Theory of Mechanism Design

Debasis Mishra[1]

October 5, 2023

[1]Economics and Planning Unit, Indian Statistical Institute, 7 Shahid Jit Singh Marg, New Delhi 110016, India, E-mail: dmishra@isid.ac.in

Contents

Chapter 1

Introduction to Mechanism Design

Consider a seller who owns an indivisible object, say a house, and wants to sell it to a set of buyers. Each buyer has a value for the object, which is her willingness to pay for the house. The seller wants to design a selling procedure, an auction for example, such that he gets the maximum possible price (revenue) by selling the house. If the seller knew the values of the buyers, then he would simply offer the house to the buyer with the highest value and give him a "take-it-or-leave-it" offer at a price equal to that value. Clearly, the (highest value) buyer has no incentive to reject such an offer. Now, consider a situation where the seller does not know the values of the buyers. What selling procedure will give the seller the maximum possible revenue? A clear answer is impossible if the seller knows nothing about the values of the buyer. However, the seller may have some information about the values of the buyers. For example, the possible range of values, the probability of having these values etc. Given these information, is it possible to design a selling procedure that guarantees maximum (expected) revenue to the seller?

In this example, the seller had a particular objective in mind - maximizing revenue. Given his objective he wanted to *design* a selling procedure such that when buyers participate in the selling procedure and try to maximize their own payoffs within the rules of the selling procedure, the seller will maximize his expected revenue over all such selling procedures.

The study of mechanism design looks at such issues. A planner (mechanism designer) needs to design a *mechanism* (a selling procedure in the above example) where strategic agents can interact. The interactions of agents result in some outcome. While there are several possible ways to design the rules of the mechanism, the planner has a particular

objective in mind. For example, the objective can be *utilitarian* (maximization of the total utility of agents) or maximization of his own utility (as was the case in the last example) or some *fairness* objective. Depending on the objective, the mechanism needs to be designed in a manner such that when strategic agents interact, the resulting outcome gives the desired objective. One can think of mechanism design as the *reverse engineering* of game theory. In game theory terminology, a mechanism induces a *game-form* whose equilibrium outcome is the objective that the mechanism designer has set.

1.0.1 Private Information and Utility Transfers

The primitives a mechanism design problem are the set of possible outcomes or alternatives and preferences of agents over the set of alternatives. These preferences are not known to the mechanism designer. Mechanism design problems can be classified based on the amount of information asymmetry present between the agents and the mechanism designer.

1. Complete Information: Consider a setting where an accident takes place on the road. Three parties (agents) are involved in the accident. Everyone knows perfectly who is at fault, i.e., who is responsible to what extent for the accident. The traffic police comes to the site but does not know the true information about the accident. The mechanism design problem is to design a set of rules where the traffic police's objective (to punish the true offenders) can be realized. The example given here falls in a broad class of problems where agents perfectly know all the information between themselves, but the mechanism designer does not know this information.

   This class of problems is usually termed as the *implementation problem*. It is usually treated separately from mechanism design because of strong requirements in equilibrium properties in this literature. We will not touch on the implementation problem in this course.

2. Private Information and Interdependence: Consider the sale of a single object. The utility of an agent for the object is his private information. This utility information may be known to him completely, but usually not known to other agents and the mechanism designer. There are instances where the utility information of an agent may not be perfectly known to him. Consider the case where a seat in a flight is

being sold by a private airlines. An agent who has never flown this airlines does not completely know his utility for the flight seat. However, there are other agents who have flown this airlines and have better utility information for the flight seat. So, the utility of an agent is influenced by the information of other agents. The mechanism designer does not know the information agents have.

Besides the type of information asymmetry, mechanism design problems can also be classified based on whether monetary transfers are involved or not. Transfers are a means to redistribute utility among agents.

1. Models without transfers. Consider a setting where a set of agents are deciding to choose a candidate in an election. There is a set of candidates in the election. Agents have preference over the candidates. Usually monetary transfers are not allowed in such voting problems. Other problems where monetary transfers are not allowed are *matching* problems - matching students to schools, kidneys to patients etc.

2. Models with transfers and quasi-linear utility. The single object auction is a classic example where monetary transfers are allowed. If an agent buys the object he is expected to pay an amount to the seller. The net utility of the agent in that case is his utility for the object minus the payment he has to make. Such net utility functions are linear in the payment component, and is referred to as the quasi-linear utility functions.

In this course, we will focus on (a) **voting models without transfers** and (b) **models with transfers and quasi-linear utility**. We will briefly touch on some models of non-quasilinearity. In voting models, we will mainly deal with **ordinal preferences**, i.e., intensities of preferences will not matter. We will mainly focus on the case where agents have **private information about their preferences over alternatives**. Note that such private information is completely known to the respective agents but not known to other agents and the mechanism designer. We will briefly touch on some models of interdependence of preferences.

1.0.2 Examples in Practice

The theory of mechanism design is probably the most successful story of game theory. Its practical applications are found in many places. Below, we will look at some of the applica-

tions.

1. *Matching.* Consider a setting where students need to be matched to schools. Students have preferences over schools and schools have preference over students. What mechanisms must be used to match students to schools? This is a model without any transfers. Lessons from mechanism design theory has been used to design centralized matching mechanisms for major US cities like Boston and New York. Such mechanisms and its variants are also used to match kidney donors to patients, doctors to hospitals, and many more.

2. *Sponsored Search Auction.* If you search for a particular keyword on Google, once the search results are displayed, one sees a list of advertisements on the top and (sometimes) on the right of the search results - see Figure 1.1. Such slots for advertisements are dynamically sold to potential buyers (advertising companies) as the search takes place. One can think of the slots on a page of search result as a set of indivisible objects. So, the sale of slots on a page can be thought of as simultaneous sale of a set of indivisible objects to a set of buyers. This is a model where buyers make payments to Google. Google uses a variant of a well studied auction in the auction theory literature (Edelman et al., 2007). Bulk of Google's revenues come from such auctions.

3. *Spectrum Auction.* Airwave frequencies are important for communication. Traditionally, Government uses these airwaves for defense communication. In late 1990s, various Governments started selling (auctioning) airwaves for private communication (Ausubel and Milgrom, 2006). Airwaves for different areas were sold simultaneously. For example, India is divided into various "circles" like Delhi, Punjab, Haryana etc. A communication company can buy the airwaves for one or more circles. Adjacent circles have synergy effects and distant circles have substitutes effects on utility. Lessons from auction theory were used to design auctions for such spectrum sale in US, UK, India, and many other European countries. The success of some of these auctions have become the biggest advertisement of game theory.

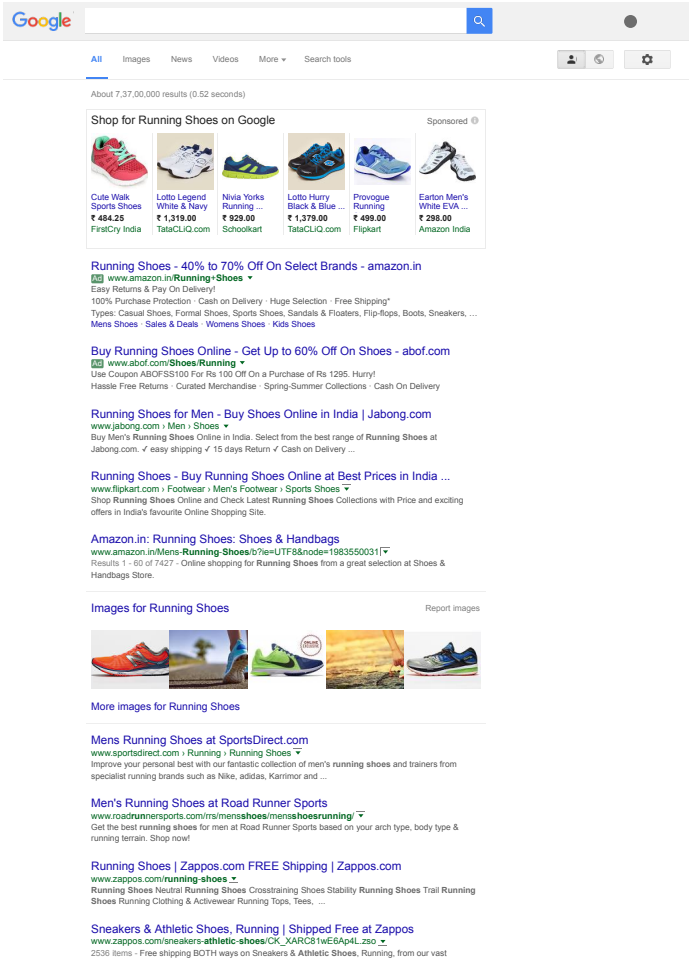

Figure 1.1: Sponsored search on Google

1.1 A General Model of Preferences

We will now formally define a model that will set up some basic ingredients to study mechanism design. The model will be general enough to cover cases where transfers are permitted and where it is excluded.

Let $N := \{1, \ldots, n\}$ be a set of n agents. Let X_i be the set of possible outcomes for agent i. These outcomes can represent a variety of situations. In particular, we will often encounter situations where X_i may be decomposable to two sets: a set of alternatives A_i and transfers (which can take any value). Hence, the set of transfers is $\mathbb{R}$. In other words, $X_i \equiv A_i \times \mathbb{R}$, where the agent i enjoys an "alternative" from A_i and a transfer amount in $\mathbb{R}$ as outcome.

There are situations where transfers are not allowed - either as an institutional constraint (as in voting) or as an ethical constraint (as in kidney donation). In that case, each X_i represents the set of alternatives (without any transfers).

In either case, the primitives of the model must allow preferences of the agents over the set of outcomes. The preference of each agent i is his preference over X_i, which is completely known to her and private information to her. This is called the **private values** model in mechanism design. The private preference over X_i is called the **type** of agent i. We will denote the preference (type) of each agent i as R_i, which will be assumed to be a **complete** and **transitive** binary relation over the set of outcomes X_i. We will denote the *symmetric or indifference* part of R_i as I_i and the *strict* part of R_i as P_i.

When there are no transfers, this is just a complete and transitive binary relation over the set of *alternatives* $A_i \equiv X_i$. A well known example is voting.

An outcome vector $x \equiv (x_1, \ldots, x_n) \in X_1 \times \ldots \times X_n$ is a vector such that $x_i \in X_i$. In many problems, we will require that $x_1 = x_2 = \ldots = x_n$. An example is the voting problem. In the voting model, X_i may represent the set of candidates in an election (and hence, $X_1 = X_2 = \ldots = X_n$) and N the set of voters. The type of agent i is a ranking of the set of candidates, which can potentially allow the agent to be indifferent between some candidates. For instance if $X_1 = X_2 = \ldots = X_n = \{a, b, c\}$ and R_i is a preference of agent i over $\{a, b, c\}$, then it can be something like $a\ R_i\ b\ R_i\ c$. Here, if agent i strictly prefers a over all candidates, then we have $a\ P_i\ b$ and $a\ P_i\ c$. An outcome vector in this model allocates the same candidate to all the voters: if a is selected then, $x_1 = x_2 = \ldots = x_n = a$. These are called **public good** problems, where the outcome vectors satisfy that they have to be the same outcome. Hence, in public good problems, we assume $X_1 = \ldots = X_n$ and each $x \in X_1 \times \ldots \times X_n$ satisfies $x_1 = \ldots = x_n$.

This need not be the case in **private good problems**. For instance, if $X_1 = X_2 = \ldots = X_n$ is a set of objects that are being allocated to n agents, then in any outcome vector $x = (x_1, \ldots, x_n)$, we will require that $x_i \neq x_j$ for each i, j pair (the same object cannot be allocated to two different agents). In the single object allocation problem, we consider sale or allocation of a single indivisible object to a set of buyers. Usually, this also involves payment or transfer. Hence, an outcome consists of two components: (a) an allocation decision and (b) a payment amount decision. Abstractly, an outcome $x_i \in X_i$ consists of (a_i, p_i) with $a_i \in \{0, 1\}$ and $p_i \in \mathbb{R}$. Each outcome vector $x \equiv (a, p)$ must satisfy $\sum_{i \in N} a_i \leq 1$.

Consider the problem of choosing a public good. The citizens of a city are deciding whether to build a public project (park, bridge, museum etc.) or not. This is usually accompanied by a decision on the amount of tax each citizen must pay to finance the project. Formally, N is the set of citizens and set of outcomes consists of outcomes which are pairs of the form (a, p) with a being the public project chosen and p being the tax amount. There are additional restrictions: total tax collected must cover the cost of the public good (if it is provided).

Summarizing, the set of possible and feasible outcomes vary from problem to problem. Abstractly, we fix the set of possible outcome vectors as some subset $\mathcal{X} \subseteq X_1 \times X_2 \times \ldots \times X_n$.

Transfers and Quasilinear preferences. When there are transfers, a preference R_i will represent a preference over the outcome space $X_i \equiv A_i \times \mathbb{R}$. We will usually restrict attention to the case where A_i is finite - but we will deal with some specific cases where A_i is not finite, in particular, where A_i is a set of lotteries over some finite set of deterministic alternatives.

We will impose some standard (and natural) restrictions on preferences. A typical outcome in X_i is a tuple (a_i, t_i), where $a_i \in A_i$ is an alternative and $t_i \in \mathbb{R}$ is the transfer **paid by** the agent. When transfers are allowed, we will mostly deal with the following kind of preferences.

Definition **1** *A classical preference R_i is* **quasilinear** *if there exists a map $v_i : A_i \to \mathbb{R}$ such that for any $(a, t), (b, t') \in X$, we have*

$$\Big[(a,t)\ R_i\ (b,t')\Big] \Leftrightarrow \Big[(v_i(a) - t) \geq (v_i(b) - t')\Big].$$

Obviously, if v_i is a map representing R_i and v_i' is such that $v_i(a) - v_i(b) = v_i'(a) - v_i'(b)$ for each $a, b \in A_i$, then it also represents the same quasilinear preference relation R_i. Often (without loss of generality), we will normalize $v_i(a) = 0$ for some a.

Quasilinear preferences rule out income effects. We do not impose this restriction on preferences in this chapter but we will do so in subsequent sections dealing with mechanism design problems with transfers.

1.2 Social Choice Functions and Mechanisms

A planner or designer has some objectives in mind. The social choice function is supposed to capture it - think of it as what the designer would have done had she known the preferences

of the agent. This will mean that if he knew the profile of preferences of the agents over the outcome space, how he would select an outcome. A social choice function is a complete description of his plan of action for every *possible* profiles of preferences.

Hence, it is important to describe what a planner thinks is the "possible" profiles of preferences. This is usually referred to as the **domain** or **type space**. Let $\mathcal{R}_i$ denote the set of all possible preferences of agent i over outcomes X_i. The domain of agent i is a subset of preferences $\mathcal{D}_i \subseteq \mathcal{R}_i$. The domains of agents are common knowledge among agents and the designer. Let

$$\mathcal{D} \equiv \mathcal{D}_1 \times \ldots \times \mathcal{D}_n,$$

be the set of all possible type profiles.

A **social choice function (SCF)** is a map $F : \mathcal{D} \to \mathcal{X}$. So, at every preference profile $R \equiv (R_1, \ldots, R_n) \in \mathcal{D}$, the scf F assigns an outcome vector in $\mathcal{X}$ for agents. We will denote the outcome assigned to an agent i at preference profile R as $F_i(R)$.

In settings where transfers are allowed, it is convenient to think of an SCF F as a collection tuples (f, p), where $f : \mathcal{D} \to \mathcal{A}$ is the allocation rule with $\mathcal{A} \subseteq (A_1 \times \ldots \times A_n)$ and $p : \mathcal{D} \to \mathbb{R}^n$ being the payment function of the n agents. At preference profile R, the allocation decision of agent i will then be denoted by $f_i(R) \in A_i$ and the payment decision by $p_i(R)$.

In settings where transfers are not permitted, $\mathcal{X} \equiv \mathcal{A}$ and $F \equiv f$. Hence, in settings where transfers are not allowed $F_i(R) = f_i(R)$ for all i and for all $R \in \mathcal{D}$.

Mechanisms. A mechanism is a more complicated object than an SCF. The main objective of a mechanism is to set up rules of interaction between agents. These rules are often designed with the objective of realizing the outcomes of a social choice function. The basic ingredient in a mechanism is a **message**. A message is a communication between the agent and the mechanism designer. You can think of it as an action chosen in various contingencies in a Bayesian game - these messages will form the actions for various contingencies of agents in a Bayesian game that the designer will set up.

A mechanism must specify the **message space** for each agent. A message space has to specify various contingencies that may arise in a mechanism and available actions at each of the contingencies. This in turn induces a Bayesian game with messages playing the role of

actions. Given a message profile, the mechanism chooses an outcome.

DEFINITION **2** *A* **mechanism** *is a collection of message spaces and a decision rule:* $\mathcal{M} \equiv (M_1, \ldots, M_n, \phi)$*, where*

- *for every* $i \in N$*,* M_i *is the message space of agent* i *and*
- $\phi : M_1 \times \ldots \times M_n \to \mathcal{X}$ *is the decision rule.*

A mechanism is a **direct mechanism** *if* $M_i = \mathcal{D}_i$ *for every* $i \in N$*.*

So, in a direct mechanism, every agent communicates a type from his type space to the mechanism designer. Hence, if F is an scf, then $((\mathcal{D}_1, \ldots, \mathcal{D}_n), F)$ is a direct mechanism - for simplicity, we will just refer to F as a (direct) mechanism.

We denote the outcome of agent i in a mechanism $(M_1, \ldots, M_n, \phi)$ by $\phi_i(m_1, \ldots, m_n)$ for each $(m_1, \ldots, m_n) \in M_1 \times \ldots \times M_n$.

The message space of a mechanism can be quite complicated. Consider the sale of a single object by a "price-based" procedure. The mechanism designer announces a price and asks every buyer to communicate if he wants to buy the object at the announced price. The price is raised if more than one buyer expresses interest in buying the object, and the procedure is repeated till exactly one buyer shows interest. The message space in such a mechanism is quite complicated. Here, a message must specify the communication of the buyer (given his type) for every contingent price.

1.3 Dominant Strategy Incentive Compatibility

We now introduce the notion of incentive compatibility. The idea of a mechanism and incentive compatibility is often attributed to the works of Hurwicz - see (Hurwicz, 1960). The goal of mechanism design is to design the message space and outcome function in a way such that when agents participate in the mechanism they have (best) actions (messages) that they can choose as a function of their private types such that the desired outcome is achieved. The most fundamental, though somewhat demanding, notion of incentive compatibility in mechanism design is the notion of dominant strategies.

A strategy is a map $s_i : \mathcal{D}_i \to M_i$, which specifies the message each agent i will choose for every realization of his preference. A strategy s_i is a **dominant strategy** for agent i in mechanism $((M_1, g_1), \ldots, (M_n, g_n))$, if for every $R_i \in \mathcal{D}_i$ we have

$$\phi_i(s_i(R_i), m_{-i}) \; R_i \; \phi_i(m_i', m_{-i}) \qquad \forall \; m_i' \in M_i, \; \forall \; m_{-i}.$$

Notice the strong requirement that s_i has to be a best strategy for *every* possible message profile of other agents at every profile. Such a strong requirement limits the settings where dominant strategies exist.

DEFINITION **3** *A social choice function F is* **implemented** *in dominant strategy equilibrium by a mechanism $(M_1, \ldots, M_n, \phi)$ if there exists strategies $(s_1, \ldots, s_n)$ such that*

1. *$(s_1, \ldots, s_n)$ is a dominant strategy equilibrium of $(M_1, \ldots, M_n, \phi)$, and*

2. *$\phi_i(s_1(R_1), \ldots, s_n(R_n)) = F_i(R_1, \ldots, R_n)$ for all $i \in N$ and for all $(R_1, \ldots, R_n) \in \mathcal{D}$.*

For direct mechanisms, we will look at equilibria where everyone tells the truth.

DEFINITION **4** *A direct mechanism (or associated social choice function) is* **strategy-proof** *or* **dominant strategy incentive compatible (DSIC)** *if for every agent $i \in N$ and every $R_i \in \mathcal{D}_i$, the* **truth-telling** *strategy $s_i(R_i) = R_i$ for all $R_i \in \mathcal{D}_i$ is a dominant strategy. In other words, F is strategy-proof if for every agent $i \in N$, every $R_{-i} \in \mathcal{D}_{-i}$, and every $R_i, R_i' \in \mathcal{D}_i$, we have*

$$F_i(R_i, R_{-i}) \; R_i \; F_i(R_i', R_{-i}),$$

So, to verify whether a social choice function is implementable or not, we need to search over infinite number of mechanisms whether any of them implements this SCF. A fundamental result in mechanism design says that one can restrict attention to the direct mechanisms.

PROPOSITION **1 (Revelation Principle, Myerson (1979))** *If a mechanism implements a social choice function F in dominant strategy equilibrium, then the direct mechanism F is strategy-proof.*

Proof: Suppose mechanism $(M_1, \ldots, M_n, \phi)$ implements F in dominant strategies. Let $s_i : \mathcal{D}_i \to M_i$ be the dominant strategy of each agent i.

Fix an agent $i \in N$. Consider two types $R_i, R_i' \in \mathcal{D}_i$. Consider R_{-i} to be the report of other agents in the direct mechanism. Let $s_i(R_i) = m_i$ and $s_{-i}(R_{-i}) = m_{-i}$. Similarly, let $s_i(R_i') = m_i'$. Then, using the fact that F is implemented by our mechanism in dominant strategies, we get

$$\phi_i(m_i, m_{-i}) \; R_i \; \phi_i(m_i', m_{-i})$$

But $F_i(R_i, R_{-i}) = \phi_i(m_i, m_{-i})$ and $F_i(R_i', R_{-i}) = \phi_i(m_i', m_{-i})$. This gives us the desired relation: $F_i(R_i, R_{-i}) \; R_i \; F_i(R_i', R_{-i})$, which establishes that F is strategy-proof. ■

Thus, a social choice function F is implementable in dominant strategies if and only if the direct mechanism F is strategy-proof. Revelation principle is a central result in mechanism design. One of its implications is that if we wish to find out what social choice functions can be implemented in dominant strategies, we can restrict attention to direct mechanisms. This is because, if some non-direct mechanism implements a social choice function in dominant strategies, revelation principle says that the corresponding direct mechanism is also strategy-proof.

Of course, a drawback is that a direct mechanism may leave out some equilibria of the main mechanism. The original mechanism may have some equilibria that may get ruled out because of restricting to the direct mechanism since it has a different strategy space. In general, this is a criticism of the mechanism design theory. Even in a direct mechanism, incentive compatibility only insists that truth-telling is an equilibrium but there may be other equilibria of the mechanism which may not implement the given social choice function. These stronger requirement that every equilibria, truth-telling or non-truth-telling, must correspond to the social choice function outcome is the cornerstone of the implementation literature.

1.4 Bayesian Incentive Compatibility

Bayesian incentive compatibility can be attributed to Harsanyi (1968a,b,c), who introduced the notion of a Bayesian game and a Bayesian equilibrium. It is a weaker requirement than

the dominant strategy incentive compatibility. While dominant strategy incentive compatibility required the equilibrium strategy to be the best strategy under all possible strategies of opponents, Bayesian incentive compatibility requires this to hold in *expectation*. This means that in Bayesian incentive compatibility, an equilibrium strategy must give the highest expected utility to the agent, where we take expectation over types of other agents. To be able to take expectation, agents must have information about the probability distributions from which types of other agents are drawn. Hence, Bayesian incentive compatibility is informationally demanding. In dominant strategy incentive compatibility the mechanism designer needed information on the type space of agents, and every agent required no prior information of other agents to compute his equilibrium. In Bayesian incentive compatibility, every agent needs to know the distribution from which agents' types are drawn.

