of step functions that satisfy (4.31)-(4.32) (each with $\tilde{a} = a^*_{FB}$) and $\hat{x} \leq \tilde{x}$. As it turns out, this maximization problem has no solution. By choosing smaller and smaller values of $\hat{x}$, the principal can approach the first-best utility (4.26) arbitrarily closely.

The near first-best result is driven by two assumptions: 2 and 5. Assumption 5 implies that for any action $\hat{a} < a^*_{FB}$, the relative likelihood that an outcome x was generated by the "inferior" distribution $\Pi(x, \hat{a})$ instead of $\Pi(x, a^*_{FB})$) tends to infinity as $x \to \underline{x}$. Heuristically, one could therefore argue that by choosing a cutoff value close to $\underline{x}$, the principal minimizes the risk of erroneously punishing an agent who picked a^*_{FB} and thereby avoids paying a costly risk premium in equilibrium. Unfortunately, the mechanism that underlies the result is more intricate. It can be best understood if we concentrate on the case $\hat{a} < a^*_{FB}$ and rewrite IC more conveniently as

$$\begin{aligned} c(a^*_{FB}) - c(\hat{a}) \quad \leq \quad & u(\underline{s})\left(\Pi(\hat{x}, a^*_{FB}) - \Pi(\hat{x}, \hat{a})\right) \\ & + u(\overline{s})\left(\Pi(\hat{x}, \hat{a}) - \Pi(\hat{x}, a^*_{FB})\right). \end{aligned} \tag{4.34}$$

For any value of $\hat{x}$ close to $\underline{x}$, assumption 5 implies that the difference $(\Pi(\hat{x}, a^*_{FB}) - \Pi(\hat{x}, \hat{a}))$ is negative. Hence, by choosing $\underline{s}$ sufficiently small, IC can always be satisfied. From (4.34) it is also clear that $\underline{s}$ must be decreasing as $\hat{x} \to \underline{x}$ because $\Pi(\hat{x}, a^*_{FB})$ and $\Pi(\hat{x}, \hat{a})$ both go to zero and, as is shown in the following proof, $\overline{s}$ is decreasing. However, as $\underline{s}$ approaches $\underline{w}$, a given decrease in $\underline{s}$ has an increasingly negative effect on $u(\underline{s})$ since by assumption 2, the slope of $u(\cdot)$ becomes negative infinite. As a consequence, $\underline{s}$ need not decrease "too fast". It turns out that $\underline{s}$ decreases sufficiently slowly so that the product $\underline{s}(\hat{x})\Pi(\hat{x}, a^*_{FB})$ in the principal's objective function $\int_X x\pi(x, a^*_{FB})\,dx - \underline{s}(\hat{x})\Pi(\hat{x}, a^*_{FB}) - \overline{s}(\hat{x})(1 - \Pi(\hat{x}, a^*_{FB}))$ always converges to zero, even if $\underline{s}(\hat{x})$ tends to $-\infty$ (which is the case if $W = \mathrm{R}$). This, together with the comparably easy to prove fact that $\overline{s}(\hat{x})(1 - \Pi(\hat{x}, a^*_{FB}))$ converges to $u^{-1}(c(a^*_{FB}))$ establishes the near first-best result.

Theorem 5 The principal's problem has no solution. By letting $\hat{x} \to \underline{x}$, she can approach the first-best utility (4.26) arbitrarily closely.

Proof (direct) By theorem 4, the principal's problem is equivalent to

$$\max_{\hat{x}} \int_X x\pi(x, a^*_{FB})\,dx - \underline{s}(\hat{x})\Pi(\hat{x}, a^*_{FB}) - \overline{s}(\hat{x})(1 - \Pi(\hat{x}, a^*_{FB})) \tag{4.35}$$

s.t.

$$\underline{s}(\hat{x}) = u^{-1}\left(c(a^*_{FB}) + \frac{c'(a^*_{FB})(1 - \Pi(\hat{x}, a^*_{FB}))}{\Pi_a(\hat{x}, a^*_{FB})}\right), \tag{4.36}$$

$$\overline{s}(\hat{x}) = u^{-1}\left(c(a^*_{FB}) - \frac{c'(a^*_{FB})\Pi(\hat{x}, a^*_{FB})}{\Pi_a(\hat{x}, a^*_{FB})}\right), \tag{4.37}$$

and

$$\hat{x} \leq \tilde{x}. \tag{4.38}$$

Let us ignore (4.38) for a moment. Inserting (4.36)-(4.37) in the objective function (4.35), the principal's problem can be written as