Since we need to compute expectations, we will represent preference over X_i by a utility function. In particular, the type of agent i is a utility function $u_i : X_i \to \mathbb{R}$. When transfers are not allowed, such utility representation is possible since $X_i \equiv A_i$ will be assumed to finite – even if it is not finite, we will assume that such a representation is always possible. When transfers are allowed, we will assume preferences will always admit a continuous (and monotonic in transfers) utility representation. So, our domain $\mathcal{D}_i$ is a domain of utility functions now.

To understand Bayesian incentive compatibility, fix a mechanism $(M_1, \ldots, M_n, \phi)$. A strategy of agent $i \in N$ for such a mechanism is a mapping $s_i : \mathcal{D}_i \to M_i$. To define a Bayesian equilibrium, let $G_i(u_{-i}|u_i)$ denote the conditional distribution of types of agents other than agent i when agent i has type u_i. A strategy profile $(s_1, \ldots, s_n)$ is a **Bayesian equilibrium** if for all $i \in N$, for all $u_i \in \mathcal{D}_i$ we have

$$\int_{u_{-i}} u_i(\phi_i(s_i(u_i), s_{-i}(u_{-i})))dG_i(u_{-i}|u_i) \geq \int_{u_{-i}} u_i(\phi_i(m_i, s_{-i}(u_{-i})))dG_i(u_{-i}|u_i) \qquad \forall\ m_i \in M_i.$$

If all u_is are drawn independently, then we need not condition in the expectation.

A direct mechanism (social choice function) F is **Bayesian incentive compatible** if $s_i(u_i) = u_i$ for all $i \in N$ and for all $u_i \in \mathcal{D}_i$ is a Bayesian equilibrium, i.e., for all $i \in N$ and for all $u_i, u_i' \in \mathcal{D}_i$ we have

$$\int_{u_{-i}} u_i(F_i(u_i, u_{-i}))dG_i(u_{-i}|u_i) \geq u_i(F_i(u_i', u_{-i}))dG_i(u_{-i}|u_i)$$

A dominant strategy incentive compatible mechanism is Bayesian incentive compatible. A mechanism $(M_1, \ldots, M_n, \phi)$ **implements** a social choice function F in Bayesian equilibrium if there exists strategies $s_i : \mathcal{D}_i \to M_i$ for each $i \in N$ such that

1. $(s_1, \ldots, s_n)$ is a Bayesian equilibrium of $(M_1, \ldots, M_n, \phi)$ and

2. $\phi_i(s_1(u_1), \ldots, s_n(u_n)) = F_i(u_1, \ldots, u_n)$ for all $i \in N$ and for all $u \in \mathcal{D}$.

Analogous to the revelation principle for dominant strategy incentive compatibility, we also have a revelation principle for Bayesian incentive compatibility.

PROPOSITION **2 (Revelation Principle)** *If a mechanism implements a social choice function F in Bayesian equilibrium, then the direct mechanism F is Bayesian incentive compatible.*

Proof: Suppose $(M_1, \ldots, M_n, \phi)$ implements F. Let $(s_1, \ldots, s_n)$ be the Bayesian equilibrium strategies of this mechanism which implements F. Fix agent i and $u_i, u_i' \in \mathcal{D}_i$. Now,

$$\begin{aligned}
\int_{u_{-i}} u_i(F_i(u_i, u_{-i}))dG_i(u_{-i}|u_i) &= \int_{u_{-i}} u_i(\phi_i(s_i(u_i), s_{-i}(u_{-i})))dG_i(u_{-i}|u_i) \\
&\geq \int_{u_{-i}} u_i(\phi_i(s_i(u_i'), s_{-i}(u_{-i})))dG_i(u_{-i}|u_i) \\
&= \int_{u_{-i}} u_i(F_i(u_i', u_{-i}))dG_i(u_{-i}|u_i),
\end{aligned}$$

where the equalities come from the fact that the mechanism implements F and the inequality comes from the fact that $(s_1, \ldots, s_n)$ is a Bayesian equilibrium of the mechanisms. ■

Like the revelation principle of dominant strategy incentive compatibility, the revelation principle for Bayesian incentive compatibility is not immune to criticisms for multiplicity of equilibria.

AN EXAMPLE OF FIRST-PRICE AUCTION.

Consider an auction of a single object. There are two buyers with quasilinear preferences. As was discussed, their preferences can be represented by (v_1, v_2) and payoff of each buyer i when he pays p_i can be represented by $v_i - p_i$.

We consider a first-price auction where buyers bid amounts (b_1, b_2) and the highest bidder wins the object and pays his bid amount - ties are broken by giving the object with equal probability to each buyer. If values of the object is uniformly distributed in $[0, 1]$, a Bayesian equilibrium of this first-price auction is that each buyer i bids $\frac{1}{2}v_i$. This is very straightforward to see. If buyer i bids b_i his payoff is $(v_i - b_i)$ times the probability of winning by bidding b_i. But he wins by bidding b_i if the other bidder has bid less than b_i. Assuming that the other bidder is following the equilibrium strategy, we see that (a) he should never bid more than $\frac{1}{2}$, and (b) the probability of winning is equivalent to the probability $\frac{v_j}{2} \leq b_i$ (where $j \neq i$). This probability is just $2b_i$. Hence, the expected payoff from bidding b_i is $(v_i - b_i)2b_i$, which is a nice concave function and it is maximized at $b_i = \frac{1}{2}v_i$.

What does the revelation principle say here? Since there is *an* equilibrium of this mechanism, the revelation principle says that there is a direct mechanism with a truth-telling Bayesian equilibrium. Such a direct mechanism is easy to construct here.

1. Ask buyers to submit their values (v_1, v_2).

2. The buyer i with the highest value wins but pays $\frac{1}{2}v_i$.

Notice that the first-price auction implements the outcome of this direct mechanism. Since the first-price auction had this outcome in Bayesian equilibrium, this direct mechanism is Bayesian incentive compatible.

1.4.1 Failure of revelation principle

There are various settings where the revelation principle need not hold. Though we do not discuss this issue in detail, we point to some literature on this. One of the important settings where revelation principle does not hold is a setting where the mechanism designer cannot "commit". An example of a setting like is the sale of a single object. Suppose today, the seller posts a price and offers the object for sale. If there are no takers for the object, there will be another posted price in the next period, but the seller cannot commit to the next posted price. In such a setting, the outcome obtained by a mechanism cannot be replicated by a direct mechanism. Bester and Strausz (2000, 2001) provide an analysis of such issues. Another issue with the revelation principle is that it ignores mixed-strategy equilibria. The issue is discussed in detail in Strausz (2003), where it is shown that in certain cases, this

is not a limitation. Revelation principle may not hold if agents are *not* rational, i.e., their actions are not consistent with a preference ordering. This issue is analyzed in Saran (2011); De Clippel (2014). In summary, revelation principle is a powerful simplification but it is not something that can be taken for granted in all settings.

Chapter 2

Mechanism Design with Transfers and Quasilinearity

This chapter will deal with some fundamental questions in mechanism design with transfers under quasilinearity. We will continue to restrict attention to private values setting, where the preference of an agent over outcomes depends on his own private information. Since transfers are allowed, then we have two decisions to make - (a) what alternative to choose (b) how much payment to make.

As discussed earlier, a preference R_i is quasilinear if there exists a **valuation function** $v_i : A \to \mathbb{R}$ such that $(a, p)\ R_i\ (b, p')$ if and only if $v_i(a) - p \geq v_i(b) - p'$. There are specific settings, where the entire v_i vector can be encoded by simpler parameters. In such cases, we assume that type of an agent i is some t_i and $v_i(a, t_i)$ is the value of agent i from an alternative a when his type is t_i. Because of quasilinearity, if he makes a payment of p_i, then his net utility is

$$U_i(t_i, a, p_i) := v_i(a, t_i) - p_i.$$

In many cases, t_i itself will represent the valuations, and hence, will be a vector in $\mathbb{R}^{|A|}$, and $v_i(a, t_i) = t_i(a)$ for each $a \in A$.

A crucial element of the quasi-linearity assumption is that we can separate the value from the alternative and the utility from the payment. The separability of value from alternative and the utility from payment allows us to formulate incentive compatibility constraints in a lucid manner.

2.1 A General Model

The set of agents is denoted by $N = \{1, \ldots, n\}$. The set of alternatives is denoted by the set A, which can be finite or infinite. For some expositions, we will assume A to be finite - but results discussed here continue to hold even if A is not finite. The type of agent $i \in N$ is denoted by t_i which lies in some set T_i, called the type space. Type t_i will be assumed to be a (potentially, a multi-dimensional) vector in $\mathbb{R}^K$, where K is some positive integer. We denote a profile of types as $t = (t_1, \ldots, t_n)$ and the set of type profiles of all agents as $T^n = \times_{i \in N} T_i$.

The quasilinear utility of agent i over outcomes $A \times \mathbb{R}$ is captured by the **valuation function** $v_i : A \times T_i \to \mathbb{R}$. Thus, $v_i(a, t_i)$ denotes the valuation of agent $i \in N$ for decision $a \in A$ when his type is $t_i \in T_i$. Note that the mechanism designer knows T_i and the form of valuation function v_i. Of course, he does not know the *realizations* of each agent's type.

We will restrict attention to this setting, called the **private values** setting, where the utility function of an agent is independent of the types of other agents, and is completely known to him.

2.1.1 Allocation Rules

A **decision rule or an allocation rule** f is a mapping $f : T^n \to A$. Hence, an allocation rule gives a decision as a function of the types of the agents. For exposition purposes, assume A is finite. From every type profile, we construct a valuation matrix with n rows (one row for every agent) and $|A|$ columns. An entry in this matrix corresponding to type profile t, agent i, and $a \in A$ has value $v_i(a, t_i)$. We show one valuation matrix for $N = \{1, 2\}$ and $A = \{a, b, c\}$ below.

$$\begin{bmatrix} v_1(a, t_1) & v_1(b, t_1) & v_1(c, t_1) \\ v_2(a, t_2) & v_2(b, t_2) & v_2(c, t_2) \end{bmatrix}$$

Choosing an allocation amounts to choosing a column vector of this matrix. Here, we give some examples of allocation rules.

- **Constant allocation:** The constant allocation rule f^c allocates some $a \in A$ for every

$t \in T^n$. In particular, there exists $a \in A$ such that for every $t \in T$ we have

$$f^c(t) = a.$$

- **Dictator allocation:** The dictator allocation rule f^d allocates the *best* decision of some **dictator** agent $i \in N$. In particular, let $i \in N$ be the dictator agent. Then, for every $t_i \in T_i$ and every $t_{-i} \in T_{-i}$,

$$f^d(t_i, t_{-i}) \in \arg\max_{a \in A} v_i(a, t_i).$$

  It picks a dictator i and always chooses the column in the valuation matrix for which the i row has the maximum value in the valuation matrix.

- **Efficient allocation:** The efficient allocation rule f^e is the one which maximizes the sum of values of agents. In particular, for every $t \in T^n$,

$$f^e(t) \in \arg\max_{a \in A} \sum_{i \in N} v_i(a, t_i).$$

  This rule first sums the entries in each of the columns in the valuation matrix and picks a column which has the maximum sum.

  Hence, efficiency implies that the total value of agents is maximized in all states of the world (i.e., for all possible type profiles of agents). We will discuss why this is *Pareto efficient* later.

  Consider an example where a seller needs to sell an object to a set of buyers. In any allocation, one buyer gets the object and the others get nothing. The buyer who gets the object realizes his value for the object, while others realize no utility. Clearly, to maximize the total value of the buyers, we need to maximize this realized value, which is done by allocating the object to the buyer with the highest value.

  This particular allocation rule is also referred to as the **utilitarian** allocation rule.

- **Weighted efficient/utilitarianism allocation:** The weighted efficient allocation rule f^w is the one which maximizes the weighted sum of values of agents. In particular, there exists $\lambda \in \mathbb{R}^n_+ \setminus \{0\}$ such that for every $t \in T^n$,

$$f^w(t) \in \arg\max_{a \in A} \sum_{i \in N} \lambda_i v_i(a, t_i).$$

This rule first does a weighted sum of the entries in each of the columns in the valuation matrix and picks a column which has the maximum weighted sum.

- **Affine maximizer allocation:** The affine maximizer allocation rule f^a is the one which maximizes the weighted sum of values of agents and a term for every allocation. In particular, there exists $\lambda \in \mathbb{R}^n_+ \setminus \{0\}$ and $\kappa : A \to \mathbb{R}$ such that for every $t \in T^n$,

$$f^a(t) \in \arg\max_{a \in A} \big[\sum_{i \in N} \lambda_i v_i(a, t_i) - \kappa(a) \big].$$

  This rule first does a weighted sum of the entries in each of the columns in the valuation matrix and subtracts κ term corresponding to this column, and picks the column which has this sum highest.

- **Max-min (Rawls) allocation:** The max-min (Rawls) allocation rule f^r picks the allocation which maximizes the minimum value of agents. In particular for every $t \in T^n$,

$$f^r(t) \in \arg\max_{a \in A} \min_{i \in N} v_i(a, t_i).$$

  This rule finds the minimum entry in each column of the valuation matrix and picks the column which has the maximum such minimum entry.

We can just keep on defining many such allocation rules. Allocation rules convey the fact that if transfers were not present and the social planner knew the valuations of the agents, how it would pick the alternatives. For instance, the weighted utilitarianism puts weights on agents and maximizes the weighted sum of values of agents.

2.1.2 Payment Functions

A payment function of agent i is a mapping $p_i : T^n \to \mathbb{R}$, where $p_i(t)$ represents the payment of agent i when type profile is $t \in T^n$. Note that $p_i(\cdot)$ can be negative or positive or zero. A positive $p_i(\cdot)$ indicates that the agent is paying money. We will refer to the collection of payment function $p \equiv (p_1, \ldots, p_n)$ as payment rule.

In many situations, we want the total payment of agents to be either non-negative (i.e., decision maker does not incur a loss) or to be zero. In particular, we will be interested in the following kinds of restrictions on payment functions.

1. A payment rule $p \equiv (p_1, \ldots, p_n)$ is **feasible** if $\sum_{i \in N} p_i(t) \geq 0$ for all $t \in T^n$.

2. A payment rule $p \equiv (p_1, \ldots, p_n)$ satisfies **no subsidy** if $p_i(t) \geq 0$ for all $i \in N$ and for all $t \in T^n$.

3. A payment rule $p \equiv (p_1, \ldots, p_n)$ is **balanced** if $\sum_{i \in N} p_i(t) = 0$ for all $t \in T^n$.

These restrictions on payment rules will depend on the context. In many cases, we will impose no restrictions on payments and study what all mechanisms can be incentive compatible. In some cases, we will put specific restrictions and ask similar questions.

2.1.3 Incentive Compatibility

A **social choice function** is a pair $F = (f, p)$, where f is an allocation rule and p is the payment functions of agents. Under a social choice function $F = (f, p)$ the utility of agent $i \in N$ with type t_i when all agents "report" $\hat{t}$ as their types is given by

$$v_i(f(\hat{t}), t_i) - p_i(\hat{t}).$$

The mechanism, as before, is a complicated object. But applying the revelation principle, we will focus on direct mechanisms. A direct mechanism is a social choice function $F = (f, p)$ with type spaces as message space. A direct mechanism (or associated social choice function) is **strategy-proof or dominant strategy incentive compatible (DSIC)** if for every agent $i \in N$, every $t_{-i} \in T_{-i}$, and every $s_i, t_i \in T_i$, we have

$$v_i(f(t_i, t_{-i}), t_i) - p_i(t_i, t_{-i}) \geq v_i(f(s_i, t_{-i}), t_i) - p_i(s_i, t_{-i}),$$

i.e., truth-telling is a dominant strategy. In this case, we will say that the payment functions $(p_1, \ldots, p_n)$ **implement** the allocation rule f (in dominant strategies) or, simply, f is implementable. Sometimes, we will also say that p makes f DSIC.

2.1.4 An Example

Consider an example with two agents $N = \{1, 2\}$ and two possible types for each agent $T_1 = T_2 = \{t^H, t^L\}$. Let $f : T_1 \times T_2 \to A$ be an allocation rule, where A is the set of alternatives. In order that f is implementable, we must find payment functions p_1 and p_2

such that the following conditions hold. For every type $t_2 \in T_2$ of agent 2, agent 1 must satisfy

$$v_1(f(t^H, t_2), t^H) - p_1(t^H, t_2) \geq v_1(f(t^L, t^2), t^H) - p_1(t^L, t_2),$$
$$v_1(f(t^L, t_2), t^L) - p_1(t^L, t_2) \geq v_1(f(t^H, t^2), t^L) - p_1(t^H, t_2).$$

Similarly, for every type $t_1 \in T_2$ of agent 1, agent 2 must satisfy

$$v_2(f(t_1, t^H), t^H) - p_2(t_1, t^H) \geq v_2(f(t^1, t^L), t^H) - p_2(t_1, t^L),$$
$$v_2(f(t_1, t^L), t^L) - p_2(t_1, t^L) \geq v_2(f(t^1, t^H), t^L) - p_2(t_1, t^H).$$

Here, we can treat p_1 and p_2 as variables. The existence of a solution to these linear inequalities guarantee f to be implementable. So, in finite type spaces, incentive compatibility and implementability are just a solutions to some linear system of inequalities.

2.1.5 Two Properties of Payments

Suppose (f, p) is strategy-proof. This means for every agent $i \in N$ and every t_{-i}, we must have

$$v_i(f(t_i, t_{-i}), t_i) - p_i(t_i, t_{-i}) \geq v_i(f(s_i, t_{-i}), t_i) - p_i(s_i, t_{-i}) \qquad \forall\ s_i, t_i \in T_i.$$

Using p, we define another payment rule. For every agent $i \in N$, we choose an arbitrary function $h_i : T_{-i} \to \mathbb{R}$. So, $h_i(t_{-i})$ assigns a real number to every type profile t_{-i} of other agents. Now, define the new payment function q_i of agent i as

$$q_i(t_i, t_{-i}) = p_i(t_i, t_{-i}) + h_i(t_{-i}). \tag{2.1}$$

We will argue the following.

LEMMA **1** *If $(f, p \equiv (p_1, \ldots, p_n))$ is strategy-proof, then (f, q) is strategy-proof, where q is defined as in Equation 2.1.*

Proof: Fix agent i and type profile of other agents at t_{-i}. To show (f, q) is strategy-proof, note that for any pair of types $t_i, s_i \in T_i$, we have

$$\begin{aligned} v_i(f(t_i, t_{-i}), t_i) - q_i(t_i, t_{-i}) &= v_i(f(t_i, t_{-i}), t_i) - p_i(t_i, t_{-i}) - h_i(t_{-i}) \\ &\geq v_i(f(s_i, t_{-i}), t_i) - p_i(s_i, t_{-i}) - h_i(t_{-i}) \\ &= v_i(f(s_i, t_{-i}), t_i) - q_i(s_i, t_{-i}), \end{aligned}$$

where the inequality followed from the fact that (f, p) is strategy-proof. ■

This shows that if we find one payment rule which implements f, then we can find an infinite set of payment rules which implements f. Moreover, these payments differ by a constant for every $i \in N$ and for every t_{-i}. In particular, the payments p and q defined above satisfy the property that for every $i \in N$ and for every t_{-i},

$$p_i(t_i, t_{-i}) - q_i(t_i, t_{-i}) = p_i(s_i, t_{-i}) - q_i(s_i, t_{-i}) = h_i(t_{-i}) \qquad \forall\ s_i, t_i \in T_i.$$

The other property that we discuss of payment rules is the fact that they depend only on allocations.

LEMMA **2 (Taxation principle)** *Suppose (f, p) is strategy-proof. For every $i \in N$ and every t_{-i}*

$$\Big[f(s_i, t_{-i}) = f(t_i, t_{-i})\Big] \Rightarrow \Big[p_i(s_i, t_{-i}) = p_i(t_i, t_{-i})\Big] \qquad \forall\ s_i, t_i$$

Proof: Let (f, p) be strategy-proof. Consider an agent $i \in N$ and a type profile t_{-i}. Let s_i and t_i be two types of agent i such that $f(s_i, t_{-i}) = f(t_i, t_{-i}) = a$. Then, the incentive constraints give us the following.

$$v_i(a, t_i) - p_i(t_i, t_{-i}) \geq v_i(a, t_i) - p_i(s_i, t_{-i})$$
$$v_i(a, s_i) - p_i(s_i, t_{-i}) \geq v_i(a, s_i) - p_i(t_i, t_{-i}).$$

This shows that $p_i(s_i, t_{-i}) = p_i(t_i, t_{-i})$. ■

So, payment is a function of types of other agents and the allocation chosen. This is sometimes referred to as the **taxation principle**.

2.1.6 Efficient Allocation Rule is Implementable

We discussed the efficient allocation rule earlier. Here, we show that there is a large class of payment functions that can implement the efficient allocation rule. First, we show that a mechanism is Pareto efficient if and only if it uses the efficient allocation rule.

DEFINITION **5** *A mechanism (f, p) is* **Pareto optimal** *if at every type profile t there exist no alternative $b \neq f(t)$ and no payment vector $(\pi_1, \ldots, \pi_n)$ such that*

$$v_i(b, t_i) - \pi_i \geq v_i(f(t), t_i) - p_i(t) \qquad \forall\ i \in N,$$
$$\sum_{i \in N} \pi_i \geq \sum_{i \in N} p_i(t).$$

with strict inequality holding for at least one inequality.

Pareto optimality of a mechanism is a tricky issue if we have **no restrictions** on payments. This is because for every mechanism, the designer can just add subsidies for all the agents (i.e., reduce their payments), and make everyone better off at every type profile. This is the necessity to add the second inequality in the above definition.

THEOREM **1** *A mechanism (f, p) is Pareto optimal if and only if f is an efficient allocation rule.*

Proof: Suppose $(f, p_1, \ldots, p_n)$ is Pareto optimal. Assume for contradiction that f is not efficient. Consider a profile t and let the outcome according to the efficient allocation rule be a. Suppose $f(t) = b$ is such that

$$\sum_{i \in N} v_i(a, t_i) > \sum_{i \in N} v_i(b, t_i).$$

Let

$$\delta = \frac{1}{n}\big[\sum_{i \in N} v_i(a, t_i) - \sum_{i \in N} v_i(b, t_i)\big].$$

Note that $\delta > 0$. Define a new payment of agent i as

$$q_i = v_i(a, t_i) - \delta - \big(v_i(b, t_i) - p_i(t)\big).$$

Notice that $\sum_{i \in N} q_i = \sum_{i \in N} p_i(t)$ and $v_i(a, t_i) - q_i = v_i(b, t_i) - p_i(t) + \delta > v_i(b, t_i) - p_i(t)$. Hence, (f, p) is not Pareto optimal, a contradiction.