$$\begin{aligned}\max_{\hat{x}} \int_X x\pi\left(x, a_{FB}^*\right) dx \qquad (4.39)\\ -u^{-1}\left(c\left(a_{FB}^*\right) + \frac{c'\left(a_{FB}^*\right)\left(1-\Pi\left(\hat{x}, a_{FB}^*\right)\right)}{\Pi_a\left(\hat{x}, a_{FB}^*\right)}\right)\Pi\left(\hat{x}, a_{FB}^*\right)\\ -u^{-1}\left(c\left(a_{FB}^*\right) - \frac{c'\left(a_{FB}^*\right)\Pi\left(\hat{x}, a_{FB}^*\right)}{\Pi_a\left(\hat{x}, a_{FB}^*\right)}\right)\left(1-\Pi\left(\hat{x}, a_{FB}^*\right)\right).\end{aligned}$$

We now show that as $\hat{x}$ approches $\underline{x}$, (4.39) converges to (4.26). Let us begin with the second term in (4.39). We can distinguish between two cases: i) $u^{-1}(\cdot)$ is bounded below, and ii) $u^{-1}(\cdot)$ is unbounded below. Case i) is trivial and implies that the second term converges to zero as $\hat{x} \to \underline{x}$ because the argument in $u^{-1}(\cdot)$ goes to $-\infty$. Case ii) is harder since the limit of the product $u^{-1}(...)\Pi\left(\hat{x}, a_{FB}^*\right)$ is " $(-\infty)\,0$" and thus undefined. A standard trick is to consider instead the product of $\Pi\left(\hat{x}, a_{FB}^*\right)$ with the tangent of $u^{-1}(\cdot)$ at some given value $\hat{x} = k$. Since $u^{-1}(\cdot)$ is increasing and strictly convex, the tangent yields a smaller value than $u^{-1}(\cdot)$ for any $\hat{x} \neq k$. Hence, if the product of $\Pi\left(\hat{x}, a_{FB}^*\right)$ with the tangent of $u^{-1}(\cdot)$ converges to zero, the product $u^{-1}(...)\Pi\left(\hat{x}, a_{FB}^*\right)$ must also converge to zero. We now prove that this is indeed the case. The tangent of $u^{-1}(\cdot)$ at k is

$$u^{-1}(k) + u^{-1\prime}(k)\, c'\left(a_{FB}^*\right)\left(\frac{1-\Pi\left(\hat{x}, a_{FB}^*\right)}{\Pi_a\left(\hat{x}, a_{FB}^*\right)} - \frac{1-\Pi\left(k, a_{FB}^*\right)}{\Pi_a\left(k, a_{FB}^*\right)}\right), \qquad (4.40)$$

and the limit of the product of $\Pi\left(\hat{x}, a_{FB}^*\right)$ with (4.40) as $\hat{x} \to \underline{x}$ is

$$u^{-1\prime}(k)\, c'\left(a_{FB}^*\right) \lim_{\hat{x}\to\underline{x}} \frac{\Pi\left(\hat{x}, a_{FB}^*\right)}{\Pi_a\left(\hat{x}, a_{FB}^*\right)}. \qquad (4.41)$$

By l'Hôpital's rule and assumption 5, (4.41) reduces to

$$u^{-1\prime}(k)\, c'\left(a_{FB}^*\right) \lim_{\hat{x}\to\underline{x}} \frac{\pi\left(\hat{x}, a_{FB}^*\right)}{\pi_a\left(\hat{x}, a_{FB}^*\right)} = 0. \qquad (4.42)$$

Consider now the third expression in (4.39). By l'Hôpital's rule and assumption 5, the limit as $\hat{x} \to \underline{x}$ is

$$\lim_{\hat{x}\to\underline{x}} u^{-1}(...)\left(1-\Pi\left(\hat{x}, a_{FB}^*\right)\right) = u^{-1}\left(c\left(a_{FB}^*\right)\right). \qquad (4.43)$$

Combining (4.42) and (4.43) implies that as $\hat{x} \to \underline{x}$, the principal's expected utility (4.39) converges to

$$\int_X x\pi\left(x, a_{FB}^*\right) dx - u^{-1}\left(c\left(a_{FB}^*\right)\right), \qquad (4.44)$$

which is equal to the benchmark solution (4.26). Finally, note that since the principal's expected utility is strictly increasing as $\hat{x} \to \underline{x}$, the ignored constraint $\hat{x} \leq \tilde{x}$ is not binding. ■

Remarks

1. Let us briefly restate the intuition which underlies the near first-best result. As $\hat{x} \to \underline{x}$, the penalty $\underline{s}$ tends to $\underline{w}$ and the payment $\overline{s}$ tends to $u^{-1}(c(a^*_{FB}))$. In the limit, the agent is punished with probability zero and receives the first-best payment $u^{-1}(c(a^*_{FB}))$ with probability one, i.e. he is completely insured. The hard part of the proof was to establish that the product $\underline{s}\Pi(\hat{x}, a^*_{FB})$ converges to zero.

2. The nonexistence result follows from the fact that the limit $\hat{x} = \underline{x}$ is not feasible since $u(\underline{s}(\underline{x})) = u(u^{-1}(-\infty)) = u(\underline{w})$ is not defined by assumption 2.

3. In the limit, the principal knows with certainty that the agent has shirked. Hence, the limit case is conceptually equivalent to the shifting support environment mentioned earlier where a large penalty in a probability 1-event also restores the first-best.

4.2 Extensions: Multiple Signals, Multiple Agents, and Multiple Tasks

Multiple Signals

Suppose that in addition to x, the principal can observe a further verifiable signal z. In principle, the sharing rule $s(\cdot)$ can then be made a function of both x and z. The question is therefore under what conditions is it optimal to include the additional signal z in the sharing rule. It turns out that the concept of a *sufficient statistic* plays a central role in answering this question.

Definition 6 (Sufficient Statistic) The variable x is a sufficient statistic for $\{x, z\}$ with respect to a if there exist functions $G(\cdot) \geq 0$ and $H(\cdot) \geq 0$ such that

$$\pi(x, z, a) = G(x, z) H(x, a) \tag{4.45}$$

for all $(x, z, a) \in X \times Z \times \Lambda$.

Thus, if x is a sufficient statistic for $\{x, z\}$ with respect to a, the joint probability distribution $\pi(x, z, a)$ can be separated into two functions $G(x, z)$ and

$H(x,a)$, where $G(x,z)$ adds pure noise and where only the second function $H(x,a)$ depends on a. In other words, x carries all the relevant information about a and the additional signal z is completely uninformative. In statistical decision theory, $G(x,z)$ is called a *garbling* or *Markov matrix*.

With a few exceptions, adding noise to the agent's pay is never beneficial, and it is therefore quite plausible that if x is a sufficient statistic for $\{x,z\}$ with respect to a, the noise term z should not be included in the optimal sharing rule. We will now prove this intuition formally. The proof rests on an extension of the first-order approach to the multi-signal case. In section 4.1, we have shown that in the case of a single signal x, MLRP and CDFC are sufficient to render the principal's first-order condition (4.20) necessary for optimality. In the multi-signal case, the FOA is valid if the joint distribution satisfies MLRP and the (marginal) distribution functions of one (!) signal satisfy FOSD and CDFC (Sinclair-Desgagné (1994)). In what follows, we simply assume that the FOA is valid.

Theorem 6 Suppose that the FOA is valid. The optimal sharing rule $s^*(x,z)$ will not depend on z if and only if x is a sufficient statistic for $\{x,z\}$ with respect to a.

Proof (direct) A straightforward extension of the FOA to many signals yields the first-order condition

$$\frac{V'(x-s^*(x,z))}{u'(s^*(x,z))} = \lambda + \delta\frac{\pi_a(x,z,a^*)}{\pi(x,z,a^*)} \tag{4.46}$$

for all $(x,z) \in X \times Z$. By assumption, the FOA is valid and the optimal allocation $(s^*(x,z), a^*)$ is characterized by (4.46). Equation (4.46) then implies that $s^*(x,z)$ is not a function of z if and only if the likelihood ratio $\pi_a(x,z,a^*)/\pi(x,z,a^*)$ is independent of z, i.e. if and only if

$$\frac{\pi_a(x,z,a^*)}{\pi(x,z,a^*)} = h(x,a^*). \tag{4.47}$$

In what follows, we need to restrict attention to distributions where (4.47) is either satisfied for all a or no a (cf. Holmström (1979), fn. 21). Given this restriction, we can solve the differential equation (4.47) and obtain

$$\pi(x,z,a) = G(x,z)\,H(x,a), \tag{4.48}$$

where $H(x,a) = \exp\left\{\int_\Lambda h(x,a)\,da\right\}$. Conversely, (4.48) also implies (4.47) (take logarithms and differentiate with respect to a). Thus, $s^*(x,z)$ is not a function of z iff x is a sufficient statistic for $\{x,z\}$ with respect to a. ■