We now show that if f is efficient, then any mechanism (f, p) with arbitrary payment rule p is Pareto optimal. Assume for contradiction some mechanism (f, p) is not Pareto optimal. So, for some profile t with $f(t) = a \in \arg\max_{x \in A} \sum_{i \in N} v_i(x, t_i)$, there exists an alternative b and some payment vector $(q_1, \ldots, q_n)$ with $\sum_{i \in N} q_i \geq \sum_{i \in N} p_i(t)$ and $v_i(b, t_i) - q_i \geq$

$v_i(a, t_i) - p_i(t)$ for all $i \in N$ with strict inequality holding for at least one of the inequalities. Then, adding it over all $i \in N$, gives $\sum_{i \in N} v_i(b, t_i) - \sum_{i \in N} q_i > \sum_{i \in N} v_i(a, t_i) - \sum_{i \in N} p_i(t)$. Using the fact that $\sum_{i \in N} q_i \geq \sum_{i \in N} p_i(t)$, we get $\sum_{i \in N} v_i(b, t_i) > \sum_{i \in N} v_i(a, t_i)$. This contradicts the definition of a. ■

We will now show that the efficient allocation rule is implementable. We know that in case of sale of a single object efficient allocation rule can be implemented by the second-price payment function. A fundamental result in mechanism design is that the efficient allocation rule is always implementable (under private values and quasi-linear utility functions). For this, a family of payment rules are known which makes the efficient allocation rule implementable. This family of payment rules is known as the *Groves* payment rules, and the corresponding direct mechanisms are known as the **Groves mechanisms** (Groves, 1973).

For agent $i \in N$, for every $t_{-i} \in T_{-i}$, the payment in the Groves mechanism is:

$$p_i^g(t_i, t_{-i}) = h_i(t_{-i}) - \sum_{j \neq i} v_j(f^e(t_i, t_{-i}), t_j),$$

where h_i is any function $h_i : T_{-i} \to \mathbb{R}$ and f^e is the efficient allocation rule.

We give an example in the case of single object auction. Let $h_i(t_{-i}) = 0$ for all i and for all t_{-i}. Let there be four buyers with values (types): 10,8,6,4. Then, efficiency requires us to give the object to the first buyer. Now, the total value of buyers other than buyer 1 in the efficient allocation is zero. Hence, the payment of buyer 1 is zero. The total value of buyers other than buyer 2 (or buyer 3 or buyer 4) is the value of the first buyer (10). Hence, all the other buyers are rewarded 10. Thus, this particular choice of h_i functions led to the auction: the highest bidder wins but pays nothing and those who do not win are awarded an amount equal to the highest bid.

THEOREM **2** *Groves mechanisms are strategy-proof.*

Proof: Consider an agent $i \in N$, $s_i, t_i \in T_i$, and $t_{-i} \in T_{-i}$. Let $f^e(t_i, t_{-i}) = a$ and

$f^e(s_i, t_{-i}) = b$. Then, we have

$$\begin{aligned} v_i(a, t_i) - p_i^g(t_i, t_{-i}) &= \sum_{j \in N} v_j(a, t_j) - h_i(t_{-i}) \\ &\geq \sum_{j \in N} v_j(b, t_j) - h_i(t_{-i}) \\ &= v_i(b, t_i) - \left[h_i(t_{-i}) - \sum_{j \neq i} v_j(b, t_j)\right] \\ &= v_i(b, t_i) - p^g(s_i, t_{-i}), \end{aligned}$$

where the inequality comes from efficiency. Hence, Groves mechanisms are strategy-proof. ■

An implication of this is that efficient allocation rule is implementable using the Groves payment rules. The natural question to ask is whether there are payment rules besides the Groves payment rules which make the efficient allocation rule DSIC. We will study this question formally later. A quick answer is that it depends on the type spaces of agents and the value function. For many reasonable type spaces and value functions, the Groves payment rules are the only payment rules which make the efficient allocation rule DSIC.

2.2 The Vickrey-Clarke-Groves Mechanism

A particular mechanism in the class of Groves mechanism is intuitive and has many nice properties. It is commonly known as the **pivotal mechanism** or the Vickrey-Clarke-Groves (VCG) mechanism (Vickrey, 1961; Clarke, 1971; Groves, 1973). The VCG mechanism is characterized by a unique $h_i(\cdot)$ function. In particular, for every agent $i \in N$ and every $t_{-i} \in T_{-i}$,

$$h_i(t_{-i}) = \max_{a \in A} \sum_{j \neq i} v_j(a, t_j).$$

This gives the following payment function. For every $i \in N$ and for every $t \in T$, the payment in the VCG mechanism is

$$p_i^{vcg}(t) = \max_{a \in A} \sum_{j \neq i} v_j(a, t_j) - \sum_{j \neq i} v_j(f^e(t), t_j). \tag{2.2}$$

Note that $p_i^{vcg}(t) \geq 0$ for all $i \in N$ and for all $t \in T^n$. Hence, the payment function in the VCG mechanism is a feasible payment function.

A careful look at Equation 2.2 shows that the second term on the right hand side is the sum of values of agents other than i in the efficient decision. The first term on the right hand side is the maximum sum of values of agents other than i (note that this corresponds to an efficient decision when agent i is excluded from the economy). Hence, the payment of agent i in Equation 2.2 is the *externality* agent i inflicts on other agents because of his presence, and this is the amount he *pays*. Thus, every agent pays his externality to other agents in the VCG mechanism.

The payoff of an agent in the VCG mechanism has a nice interpretation too. Denote the payoff of agent i in the VCG mechanism when his true type is t_i and other agents report t_{-i} as $\pi_i^{vcg}(t_i, t_{-i})$. By definition, we have

$$\begin{aligned}
\pi_i^{vcg}(t_i, t_{-i}) &= v_i(f^e(t_i, t_{-i}), t_i) - p_i^{vcg}(t_i, t_{-i}) \\
&= v_i(f^e(t_i, t_{-i}), t_i) - \max_{a \in A} \sum_{j \neq i} v_j(a, t_j) + \sum_{j \neq i} v_j(f^e(t_i, t_{-i}), t_j) \\
&= \max_{a \in A} \sum_{j \in N} v_j(a, t_j) - \max_{a \in A} \sum_{j \neq i} v_j(a, t_j),
\end{aligned}$$

where the last equality comes from the definition of efficiency. The first term is the total value of *all* agents in an efficient allocation rule. The second term is the total value of *all agents except agent* i in an efficient allocation rule of the economy in which agent i is absent. Hence, payoff of agent i in the VCG mechanism is his **marginal contribution** to the economy.

2.2.1 Illustration of the VCG (Pivotal) Mechanism

Consider the sale of a single object using the VCG mechanism. Fix an agent $i \in N$. Efficiency says that the object must go to the bidder with the highest value. Consider the two possible cases. In one case, bidder i has the highest value. So, when bidder i is present, the sum of values of other bidders is zero (since no other bidder wins the object). But when bidder i is absent, the maximum sum of value of other bidders is the second highest value (this is achieved when the second highest value bidder is awarded the object). Hence, the externality of bidder i is the second-higest value. In the case where bidder $i \in N$ does not have the highest value, his externality is zero. Hence, for the single object case, the VCG mechanism

	$\emptyset$	$\{1\}$	$\{2\}$	$\{1,2\}$
$v_1(\cdot)$	0	8	6	12
$v_2(\cdot)$	0	9	4	14

Table 2.1: An Example of VCG Mechanism with Multiple Objects

is simple: award the object to the bidder with the highest (bid) value and the winner pays the amount equal to the second highest (bid) value but other bidders pay nothing. This is the well-known second-price auction or the Vickrey auction. By Theorem 2, it is strategy-proof.

Consider the case of choosing a public project. There are three possible projects - an opera house, a park, and a museum. Denote the set of projects as $A = \{a, b, c\}$. The citizens have to choose one of the projects. Suppose there are three citizens, and the values of citizens are given as follows (row vectors are values of citizens and columns have three alternatives, a first, b next, and c last column):

$$\begin{bmatrix} 5 & 7 & 3 \\ 10 & 4 & 6 \\ 3 & 8 & 8 \end{bmatrix}$$

It is clear that it is efficient to choose alternative b. To find the payment of agent 1 according to the VCG mechanism, we find its externality on other agents. Without agent 1, agents 2 and 3 can get a maximum total value of 14 (on project c). When agent 1 is included, their total value is 12. So, the externality of agent 1 is 2, and hence, its VCG payment is 2. Similarly, the VCG payments of agents 2 and 3 are respectively 0 and 4.

We illustrate the VCG mechanism for the sale of multiple objects by an example. Consider the sale of two objects, with values of two agents on bundles of goods given in Table 2.1. The efficient allocation in this example is to give bidder 1 object 2 and bidder 2 object 1 (this generates a total value of $6 + 9 = 15$, which is higher than any other allocation). Let us calculate the externality of bidder 1. The total value of bidders other than bidder 1, i.e. bidder 2, in the efficient allocation is 9. When bidder 1 is removed, bidder 2 can get a maximum value of 14 (when he gets both the objects). Hence, externality of bidder 1 is $14 - 9 = 5$. Similarly, we can compute the externality of bidder 2 as $12 - 6 = 6$. Hence, the payments of bidders 1 and 2 are 5 and 6 respectively.

Another simpler combinatorial auction setting is when agents or bidders are interested (or can be allocated) in at most one object - this is the case in job markets or housing markets.

	$\emptyset$	$\{1\}$	$\{2\}$
$v_1(\cdot)$	0	5	3
$v_2(\cdot)$	0	3	4
$v_3(\cdot)$	0	2	2

Table 2.2: An Example of VCG Mechanism with Multiple Objects

Then, every bidder has a value for every object but wants at most one object. Consider an example with three agents and two objects. The valuations are given in Table 2.2. The total value of agents in the efficient allocation is $5+4=9$ (agent 1 gets object 1 and agent 2 gets object 2, but agent 3 gets nothing). Agents 2 and 3 get a total value of $4+0=4$ in this efficient allocation. When we maximize over agents 2 and 3 only, the maximum total value of agents 2 and 3 is $6=4+2$ (agent 2 gets object 2 and agent 3 gets object 1). Hence, externality of agent 1 on others is $6-4=2$. Hence, VCG payment of agent 1 is 2. Similarly, one can compute the VCG payment of agent 2 to be 2.

2.2.2 The VCG Mechanism in the Combinatorial Auctions

We have already shown that the VCG mechanism has several interesting properties: (a) it is dominant strategy incentive compatible, (b) the allocation rule is efficient, and (c) payments are non-negative, and hence, feasible. We discuss below a specific model and show that stronger properties than these are also true in this model.

The particular model we discuss is the combinatorial auction problem. We now describe the formal model. There is a set of objects $M=\{1,\ldots,m\}$. The set of *bundles* is denoted by $\Omega=\{S:S\subseteq M\}$. The type of an agent $i\in N$ is a vector $t_i\in\mathbb{R}_+^{|\Omega|}$. Hence, $T_1=\ldots=T_n=\mathbb{R}_+^{|\Omega|}$. Here, $t_i(S)$ denotes the value of agent (bidder) i on bundle S - in particular, $t_i(S)\geq 0 \forall S\in\Omega, \forall i\in N$ (i.e. all bundles are goods). An allocation in this case is a partitioning of the set of objects: $X=(X_0,X_1,\ldots,X_n)$, where $X_i\cap X_j=\emptyset$ and $\cup_{i=0}^n X_i=M$. Here, X_0 is the unallocated set of objects and X_i $(i\neq 0)$ is the bundle allocated to agent i, where X_i can be empty set also. It is natural to normalize $t_i(\emptyset)=0$ for all t_i and for all i.

Let f^e be the efficient allocation rule. Another crucial feature of the combinatorial auction setting is it is *externality free*. Suppose $f^e(t)=X$. Then $v_i(X,t_i)=t_i(X_i)$, i.e., utility of agent i depends on the bundle allocated to agent i only, but not on the bundles allocated to

other agents.

The first property of the VCG mechanism we note in this setting is that the *losers* pay zero amount. Suppose i is a *loser* (i.e., gets empty bundle in efficient allocation) when the type profile is $t = (t_1, \ldots, t_n)$. Let $f^e(t) = X$. By assumption, $v_i(X_i, t_i) = t_i(\emptyset) = 0$. Let $Y \in \arg\max_a \sum_{j \neq i} v_j(a, t_j)$. We need to show that $p_i^{vcg}(t_i, t_{-i}) = 0$. Since the VCG mechanism is feasible, we know that $p_i^{vcg}(t_i, t_{-i}) \geq 0$. Now,

$$\begin{aligned}
p_i^{vcg}(t_i, t_{-i}) &= \max_{a \in A} \sum_{j \neq i} v_j(a, t_j) - \sum_{j \neq i} v_j(f^e(t_i, t_{-i}), t_j) \\
&= \sum_{j \neq i} t_j(Y_j) - \sum_{j \neq i} t_j(X_j) \\
&\leq \sum_{j \in N} t_j(Y_j) - \sum_{j \in N} t_j(X_j) \\
&\leq 0,
\end{aligned}$$

where the first inequality followed from the facts that $t_i(Y_i) \geq 0$ and $t_i(X_i) = 0$, and the second inequality followed from the efficiency of X. Hence, $p_i^{vcg}(t_i, t_{-i}) = 0$.

An important property of a mechanism is **individual rationality** or **voluntary participation**. Suppose by not participating in a mechanism an agent gets zero payoff. Then the mechanism must give non-negative payoff to the agent in every state of the world (i.e., in every type profile of agents). The VCG mechanism in the combinatorial auction setting satisfies individual rationality. Consider a type profile $t = (t_1, \ldots, t_n)$ and an agent $i \in N$. Let $Y \in \arg\max_a \sum_{j \neq i} v_j(a, t_j)$ and $X \in \arg\max_a \sum_{j \in N} v_j(a, t_j)$. Now,

$$\begin{aligned}
\pi_i^{vcg}(t) &= \max_a \sum_{j \in N} v_j(a, t_j) - \max_a \sum_{j \neq i} v_j(a, t_j) \\
&= \sum_{j \in N} t_j(X_j) - \sum_{j \neq i} t_j(Y_j) \\
&\geq \sum_{j \in N} t_j(X_j) - \sum_{j \in N} t_j(Y_j) \\
&\geq 0,
\end{aligned}$$

where the first inequality followed from the fact that $t_j(Y_j) \geq 0$ and the second inequality followed from efficiency of X. Hence, $\pi_i^{vcg}(t) \geq 0$, i.e., the VCG mechanism is individual rational.

2.2.3 The Sponsored Search Auctions

Google sells advertisement slots to advertisers via auctions. The auctions are run for **every** search phrase. Fix a particular search phrase, say, "hotels in New Delhi". Once this phrase is searched on Google, bidders (computer programmed agents of different companies) participate in this auction. An advertisement that can appear along side a search page is called a **slot**. For every search phrase, there is a fixed number of slots available and fixed number of bidders interested. Suppose there are m slots and n bidders for the phrase "hotels in New Delhi". Assume $n > m$. The type of each bidder is a single number - θ_i for bidder i. Type of an agent represents the value that agent derives when his advertisement is **clicked**. Every slot has a probability of getting clicked. This is called the **clickthrough rate (CTR)**. The CTR of slot i is α_i. The CTR vector $\alpha = (\alpha_1, \ldots, \alpha_m)$ is known to everyone. The slots are naturally ordered top to bottom, and assume that, let $\alpha_1 > \alpha_2 > \ldots > \alpha_m$. CTR is assumed to be common knowledge among bidders - apparently, Google estimates CTR from data and gives it to bidders.

An alternative in this model represents an assignment of agents to slots (with some agents not receiving any slot). Let A be the set of all alternatives. An alternative $a \in A$ can be described by a n dimensional vector integers in $\{0, 1, \ldots, m\}$, where a_i indicates the slot to which agent i is assigned, and $a_i = 0$ means agent i is not assigned to any slot. The value function of agent i is his expected value $v_i(a, \theta_i) = \theta_i \alpha_{a_i}$, where $\alpha_0 = 0$.

Suppose $n = 4$ and $m = 3$. Let $\theta_1 = 10, \theta_2 = 8, \theta_3 = 6, \theta_4 = 5$. Let $\alpha_1 = 0.8, \alpha_2 = 0.6, \alpha_3 = 0.5$. In efficiency, the slots should go to agents with top 3 values of θ, who are agents $1, 2$, and 3.

The total value obtained in the efficient allocation is $10(0.8) + 8(0.6) + 6(0.5) = 15.8$. So, agents other than agent 1 get a total value of $8(0.6) + 6(0.5) = 7.8$. If agent 1 was not there, then the total value obtained in the efficient allocation is $8(0.8) + 6(0.6) + 5(0.5) = 12.5$. Hence, his externality is $12.5 - 7.8 = 4.7$, and his VCG payment is thus 4.7. Similarly, VCG payments of agents 2 and 3 are 3.1 and 2.5 respectively.

Generally, agents with top m values of θ get the top m slots with ith ($i \leq m$) highest θ value agent getting the ith slot. Without loss of generality, assume that $\theta_1 \geq \theta_2 \geq \ldots \theta_n$ - assume also that ties are broken in favor of lower indexed agent. In efficiency, agents 1 to m get a slot. In particular, agent j ($j \leq m$) gets slot j with clickthrough rate α_j. Any

agent j pays zero if he is not allocated a slot, i.e., $j > m$. For any agent $j \leq m$, we need to compute his externality. Note that the total value of agents other than agent j in an efficient allocation is

$$\sum_{i=1}^{j-1} \theta_i \alpha_i + \sum_{i=j+1}^{m} \theta_i \alpha_i.$$

If agent j is removed, then the total value of agents other than agent j in an efficient allocation is

$$\sum_{i=1}^{j-1} \theta_i \alpha_i + \sum_{i=j+1}^{m+1} \theta_i \alpha_{i-1}.$$

So, the externality of agent j is

$$\theta_{m+1}\alpha_m + \sum_{i=j+1}^{m} \theta_i(\alpha_{i-1} - \alpha_i),$$

where we assume that the summation term for $j = m$ is zero.

Google uses something called a **Generalized Second Price (GSP)** auction: (a) agents with top m values of θ are given the slots with highest agent getting the top slot (i.e., slot with highest CTR), second highest agent getting the next top slot, and so on, (b) if an agent wins slot k with CTR α_k, he pays $\theta_{m+1}\alpha_k$ (where θ_{m+1} is the highest losing type).

In the previous example, agent 1 will pay $5(0.8) = 4$ in the GSP. This is clearly different from what he should pay in the VCG mechanism. In the example above, fix the bids of agents other than agent 2 as follows: (agent 1: 10, agent 3: 6, agent 4: 5). Now, let agent 2 not bid truthfully, and bid $10 + \epsilon$ ($\epsilon > 0$) to become the highest bidder. So, he gets the top slot with clickthrough rate 0.8. So, his value is now $8 \times 0.8 = 6.4$ (remember, his true type is $\theta_2 = 8$). He pays $5 \times 0.8 = 4$. So, his net utility is 2.4. If he is truthful, he pays $5 \times 0.6 = 3$, and gets a value of $8 \times 0.6 = 4.8$. So, his net utility of being truthful is 1.8. So, deviation is profitable, and truth-telling is not a dominant strategy.

2.3 Affine Maximizer Allocation Rules are Implementable

The previous discussions showed that Groves mechanisms can implement efficient allocation rules if there are no constraints on payments. However, in many instances, we may be

interested in exploring a non-efficient allocation rule. There are many reasons for this. One important reason is that Groves mechanisms may not be feasible if there are restrictions on payments - for instance, budget-balance. In such cases, if we know what other allocation rules are implementable, we can explore if they can be implementable with restricted payments. Another reason is that efficiency treats all equal (in terms of preferences) individuals in a society equally. This need not be the case in many settings, where we may be asked to do "affirmative action", and give more weightage to certain individuals than others. Another reason may be that the planner himself may not be a utilitarian planner, and depending on his objective, he may choose a non-efficient allocation rule.

Below, we discuss another general class of allocation rules which are implementable. These are the affine maximizer class discussed earlier. As discussed earlier, an affine maximizer allocation rule is characterized by a vector of non-negative weights $\lambda \equiv (\lambda_1, \ldots, \lambda_n)$, not all equal to zero, for agents and a mapping $\kappa : A \to \mathbb{R}$. Then, at any type profile t,

$$f^a(t) \arg\max_{a \in A} \Big[\sum_{i \in N} \lambda_i v_i(a, t_i) - \kappa(a) \Big]$$

If $\lambda_i = \lambda_j$ for all $i, j \in N$ and $\kappa(a) = 0$ for all $a \in A$, we recover the efficient allocation rule. When $\lambda_i = 1$ for some $i \in N$ and $\lambda_j = 0$ for all $j \neq i$, and $\kappa(a) = 0$ for all $a \in A$, we get the dictatorial allocation rule. Thus, the affine maximizer is a general class of allocation rules. We show that there exists payment rules which implements the affine maximizer allocation rule. For this we only consider a particular class of affine maximizers.

DEFINITION **6** *An affine maximizer allocation rule f^a with weights $\lambda_1, \ldots, \lambda_n$ and $\kappa : A \to \mathbb{R}$ satisfies* **independence of irrelevant agents (IIA)** *if for all $i \in N$ with $\lambda_i = 0$, we have that for all t_{-i} and for all s_i, t_i, $f(s_i, t_{-i}) = f(t_i, t_{-i})$.*

The IIA property is a consistent tie-breaking requirement. For instance, consider the dictatorship allocation rule with two agents $\{1, 2\}$. Suppose agent 1 is the dictator: $\lambda_1 = 1, \lambda_2 = 0$ and suppose there are three alternatives $\{a, b, c\}$. Since the allocation rule is a dictatorship, $\kappa(a) = \kappa(b) = \kappa(c) = 0$. The type of each agent is a vector in $\mathbb{R}^3$ describing the value for each alternative. For instance $t_1 = (5, 5, 3)$ means, agent 1 has value 5 for alternatives a and b and value 3 for alternative c. Since values on alternatives can be the

same, we can break the ties in this dictatorship by considering values of agent 2. In particular, if there are more than one alternatives that maximize the value for agent 1, then we choose an alternative that is the worst for agent 2. For instance, if $t_1 = (5, 5, 3)$ and $t_2 = (4, 3, 2)$, then $f(t_1, t_2) = b$ (since $t_2(b) = 3 < t_2(a) = 4$). But then, consider $t_2' = (3, 4, 2)$ and note that $f(t_1, t_2') = a$. This is a violation of IIA - in fact, agent 2 can manipulate dictatorship by reporting t_2' when his true type is t_2.

Allocation rules violating IIA may not be implementable (i.e., there may not exist payment rules that make the resulting mechanism strategy-proof). However, we show that every IIA affine maximizer is implementable. Fix an IIA affine maximizer allocation rule f^a, characterized by λ and κ. We generalize Groves payments for this allocation rule.

For agent $i \in N$, for every $t_{-i} \in T_{-i}$, the payment in the *generalized* Groves mechanism is:

$$p_i^{gg}(t_i, t_{-i}) = \begin{cases} h_i(t_{-i}) - \frac{1}{\lambda_i}\big[\sum_{j \neq i} \lambda_j v_j(f^a(t_i, t_{-i}), t_j) - \kappa(f^a(t_i, t_{-i}))\big] & \text{if } \lambda_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

where h_i is any function $h_i : T_{-i} \to \mathbb{R}$ and f^a is the IIA affine maximizer allocation rule.