Remarks

1. An important implication of theorem 6 is that randomization (i.e. adding noise to the optimal sharing rule) does not pay. This can be neatly illustrated if we assume that the signals z and x are independently distributed and that z is pure noise. The joint density $\pi(x,z,a)$ can then be written as the product of the two densities $\pi_1(z)$ and $\pi_2(x,a)$. Suppose that contrary to our assertion, the optimal sharing rule $s^*(x,z)$ is a function of both x and z. Consider now an alternative sharing rule $s(x)$ that depends only on x and that is defined by

$$u(s(x)) = \int_Z u(s^*(x,z))\,\pi_1(z)\,dz \tag{4.49}$$

holds for all x. Thus, for each x, the certain payment $s(x)$ gives the agent the same utility as the lottery $\pi_1(z)$ with outcomes $s^*(x,z)$. Note that $u(s(x)) = \int_Z u(s(x))\,\pi_1(z)\,dz$ since $u(s(x))$ is a constant with respect to z. Integrating (4.49) with respect to x gives

$$\begin{aligned} &\int_X \int_Z u(s(x))\,\pi_1(z)\,\pi_2(x,a)\,dzdx \\ = &\int_X \int_Z u(s^*(x,z))\,\pi_1(z)\,\pi_2(x,a)\,dzdx, \end{aligned} \tag{4.50}$$

i.e. for any value of a, the agent receives the same expected utility under $s(x)$ and $s^*(x,z)$, which implies that $s(x)$ satisfies IR and implements the same actions as $s^*(x,z)$. Suppose the agent is risk averse. By Jensen's inequality and $u''(\cdot) < 0$, (4.49) implies

$$(x - s(x)) > \int_Z (x - s^*(x,z))\,\pi_1(z)\,dz \tag{4.51}$$

for all x, and therefore

$$\begin{aligned} V(x - s(x)) &> V\left(\int_Z (x - s^*(x,z))\,\pi_1(z)\,dz\right) \\ &\geq \int_Z V(x - s^*(x,z))\,\pi_1(z)\,dz, \end{aligned} \tag{4.52}$$

where the second inequality follows from Jensen's inequality and concavity of $V(\cdot)$. Finally, using $V(x - s(x)) = \int_Z V(x - s(x))\,\pi_1(z)\,dz$ and integrating with respect to x, we obtain

$$\begin{aligned} &\int_X \int_Z V(x - s(x))\,\pi_1(z)\,\pi_2(x,a)\,dzdx \\ > &\int_X \int_Z V(x - s^*(x,z))\,\pi_1(z)\,\pi_2(x,a)\,dzdx, \end{aligned} \tag{4.53}$$

i.e. the principal is strictly better off under $s(x)$, which contradicts optimality of $s^*(x,z)$. As a general rule, randomization is never optimal when the agent is risk averse and has either additively or multiplicatively separable utility (Gjesdal (1982), Grossman and Hart (1983)).

2. A second implication of theorem 6 is that the optimal sharing rule should condition on *any* signal that contains information about a, regardless of how noisy it is.

3. In the face of these results, one is once again tempted to believe that the principal faces a statistical inference problem. Remember that this is only seemingly the case since in equilibrium, she knows exactly which action was taken. Hence, no information is extracted from $\{x,z\}$.

Multiple Agents I: Relative Performance Evaluation

The analysis of the multi-agent problem depends crucially on whether the principal can observe the agent's individual contributions or merely joint output. The former case is examined in this section and constitutes a straightforward extension of the multi-signal model. The latter case has a completely different flavor and yields important implications for the separation of ownership and control in organizations. It is studied in the next section.

Consider the following extension of the canonical agency setting.

1. The principal is risk neutral, i.e. $V(x-s(x)) = x - s(x)$.

2. There are n risk averse agents. Each agent takes an unobservable action a_i which affects the distribution function $\Pi_i(x_i, a_i)$ of a monetary outcome x_i.

3. The principal can observe the vector of outcomes $x = (x_1, ..., x_n)$. Total output x is distributed with distribution function $\Pi(x,a)$ and density $\pi(x,a)$ where $a = (a_1, ..., a_n)$ is a profile of the agents' actions.

Given that x is verifiable, an agent's compensation can depend on the entire vector x. We will now derive conditions under which this is optimal. The principal's multi-agent problem is to determine a vector of actions a and a profile of sharing rules $s(x) = (s_1(x), ..., s_n(x))$ that maximize her expected utility subject to the constraints that i) IR is satisfied for each agent, and ii) given $s(x)$, the strategy profile a is a Nash equilibrium. Although the specific equilibrium concept is inessential, it is clearly reasonable to let the agents' action choices depend on one another.