THEOREM **3** *An IIA affine maximizer allocation rule is implementable using the generalized Groves mechanism.*

Proof: Consider an agent $i \in N$, $s_i, t_i \in T_i$, and $t_{-i} \in T_{-i}$. Suppose $\lambda_i > 0$. Then, we have

$$\begin{aligned} v_i(f^a(t_i, t_{-i}), t_i) - p_i^{gg}(t_i, t_{-i}) &= \frac{1}{\lambda_i}\big[\sum_{j \in N} \lambda_j v_j(f^a(t_i, t_{-i}), t_j) - \kappa(f^a(t_i, t_{-i}))\big] - h_i(t_{-i}) \\ &\geq \frac{1}{\lambda_i}\big[\sum_{j \in N} \lambda_j v_j(f^a(s_i, t_{-i}), t_j) - \kappa(f^a(s_i, t_{-i}))\big] - h_i(t_{-i}) \\ &= v_i(f^a(s_i, t_{-i}), t_i) - h_i(t_{-i}) + \frac{1}{\lambda_i}\big[\sum_{j \neq i} \lambda_j v_j(f^a(s_i, t_{-i}), t_j) + \kappa(f^a(s_i, t_{-i}))\big] \\ &= v_i(f^a(s_i, t_{-i}), t_i) - p^{gg}(s_i, t_{-i}), \end{aligned}$$

where the inequality comes from the definition of affine maximization. If $\lambda_i = 0$, then $f^a(t_i, t_{-i}) = f^a(s_i, t_{-i})$ for all $s_i, t_i \in T_i$ (by IIA). Also $p_i^{gg}(t_i, t_{-i}) = p_i^{gg}(s_i, t_{-i}) = 0$ for all $s_i, t_i \in T_i$. Hence, $v_i(f^a(t_i, t_{-i}), t_i) - p_i^{gg}(t_i, t_{-i}) = v_i(f^a(s_i, t^{-i}), t_i) - p_i^{gg}(s_i, t_{-i})$. So, the generalized Groves payment rule implements the affine maximizer allocation rule. ■

2.3.1 Public Good Provision

The public good provision problem is a classic problem. There are two alternatives: a_1 is the alternative to provide the public good and a_0 is the alternative of not providing the public good. The value from a_0 is zero to all the agents. Agents derive value from a_1 which is private information. Denote the value of agent i for a_1 as θ_i ($\theta_i \geq 0 \forall i \in N$). There is a cost of C providing the public good.

The "first-best" allocation rule in this case is to provide the public good when the sum of values of agents is greater than or equal to C. This can be written as an affine maximizer rule. Choose $\kappa(a_0) = 0, \kappa(a_1) = -C$ and $\lambda_i = 1$ for all $i \in N$, where N is the set of agents.

The pivotal mechanism corresponding to this allocation rule is the first one that Clarke called the pivotal mechanism. An agent i is **pivotal** if his inclusion in the decision process changes the decision for the other $N \setminus \{i\}$ agents. In particular, if agents in $N \setminus \{i\}$ chose not to be provided the public good using the first-best rule, and when agent i was added, agents in N chose to get the public good using the first-best rule. Here, agent i is pivotal. Note that if agents in $N \setminus \{i\}$ chose to get the public good using the first-best rule, and when agent i is added, agents in N will always choose to get the public good using the first-best rule. Hence, agent i cannot be pivotal here.

The pivotal mechanism in this problem states that an agent i pays zero if he is not pivotal and pays an amount equal to his externality if he is pivotal. The externality can be computed easily. Note that at a type profile $\theta \equiv (\theta_1, \ldots, \theta_n)$, if the public good is not provided, then it will not be provided without any agent. Hence, no agent is pivotal and payment of all the agents are zero. But if the public good is provided at θ and agent i is pivotal, then removing agent i changes the decision to not provide the public good. This implies that $\sum_{j \neq i} \theta_j < C$. Hence, without agent i, the total utility to all the agents in $N \setminus \{i\}$ is zero. Once, agent i arrives, their total utility is $\sum_{j \neq i} \theta_j - C$. Hence, his payment is $C - \sum_{j \neq i} \theta_j$.

Now, it is easy to verify that this corresponds to the payment we described in the previous section, where we take $h_i(\theta_{-i})$ to be the maximum sum of values without agent i in the first-best allocation rule.

2.3.2 Restricted and Unrestricted Type Spaces

Consider a simple model where $t_i \in \mathbb{R}^{|A|}$, where A is finite and $v_i(a, t_i) = t_i(a)$ for all $i \in N$. So, the type space of agent i is now $T_i \subseteq \mathbb{R}^{|A|}$. We say type space T_i of agent i is **unrestricted** if $T_i = \mathbb{R}^{|A|}$. So, all possible vectors in $\mathbb{R}^{|A|}$ is likely to be the type of agent i if its type space is unrestricted. Notice that it is an extremely restrictive assumption. We give two examples where unrestricted type space assumption is **not** natural.

- Choosing a public project. Suppose we are given a set of public projects to choose from. Each of the possible public projects (alternatives) is a "good" and not a "bad". In that case, it is natural to assume that the value of an agent for any alternative is non-negative. Further, it is reasonable to assume that the value is bounded. Hence, $T_i \subseteq \mathbb{R}_+^{|A|}$ for every agent $i \in N$. So, unrestricted type space is not a natural assumption here.

- Auction settings. Consider the sale of a single object. The alternatives in this case are $A = \{a_0, a_1, \ldots, a_n\}$, where a_0 denote the alternative that the object is not sold to any agent and a_i with $i > 0$ denotes the alternative that the object is sold to agent i. Notice here that agent i has **zero** value for all the alternatives except alternative a_i. Hence, the unrestricted type space assumption is not valid here.

Are there problems where the unrestricted type space assumption is natural? Suppose the alternatives are such that it can be a "good" or "bad" for the agents, and any possible value is plausible. If we accept the assumption of unrestricted type spaces, then the following is an important theorem. We skip the long proof.

Theorem **4 (Roberts' theorem)** *Suppose A is finite and $|A| \geq 3$. Further, type space of every agent is unrestricted. Then, every onto and implementable allocation rule is an affine maximizer.*

We have already shown that IIA affine maximizers are implementable by constructing generalized Groves payments which make them DSIC. Roberts' theorem shows that these are almost the entire class. The assumptions in the theorem are crucial. If we relax unrestricted type spaces or let $|A| = 2$ or allow randomization, then the set of DSIC allocation rules are larger.

It is natural to ask why restricted type spaces allow for larger class of allocation rules to be DSIC. The answer is very intuitive. Remember that the type space is something that the mechanism designer knows (about the range of private types of agents). If the type space is restricted then the mechanism designer has more precise information about the types of the agents. So, there is *less opportunity* for an agent to lie. Given an allocation rule f if we have two type spaces T and $\bar{T}$ with $T \subsetneq \bar{T}$, then it is possible that f is DSIC in T but not in $\bar{T}$ since $\bar{T}$ allows an agent a larger set of type vectors where it can deviate. In other words, the set of constraints in the DSIC definition is larger for $\bar{T}$ then for T. So, finding payments to make f DSIC is difficult for larger type spaces but easier for smaller type spaces. Hence, the set of DSIC allocation rules becomes larger as we shrink the type space of agents.

Chapter 3

Mechanism Design for Selling a Single Object

3.1 The Single Object Auction Model

In the single object auction case, the type set of an agent is one dimensional, i.e., $T_i \subseteq \mathbb{R}^1$ for all $i \in N$. This reflects the value of an agent if he wins the object. An allocation gives a probability of winning the object. Let A denote the set of all **deterministic** allocations (i.e., allocations in which the object either goes to a single agent or is unallocated). Let ΔA denote the set of all probability distributions over A. An allocation rule is now a mapping $f : T^n \to \Delta A$.

Given an allocation, $a \in \Delta A$, we denote by a_i the allocation probability of agent i. It is standard to have $v_i(a, s_i) = a_i \times s_i$ for all $a \in \Delta A$ and $s_i \in T_i$ for all $i \in N$. Such a form of v_i is called a **product form**.

For an allocation rule f, we denote $f_i(t_i, t_{-i})$ as the probability of winning the object of agent i when he reports t_i and others report t_{-i}.

3.1.1 The Vickrey Auction

Before analyzing a single object sale mechanism, we first take a look at the Vickrey auction. Consider the Vickrey auction and an agent i. Denote the highest valuation among agents in $N \setminus \{i\}$ as $v^{(2)}$. Suppose the valuation of agent i is v_i. Then, according to the rules of the

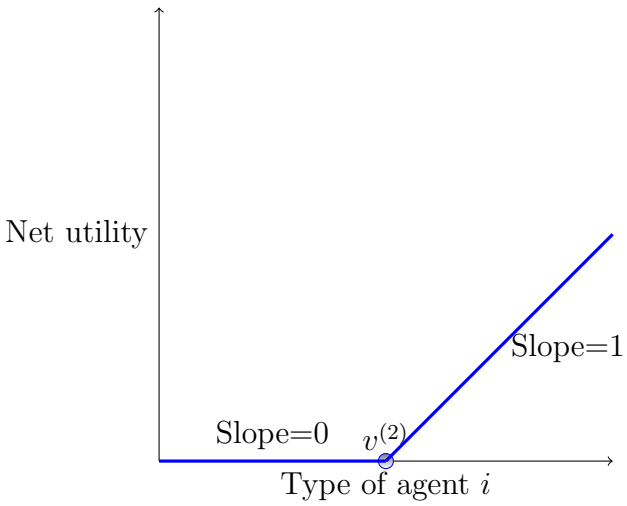

Figure 3.1: Net utility as a function of type of agent i

Vickrey auction, agent i does not win the object if $v_i < v^{(2)}$ and wins the object if $v_i > v^{(2)}$. Further, his net utility from the Vickrey auction is zero if he does not win the object. If he wins the object, then his net utility is $v_i - v^{(2)}$, i.e., increases linearly with v_i.

If we draw the net utility as a function of v_i, it will look something like in Figure 3.1. Notice that this function is *convex* and its derivative is zero if $v_i < v^{(2)}$ and 1 if $v_i > v^{(2)}$. This function is not differentiable at $v_i = v^{(2)}$. Hence, the derivative of the net utility function (wherever it exists) coincides with the probability of winning the object - see Figure 3.1. Since a convex function is differentiable *almost everywhere*, this fact is true almost everywhere.

These observations hold in general, and it is true for *any* dominant strategy incentive compatible mechanism. To show this, we first record some elementary facts from convex analysis.

3.1.2 Facts from Convex Analysis

We will state some basic facts about convex functions. We will only be interested in functions of the form $g : I \to \mathbb{R}$, where $I \subseteq \mathbb{R}$ is an interval.

DEFINITION **7** *A function $g : I \to \mathbb{R}$ is* **convex** *if for every $x, y \in I$ and for every $\lambda \in (0, 1)$, we have*

$$\lambda g(x) + (1 - \lambda)g(y) \geq g(\lambda x + (1 - \lambda)y).$$

Convex functions are continuous in the interior of its domain. So, if $g : I \to \mathbb{R}$ is convex, then g is continuous in the interior of I. Further, g is differentiable *almost everywhere* in I. More formally, there is a subset of $I' \subseteq I$ such that I' is dense in I, $I \setminus I'$ has measure zero [1], and g is differentiable at every point in I'. If g is differentiable at $x \in I$, we denote the derivative of g at x as $g'(x)$. The following notion extends the idea of a derivative.

DEFINITION **8** *For any $x \in I$, x^* is a* **subgradient** *of the function $g : I \to \mathbb{R}$ at x if*

$$g(z) \geq g(x) + x^*(z - x) \qquad \forall\ z \in I.$$

LEMMA **3** *Suppose $g : I \to \mathbb{R}$ is a convex function. Suppose x is in the interior of I and g is differentiable at x, then $g'(x)$ is the unique subgradient of g at x.*

Proof: Consider any $x \in I$ in the interior of I such that the convex function $g : I \to \mathbb{R}_+$ is differentiable at x. Now, pick any $z \in I$. Assume that $z > x$ (a similar proof works if $z < x$). For any $(z - x) \geq h > 0$, we note that $x + h = \frac{h}{(z-x)}z + (1 - \frac{h}{(z-x)})x$. As a result, convexity of g ensures that

$$\frac{h}{(z-x)}g(z) + (1 - \frac{h}{(z-x)})g(x) \geq g(x + h).$$

Simplifying, we get

$$\frac{g(z) - g(x)}{(z - x)} \geq \frac{g(x + h) - g(x)}{h}.$$

Since this is true for any $h > 0$, it is also true that

$$\frac{g(z) - g(x)}{(z - x)} \geq \lim_{h \to 0} \frac{g(x + h) - g(x)}{h} = g'(x).$$

Hence, $g'(x)$ is a subgradient of g at x. This also shows that there is at least one subgradient of g at x.

To show uniqueness, suppose there is another subgradient $x^* \neq g'(x)$ at x. Suppose $x^* > g'(x)$. Then, for all $h > 0$, we know that

$$\frac{g(x + h) - g(x)}{h} \geq x^* > g'(x).$$

But since this is true for all $h > 0$, we have that

$$g'(x) = \lim_{h \to 0} \frac{g(x + h) - g(x)}{h} \geq x^* > g'(x),$$

[1] This does not necessarily mean that $I \setminus I'$ is countable - an example is the Cantor set.

which is a contradiction.

Suppose $x^* < g'(x)$. Then, for all $h > 0$, we know that

$$g(x-h) \geq g(x) - x^* h.$$

Equivalently,

$$\frac{g(x) - g(x-h)}{h} \leq x^*.$$

Since this is true for all $h > 0$, we have that

$$g'(x) = \lim_{h \to 0} \frac{g(x) - g(x-h)}{h} \leq x^*.$$

This is a contradiction. ■

Lemma 3 extends in the following natural way.

LEMMA **4** *Suppose $g : I \to \mathbb{R}$ is a convex function. Then for every $x \in I$, the subgradient of g at x exists.*

We skip the proof of Lemma 4. Lemma 3 showed that if g is differentiable at x and x is in the interior, then $g'(x)$ is the unique subgradient. For all other points in $I \setminus I'$ (which is a set of measure zero), the set of subgradients can be shown to be a convex set. In particular, if x is an interior point of I where g is not differentiable, then we can define $g'_+(x) = \lim_{z \to x : z \in I', z > x} g'(z)$ and $g'_-(x) = \lim_{z \to x : z \in I', z < x} g'(z)$, where I' is the set of points where g is differentiable. These limits exist since the set of points where g is differentiable is dense in I. One can easily show that $g'_+(x) \geq g'_-(x)$. We can then show that the set of subgradients of g at x is $[g'_-(x), g'_+(x)]$.

The set of subgradients of g at a point $x \in I$ will be denoted by $\partial g(x)$. By Lemma 3, $\partial g(x)$ is equal to $\{g'(x)\}$ if $x \in I'$ and by Lemma 4, it is non-empty otherwise. The following lemma is crucial.

LEMMA **5** *Suppose $g : I \to \mathbb{R}$ is a convex function. Let $\phi : I \to \mathbb{R}$ such that $\phi(z) \in \partial g(z)$ for all $z \in I$. Then, for all $x, y \in I$ such that $x > y$, we have $\phi(x) \geq \phi(y)$.*

Proof: By definition, $g(x) \geq g(y) + \phi(y)(x-y)$ and $g(y) \geq g(x) + \phi(x)(y-x)$. Adding these two inequalities, we get $(x-y)(\phi(x) - \phi(y)) \geq 0$. Since $x > y$, we get $\phi(x) \geq \phi(y)$. ■

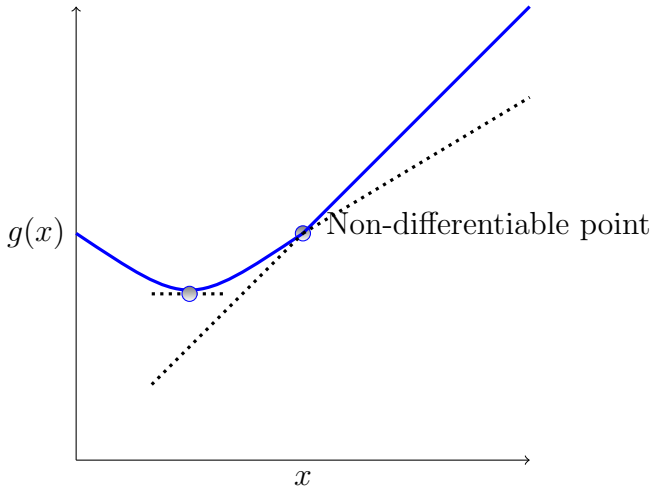

Figure 3.2: A convex function and its subgradients

As a corollary to Lemma 5, we get that if g is differentiable at x and y and $x > y$, then we have $g'(x) \geq g'(y)$. This also shows that for any $x \in I$, $g'_+(x) \geq g'_-(x)$.

Figure 3.2 illustrates the idea. It shows a convex function g and two points in its domain. The left one is a point where g is differentiable and its unique subgradient is shown in Figure 3.2. On the other hand, the right one is a point where g is not differentiable. Figure 3.2 shows the least subgradient and the maximum subgradient at that point. Any selection from that cone will be a suitable subgradient at that point.

If g is differentiable everywhere, then g can be written as the definite integral of its derivative. In particular, if $x, y \in I$, then $g(x) = g(y) + \int_y^x g'(z)dz$. However, this can be extended easily to convex functions since a convex function is differentiable almost everywhere. The following lemma establishes that. We skip its proof.

LEMMA **6** *Let $g : I \to \mathbb{R}$ be a convex function. Then, for any $x, y \in I$,*

$$g(x) = g(y) + \int_y^x \phi(z)dz,$$

where $\phi : I \to \mathbb{R}$ is a map satisfying $\phi(z) \in \partial g(z)$ for all $z \in I$.

3.1.3 Monotonicity and Revenue Equivalence

We now use the facts from convex analysis to establish a fundamental theorem in single object auction analysis. A crucial property that we will use is the following monotonicity

property of allocation rules.

DEFINITION **9** *An allocation rule f is called* **non-decreasing** *if for every agent $i \in N$ and every $t_{-i} \in T_{-i}$ we have $f_i(t_i, t_{-i}) \geq f_i(s_i, t_{-i})$ for all $s_i, t_i \in T_i$ with $s_i < t_i$.*

A non-decreasing allocation rule satisfies a simple property. For every agent and for every report of other agents, the probability of winning the object does not decrease with increase in type of this agent. Figure 3.3 shows a non-decreasing allocation rule.

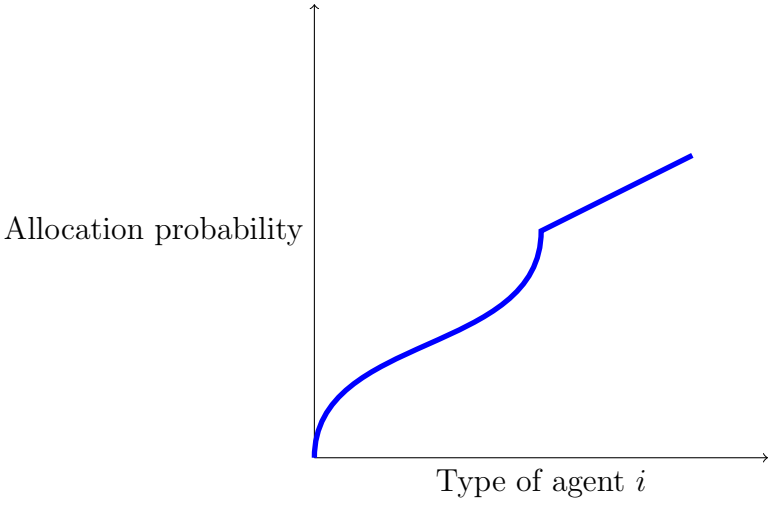

Figure 3.3: Non-decreasing allocation rule

This property characterizes the set of implementable allocation rules in this case.

THEOREM **5** *Suppose T_i is an interval $[0, b_i]$ for all $i \in N$ and v is in product form. An allocation rule $f : T^n \to \Delta A$ and a payment rule $(p_1, \ldots, p_n)$ is DSIC if and only if f is non-decreasing and for all $i \in N$, for all $t_{-i} \in T^{n-1}$, and for all $t_i \in T_i$*

$$p_i(t_i, t_{-i}) = p_i(0, t_{-i}) + t_i f_i(t_i, t_{-i}) - \int_0^{t_i} f_i(x_i, t_{-i}) dx_i.$$

Proof: Given a mechanism $M \equiv (f, p_1, \ldots, p_n)$, the indirect utility function of agent i from the mechanism M when other agents report t_{-i} is defined as

$$\mathcal{U}_i^M(t_i, t_{-i}) = t_i f_i(t_i, t_{-i}) - p_i(t_i, t_{-i}) \quad \forall\ t_i \in T_i.$$

The indirect utility is the net utility of agent i by reporting his true type (given the reports of other agents). Using $\mathcal{U}^M$, we can rewrite the incentive constraints as follows. Mechanism M

is dominant strategy incentive compatible if and only if for all $i \in N$ and for all $t_{-i} \in T^{n-1}$, we have

$$\mathcal{U}_i^M(t_i, t_{-i}) \geq \mathcal{U}_i^M(s_i, t_{-i}) + f_i(s_i, t_{-i})(t_i - s_i) \qquad \forall\ s_i, t_i \in T_i.$$

Now, fix an agent $i \in N$ and t_{-i}. Suppose mechanism $M \equiv (f, p_1, \ldots, p_n)$ is DSIC. We do the proof in some steps.

Step 1 - Subgradient. Define $g(t_i) := \mathcal{U}_i^M(t_i, t_{-i})$ and $\phi(t_i) := f_i(t_i, t_{-i})$ for all $t_i \in T_i$. Then, DSIC implies that for all $s_i, t_i \in T_i$, we have

$$g(t_i) \geq g(s_i) + \phi(t_i)(t_i - s_i).$$

Hence, $\phi(t_i)$ is a subgradient of g at t_i.

Step 2 - Convexity of $\mathcal{U}_i^M$. Next, we show that g is convex. To see this, pick $x_i, z_i \in T_i$ and consider $y_i = \lambda x_i + (1 - \lambda) z_i$ for some $\lambda \in (0, 1)$. Due to DSIC, we know that

$$g(x_i) \geq g(y_i) + (x_i - y_i)\phi(y_i)$$
$$g(z_i) \geq g(y_i) + (z_i - y_i)\phi(y_i)$$

Multiplying the first inequality by λ and the second by $(1 - \lambda)$ and adding them together gives

$$\lambda g(x_i) + (1 - \lambda) g(z_i) \geq g(y_i).$$

In fact, this shows that any g which satisfies the subgradient inequalities must be convex.

Step 3 - Apply Lemmas 5 and 6. By Lemma 5, ϕ is non-decreasing. By Lemma 6, for any $t_i \in T_i$,

$$g(t_i) = g(0) + \int_0^{t_i} \phi(x_i) dx_i.$$

Substituting for g, we get

$$\mathcal{U}_i^M(t_i, t_{-i}) = \mathcal{U}_i^M(0, t_{-i}) + \int_0^{t_i} f_i(x_i, t_{-i}) dx_i.$$

Substituting for $\mathcal{U}^M$, we get

$$p_i(t_i, t_{-i}) = p_i(0, t_{-i}) + t_i f_i(t_i, t_{-i}) - \int_0^{t_i} f_i(x_i, t_{-i}) dx_i.$$

This proves one direction.

Now, for the converse. If f is non-decreasing and p_i for all i is of the form described, then we have to show that the mechanism is DSIC. To show this, fix, $i \in N$, t_{-i}, and consider s_i, t_i. Now, substituting for p_i, we get

$$\begin{aligned}\big[t_i f_i(t_i, t_{-i}) - p_i(t_i, t_{-i})\big] - \big[t_i f_i(s_i, t_{-i}) - p_i(s_i, t_{-i})\big] &= (s_i - t_i) f_i(s_i, t_{-i}) - \int_{t_i}^{s_i} f_i(x_i, t_{-i}) dx_i \\ &\geq 0,\end{aligned}$$

where the inequality followed from the fact that f is non-decreasing. ■

An implication of this result is the following. Take two payment functions p and q that make f DSIC. Then, for every $i \in N$ and every t_{-i}, we know that for every $s_i, t_i \in T_i$,

$$p_i(s_i, t_{-i}) - p_i(t_i, t_{-i}) = \big[s_i f_i(s_i, t_{-i}) - \int_0^{s_i} f_i(x_i, t_{-i}) dx_i\big] - \big[t_i f_i(t_i, t_{-i}) - \int_0^{t_i} f_i(x_i, t_{-i}) dx_i\big]$$

and

$$q_i(s_i, t_{-i}) - q_i(t_i, t_{-i}) = \big[s_i f_i(s_i, t_{-i}) - \int_0^{s_i} f_i(x_i, t_{-i}) dx_i\big] - \big[t_i f_i(t_i, t_{-i}) - \int_0^{t_i} f_i(x_i, t_{-i}) dx_i\big]$$

Hence,

$$\begin{aligned} p_i(s_i, t_{-i}) - p_i(t_i, t_{-i}) &= q_i(s_i, t_{-i}) - q_i(t_i, t_{-i}), \\ \text{or } p_i(s_i, t_{-i}) - q_i(s_i, t_{-i}) &= p_i(t_i, t_{-i}) - q_i(t_i, t_{-i}).\end{aligned}$$

This result is also known as the revenue equivalence result in single object auction.

One important difference between the characterization in Theorem 5 and the characterization of Roberts' theorem is worth pointing out. The latter characterizations are very specific about the parameters to be used in the mechanism - Roberts' theorem asks us to design mechanisms by identifying weights for agents and alternatives and then doing a maximization of weighted values. However, the characterization in Theorem 5 is implicit. It only identifies *properties* of a mechanism that is necessary and sufficient for DSIC. It is still useful for verifying if a given mechanism is DSIC or not.

An immediate corollary of Theorem 5 is the following.

COROLLARY **1** *An allocation rule is implementable in dominant strategies if and only if it is non-decreasing.*

Proof: Suppose f is an implementable allocation rule. Then, there exists $(p_1, \ldots, p_n)$ such that $(f, p_1, \ldots, p_n)$ is DSIC. One direction of Theorem 5 showed that f must be non-decreasing. For the converse, Theorem 5 identified payment rules that make a non-decreasing allocation rule implementable. ■

The fact that any non-decreasing allocation rule can be implemented in the single object auction rule is insightful. Many allocation rules can be verified if they are implementable or not by checking if they are non-decreasing. The constant allocation rule is clearly non-decreasing (it is constant in fact). The dicatorial allocation rule is also non-decreasing. The efficient allocation rule is non-decreasing because if you are winning the object by reporting some type, efficiency guarantees that you will continue to win it by reporting a higher type (remember that efficient allocation rule in the single object case awards the object to an agent with the highest type).

Efficient allocation rule with a *reserve price* is the following allocation rule. If types of all agents are below a threshold level r, then the object is not sold, else all agents whose type is above r are considered and sold to one of these agents who has the highest type. It is clear that this allocation rule is also DSIC since it is non-decreasing. We will encounter this allocation rule again when we study optimal auction design.

Consider an agent $i \in N$ and fix the types of other agents at t_{-i}. Figure 3.3 shows how agent i's probability of winning the object can change in a DSIC allocation rule. If we restrict attention to DSIC allocation rules which either do not give the object to an agent or gives it to an agent with probability 1, then the shape of the curve depicting probability of winning the object will be a step function. We call such allocation rules **deterministic** allocation rules.

3.1.4 The Efficient Allocation Rule and the Vickrey Auction

We start off by deriving a deterministic mechanism using Theorem 5. The mechanism we focus is the Vickrey auction that uses the efficient allocation rule. Though the efficient allocation rule may break ties using randomization, we assume that ties are broken deterministically, i.e., each agent gets the object either with probability 1 or 0.

Suppose f is the efficient allocation. We know that the class of Groves payments make f DSIC. Suppose we impose the restriction that $p_i(0, t_{-i}) = 0$ for all $i \in N$ and for all t_{-i}.

Note that if t_i is not the highest type in the profile, then $f_i(x_i, t_{-i}) = 0$ for all $x_i \leq t_i$. Hence, by Theorem 5, $p_i(t_i, t_{-i}) = 0$. If t_i is the highest type and t_j is the second highest type in the profile, then $f_i(x_i, t_{-i}) = 0$ for all $x_i \leq t_j$ and $f_i(x_i, t_{-i}) = 1$ for all $t_i \geq x_i > t_j$. So, using Theorem 5, $p_i(t_i, t_{-i}) = t_i - [t_i - t_j] = t_j$. This is indeed the Vickrey auction. The revenue equivalence result says that any other DSIC auction must have payments which differ from the Vickrey auction by the amount a bidder pays at type 0, i.e., $p_i(0, t_{-i})$.

3.1.5 Deterministic Allocations Rules

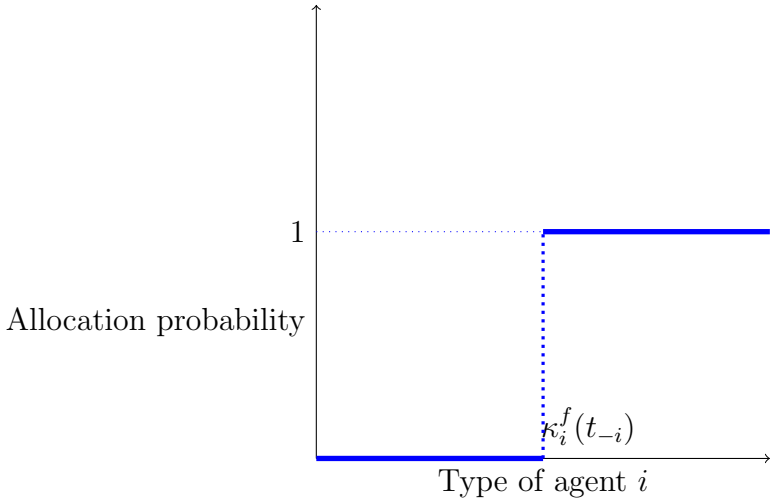

Figure 3.4: A deterministic implementable allocation rule

Call an allocation rule f **deterministic** (in single object setting) if for all $i \in N$ and every type profile t, we have $f_i(t) \in \{0, 1\}$. The aim of this section is to show the simple nature of payment rules for a deterministic allocation rule to be DSIC. We assume that set of types of agent i is $T_i = [0, b_i]$. Suppose f is a deterministic allocation rule which is DSIC. Hence, it is non-decreasing. For every $i \in N$ and every t_{-i}, the shape of $f_i(\cdot, t_{-i})$ is a step function (as in Figure 3.4). Now, define,

$$\kappa_i^f(t_{-i}) = \begin{cases} \inf\{t_i \in T_i : f_i(t_i, t_{-i}) = 1\} & \text{if } f_i(t_i, t_{-i}) = 1 \text{ for some } t_i \in T_i \\ 0 & \text{otherwise} \end{cases}$$

If f is DSIC, then it is non-decreasing, which implies that for all $t_i > \kappa_i^f(t_{-i})$, i gets the object and for all $t_i < \kappa_i^f(t_{-i})$, i does not get the object.

Consider a type $t_i \in T_i$. If $f_i(t_i, t_{-i}) = 0$, then using revenue equivalence, we can compute any payment which makes f DSIC as $p_i(t_i, t_{-i}) = p_i(0, t_{-i})$. If $f_i(t_i, t_{-i}) = 1$, then $p_i(t_i, t_{-i}) = p_i(0, t_{-i}) + t_i - [t_i - \kappa_i^f(t_{-i})] = p_i(0, t_{-i}) + \kappa_i^f(t_{-i})$. Hence, if p makes f DSIC, then for all $i \in N$ and for all t

$$p_i(t) = p_i(0, t_{-i}) + \kappa_i^f(t_{-i}).$$

The payments when $p_i(0, t_{-i}) = 0$ has special interpretation. If $f_i(t) = 0$, then agent i pays nothing (losers pay zero). If $f_i(t) = 1$, then agent i pays the minimum amount required to win the object when types of other agents are t_{-i}. If f is the efficient allocation rule, this reduces to the second-price Vickrey auction.

We can also apply this to other allocation rules. Suppose $N = \{1, 2\}$ and the allocations are $A = \{a_0, a_1, a_2\}$, where a_0 is the allocation where the seller keeps the object, a_i $(i \neq 0)$ is the allocation where agent i keeps the object. Given a type profile $t = (t_1, t_2)$, the seller computes, $U(t) = \max(2, t_1^2, t_2^3)$, and allocation is a_0 if $U(t) = 2$, it is a_1 if $U(t) = t_1^2$, and a_2 if $U(t) = t_2^3$. Here, 2 serves as a (pseudo) *reserve price* below which the object is unsold. It is easy to verify that this allocation rule is non-decreasing, and hence DSIC. Now, consider a type profile $t = (t_1, t_2)$. For agent 1, the minimum he needs to bid to win against t_2 is $\sqrt{\max\{2, t_2^3\}}$. Similarly, for agent 2, the minimum he needs to bid to win against t_1 is $(\max\{2, t_1^2\})^{\frac{1}{3}}$. Hence, the following is a payment scheme which makes this allocation rule DSIC. At any type profile $t = (t_1, t_2)$, if none of the agents win the object, they do not pay anything. If agent 1 wins the object, then he pays $\sqrt{\max\{2, t_2^3\}}$, and if agent 2 wins the object, then he pays $(\max\{2, t_1^2\})^{\frac{1}{3}}$.

3.1.6 Individual Rationality

We can find out conditions under which a mechanism is individually rational. We use *ex post* individual rationality here.

DEFINITION **10** *A mechanism (f, p) is* **ex-post individually rational** *if for all $i \in N$ and for all t_{-i},*

$$t_i f_i(t_i, t_{-i}) - p_i(t_i, t_{-i}) \geq 0 \qquad \forall\ t_i.$$

LEMMA **7** *Suppose a mechanism (f,p) is strategy-proof. The mechanism (f,p) is ex-post individually rational if and only if for all $i \in N$ and for all t_{-i},*

$$p_i(0,t_{-i}) \leq 0.$$

Further, a mechanism (f,p) is ex-post individually rational and $p_i(t_i,t_{-i}) \geq 0$ for all $i \in N$ and for all t_{-i} if and only if for all $i \in N$ and for all t_{-i},

$$p_i(0,t_{-i}) = 0.$$

Proof: Suppose (f,p) is ex-post individually rational. Then $0 - p_i(0,t_{-i}) \geq 0$ for all $i \in N$ and for all t_{-i}. For the converse, suppose $p_i(0,t_{-i}) \leq 0$ for all $i \in N$ and for all t_{-i}. In that case, $t_i - p_i(t_i,t_{-i}) = t_i - p_i(0,t_{-i}) - t_i f_i(t_i,t_{-i}) + \int_0^{t_i} f_i(x_i,t_{-i})dx_i \geq 0$.

Ex-post individual rationality says $p_i(0,t_{-i}) \leq 0$ and the requirement $p_i(0,t_{-i}) \geq 0$ ensures $p_i(0,t_{-i}) = 0$. For the converse, $p_i(0,t_{-i}) = 0$ ensures ex-post individual rationality. ■

Hence, ex-post individual rationality along with the requirement that payments are always non-negative pins down $p_i(0,t_{-i}) = 0$ for all $i \in N$ and for all t_{-i}.

3.1.7 Beyond Vickrey auction: examples

We have seen that Vickrey auction is DSIC. We give some more examples of DSIC mechanisms below. We will be informal in the description of these mechanisms to avoid cumbersome notations.

The first one is a Groves mechanism but not the Vickrey auction. Consider the mechanism where the highest valued agent wins the object (ties broken in some way) but payments are slightly different. In particular, winner pays an amount equal to the second highest value but everyone is **compensated** some amounts: highest and second highest valued agents are compensated an amount equal to $\frac{1}{n}$ of the third highest value and other agents are compensated an amount equal to $\frac{1}{n}$ of the second highest value.

To show that this mechanism is DSIC, we observe that the allocation rule (efficient rule) is non-decreasing. Then, we need to show that the payment respects revenue equivalence. To see this consider, for all $i \in N$, for all t_{-i},

$$p_i(0,t_{-i}) = -\frac{1}{n}\text{second highest in } \{t_j\}_{j \neq i}.$$

Notice that the Vickrey auction has $p_i(0, t_{-i}) = 0$ for all i and for all t_{-i}. Here, agents are compensated some amount. To see the effect of choice of such a payment rule, consider a profile $t_1 > t_2 > \ldots > t_n$ (here, we are not considering ties for convenience). Agent 1 wins the object and pays

$$p_1(0, t_{-1}) + t_1 - \int_0^{t_1} f_1(x_1, t_{-1}) = -\frac{1}{n} t_3 + t_2.$$

Agent 2 does not get the object but pays

$$p_2(0, t_{-2}) = -\frac{1}{n} t_3.$$

Agent $j > 2$ does not get the object but pays

$$p_j(0, t_{-j}) = -\frac{1}{n} t_2.$$

By Theorem 5, such a mechanism is DSIC. Further, $p_i(0, t_{-i}) \leq 0$ for all i and for all t_{-i}. Hence, by Lemma 7, such a mechanism also satisfies individual rationality. Notice that the total sum of payments in this mechanism is $\frac{2}{n}(t_2 - t_3)$, which approaches zero as number of agents increase. Hence, this mechanism gives back *all surplus* (t_1 in this case) to agents as n increases.

Another example of a Groves mechanism is the choice of $p_i(0, t_{-i}) = \max_{j \neq i} t_j$. This choice leads to the mechanism that we have discussed earlier - winner does not pay anything and everyone else is compensated an amount equal to the highest value.

We will now discuss a non-Groves mechanism. Again, we will keep the discussions informal here - which means, we will ignore profiles where there are ties. This is a mechanism proposed by Green and Laffont, and famously referred to as the Green-Laffont mechanism. The Green-Laffont mechanism gives the object with probability $(1 - \frac{1}{n})$ to the highest valued agent. It gives the object to the second highest valued agent with probability $\frac{1}{n}$. Clearly, the allocation rule is non-decreasing. We will specify the $p_i(0, t_{-i})$ term to complete the description of the mechanism. In fact, it is the same term we used for the first mechansim. For all $i \in N$, for all t_{-i},

$$p_i(0, t_{-i}) = -\frac{1}{n} \text{second highest in } \{t_j\}_{j \neq i}.$$

Since the object is not given to the highest valued agent with probability 1, this is **not** a Groves mechanism. However, it has some other desirable properties. By Lemma 7, it is a

desirable mechanism. Consider a profile $t_1 > t_2 > \ldots > t_n$. Note that the payment of agent 1 is

$$p_1(0, t_{-1}) + (1 - \frac{1}{n})t_1 - \frac{1}{n}(t_2 - t_3) - (1 - \frac{1}{n})(t_1 - t_2) = \Big(1 - \frac{2}{n}\Big)t_2.$$

Payment of agent 2 is

$$p_2(0, t_{-2}) + \frac{1}{n}t_2 - \frac{1}{n}(t_2 - t_3) = 0.$$

For every other agent $j \notin \{1, 2\}$, payment of agent j is

$$p_j(0, t_{-j}) = -\frac{1}{n}t_2.$$

As a result, we see that this mechanism is **budget-balanced**:

$$\sum_{i \in N} p_i(t) = \Big(1 - \frac{2}{n}\Big)t_2 + 0 - (n-2)\frac{1}{n}t_2 = 0.$$

Again, notice that the for large n, the Green-Laffont mechanism gives the object with very high probability to agent 1. Since the mechanism is budget-balanced, it distributes the *almost the entire* surplus as n becomes large.

More familiar DSIC mechanisms which are variants of Vickrey auction are Vickrey auctions with reserve prices. Here, the seller announces a reserve price r and the object is sold to the highest valuation agent if and only if its valuation is above the reserve price r. Using Theorem 5, we see that if we set $p_i(0, t_{-i}) = 0$ for all i and for all t_{-i}, we get the following payment function. Take a type profile t - for simplicity, assume $t_1 > t_2 \geq \ldots \geq t_n$. If $t_1 \geq r \geq t_2$, then we get

$$p_1(t) = 0 + t_1 - (t_1 - r) = r.$$

If $r \leq t_2$, then we get $p_1(t) = t_2$. Other agents pay zero. Hence, the winner of the object pays $\max(\max_{j \neq i} t_j, r)$ and everyone else pays zero.

3.1.8 Bayesian incentive compatibility

We now investigate the analogue of our incentive compatibility characterization when we consider Bayesian incentive compatibility. For this, we need to specify the common prior belief of the agents. We assume that the value of agent $i \in N$ is drawn from $[0, b_i]$. We assume that there is a common prior which is a joint probability distribution of all the types - we denote the cumulative distribution by G. For every agent i, we denote by $G_{-i}(\cdot|s_i)$ the

conditional distribution of types of agents other than agent i when agent i has type s_i. We assume that $G_{-i}(\cdot|s_i)$ admits a density function $g_{-i}(\cdot|s_i)$ which is positive everywhere.

Let $T_i := [0, b_i]$ and $T^n := [0, b_1] \times \ldots \times [0, b_n]$. Similarly, let $T_{-i} = \times_{j \in N \setminus \{i\}} T_j$. A typical valuation of bidder i will be denoted as $t_i \in T_i$, a valuation profile of bidders will be denoted as $t \in T^n$, and a valuation profile of bidders in $N \setminus \{i\}$ will be denoted as $t_{-i} \in T_{-i}$. The valuation profile $t = (t_1, \ldots, t_i, \ldots, t_n)$ will sometimes be denoted as (t_i, t_{-i}).

Every mechanism $(f, p_1, \ldots, p_n)$ induces an expected allocation rule and an expected payment rule (α, π), defined as follows. The expected allocation of agent i with type t_i when he reports $s_i \in T_i$ in allocation rule f is

$$\alpha_i(s_i|t_i) = \int_{T_{-i}} f_i(s_i, s_{-i}) g_{-i}(s_{-i}|t_i) ds_{-i}.$$

Similarly, the expected payment of agent i with type t_i when he reports $s_i \in T_i$ in payment rule p_i is

$$\pi_i(s_i|t_i) = \int_{T_{-i}} p_i(s_i, s_{-i}) g_{-i}(s_{-i}|t_i) ds_{-i}.$$

So, the expected utility from a mechanism $M \equiv (f, p_1, \ldots, p_n)$ to an agent i with true value t_i by reporting a value s_i is $\alpha_i(s_i|t_i)t_i - \pi_i(s_i|t_i)$.

DEFINITION **11** *A mechanism $(f, p_1, \ldots, p_n)$ is* **Bayesian incentive compatible (BIC)** *if for every agent $i \in N$ and for every possible values $s_i, t_i \in T_i$ we have*

$$\alpha_i(t_i|t_i)t_i - \pi_i(t_i|t_i) \geq \alpha_i(s_i|t_i)t_i - \pi_i(s_i|t_i). \qquad \textbf{(BIC)}$$

Equation **BIC** says that a bidder maximizes his expected utility by reporting true value. Given that other bidders report truthfully, when bidder i has value t_i, he gets more expected utility by reporting t_i than by reporting any other value $s_i \in T_i$.

We say an allocation rule f is **Bayes-Nash implementable** if there exists payment rules $(p_1, \ldots, p_n)$ such that $(f, p_1, \ldots, p_n)$ is a Bayesian incentive compatible mechanism.

3.1.9 Independence and characterization of BIC

We now give a characterization of BIC mechanisms, analogous to Theorem 5, when the priors are independent. The independent prior assumption means that each agent i's value

for the object is drawn using a distribution G_i (with density g_i) and this in turn means that $G_{-i}(t_{-i}|t_i) = \times_{j \neq i} G_j(t_j)$.

Because of independence, the conditional term can be dropped from all the notations: so, $\alpha_i(s_i|t_i)$ and $\pi_i(s_i|t_i)$ will just be written $\alpha_i(s_i)$ and $\pi_i(s_i)$ respectively.

We say that an allocation rule f is **non-decreasing in expectation (NDE)** if for all $i \in N$ and for all $s_i, t_i \in T_i$ with $s_i < t_i$ we have $\alpha_i(s_i) \leq \alpha_i(t_i)$. Similar to the characterization in the dominant strategy case, we have a characterization in the Bayesian incentive compatible mechanisms.

THEOREM **6** *A mechanism $(f, p_1, \ldots, p_n)$ is Bayesian incentive compatible if and only if f is NDE and for every $i \in N$, p_i satisfies*

$$\pi_i(t_i) = \pi_i(0) + t_i \alpha_i(t_i) - \int_0^{t_i} \alpha_i(s_i) ds_i \qquad \forall\ t_i \in [0, b_i].$$

The proof is a replication of the arguments we did for dominant strategy case in Theorem 5. We skip the proof (but you are encouraged to reconstruct the arguments).

A BIC allocation rule need not be DSIC. We give an example to illustrate this. Consider a setting with two agents $N = \{1, 2\}$. Suppose the values of both the agents are drawn uniformly and independently from $[0, 1]$. Figure 3.5 shows an allocation rule f.

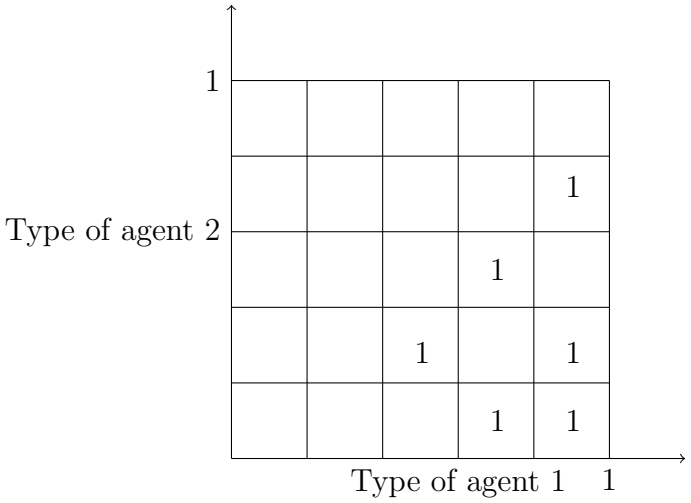

Figure 3.5: A BIC allocation rule which is not DSIC

The type profiles are divided into cells of equal size (25 of them in total). Some of the cells are assigned some numbers - this is the probability with which agent 1 gets the object

in f. The cells in which no number is written, the probability of agent 1 getting the object at those profiles is zero. For our purpose, the probability of agent 2 getting the object is irrelevant - for simplicity, we can assume it to be zero (hence, in all other cells the seller keeps the object).

An easy calculation reveals that the expected probability of agent 1 winning the object is non-decreasing: it is zero if $t_1 \leq \frac{2}{5}$, it is $\frac{1}{5}$ if $t_1 \in (\frac{2}{5}, \frac{3}{5}]$, it is $\frac{2}{5}$ if $t_1 \in (\frac{3}{5}, \frac{4}{5}]$, and it is $\frac{3}{5}$ if $t_1 > \frac{4}{5}$. Hence, the allocation rule a is BIC but not DSIC.

Theorem 6 says that the (expected) payment of a bidder in a mechanism is uniquely determined by the allocation rule once we fix the expected payment of a bidder with the lowest type. Hence, a mechanism is uniquely determined by its allocation rule and the payment of a bidder with the lowest type.

It is instructive to examine the payment function when $\pi_i(0) = 0$. Then payment of agent i at type t_i becomes

$$\pi_i(t_i) = \alpha_i(t_i)t_i - \int_0^{t_i} \alpha_i(x_i)dx_i.$$

Because of non-decreasing $\alpha_i(\cdot)$ this is always greater than or equal to zero - it is the difference between area of the rectangle with sides $\alpha_i(x_i)$ and x_i and the area under the curve $\alpha_i(\cdot)$ from 0 to x_i.