The following definition is a generalization of definition 6.

Definition 7 (Sufficient Statistic) The function $T_i(x)$ is a sufficient statistic for x with respect to a_i if there exist functions $G_i(\cdot) \geq 0$ and $H_i(\cdot) \geq 0$ such that

$$\pi(x,a) = G_i(x,a_{-i})\, H_i(T_i(x),a) \tag{4.54}$$

for all $(x,a) \in X \times \Lambda$, where $X = \times X_i$. Moreover, the vector $T(x) = (T_1(x),...,T_n(x))$ is a sufficient statistic for x with respect to a if each component $T_i(x)$ is a sufficient statistic with respect to a_i.

For example, $T_i(x)$ could only comprise a single element x_i, or it could be some average of the individual components $x_1,...,x_n$. We now derive the analogue of the sufficiency part of theorem 6 for the multi-agent setting.

Theorem 7 If $T(x)$ is a sufficient statistic for x with respect to a, then for any profile of sharing rules $s(x) = (s_1(x),...,s_n(x))$, there exists a profile $s(T(x)) = \{s_1(T_1(x)),...,s_n(T_n(x))\}$ that conditions only on $T(x)$ and makes the principal strictly better off.

Proof (direct) First, we show that if a is a Nash equilibrium under $s(x)$, it is also a Nash equilibrium under $s(T(x))$. Subsequently, we show that the cost of implementing a under $s(T(x))$ is strictly less than under $s(x)$.

Suppose a is a Nash equilibrium under $s(x)$. Consider a particular agent i and take the actions of the remaining $n-1$ agents as given. For any T_i in the range of $T_i(x)$, define the value of $s_i(T_i(x))$ at $T_i(x) = T_i$ by

$$u_i(s_i(T_i)) = \frac{\int_{\{x|T_i(x)=T_i\}} u_i(s_i(x))\, G_i(x,a_{-i})\, dx}{\int_{\{x|T_i(x)=T_i\}} G_i(x,a_{-i})\, dx}. \tag{4.55}$$

Note that $G(\cdot)$ and therefore also $s_i(T_i(x))$ are independent of a_i. Moreover,

$$\begin{aligned} \frac{G_i(x,a_{-i})}{\int_{\{x|T_i(x)=T_i\}} G_i(x,a_{-i})\, dx} &= \frac{G_i(x,a_{-i})\, H_i(T_i,a)}{\int_{\{x|T_i(x)=T_i\}} G_i(x,a_{-i})\, H_i(T_i,a)\, dx} \\ &= \frac{\pi(x,a)}{\int_{\{x|T_i(x)=T_i\}} \pi(x,a)\, dx} \end{aligned} \tag{4.56}$$

is a probability measure on the set $\{x|\, T_i(x) = T_i\}$, which implies that the right-hand side in (4.55) is agent i's expected utility under $s_i(x)$ conditional on the event that $T_i(x) = T_i$. Thus, for any T_i, $s_i(T_i)$ is the certain payment that yields the same utility as a lottery on the set $\{x|\, T_i(x) = T_i\}$ with density $G_i(x,a_{-i}) / \int_{\{x|T_i(x)=T_i\}} G_i(x,a_{-i})\, dx$ and outcomes $s_i(x)$. By (4.55),

$$\begin{aligned} &\int_{\{x|T_i(x)=T_i\}} u_i(s_i(T_i))\, G_i(x,a_{-i})\, dx \\ = &\int_{\{x|T_i(x)=T_i\}} u_i(s_i(x))\, G_i(x,a_{-i})\, dx. \end{aligned} \tag{4.57}$$

Multiplying both sides with $H_i(T_i, a)$ and integrating over T_i, we have

$$\int_X u_i(s_i(T_i(x)))\,\pi(x,a)\,dx = \int_X u_i(s_i(x))\,\pi(x,a)\,dx, \tag{4.58}$$

i.e. the agent enjoys the same expected utiliy for any profile a (and consequently for any action a_i since a_{-i} is taken as given). This argument can now be repeated for any of the other $n-1$ agents, which implies that $s(T(x))$ and $s(x)$ implement the same actions.

Finally, let us show that the principal's expected utility under $s_i(T_i(x))$ is at least as great as under $s_i(x)$. From (4.55), Jensen's inequality and strict concavity of $u(\cdot)$, it follows that

$$s_i(T_i) < \frac{\int_{\{x|T_i(x)=T_i\}} s_i(x)\,G_i(x,a_{-i})\,dx}{\int_{\{x|T_i(x)=T_i\}} G_i(x,a_{-i})\,dx}, \tag{4.59}$$

for all T_i. As in the derivation of (4.57)-(4.58), multiplying both sides of (4.59) with $H_i(T_i,a)\int_{\{x|T_i(x)=T_i\}} G_i(x,a_{-i})\,dx$ and integrating over the set of T_i's yields

$$\int_X s_i(T_i(x))\,\pi(x,a)\,dx < \int_X s_i(x)\,\pi(x,a)\,dx. \tag{4.60}$$

Since this is true for any of the n agents, (4.60) implies

$$\int_X (x - s(T(x)))\,\pi(x,a)\,dx > \int_X (x - s(x))\,\pi(x,a)\,dx, \tag{4.61}$$

i.e. the principal is strictly better off with $s(T(x))$. ■

Remarks

1. An immediate implication of theorem 7 is that the optimal sharing rule must have the form $s^*(T(x))$.

2. The reverse of theorem 7 also holds when the notion of a sufficient statistic is replaced with that of a *globally sufficient statistic*. $T(x)$ is globally sufficient if for all a, i, T_i, and $x', x'' \in \{x|\,T_i(x) = T_i\}$,

$$\frac{\pi_{a_i}(x',a)}{\pi(x',a)} = \frac{\pi_{a_i}(x'',a)}{\pi(x'',a)}. \tag{4.62}$$

   See Holmström (1982), theorem 6, for details.

We will now illustrate theorem 7 with two examples. In the first example, outputs are independently distributed. Hence, agent j's output contains no information about the action taken by agent i. In the light of our results, we would then expect that rewarding agents on the basis of peer performance is pointless. The following corollary confirms this intuition.

Corollary 4 If outputs are independently distributed, the optimal sharing rule for agent i depends on x_i alone, i.e. $s_i^*(x) = s_i^*(x_i)$.

Proof (direct) By independence,

$$\pi(x, a) = \prod_{i=1}^{n} \pi_i(x_i, a_i). \tag{4.63}$$

Thus, $T_i(x) = x_i$ is a sufficient statistic for x with respect to a_i. By theorem 7, the optimal sharing rule must then be of the form $s_i^*(x_i)$. ■

Remarks

1. With a little more structure, it can be shown that the reverse of corollary 4 is also true (see Holmström (1982), theorem 7). Thus, when output is not independently distributed, *relative performance evaluation* is generally optimal.

2. An important implication of corollary 4 is that competition (i.e. rewarding agents according to their relative perfomance) is valueless when there is no common underlying uncertainty. This implies that *rank-order tournaments* which award prizes (e.g. promotions, pay raises, etc.) on the basis of ordinal rankings are suboptimal in the presence of idiosyncratic risks. But even if outputs are interdependent, rank-order tournaments are informationally wasteful and perform worse than incentive schemes based on cardinal measures since ordinal rankings are typically not a sufficient statistic.

3. In some situations, the principal may want to base wages on peer performance even if output is independently distributed. For instance, if actions are complementary, i.e. if a higher action by one agent improves the marginal productivity of his co-workers, rewarding agents on the basis of group performance fosters teamwork and can be overall beneficial (Itoh (1991)).

The second example illustrates that it can be sometimes sufficient to relate individual output to some aggregate measure of team performance. Consider the following simple technology:

$$x_i = a_i + \eta + \epsilon_i, \tag{4.64}$$

where η is a common uncertainty parameter and the ϵ_i's are idiosyncratic shocks. In addition, assume that $\eta, \epsilon_1, ..., \epsilon_n$ are all independent and normally distributed. Define $\tau_i = \frac{1}{\text{Var}(\epsilon_i)}$ as the *precision* of ϵ_i and let $\overline{x} = \sum_i \alpha_i x_i$ be a weighted average of the agent's outcomes with weights defined as $\alpha_i = \frac{\tau_i}{\sum_i \tau_i}$.

Corollary 5 Given the technology $x_i = a_i + \eta + \epsilon_i$, the optimal sharing rule for agent i is of the form $s_i^* (x_i, \overline{x})$.

For a proof, see Holmström (1982), theorem 8. As in corollary 4, the proof shows that the density $\pi (x, a)$ is multiplicatively separable into functions that allow the interpretation that for all i, $T_i (x) = (x_i, \overline{x})$ is a sufficient statistic for x with respect to a_i.

Remarks

1. Note that corollary 5 does not claim that the optimal sharing rule should depend on $x_i - \overline{x}$ (for instance, by paying agent i less if x_i falls short of $\overline{x}$ and more if it exceeds $\overline{x}$). Also, the fact that $\overline{x}$ captures all the relevant information about the common uncertainty η is a pure artefact of the normal distribution.