We next impose a the analogue of individual rationality in the Bayesian set up.

DEFINITION **12** *A mechanism $(f, p_1, \ldots, p_n)$ is* **interim individually rational (IIR)** *if for every bidder $i \in N$ we have*

$$\alpha_i(t_i)t_i - \pi_i(t_i) \geq 0 \qquad \forall\ t_i \in T_i.$$

IIR is weaker than the (ex post) individual rationality we had discussed earlier since IIR only requires *interim* expected utility from truthtelling to be non-negative. The set of BIC and IIR mechanisms can now be characterized as follows.

LEMMA **8** *A mechanism $(f, p_1, \ldots, p_n)$ is BIC and IIR if and only if*

(1) f is NDE.

(2) For all $i \in N$,

$$\pi_i(t_i) = \pi_i(0) + t_i\alpha_i(t_i) - \int_0^{t_i} \alpha_i(s_i)ds_i \qquad \forall\ t_i \in [0, b_i].$$

(3) For all $i \in N$, $\pi_i(0) \leq 0$.

Proof: Suppose $(f, p_1, \ldots, p_n)$ is BIC. By Theorem 6, (1) and (2) follows. Applying IIR at $t_i = 0$, we get $\pi_i(0) \leq 0$, which is (3).

Now, suppose (1),(2), and (3) holds for a mechanism $(f, p_1, \ldots, p_n)$. By Theorem 6, the mechanism is BIC. At any type t_i,

$$\begin{aligned} t_i\alpha_i(t_i) - \pi_i(t_i) &= \int_0^{t_i} \alpha_i(s_i)ds_i - \pi_i(0) \\ &\geq \int_0^{t_i} \alpha_i(s_i)ds_i \\ &\geq 0, \end{aligned}$$

where the first inequality follows from (3). Hence, the mechanism satisfies IIR. ■

Revenue equivalence of standard auctions. Theorem 6 has some immediate applications. When agents are symmetric, i.e., $G_i = G_j$, $[0, b_i] = [0, b_j]$ for all $i, j \in N$, then the first price auction has a **monotone and symmetric** equilibrium, where the highest valued agent submits the highest bid. As a result, the first-price auction implements the efficient allocation rule in Bayes-Nash equilibrium. The Vickrey auction also implements the efficient allocation rule in Bayes-Nash (dominant strategy, in fact) equilibrium. Further, the payment of an agent with zero value is zero in both the auctions. Hence, Theorem 6 implies that the interim payment of a bidder is the same in the first-price and the Vickrey auction. As a result, both the first-price and the Vickrey auction generates the same expected revenue.

Such revenue equivalence results have been established for many auction formats in auction theory. Many of these results are corollary of Theorem 6 once we establish that (a) they implement the same allocation rule in Bayes-Nash equilibrium and (b) the interim payment of the lowest type is the same.

3.2 The One Agent Problem

It is instructive to look at the one agent/buyer problem and study it further. Since there is only one agent, the notion of Bayesian and dominant strategy incentive compatibility are the same. Suppose there is a single agent whose type is distributed in some interval $T \equiv [0, \beta]$

with a distribution G whose density is g. A mechanism consists of an allocation rule and a payment rule

$$f : [0, \beta] \to [0, 1], \qquad p : [0, \beta] \to \mathbb{R}.$$

A mechanism (f, p) is **incentive compatible** if for every $s, t \in T$, we have

$$tf(t) - p(t) \geq tf(s) - p(s).$$

A mechanism (f, p) is **individually rational** if for every $t \in T$,

$$tf(t) - p(t) \geq 0.$$

Let $\mathcal{M}$ be the set of all incentive compatible and individually rational mechanism. For any mechanism $M \equiv (f, p)$, the expected revenue from M is given by

$$\Pi^M := \int_0^\beta p(t)g(t)dt.$$

A mechanism M is an **optimal** mechanism if $M \in \mathcal{M}$ and for all $M' \in \mathcal{M}$, we have

$$\Pi^M \geq \Pi^{M'}.$$

How does an optimal mechanism look like?

Consider an incentive compatible and individually rational mechanism $(f, p) \equiv M \in \mathcal{M}$. By revenue equivalence, we can write for every $t \in T$,

$$p(t) = p(0) + tf(t) - \int_0^t f(x)dx.$$

Individual rationality implies that $p(0) \leq 0$. If we want to maximize revenue, then clearly $p(0) = 0$. So, we have proved the following lemma.

LEMMA **9** *Suppose (f, p) is an optimal mechanism. Then, for every $t \in T$,*

$$p(t) = tf(t) - \int_0^t f(x)dx.$$

Hence, we focus attention to the set of incentive compatible and individually rational mechanisms $(f, p) \in \mathcal{M}$ such that $p(0) = 0$, and denote it as $\mathcal{M}^0$. Note that the payment in these mechanisms is uniquely determined by the allocation rule: for every $t \in T$, $p(t) = tf(t) - \int_0^t f(x)dx$. We call such a payment rule the **benchmark** payment rule of f, and the

corresponding mechanism, the **benchmark** mechanism of f. Hence, the optimization problem reduces to searching over the set of all non-decreasing allocation rules. We denote the set of all non-decreasing allocation rules $\mathcal{F}$. For every $f \in \mathcal{F}$, the expected revenue from the benchmark mechanism of f is given by

$$\Pi^f := \int_0^\beta \big[tf(t) - \int_0^t f(x)dx\big]g(t)dt.$$

So, we have shown the following lemma.

LEMMA **10** *Suppose (f, p) is an optimal mechanism. Then, the following are true:*

- *for every $f' \in \mathcal{F}$, $\Pi^f \geq \Pi^{f'}$.*
- *for every $t \in T$, $p(t) = tf(t) - \int_0^t f(x)dx$.*

We now prove a crucial lemma.

LEMMA **11** *For any implementable allocation rule f, we have*

$$\Pi^f = \int_0^\beta w(t)f(t)g(t)dt,$$

where $w(t) = \Big(t - \frac{1-G(t)}{g(t)}\Big)$ for all t.

Proof: Note that if f is implementable, then

$$\begin{aligned}
\Pi^f &= \int_0^\beta \Big(tf(t) - \int_0^t f(x)dx\Big)g(t)dt \\
&= \int_0^\beta tf(t)g(t)dt - \int_0^\beta \Big(\int_0^t f(x)dx\Big)g(t)dt \\
&= \int_0^\beta tf(t)g(t)dt - \int_0^\beta \Big(\int_t^\beta g(x)dx\Big)f(t)dt \\
&= \int_0^\beta tf(t)g(t)dt - \int_0^\beta \big(1 - G(t)\big)f(t)dt \\
&= \int_0^\beta \Big(t - \frac{1-G(t)}{g(t)}\Big)f(t)g(t)dt.
\end{aligned}$$

■

By Lemma 10 and Lemma 11, our optimization problem is the following:

$$\max_{f\in\mathcal{F}} \Pi^f = \max_{f\in\mathcal{F}} \int_0^\beta w(t)f(t)g(t)dt.$$

We first optimize a **relaxed** problem where we consider an arbitrary f - not necessarily belonging to $\mathcal{F}$. For this, we impose a condition on the distribution. The distribution G satisfies the **monotone hazard rate (MHR)** condition if $\frac{g(x)}{1-G(x)}$ is non-decreasing in x. Standard distributions such as uniform and exponential satisfy MHR. Consider the uniform distribution on $[0, b]$. This means $G(x) = \frac{x}{b}$. So, $\frac{g(x)}{1-G(x)} = \frac{1}{b-x}$, which is non-decreasing in x. If G satisfies MHR, then there is a **unique** solution to $x = \frac{1-G(x)}{g(x)}$. Denote this unique solution as κ^* - so, κ^* uniquely solves

$$w(\kappa^*) = 0.$$

One notices that since w is strictly increasing, $w(t) < 0$ for all $t < \kappa^*$ and $w(t) > 0$ for all $t > \kappa^*$. Then, by the expression in Lemma 11, the maximum value of Π^f is achieved for the f which sets $f(t) = 0$ for all $t < \kappa^*$ (since $w(t) < 0$ for all $t < \kappa^*$) and $f(t) = 1$ for all $t > \kappa^*$ (since $w(t) > 0$ for all $t > \kappa^*$). The value of $f(t)$ at $t = \kappa^*$ does not matter. Since such an f is monotone, it is implementable.

THEOREM **7** *An optimal mechanism (f, p) satisfies the following under the MHR condition:*

- *there is a unique κ^* that solves $x = \frac{1-G(x)}{g(x)}$ and for all $t < \kappa^*$, we have $f(t) = 0$ and for all $t > \kappa^*$ we have $f(t) = 1$,*
- *for all $t \in T$, $p(t) = f(t)\kappa^*$.*

3.2.1 Monopolist problem

In this section, we consider a related problem where a monopolist is trying to sell a good to a buyer with unknown value θ for "quality". The monopolist can produce the good with quality $q \in [0, 1]$ at cost $C(q)$ and the buyer with value θ values this quality at θq. We assume that C is strictly convex, increasing, and continuously differentiable. So, the type of the buyer is θ, which we assume to lie in some interval $[0, \beta]$. This model was studied in the seminal work of Mussa and Rosen (1978).

As before, a **mechanism** is a pair of maps $q : \Theta \to \mathbb{R}_{++}$ and $p : \Theta \to \mathbb{R}$. So, the mechanism here commits a quality level and payment for every possible type. The timing of the game is as follows.

- The seller announces a mechanism (and he commits to it).
- The buyer realizes her type.
- The buyer announces a type.
- The buyer gets an outcome (quality, payment) pair according to the announced type and mechanism.
- The buyer and the seller realize their payoffs.

We assume that agent's utility is *linear*: for consuming quality q at price p, he gets utility equal to $q\theta - p$. This can be generalized by a function $u(q,\theta) - p$, where u is increasing in each argument and satisfies increasing differences property (or, single crossing).

Fix a mechanism (q,p). Denote the net utility of agent of type θ by reporting θ' to the mechanism (q,p) as:

$$U^{q,p}(\theta'|\theta) := q(\theta')\theta - p(\theta').$$

DEFINITION **13** *A mechanism (q,p) is* **incentive compatible** *if for every θ,*

$$U^{q,p}(\theta|\theta) \geq U^{q,p}(\theta'|\theta).$$

Notice that

$$U^{q,p}(\theta'|\theta) = U(\theta'|\theta') + q(\theta')(\theta - \theta').$$

Incentive constraints say that for all $\theta, \theta' \in \Theta$,

$$U^{q,p}(\theta|\theta) \geq U^{q,p}(\theta'|\theta) = U^{q,p}(\theta'|\theta') + q(\theta')(\theta - \theta').$$

For simplicity of notation, we denote $U^{q,p}(\theta|\theta)$ as $U^{q,p}(\theta)$. Hence, we can write the IC constraints as

$$U^{q,p}(\theta) \geq U^{q,p}(\theta') + q(\theta')(\theta - \theta'). \tag{3.1}$$

As before, it is routine to verify that in every incentive compatible mechanism (q, p), we must have that for every $\theta \in [0, 1]$,

$$U^{q,p}(\theta) = U^{q,p}(0) + \int_0^\theta q(\theta')d\theta'. \tag{3.2}$$

This is the **payoff equivalence** formula. The payoff equivalence formula in Equation (3.2) also gives us a **revenue equivalence formula** by expanding the U terms: for all $\theta \in [0, 1]$,

$$p(\theta) = p(0) + q(\theta)\theta - \int_0^\theta q(\theta')d\theta'. \tag{3.3}$$

This gets us to a characterization of IC and IR constraints as before.

PROPOSITION **3** *A mechanism (q, p) is incentive compatible and individually rational if and only if*

1. *q is non-decreasing.*
2. *revenue equivalence formula in (3.3) holds.*
3. *$p(0) \leq 0$.*

Now, we return to the objective function of the monopolist. Suppose F is the cdf of types. We assume that F is strictly increasing, differentiable with density f. The seller seeks to maximize the following expression over all IC and IR mechanisms:

$$\int_0^\beta \Big[p(\theta) - C(q(\theta))\Big] f(\theta)d\theta.$$

Using, revenue equivalence formula (3.3), we simplify this to

$$\int_0^\beta \Big[p(0) + q(\theta)\theta - \int_0^\theta q(\theta')d\theta' - C(q(\theta))\Big] f(\theta)d\theta.$$

Since IR implies $p(0) \leq 0$, in any optimal mechanism, we must therefore have $p(0) = 0$. Hence, the objective function becomes

$$\int_0^\beta \Big[q(\theta)\theta - \int_0^\theta q(\theta')d\theta' - C(q(\theta))\Big] f(\theta)d\theta.$$

Since this is only a function of q, we only need the constraint that q is non-decreasing. We make a some simplification to this term.

$$\begin{aligned}
&\int_0^\beta \Big[q(\theta)\theta - \int_0^\theta q(\theta')d\theta' - C(q(\theta))\Big] f(\theta)d\theta \\
&= \int_0^\beta \big[q(\theta)\theta - C(q(\theta))\big] f(\theta)d\theta - \int_0^1 \Big(\int_0^\theta q(\theta')d\theta'\Big) f(\theta)d\theta \\
&= \int_0^\beta \big[q(\theta)\theta - C(q(\theta))\big] f(\theta)d\theta - \int_0^1 \Big(\int_\theta^1 f(\theta')d\theta'\Big) q(\theta)d\theta \\
&= \int_0^\beta \big[q(\theta)\theta - C(q(\theta))\big] f(\theta)d\theta - \int_0^1 \big(1 - F(\theta)\big) q(\theta)d\theta \\
&= \int_0^\beta \Big(\theta q(\theta) - C(q(\theta)) - \frac{1-F(\theta)}{f(\theta)} q(\theta)\Big) f(\theta)d\theta.
\end{aligned}$$

Forgetting the fact that q needs to be non-decreasing, we solve this unconstrained objective function. We find the point-wise maximum and that should maximize the overall expression. Point-wise maximum gives a first order condition for each θ as:

$$\theta - C'(q) - \frac{1-F(\theta)}{f(\theta)} = 0.$$

Denoting the **virtual value** at θ as $v(\theta) := \theta - \frac{1-F(\theta)}{f(\theta)}$, we se that the optimal quality at type θ must satisfy

$$C'(q(\theta)) = v(\theta).$$

Since C is strictly convex, the objective function at each point θ is strictly concave in q. Hence, this is also a global optimal. However, the optimal solution may not satisfy $q(\theta) \geq 0$. To ensure this, strict concavity implies that if the optimum lies to the left of 0, then under non-negativity constraint, we must have $q(\theta) = 0$ as optimal. So, optimal solution can be described as follows. Let $\hat{q}(\theta)$ be the solution to $C'(\hat{q}(\theta)) = v(\theta)$. Then, the optimal quality contract is: for all θ,

$$q^*(\theta) = \max(0, \hat{q}(\theta))$$

with price

$$p^*(\theta) = \theta q^*(\theta) - \int_0^\theta q^*(\theta')d\theta'.$$

Now, this point-wise optimal solution need not satisfy the fact q is non-decreasing. However, if virtual value is increasing, then it ensures that q is non-decreasing. To see this,

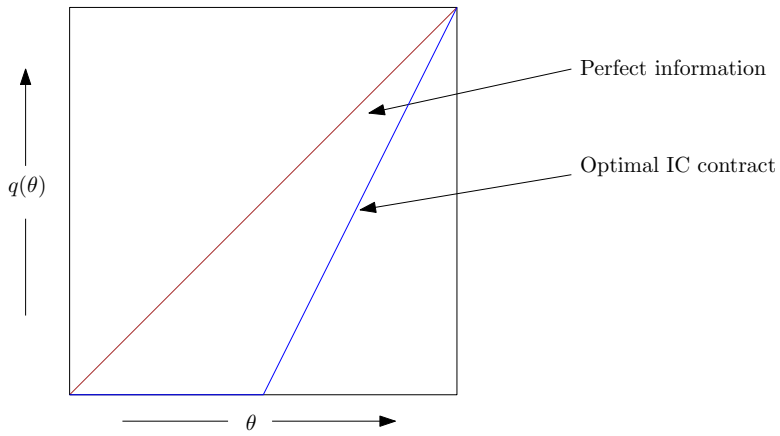

Figure 3.6: Adverse selection

assume for contradiction for some $\theta > \theta'$, we have $q(\theta) < q(\theta')$. Then, $q(\theta') > 0$. Further, $\hat{q}(\theta) \leq q(\theta)$ implies $\hat{q}(\theta) < \hat{q}(\theta')$. Then, convexity of C implies $C'(\hat{q}(\theta)) \leq C'(\hat{q}(\theta'))$. But then, $v(\theta) \leq v(\theta')$, which contradicts the fact that v is increasing. Notice that virtual value is increasing can be satisfied if inverse hazard rate $\frac{f(\theta)}{1-F(\theta)}$ is non-decreasing - an assumption satisfied by many distribution including the uniform distribution.

As an exercise, suppose $C(q) = \frac{1}{2}q^2$ with $q \in [0,1]$ and F is the uniform distribution in $[0,1]$. Then, we see that for each θ, $v(\theta) = 2\theta - 1$. Hence, $C'(q(\theta)) = q$ must be equal to $2\theta - 1$. Hence, we get $q^*(\theta) = \max(0, 2\theta - 1)$. Notice that in the perfect information case, the seller should ensure $C'(q(\theta)) = \theta$, which gives $q(\theta) = \theta$. So, there is under-provision to lower types due to incentive constraint. This is shown in Figure 3.6.

Constant marginal cost

If marginal cost is constant, then the optimal contract exhibits extreme *pooling*. To see this, suppose that q can take any value in $[0,1]$ and $C(q) = cq$ for some $c > 0$. Then the

optimization program is

$$\begin{aligned}
&\max_{q \text{ non-decreasing}} \int_0^1 \Big(\theta q(\theta) - cq(\theta) - \frac{1-F(\theta)}{f(\theta)} q(\theta)\Big) f(\theta) d\theta \\
&= \max_{q \text{ non-decreasing}} \int_0^1 \Big(\theta - c - \frac{1-F(\theta)}{f(\theta)}\Big) q(\theta) f(\theta) d\theta \\
&= \max_{q \text{ non-decreasing}} \int_0^1 \Big(v(\theta) - c\Big) q(\theta) f(\theta) d\theta
\end{aligned}$$

This has a simple optimal solution: whenever $v(\theta) < c$, set $q(\theta) = 0$ and whenever $v(\theta) > c$, set $q(\theta) = 1$. Monotonicity of v ensures monotonicity of q. Notice that if $q(\theta) = 0$, we have $p(\theta) = 0$. By the revenue equivalence formula, if $q(\theta) = 1$ (which implies that $\theta \geq v^{-1}(c)$)

$$p(\theta) = \theta - \int_{v^{-1}(c)}^{\theta} q(\theta') d\theta' = \theta - (\theta - v^{-1}(c)) = v^{-1}(c).$$

Hence, every buyer who gets the maximum possible quality pays the "posted-price" $v^{-1}(c)$. Thus, the optimal contract is equivalent to saying that the seller announces a **posted-price** equal to $v^{-1}(c)$ and the buyer with type greater than the posted price gets maximum quality and those below the posted price get zero quality.

3.3 Optimal Auction Design

We now go back to our model with a single seller who is selling a single object and we ignore cost of production now. The only difference from previous sections is that there are $n \geq 1$ buyers. We will describe the design of optimal auction for selling a single indivisible object to such a set of bidders (buyers) who have quasi-linear utility functions. The seminal paper in this area is (Myerson, 1981). We present a detailed analysis of this work.

The optimal mechanism is a mechanism that maximizes the expected revenue of the seller over all mechanisms identified in Theorem 8. To compute expected revenue from a mechanism $(f, p \equiv (p_1, \ldots, p_n))$, we note that the expected payment of agent i with type t_i is $\pi_i(t_i)$. Hence, (ex-ante) expected payment of agent i to this mechanism is

$$\int_0^{b_i} \pi_i(t_i) g_i(t_i) dt_i.$$

Hence, the expected revenue from the mechanism $(f, p \equiv (p_1, \ldots, p_n))$ is

$$\Pi(f,p) = \sum_{i \in N} \int_0^{b_i} \pi_i(t_i) g_i(t_i) dt_i.$$

We say a mechanism (f,p) is an **optimal mechanism** if

- (f,p) is Bayesian incentive compatible and individually rational,
- and $\Pi(f,p) \geq \Pi(f',p')$ for any other Bayesian incentive compatible and individually rational mechanism (f',p').

Theorem 8 will play a crucial role since it has identified the entire class of BIC and IIR mechanisms, over which we are optimizing.

Fix a mechanism (f,p) which is Bayesian incentive compatible and individually rational. For any bidder $i \in N$, the expected payment of bidder $i \in N$ is given by

$$\int_0^{b_i} \pi_i(t_i) g_i(t_i) dx_i = \pi_i(0) + \int_0^{b_i} \alpha_i(t_i) t_i g_i(t_i) dt_i - \int_0^{b_i} \int_0^{t_i} \big(\alpha_i(s_i) ds_i\big) g_i(t_i) dt_i,$$

where the last equality comes by using revenue equivalence (Theorem 6). By interchanging the order of integration in the last term, we get

$$\begin{aligned} \int_0^{b_i} \int_0^{t_i} \big(\alpha_i(s_i) ds_i\big) g_i(t_i) dt_i &= \int_0^{b_i} \big(\int_{t_i}^{b_i} g_i(s_i) ds_i\big) \alpha_i(t_i) dt_i \\ &= \int_0^{b_i} (1 - G_i(t_i)) \alpha_i(t_i) dt_i. \end{aligned}$$

Hence, we can write

$$\Pi(a,p) = \sum_{i \in N} \pi_i(0) + \sum_{i \in N} \int_0^{b_i} \big(t_i - \frac{1 - G_i(t_i)}{g_i(t_i)}\big) \alpha_i(t_i) g_i(t_i) dt_i.$$

We now define the **virtual valuation** of bidder $i \in N$ with valuation $t_i \in T_i$ as

$$w_i(t_i) = t_i - \frac{1 - G_i(t_i)}{g_i(t_i)}.$$

Note that since $g_i(t_i) > 0$ for all $i \in N$ and for all $t_i \in T_i$, the virtual valuation $w_i(t_i)$ is well defined. Also, virtual valuations can be negative. Using this and the definition of $\alpha_i(\cdot)$, we

can write

$$\begin{aligned}
\Pi(f,p) &= \sum_{i\in N}\pi_i(0) + \sum_{i\in N}\int_0^{b_i} w_i(t_i)\alpha_i(t_i)g_i(t_i)dt_i \\
&= \sum_{i\in N}\pi_i(0) + \sum_{i\in N}\int_0^{b_i} \Big(\int_{T_{-i}} f_i(t_i,t_{-i})g_{-i}(t_{-i})dt_{-i}\Big)w_i(t_i)g_i(t_i)dt_i \\
&= \sum_{i\in N}\pi_i(0) + \sum_{i\in N}\int_{T^n} w_i(t_i)f_i(t)g(t)dt \\
&= \sum_{i\in N}\pi_i(0) + \int_{T^n}\left[\sum_{i\in N} w_i(t_i)f_i(t)\right]g(t)dt.
\end{aligned}$$

Since IIR requires $\pi_i(0) \leq 0$ for all $i \in N$, if we want to maximize $\Pi(f,p)$, we must set $\pi_i(0) = 0$ for all $i \in N$. As a result, the optimization problem only involves finding the allocation rule, and the payment rule can be computed using Theorem 6 and setting $\pi_i(0) = 0$ for all $i \in N$. So, we can succinctly write down the optimal mechanism optimization problem.

$$\max_f \int_{T^n}\left[\sum_{i\in N} w_i(t_i)f_i(t)\right]g(t)dt$$
$$\text{subject to} \quad f \text{ is NDE.}$$

The term in the objective function is exactly the **total expected virtual valuation** from an allocation rule. This is because, the term $\sum_{i\in N} w_i(t_i)f_i(t)$ is the total *realized* virtual valuation of all bidders at type profile t from allocation rule f. This observation leads to the following important result.