2. If a particular noise term ϵ_j has low variance (i.e. high precision), then x_j is a fairly good predictor of the common parameter η (recall that the a_i's are known in equilibrium) and receives more weight in the average $\overline{x}$. Thus, by including $\overline{x}$ in the optimal sharing rule, the principal "filters out" as much as possible of the systematic risk η, which in turn reduces the risk premia she must pay to the agents. Using the strong law of large numbers, Holmström also shows that as $n \to \infty$, the parameter η can be determined with arbitrary precision.

Multiple Agents II: Team Incentives

In the previous section, our starting point was a setting with n agents and one principal. In many multi-agent situations, however, a principal does not exist. A group of agents is called a *team* or *partnership* if the joint output x is fully distributed among the agents themselves. That is, a team is defined by the *budget-balancing* condition

$$\sum_i s_i (x) = x \tag{4.65}$$

for all $x \in X$. We will now investigate a simple model of team production in which the efficient outcome x_{FB}^* cannot be attained. Subsequently, we show

that by relaxing the budget-balancing condition and introducing a residual claimant (the principal), the first-best can be restored. Hence, we provide a rationale for the existence of a principal in a multi-agent framework. Consider the following model:

1. There are n risk neutral agents with preferences over wealth and actions defined by $U_i(w_i, a_i) = w_i - c_i(a_i)$, where $c_i'(\cdot) > 0$, $c_i''(\cdot) > 0$, $c_i(0) = 0$, and $\lim_{a_i \to 0_+} c_i'(a_i) = 0$.

2. Each agent takes an unobservable action $a_i \geq 0$ which affects a deterministic monetary outcome $x(a)$, where $a = (a_1, ..., a_n)$. The function $x(\cdot)$ is strictly increasing, concave, and differentiable with $x(0) = 0$.

If a is observable and verifiable, the first-best problem is

$$\max_a x(a) - \sum\nolimits_i c(a_i) \tag{4.66}$$

with first order condition

$$\frac{\partial x(a_{FB}^*)}{\partial a_i} = c'(a_{iFB}^*) \tag{4.67}$$

for all $i = 1, ..., n$. Thus, the first-best equates marginal productivity and marginal cost for each agent. Note that the first-best problem is strictly concave and has a unique interior solution $a_{iFB}^* > 0$ for all i.

When a is unobservable, the fact that x cannot be split up into individual contributions creates a free-rider problem: For each additional unit of a_i, agent i bears the full cost but must share part of the marginal output with his co-workers. This suggests that in the presence of externalities, the second-best solution entails an underprovision of effort.

Theorem 8 There exists no profile of differentiable sharing rules that is budget-balanced and yields a_{FB}^* as a Nash equilibrium.

Proof (by contradiction) Suppose such a profile exists. Since a_{FB}^* is a Nash equilibrium, a_{iFB}^* satisfies

$$a_{iFB}^* \in \arg\max_{a_i} s_i\left(x\left(a_i, a_{-iFB}^*\right)\right) - c_i(a_i) \tag{4.68}$$

for all i. Moreover, since $a_{iFB}^* > 0$ and $s_i(\cdot)$ is differentiable, a_{iFB}^* also satisfies the first-order necessary condition

$$s_i'(x(a_{FB}^*)) \frac{\partial x(a_{FB}^*)}{\partial a_i} = c_i(a_{iFB}^*). \tag{4.69}$$

This is consistent with (4.67) if and only if

$$s_i'\left(x\left(a_{FB}^*\right)\right) = 1 \tag{4.70}$$

for all i. But budget-balancing implies $\sum_i s_i'(x) \equiv 1$, a contradiction. ■

Remarks

1. If the differentiability assumption is dropped, theorem 8 continues to hold, albeit the proof becomes more complicated. See Holmström (1982), theorem 1.

2. Under risk aversion, theorem 8 breaks down. As is shown by Rasmusen (1987), a "scapegoat" contract in which a randomly selected agent is punished if $x < x_{FB}^*$, and a "massacre" contract in which all but a randomly selected agent are punished if $x < x_{FB}^*$ both implement a_{FB}^* and are budget-balancing (in either case, the penalty is distributed to the remaining agent(s)). This is at odds with our results from the single-agent model where the first-best is implementable if the agent is risk neutral but not if he is risk averse.