THEOREM **8** *The allocation rule in an optimal mechanism maximizes the total expected virtual valuation among all Bayes-Nash implementable (NDE) allocation rules.*

We will first investigate what happens to the optimal mechanism without the NDE constraint on the allocation rule. We will call this the **unconstrained optimal mechanism**. There are two immediate corollaries to Theorem 8.

COROLLARY **2** *Suppose the NDE constraint is ignored. Then, an unconstrained optimal mechanism is deterministic: at every type profile, it allocates the object with probability to one to the highest virtual valuation buyer if the highest virtual valuation is non-negative; else it does not allocate the object.*

Proof: Without the constraint that f has to be NDE, we can maximize our objective function by doing a *point-wise* maximization. In particular, at every type profile t, we assign $f_i(t) = 0$ for all $i \in N$ if $w_i(t_i) < 0$ for all $i \in N$; else we assign $f_i(t) = 1$ for some $i \in N$ such that $w_i(t_i) \geq w_j(t_j)$ for all $j \neq i$. In other words, the highest virtual valuation agent wins the object if he has non-negative virtual valuation, else the object is unsold. Clearly, this maximizes the objective function without the NDE constraint. ■

There is also another interesting corollary to Theorem 8. Call an allocation rule f satisfies **no wastage** if $\sum_{i \in N} f_i(t) = 1$ for every type profile t. We will consider the revenue maximization over the class of all no wastage allocation rules.

COROLLARY **3** *Suppose the NDE constraint is ignored. Then, an unconstrained optimal mechanism is over the class of no wastage allocation rules is deterministic: it allocates the object with probability to one to the highest virtual valuation buyer at every type profile.*

The proof of Corollary 3 is identical to the proof of Corollary 2 except that we have to allocate at every type profile. Corollary 3 leads to the following corollary.

DEFINITION **14** *A virtual valuation w_i of agent i is* **regular** *if for all $s_i, t_i \in T_i$ with $s_i > t_i$, we have $w_i(s_i) > w_i(t_i)$.*

Regularity requires that the virtual valuation functions are strictly increasing. The following condition on distributions ensures that regularity holds. The hazard rate of a distribution g_i is defined as $\lambda_i(t_i) = \frac{g_i(t_i)}{1-G_i(t_i)}$ for all $i \in N$.

LEMMA **12** *If the hazard rate λ_i is non-decreasing, then the virtual valuation w_i is regular.*

Proof: Consider $s_i, t_i \in T_i$ such that $s_i > t_i$. Then,

$$w_i(s_i) = s_i - \frac{1}{\lambda_i(s_i)} > t_i - \frac{1}{\lambda_i(t_i)} = w_i(t_i).$$

■

The uniform distribution satisfies the non-decreasing hazard rate condition. Because $\frac{1-G_i(t_i)}{g_i(t_i)} = b_i - t_i$, which is non-increasing in t_i. For the exponential distribution, $g_i(t_i) = \mu e^{-\mu t_i}$ and $G_i(t_i) = 1 - e^{-\mu t_i}$. Hence, $\frac{1-G_i(t_i)}{g_i(t_i)} = \frac{1}{\mu}$, which is a constant. So, the exponential distribution also satisfies the non-decreasing hazard rate condition.

We say buyers are **symmetric** if they draw their values from identical distributions: $b_i = b_j$ and $G_i = G_j$ for each $i, j \in N$. We call the common distribution G.

COROLLARY **4** *Suppose buyers are symmetric and distribution is regular. Then, the Vickrey auction is the optimal mechanism over the class of no wastage allocation rules.*

Proof: We already know from Corollary 3 that the unconstrained optimal mechanism over no wastage allocation rules allocates the object to the highest virtual valuation buyer at every type profile. If buyers are symmetric, then each buyer has the same virtual valuation function: denote it as w. Since the distribution is regular, we see that the highest virtual valuation buyer is also the highest valuation buyer. So, the allocation rule is efficient. We know that the lowest type payment is zero. By revenue equivalence, the payment corresponds to the Vickrey auction. ■

In general, we are not interested in imposing no wastage. Then, we start from Corollary 2. It may so happen that the optimal solution obtained may not satisfy the NDE constraint. Below, we impose conditions on the distributions of agents that ensure that the unconstrained optimal solution satisfies the constraints, and hence, a constrained optimal solution.

This leads to our main observation.

LEMMA **13** *Suppose regularity holds for the virtual valuation of each agent. Then, the allocation rule in the optimal mechanism solves the following unconstrained optimization problem.*

$$\max_{f} \int_{T^n} \left[\sum_{i \in N} w_i(t_i) f_i(t) \right] g(t) dt.$$

Proof: We have already seen that the optimal solution to the unconstrained optimization problem is done as follows: for every type profile t, $f_i(t) = 0$ for all $i \in N$ if $w_i(t_i) < 0$ for all $i \in N$ and $f_i(t) = 1$ for some $i \in N$ if $w_i(t_i) \geq 0$ and $w_i(t_i) \geq w_j(t_j)$ for all $j \in N$. If the regularity condition holds, then f is NDE. To see this, consider a bidder $i \in N$ and $s_i, t_i \in T_i$ with $s_i > t_i$. Regularity gives us $w_i(s_i) > w_i(t_i)$. By the definition of the allocation rule, for all $t_{-i} \in T_{-i}$, we have $f_i(s_i, t_{-i}) \geq f_i(t_i, t_{-i})$. Hence, f is non-decreasing, and hence, it is NDE. ■

Our discussions to the main theorem of this section.

THEOREM **9** *Suppose the regularity holds for each agent. Consider the following allocation rule f^*. For every type profile $t \in T^n$, $f_i^*(t) = 0$ if $w_i(t_i) < 0$ for all $i \in N$ and else, $f_i^*(t) = 1$ for some $i \in N$ such that $w_i(t_i) \geq 0$, $w_i(t_i) \geq w_j(t_j)\ \forall\ j \in N$. There exists payments $(p_1, \ldots, p_n)$ such that $(f^*, p_1, \ldots, p_n)$ is an optimal mechanism.*

We now come back to the payments. To remind, we need to ensure that payments satisfy the revenue equivalence and $\pi_i(0) = 0$ for all $i \in N$. Since f^* can be implemented in dominant strategies and it is a deterministic allocation rule, we can ensure this by satisfying the revenue equivalence formulae for the dominant strategy case (which simplifies if the allocation rule is deterministic) and setting $p_i(0, t_{-i}) = 0$ for all i and for all t_{-i}. From our earlier analysis, the payment then is uniquely determined as the following (from Theorem 5).

For every $i \in N$ and for every t_{-i}, let $\kappa_i^{f^*}(t_{-i}) = \inf\{t_i : f_i^*(t_i, t_{-i}) = 1\}$. If $f^*(t_i, t_{-i}) = 0$ for all $t_i \in T_i$, then set $\kappa_i^{f^*}(t_{-i}) = 0$.

THEOREM **10** *Suppose the regularity holds for each agent. Consider the following allocation rule f^*. For every type profile $t \in T^n$, For every type profile $t \in T^n$, $f_i^*(t) = 0$ if $w_i(t_i) < 0$ for all $i \in N$ and else, $f_i^*(t) = 1$ for some $i \in N$ such that $w_i(t_i) \geq 0$, $w_i(t_i) \geq w_j(t_j)\ \forall\ j \in N$. For every agent $i \in N$, consider the following payment rule. For every $(t_i, t_{-i}) \in T^n$,*

$$p_i^*(t_i, t_{-i}) = \begin{cases} 0 & \text{if } f_i^*(t_i, t_{-i}) = 0 \\ \kappa_i^{f^*}(t_{-i}) & \text{if } f_i^*(t_i, t_{-i}) = 1 \end{cases}$$

The mechanism $(f^*, p_1^*, \ldots, p_n^*)$ is an optimal mechanism.

Proof: By Theorem 9, there is an optimal mechanism involving f^*. Under regularity, f^* is non-decreasing, and hence, dominant strategy implementable. For the mechanism to be optimal, we only need to show that $(p_1^*, \ldots, p_n^*)$ satisfy the revenue equivalence formulae in Theorem 6 with $\pi_i^*(0) = 0$ for all $i \in N$.

The payments $(p_1^*, \ldots, p_n^*)$ satisfy the revenue equivalence formula in Theorem 5. Hence, by Theorem 5, $(f^*, p_1^*, \ldots, p_n^*)$ is dominant strategy incentive compatible, and hence, BIC. So, they satisfy the revenue equivalence formula in Theorem 6. Since $p_i^*(0, t_{-i}) = 0$ for all $i \in N$ and for all t_{-i}, we have $\pi_i^*(t_i) = 0$ for all $i \in N$ and for all $t_i \in T_i$. This shows that $(f^*, p_1^*, \ldots, p_n^*)$ is an optimal mechanism. ■

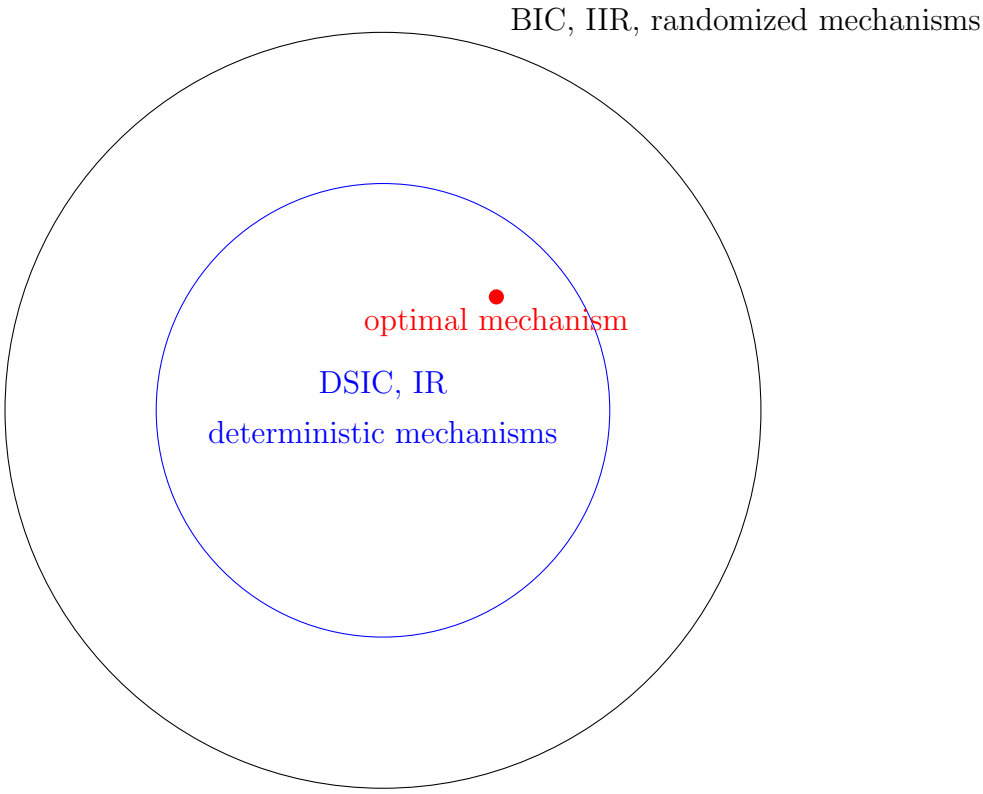

Figure 3.7: Optimal mechanism is DSIC, IR, and deterministic

Figure 3.7 highlights the fact that we started out searching for an optimal mechanism in a large family of BIC, IIR, and randomized mechanisms. But the optimal mechanism turned out to be DSIC, IR, and deterministic.

If the regularity condition does not hold, the optimal mechanism is more complicated, and you can refer to Myerson's paper for a complete treatment.

Symmetric Bidders

Finally, we look at the special case where the buyers are **symmetric**, i.e., they draw the valuations using the same distribution - $g_i = g$ and $T_1 = T_2 = \ldots = T_n$ for all $i \in N$. So, virtual valuations are the same: $w_i = w$ for all $i \in N$. In this case $w(t_i) > w(t_j)$ if and only if $t_i > t_j$ by regularity. Hence, maximum virtual valuation corresponds to the maximum valuation.

Thus, $\kappa_i(t_{-i}) = \max\{w^{-1}(0), \max_{j \neq i} t_j\}$. This is exactly, the second-price auction with the reserve price of $w^{-1}(0)$. Hence, when the buyers are symmetric, then the second-price

auction with a reserve price equal to $w^{-1}(0)$ is optimal.

An Example

Consider a setting with two buyers whose values are distributed uniformly in the intervals $T_1 = [0, 12]$ (buyer 1) and $T_2 = [0, 18]$ (buyer 2). Virtual valuation functions of buyer 1 and buyer 2 are given as:

$$w_1(t_1) = t_1 - \frac{1 - G_1(t_1)}{g_1(t_1)} = t_1 - (12 - t_1) = 2t_1 - 12$$
$$w_2(t_2) = t_2 - \frac{1 - G_2(t_2)}{g_2(t_2)} = t_2 - (18 - t_2) = 2t_2 - 18.$$

Hence, the reserve prices for both the bidders are respectively $r_1 = 6$ and $r_2 = 9$. The optimal auction outcomes are shown for some instances in Table 3.1.

Valuations	Allocation (who gets object)	Payment of Buyer 1	Payment of Buyer 2
$(t_1 = 4, t_2 = 8)$	Object not sold	0	0
$(t_1 = 2, t_2 = 12)$	Buyer 2	0	9
$(t_1 = 6, t_2 = 6)$	Buyer 1	6	0
$(t_1 = 9, t_2 = 9)$	Buyer 1	6	0
$(t_1 = 8, t_2 = 15)$	Buyer 2	0	11

Table 3.1: Description of Optimal Mechanism

Efficiency and Optimality

One of the conclusions that we can draw from the previous analysis is that the optimal mechanism is not efficient. We illustrate this with an example. Suppose there are two agents $N = \{1, 2\}$. Suppose $T_1 = [0, 10]$ and $T_2 = [0, 6]$. To compute the optimal mechanism, we need to compute the virtual valuation functions. For agent 1, for every $t_1 \in T_1$, we have

$$w_1(t_1) = 2t_1 - 10.$$

For agent 2, for every $t_2 \in T_2$, we have

$$w_2(t_2) = 2t_2 - 6.$$

The optimal mechanism is shown in Figure 3.8. Notice that the object is unsold if $t_1 < 5$ and $t_2 < 3$. This is inefficient. This inefficiency occurs because of the reserve prices in the optimal mechanism. There is another source of inefficiency. Efficiency requires that agent 1 wins the object if $t_1 > t_2$. However, the optimal mechanism requires that $2t_1 - 10 \geq 0$ and $2t_1 - 10 \geq 2t_2 - 6$. This means, agent 2 wins the object in some cases where agent 1 should have won - this is shown in Figure 3.8. For instance, at the type profile $(5, 4)$, we have $2t_2 - 6 = 2 > 0 = 2t_1 - 5$. Hence, agent 2 wins the object, but efficiency requires agent 1 must win the object here. This inefficiency occurs because the virtual valuation function of both the agents is not the same, which happens because the distribution of values is asymmetric. When bidders are symmetric, this source of inefficiency disappears. So, with symmetric bidders, whenever the object is allocated, it is allocated efficiently.

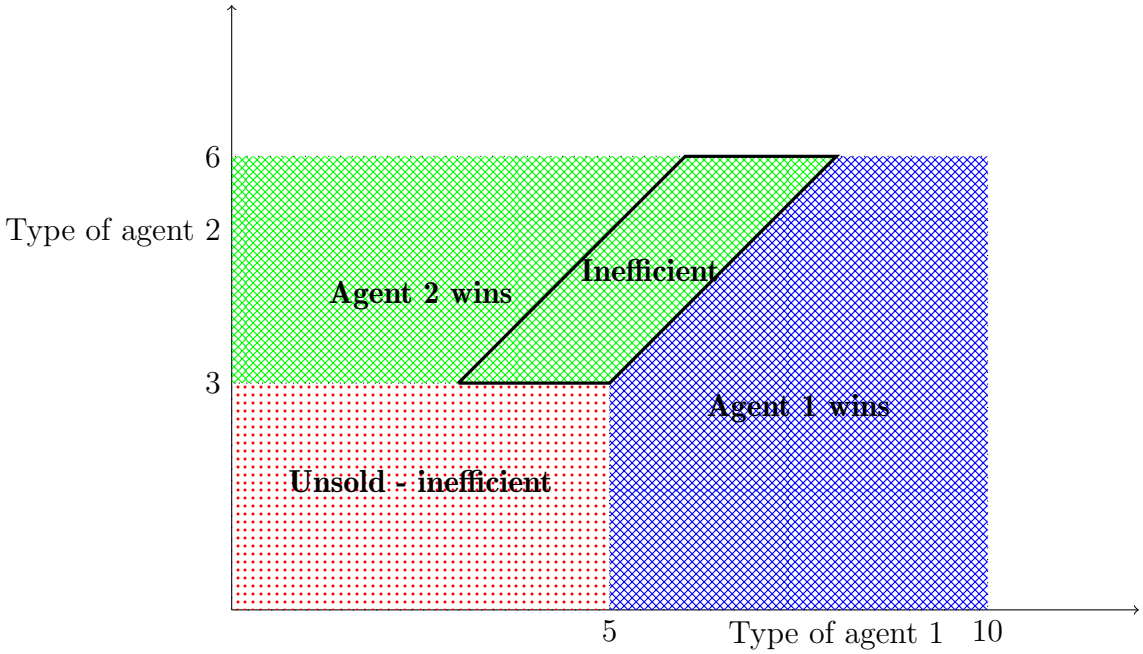

Figure 3.8: Inefficiency of optimal mechanism

Auction versus negotiation

Our analysis also leads to an interesting observation. This observation was noted in greater generality in Bulow and Klemperer (1996). Consider the symmetric buyer case. The optimal auction is a Vickrey auction with a *reserve price*. The reserve price depends on the prior

information. If the seller did not know the prior well, then it is very difficult to set the correct reserve price.

Now, consider a model with n symmetric bidders with a regular distribution. Suppose we could **hire** another symmetric bidder for "free". Then, we make the following claim.

THEOREM **11** *Suppose buyers are symmetric and distribution is regular. The Vickrey auction (without any reserve price) for* $(n+1)$ *bidders generates more expected revenue than the optimal mechanism with* n *bidders.*

Proof: Consider the following mechanism for $(n+1)$ bidders. Pick some bidder, say $(n+1)$, and conduct the optimal n-bidder mechanism for bidders $1, \ldots, n$. If the object is not sold in this n-bidder optimal mechanism, give it to bidder $(n+1)$ for free. Note that this mechanism is BIC and satisfies IR. Further, its revenue is at least as much as the revenue from the n-bidder optimal mechanism. Finally, this mechanism satisfies no wastage. By Corollary 4, the $(n+1)$ bidder Vickrey auction generates more expected revenue than this mechanism. This concludes the proof. ■

The result in Bulow and Klemperer (1996) is more general than this. But the result hints that if the cost of hiring an extra bidder is less than the optimal and prior dependent mechanism can be replaced by a prior-free mechanism.

Approximation for asymmetric case

We saw that the Vickrey auction with a reserve price equal to inverse of virtual value at zero is an optimal mechanism in the symmetric setting. However, the optimal mechanism is complex in the asymmetric case. The seller needs to know the exact virtual value function to figure out the optimal mechanism. In contrast, in the symmetric case, the seller only needs to know the value at which the buyer's virtual value crosses zero.

Hartline and Roughgarden ask the extent of loss in expected revenue if one restricts attention to Vickrey auction with *monopoly reserves*. They assume that the distributions of each bidder satisfies **monotone hazard rate** property, which implies that the virtual value function is increasing.

They consider Vickrey auction with non-anonymous reserve prices: denote this as (q, p).

In particular, the reserve price of bidder i is

$$r_i = w_i^{-1}(0)$$

where w_i^{-1} is the inverse of the virtual value function of bidder i. Note that the Vickrey auction with reserve prices $(r_1, \ldots, r_n)$ is the following mechanism. At every type profile t, let $E(t)$ be the set of bidders whose value is greater than the reserve price:

$$E(t) = \{i \in N : t_i \geq r_i\}$$

At any valuation profile t, if $E(t) = \emptyset$, then the object is not allocated. Else, if $E(t) \neq \emptyset$, then the bidder with the highest value in $E(t)$ is allocated the object: i.e., $q_i(t) = 1$ implies $t_i \geq r_i$ and $t_i \geq t_j$ for all $j \in E(t)$.

Two remarks about the mechanism (q, p). First, q is a deterministic and increasing allocation rule: $q_i(t_i, t_{-i}) \geq q_i(t_i', t_{-i})$ for all i, for all t_{-i}, and for all $t_i > t_i'$. Second, the payment can be uniquely determined as follows. If i is not allocated, she pays zero. If $q_i(t_i, t_{-i}) = 1$, then the least type where i wins the object is her payment. To determine this, note that for any $j \neq i$, if $j \in E(t_i, t_{-i})$, then $j \in E(t_i', t_{-i})$ for all t_i'. Hence, as bidder i varies (lowers) its value, the set of other bidders in E do not change. So, to be the winner, value of bidder i has to be greater than r_i and higher than values of other bidders in $E(t_i, t_{-i})$. Hence, payment of bidder i is $\max(r_i, \max_{j \in E(t_i, t_{-i}), j \neq i} t_j)$.

The second observation is that the set of type profiles where the objects is not allocated is the same in both the optimal mechanism and in (q, p): these are the valuation profiles where every bidder has negative virtual value, i.e., the valuation profiles $t \equiv (t_1, \ldots, t_n)$ such that $t_i < r_i$ for all $i \in N$. Further, whenever the object is allocated to a bidder in both the mechanisms, mechanism (q, p) allocates efficiently.

THEOREM **12** *The expected revenue from Vickrey auction with reserve prices $(r_1, \ldots, r_n)$ is at least $\frac{1}{2}$ of the expected revenue from an optimal mechanism.*

Proof: Let (q, p) be the Vickrey auction with reserve prices $(r_1, \ldots, r_n)$ and (q^*, p^*) be the optimal mechanism. Let T^+ be the set of type profiles where the object is allocated in mechanism (q, p). By construction, T^+ is also the set of type profiles where the object is allocated in (q^*, p^*).