3. Theorem 8 also breaks down if the action space Λ is finite and if $a \neq \hat{a}$ implies $x(a) \neq x(\hat{a})$ (Legros and Matthews (1993)). This case is trivial as it implies that actions are not perfect substitutes. For example, if there are three agents, $x(3,1,1) \neq x(1,1,3)$. Thus, shirkers can be unambiguously identified and a heavy penalty levied upon a shirker restores the first-best. However, Legros and Matthews also show that in some less trivial cases, the first-best can be approximated arbitrarily closely if unbounded penalties are feasible.

Let us now replace budget-balancing with the *feasibility* condition

$$\sum\nolimits_i s_i(x) \leq x \tag{4.71}$$

and introduce a principal who receives the surplus $x - \sum_i s_i(x)$ (note that the principal does not engage in any productive activity). Unlike in the pure partnership model, a_{FB}^* can now be sustained as a Nash equilibrium.

Theorem 9 There exist profiles of sharing rules that are feasible and yield a_{FB}^* as a Nash equilibrium.

Proof (direct) Consider the following family of profiles indexed by the vector $(\overline{s}_1, ..., \overline{s}_n)$:

$$s_i(x) = \begin{cases} \overline{s}_i \text{ if } x \geq x\left(a_{FB}^*\right) \\ 0 \text{ otherwise,} \end{cases} \tag{4.72}$$

where $\sum_i \overline{s}_i = x(a^*_{FB})$ and $\overline{s}_i > c_i(a^*_{iFB}) > 0$ for all $i = 1, ..., n$ (note that such $\overline{s}_i$'s exist since $\sum_i \overline{s}_i = x(a^*_{FB}) > \sum_i c_i(a^*_{iFB})$).

Consider agent i and take the actions of the remaining agents a^*_{-iFB} as given. If agent i selects $a_i < a^*_{iFB}$, his utility is $-c_i(a_i) \leq 0$, if he selects $a_i = a^*_{iFB}$, his utility is $\overline{s}_i - c_i(a^*_{iFB}) > 0$, and if he selects $a_i > a^*_{iFB}$, his utility is $\overline{s}_i - c_i(a_i)$, which is less than under $a_i = a^*_{iFB}$. Repeating the argument for all i proves that a^*_{FB} is a Nash equilibrium. ■

Remarks

1. Note that the equilibrium a^*_{FB} is not unique. Clearly, the profile $a = (0, ..., 0)$ is another Nash equilibrium under the sharing rule (4.72).

2. Under uncertainty (i.e. if x is generated by a probability distribution $\pi(x, a)$), the first-best solution can be attained by "selling the firm" to each (!) agent, i.e. $s_i(x) = x - K$, where the constant K ensures that budget-balancing is satisfied (in expectations) on the equilibrium path.

3. What is the role of the principal? If the agents can commit to the sharing rules (4.72), the principal is not needed. Here, commitment means that off the equilibrium path, any output that is not distributed (i.e. $x < x(a^*_{FB})$) must be destroyed. Clearly, this is a very strong assumption. However, if we allow for renegotiation, the scheme (4.72) is no longer credible. Once an agent deviates and chooses $a_i < a^*_{iFB}$, it is strictly Pareto-improving to distribute x among the agents. This argument can be stated more formally: From earlier discussions of renegotiation, we know that the optimal renegotiation-proof contract can be found using backwards induction. That is, for any x, we must first ascertain the set of contracts that are ex-post efficient. But ex-post efficiency implies $\sum_i s_i(x) = x$, which in turn implies that a^*_{FB} cannot be sustained as a Nash equilibrium. These considerations are no longer true once we introduce a principal. Since the principal receives the remaining output $x - \sum_i s_i(x)$, *any* distribution of x is ex-post efficient. In particular, the sharing rule (4.72) is now renegotiation-proof and a^*_{FB} can be supported as a Nash equilibrium. Hence, the sole role of the principal is to *break the budget*, i.e. to make $\sum_i s_i(x) \leq x$ credible in the presence of renegotiation.

4. In a seminal paper, Alchian and Demsetz (1972) suggest that the free-rider problem can be alleviated by hiring a principal to monitor the agents' actions. But this creates a new problem: Who monitors the monitor? Alchian and Demsetz's solution is to align the principal's

interests with that of the firm by making her the residual claimant to the firm's net earnings. Hence, the partnership is transformed into a capitalistic firm with the monitor acting effectively as the owner. What makes this story unappealing is that in large corporations, the owners are typically widely dispersed shareholders who delegate the monitoring to the board of directors. In this regard, the present story of the principal as a budget breaker is more appealing as it allows for the separation of ownership and control.