By Theorem 8, we know that the expected payment from q (denoted as $\textsc{Rev}(q,p)$) is

$$\textsc{Rev}(q,p) = \int_{t\in T^+} \big[\sum_{i=1}^{n} w_i(t_i)q_i(t)\big]g(t)dt$$

But in this auction, each bidder i pays at least r_i when she wins. So,

$$\textsc{Rev}(q,p) \geq \int_{t\in T^+} \big[\sum_{i=1}^{n} r_i q_i(t)\big]g(t)dt$$

Combining these two, we get

$$2\textsc{Rev}(q,p) \geq \int_{t\in T^+} \big[\sum_{i=1}^{n} (w_i(t_i)+r_i)q_i(t)\big]g(t)dt \tag{3.4}$$

Now, if $q_i(t) > 0$, then $t_i \geq r_i$. Hence, $r_i + w_i(t_i) = r_i + t_i - \frac{1-G_i(t_i)}{g_i(t_i)} \geq r_i + t_i - \frac{1-G_i(r_i)}{g_i(r_i)}$, where the inequality follows from monotone hazard rate and $t_i \geq r_i$. But, by definition $r_i - \frac{1-G_i(r_i)}{g_i(r_i)} = 0$. Hence, we get $r_i + w_i(t_i) \geq t_i$ for all t_i such that $q_i(t) > 0$. Hence,

$$2\textsc{Rev}(q,p) \geq \int_{t\in T^+} \big[\sum_{i=1}^{n} t_i q_i(t)\big]g(t)dt \tag{3.5}$$

But at any type profile $t \in T^+$ we know that the Vickrey auction is efficient (but the optimal mechanism may not be efficient). Hence, we know for any type profile $t \in T^+$

$$\sum_{i=1}^{n} t_i q_i(t) \geq \sum_{i=1}^{n} t_i q_i^*(t) \geq \sum_{i=1}^{n} p_i^*(t),$$

The second inequality follows because there is an optimal mechanism which is ex-post individually rational, and hence, $t_i q_i^*(t) \geq p_i^*(t)$ for every i and every v. Further, for all $t \notin T^+$, $p_i^*(t) = 0$ for all $i \in N$. Hence, the RHS of the above expression is just $\textsc{Rev}(q^*, p^*)$.

Using (3.5), we conclude that

$$2\textsc{Rev}(q,p) \geq \int_{t\in T^+} \big[\sum_{i=1}^{n} p_i^*(t))\big]g(t)dt = \textsc{Rev}(q^*,p^*)$$

This completes the proof. ■

3.4 Correlation and full surplus extraction

We study the single object auction problem again but with correlated types. The model is the same. There are n buyers and one seller with an object. The seller has zero value for the object. Value of each buyer i is a random variable $v_i \in V_i$, where V_i is a finite set. Let $V := V_1 \times \ldots \times V_n$ and for each $i \in N$ let $V_{-i} := \times_{j \neq i} V_j$. The joint distribution of value of buyers is given by a probability distribution function π, where we denote the probability that valuation profile $(v_1, \ldots, v_n)$ being realized as $\pi(v_1, \ldots, v_n)$. Each buyer i observes his valuation v_i but not the valuations of others. Hence, conditional probability distributions are relevant in this context. We will denote by $\pi(v_{-i}|v_i) := \frac{\pi(v_i, v_{-i})}{\sum_{v'_{-i} \in V_{-i}} \pi(v_i, v'_{-i})}$ the conditional probability of valuation profile v_{-i} appearing when buyer i has value v_i. An **information structure** is (V, π). Throughout, we will assume that the information structure satisfies $\pi_i(v_{-i}|v_i) > 0$ for all v_{-i} and for all v_i, for all $i \in N$.

We are interested in dominant strategy incentive compatible mechanism. Due to revelation principle, a mechanism is specified by two rules: (a) **allocation rule**, $Q_i : V \to [0, 1]$ for each i, where $Q_i(v)$ is the probability of getting object for buyer i at valuation profile v - feasibility is assumed, i.e., $\sum_{i \in N} Q_i(v) \leq 1$ for all $v \in V$; (b) **payment rule**, $P_i : V \to \mathbb{R}$ for each i.

Definition **15** *An auction* $(\{Q_i, P_i\}_{i \in N})$ *is* **dominant strategy incentive compatible (DSIC)** *if for every* $i \in N$*, every* $v_{-i} \in V_{-i}$*, and every* $v_i, v'_i \in V_i$*, we have*

$$v_i Q_i(v_i, v_{-i}) - P_i(v_i, v_{-i}) \geq v_i Q_i(v'_i, v_{-i}) - P_i(v'_i, v_{-i}).$$

The notion of participation we use is interim individual rationality.

Definition **16** *An auction* $(\{Q_i, P_i\}_{i \in N})$ *is* **interim individually rational (IIR)** *if for every* $i \in N$ *and every* $v_i \in V_i$*, we have*

$$\sum_{v_{-i} \in V_{-i}} \Big[v_i Q_i(v_i, v_{-i}) - P_i(v_i, v_{-i}) \Big] \pi(v_{-i}|v_i) \geq 0.$$

The payoff to the seller in an auction is her expected revenue. We are interested in designing a DSIC mechanism that can *extract entire surplus* from the buyers.

DEFINITION **17** *An information structure* (V, π) *guarantees* **full surplus extraction** *in DSIC if there exists a DSIC auction* $(\{Q_i, P_i\}_{i \in N})$ *satisfying IIR such that*

$$\sum_{v \in V} \pi(v) \sum_{i \in N} P_i(v) = \sum_{v \in V} \pi(v) \max_{i \in N} v_i.$$

The interest in full surplus extraction is natural. If the seller knew the valuations of all the agents, then it can go the highest valued agent and offer her the good at the value, and hence, extracting full surplus in that case is always possible. On the other hand, if full surplus can be extracted from the agents, when they have their value information private is truely remarkable. This means that agents need not be assigned any *information rent*.

Why do we expect that full surplus extraction is possible in this setting? First, observe that for full surplus extraction, we need to assign the object to the highest valued agent. Hence, the allocation rule in any such mechanism must be efficient. Then, we can use some Groves mechanism. To get some intuition why we can construct a Groves mechanism, let us consider a two bidder model: $N = \{1, 2\}$. Suppose both bidders draw values from a set $\{v^1, \ldots, v^K\}$.

Let us focus on bidder 1 and let Π^1 denote her conditional probability matrix. So, the (i, j)-th entry in Π^1 gives the conditional probability that bidder 2 has value v^j conditional on bidder 1 has value v^i. So, the row sum of Π^1 is one.

Cremer and McLean (1988) propose the following mechanism. First, before the bidders participate in the mechanism, they accept a "lottery" over other bidder's values. You can view it as a participation fee for each bidder which is conditioned on the value of the other bidder. Second, the bidders participate in a Vickrey auction. Notice that the first phase participation fee does not distort incentives as those fees depend on the values of the other bidders. Now, conditional on value v^k being realized, suppose the expected payoff of bidder 1 from the Vickrey auction is $U^*(v^k)$. Then, given a participation fee map $c_1 : \{v^1, \ldots, v^K\} \to \mathbb{R}$, we need to ensure that

$$\sum_{k'=1}^{K} c_1(v^{k'}) \Pi^1_{k',k} = U^*(v^k) \qquad \forall\ k \in \{1, \ldots, K\}.$$

If we treat c_1 as some kind of variable in this equation, we can ensure a solution to this system. Suppose the Π^1 matrix is invertible, then such participation fees can be constructed. The invertibility will require a full rank condition on Π^1. This puts some restriction on the

extent of correlation. For instance, each row in Π^1 is not in the span of other rows in the matrix. Notice that this is not possible with independence as every row is *identical* in Π^1 if values are independent.

The following theorem characterizes information structures that allow full surplus extraction.

The proof uses Farkas lemma, which is a useful tool to remember - see any standard textbook on linear programming for a proof.

LEMMA **14 (Farkas Lemma)** *Let a_{ij} and b_i be real numbers for each $i \in \{1, \ldots, m\}$ and for each $j \in \{1, \ldots, n\}$. Consider the following two sets.*

$$F = \{x \in \mathbb{R}^n_+ : \sum_{j=1}^{n} a_{ij} x_j = b_i \ \forall \ i \in \{1, \ldots, m\}\}$$

$$G = \{y \in \mathbb{R}^m : \sum_{i=1}^{m} a_{ij} y_i \geq 0 \ \forall \ j \in \{1, \ldots, n\}, \sum_{i=1}^{m} b_i y_i < 0\}.$$

Then, $F \neq \emptyset$ if and only if $G = \emptyset$. The claim continues to hold if x is not restricted to be non-negative in F if we make the inequalities in G equalities.

The set G is often called the **Farkas alternative** of F. The Cremer-McLean full surplus extraction is the following.

THEOREM **13** *An information structure (V, π) guarantees full surplus extraction in DSIC if for all $i \in N$, there exists no $\rho_i : V_i \to \mathbb{R}$ with $\rho_i(v_i) \neq 0$ for some $v_i \in V_i$ such that*

$$\sum_{v_i \in V_i} \rho_i(v_i) \pi(v_{-i}|v_i) = 0 \qquad \forall \ v_{-i} \in V_{-i}.$$

Proof: Consider the Vickrey auction - a DSIC auction. Let $(\{Q_i^*, P_i^*\}_{i \in N})$ be the allocation and payment rules in the Vickrey auction, where Q^* is the efficient allocation rule. For every $i \in N$ and for every $v_i \in V_i$, denote the net utility of agent i with value v_i in the Vickrey auction as

$$U_i^*(v_i) := \sum_{v_{-i}} \Big[Q_i^*(v_i, v_{-i}) v_i - P_i^*(v_i, v_{-i}) \Big] \pi(v_{-i}|v_i).$$

Denoting $\pi(v_i) = \sum_{v'_{-i}} \pi(v_i, v'_{-i})$, we rewrite the above as

$$\sum_{v_i \in V_i} U_i^*(v_i) \pi(v_i) = \sum_{v \in V} \Big[Q_i^*(v) v_i - P_i^*(v) \Big] \pi(v). \tag{3.6}$$

By definition of the Vickrey auction for every $i \in N$, we have $U_i^*(v_i) \geq 0$ for all $v_i \in V_i$ with strict inequaltiy holding for some v_i. Further, since $Q_i^*(v_i, v_{-i}) = 0$ if $v_i < \max_{j \in N} v_j$ and $Q_i^*(v_i, v_{-i}) = 1$ for some $i \in \arg\max_{j \in N} v_j$, we can write

$$\begin{aligned}\sum_{v \in V} \pi(v) \max_{i \in N} v_i &= \sum_{v \in V} \pi(v) \sum_{i \in N} Q_i^*(v_i, v_{-i}) v_i \\ &= \sum_{i \in N} \Big[\sum_{v_i \in V_i} U_i^*(v_i)\pi(v_i) + \sum_{v \in V} P_i^*(v)\pi(v) \Big].\end{aligned}$$

Hence, full surplus extraction is possible if and only if there exists a DSIC auction $(\{Q_i, P_i\}_{i \in N})$ such that

$$\sum_{v \in V} \pi(v) \sum_{i \in N} P_i(v) = \sum_{i \in N} \Big[\sum_{v_i \in V_i} U_i^*(v_i)\pi(v_i) + \sum_{v \in V} P_i^*(v)\pi(v) \Big].$$

Now, we construct a function $c_i : V_{-i} \to \mathbb{R}$ for every agent i such that

$$\sum_{v_{-i}} c_i(v_{-i})\pi(v_{-i}|v_i) = U_i^*(v_i) \qquad \forall\ v_i \in V_i. \tag{3.7}$$

Notice that if such a function can be constructed for every i, then Equation 3.7 guarantees

$$\sum_{v \in V} c_i(v_{-i})\pi(v) = \sum_{v_i \in V_i} U_i^*(v_i)\pi(v_i).$$

Thus, existence of c_i for each $i \in N$ satisfying Equation 3.7 guarantees another mechanism $(\{Q_i^*, P_i\}_{i \in})$, where for every $i \in N$ and every v, we have $P_i(v) := c_i(v_{-i}) + P_i^*(v)$ such that

$$\begin{aligned}\sum_{v \in V} \pi(v) \sum_{i \in N} P_i(v) &= \sum_{v \in V} \pi(v) \sum_{i \in N} \big[c_i(v_{-i}) + P_i^*(v) \big] \\ &= \sum_{i \in N} \Big[\sum_{v_i \in V_i} U_i^*(v_i)\pi(v_i) + \sum_{v \in V} P_i^*(v)\pi(v) \Big],\end{aligned}$$

Note that the interim payoff of every agent at every type is zero in the new mechanism. Hence, the new mechanism satisfies IIR. This along with the fact that $(\{Q_i^*, P_i\}_{i \in N})$ is DSIC (this is because we just added payment terms to a DSIC mechanism Vickrey that does not depend on every agent's valuation) ensures full surplus extraction is possible. We show that Equation 3.7 has a solution if the condition in the theorem holds. To see this, construct the Farkas alternative for each $i \in N$:

$$\sum_{v_i \in V_i} \rho_i(v_i) U_i^*(v_i) < 0 \tag{3.8}$$

$$\sum_{v_i \in V_i} \rho_i(v_i)\pi(v_{-i}|v_i) = 0 \qquad \forall\ v_{-i} \in V_{-i}. \tag{3.9}$$

By the condition in the theorem, the only solution to Equation 3.9 is $\rho_i(v_i) = 0$ for all $v_i \in V_i$. But this ensures that Inequality (3.8) is not satisfied. Hence, Farkas alternative has no solution and Equation 3.7 has a solution. ■

Theorem 13 uses finite type space. This is not really necessary. Cremer and McLean (1988) contains discussions about how this can be extended to infinite type spaces for dominant strategy mechanisms. Cremer and McLean (1988) also provide weaker conditions on information structures (with finite type space) under which a Bayesian incentive compatible mechanism can extract full surplus. This is extended to infinite type spaces in McAfee and Reny (1992).

Though remarkable, the Cremer-McLean full surplus extraction result has its own critics. First, the participation fees before the Vickrey auction in their mechanism may be quite high. This means, the mechanism will fail ex-post IR. Further, it is not clear how "generic" the set of beliefs is, the extent of correlation it allows, and so on. Thus, the Cremer-McLean result, though an important benchmark, remains a paradox in the mechanism design literature.

Chapter 4

Redistribution mechanisms

Redistribution is a fundamental objective in many models of mechanism design. A Government wants to redistribute the surplus it generates from selling public assets; buyers and sellers trade to redistribute surplus gains; firms are merged to redistribute gains from synergies etc. Redistribution is different from the objectives we have seen earlier, where the mechanism designer (seller) wanted to maximize her expected revenue. An objective of redistribution is to efficiently allocate the resource *without* wasting any transfer, i.e., the payments have to balance. The balancing of payment makes the redistribution problem quite different from others. Of course, the incentive and participation constraints continue to be there. We consider the problem of *redistributing* the surplus from a single object. Most of the machinery developed in the optimal mechanism design will be useful because the incentive and participation constraints remain the same.

The problem is of great practical interest. Land acquisition and redistribution has been one of the major problems in many countries. In such problems, many stakeholders "own" various portions of a land. For efficiency, we would like to allocate the entire land to one stakeholder (of course, in some settings, this may not be legally possible). Because of property rights, the other owners need to be properly compensated. Often Government allocated resources like Spectrum or Mines need to be reallocated. How should heirs divide an estate? Sometimes, there is a will and sometimes there is no will. Apparently, a commonly used method by estate agents is an auction whose revenue is "shared" by heirs.

4.1 A model of redistributing a single object

There is one unit of a divisible good which is jointly owned by a set of agents $N := \{1, \ldots, n\}$. The share of agent i of the good is denoted by $r_i \in [0, 1]$ and these shares add up to one - $\sum_{j \in N} r_j = 1$. Two particular configuration of shares are worth pointing out.

- **One seller many buyers model.** Here, $r_i = 1$ for some $i \in N$ and $r_j = 0$ for all $j \neq i$.
- **Equal partnership model.** Here, $r_i = \frac{1}{n}$ for all $i \in N$.
- **One buyer many sellers model.** Here $r_i = 0$ for some $i \in N$ and $r_j = \frac{1}{n-1}$ for all $j \neq i$.

The *per unit* value of each agent for the good is denoted by v_i, which is independently distributed in $V \equiv [0, \beta]$ with an absolutely continuous distribution F with density f. So, we only consider a very symmetric setting where agents are identical ex-ante. The extension to asymmetric case is not very clean. Let $G(x) = [F(x)]^{n-1}$ for all $x \in [0, \beta]$. Notice that G is the cdf of the random variable which is the maximum of $(n-1)$ independent draws using F.

A **mechanism** is a collection of pair of maps $\{Q_i, T_i\}_{i \in N}$, where for each $i \in N$, $Q_i : V^n \to [0, 1]$ is the share allocation rule of agent i and $T_i : V^n \to \mathbb{R}$ is the transfer rule (amount *paid to*) of agent i. A mechanism is **feasible** if for all $v \in V^n$

(a) the allocation rules are feasible, i.e., $\sum_{i \in N} Q_i(v) \leq 1$ and

(b) transfer rules are budget balanced, i.e., $\sum_{i \in N} T_i(v) = 0$.

We will be interested in interim allocation probabilities and interim payments of agents for a given feasible mechanism. Fix a feasible mechanism $\{Q_i, T_i\}_{i \in N}$. Define for every $i \in N$ and every $v_i \in V$,

$$q_i(v_i) = \int_{v_{-i} \in V^{n-1}} Q_i(v_i, v_{-i}) d(F_{N-i}(v_{-i}))$$
$$t_i(v_i) = \int_{v_{-i} \in V^{n-1}} T_i(v_i, v_{-i}) d(F_{N-i}(v_{-i})),$$

where $F_{v_{N-i}} = \prod_{j \in N \setminus \{i\}} F(v_j)$. So, every feasible mechanism $\{Q_i, T_i\}_{i \in N}$ generates interim rules $\{q_i, t_i\}_{i \in N}$. With this, we can define the notion of Bayesian incentive compatibility.

DEFINITION **18** *A mechanism* $\{Q_i, T_i\}_{i \in N}$ *is* **Bayesian incentive compatible (BIC)** *if for every* $i \in N$

$$v_i q_i(v_i) + t_i(v_i) \geq v_i q_i(v_i') + t_i(v_i') \qquad \forall\ v_i, v_i' \in V.$$

The standard notion of individual rationality needs to be modified to ensure the fact that each agent has some property right.

DEFINITION **19** *A mechanism* $\{Q_i, T_i\}_{i \in N}$ *is* **individually rational (IR)** *if for every* $i \in N$

$$v_i q_i(v_i) + t_i(v_i) \geq r_i v_i \qquad \forall\ v_i \in V.$$

The individual rationality is the main point of departure from the earlier optimal mechanism model, where we just had to ensure non-negative payoffs to agents in the mechanism. Here, because of property rights, we need to ensure larger payoff share to some agents.

We are interested in knowing when we can redistribute the entire surplus.

DEFINITION **20** *A parternship* $\{r_i\}_{i \in N}$ *can be* **dissolved efficiently** *if there exists a feasible, efficient, Bayesian incentive compatible and individually rational mechanism* $\{Q_i, T_i\}_{i \in N}$ *for this partnership.*

4.2 Characterizations of IC and IR constraints

The first step involves characterizing BIC and IR constraints. This mirrors Myerson (1981) but with some minor differences to account for the difference in IR constraint and also keeping in mind the objective. For some of the proofs, we will use the following notation. Given a mechanism $\{Q_i, T_i\}_{i \in N}$, we will denote the interim utility of agent i with type v_i from this mechanism as

$$U_i(v_i) = v_i q_i(v_i) + t_i(v_i),$$

where we have supresed the notation indicating U_i depends on the mechanism $\{Q_i, T_i\}_{i \in N}$. Notice that IC is equivalently stated as for all $i \in N$

$$U_i(v_i) - U_i(v_i^*) \geq (v_i - v_i^*) q_i(v_i^*) \qquad \forall\ v_i, v_i^* \in V. \tag{4.1}$$

LEMMA **15 (IC Characterization)** *A mechanism $\{Q_i, T_i\}_{i\in N}$ is Bayesian incentive compatible if and only if for every $i \in N$*

- *q_i is non-decreasing*
- *$t_i(v_i^*) - t_i(v_i) = v_i q_i(v_i) - v_i^* q_i(v_i^*) - \int_{v_i^*}^{v_i} q_i(x)dx$ for all $v_i, v_i^* \in V$.*

Proof: The proof is same as earlier. **Sufficiency.** For every $i \in N$ and every $v_i, v_i^* \in V$, we have

$$\begin{aligned} U_i(v_i) - U_i(v_i^*) &= v_i q_i(v_i) - v_i^* q_i(v_i^*) + t_i(v_i) - t_i(v_i^*) \\ &= \int_{v_i^*}^{v_i} q_i(x)dx \\ &\geq (v_i - v_i^*) q_i(v_i^*), \end{aligned}$$

where the last inequality follows from the fact that q_i is non-decreasing. This is the relevant BIC constraint (4.1).

Necessity. BIC constraints (4.1) imply that for every $i \in N$, the function U_i is convex. Further, the BIC constraints (4.1) imply that q_i is subgradient of U_i at every point in V. Hence, q_i is non-decreasing and for every v_i, v_i^*, we have

$$\begin{aligned} & U_i(v_i) - U_i(v_i^*) = \int_{v_i^*}^{v_i} q_i(x)dx \\ \Leftrightarrow\ & t_i(v_i^*) - t_i(v_i) = v_i q_i(v_i) - v_i^* q_i(v_i^*) - \int_{v_i^*}^{v_i} q_i(x)dx. \end{aligned}$$

■

We now turn our attention to *efficient* mechanisms. An allocation rule $\{Q_i\}_{i\in S}$ is **efficient** if for every type profile $v \in V^n$, $Q_i(v) > 0$ implies $i \in \arg\max_{j\in N} v_j$ and $\sum_{j\in N} Q_j(v) = 1$. We will denote an allocation rule by $\{Q_i^e\}_{i\in N}$.

If $\{Q_i^e, T_i\}_{i\in N}$ is an efficient mechanism, then for every $i \in N$ and every $v_i \in V_i$, we have

$$q_i^e(v_i) = G(v_i) = [F(v_i)]^{n-1}.$$

Notice the following properties: (a) q_i^e is strictly increasing; (b) $q_i^e(0) = 0$ and $q_i^e(\beta) = 1$; (c) q_i^e is differentiable. We will use these properties repeatedly in the proofs.

The next lemma establishes an important property about efficient mechanisms.

LEMMA **16** *Suppose* $\{Q_i^e, T_i\}_{i\in N}$ *is an efficient and BIC mechanism. Let* v_i^* *be such that* $G(v_i^*) = r_i$*. Then, the following holds:*

$$U_i(v_i) - r_i v_i \geq U_i(v_i^*) - r_i v_i^* \qquad \forall\ v_i \in V.$$

Proof: Fix an agent $i \in N$. Since $q_i^e(0) = 0$ and $q_i^e(\beta) = 1$, and q_i^e is a strictly increasing continuous function, there is a unique v_i^* such that $q_i^e(v_i^*) = r_i$ or $G(v_i^*) = r_i$. By monotonicity, if $v_i > v_i^*$, we have $q_i^e(v_i) > r_i$ and if $v_i < v_i^*$, we have $q_i^e(v_i) < r_i$. Using this, we immediately get the following:

$$U_i(v_i) - r_i v_i - U_i(v_i^*) + r_i v_i^* = \int_{v_i^*}^{v_i} q_i^e(x)dx - r_i(v_i - v_i^*) \geq 0,$$

where the last inequality follows since $q_i^e(x) > r_i$ if $x > v_i^*$ and $q_i^e(x) < r_i$ if $x < v_i^*$. ■

The next lemma characterizes IR mechanisms.

LEMMA **17** *Suppose* $\{Q_i^e, T_i\}_{i\in N}$ *is an efficient and BIC mechanism. Then,* $\{Q_i^e, T_i\}_{i\in N}$ *is IR if and only if* $t_i(v_i^*) \geq 0$ *for all* $i \in N$*, where* v_i^* *is as defined in Lemma 16.*

Proof: By Lemma 16, an efficient mechanism $\{Q_i^e, T_i\}_{i\in N}$ is IR if and only if for every $i \in N$ and every $v_i \in V$, we have $U_i(v_i^*) - r_i v_i^* \geq 0$. This is equivalent to requiring $q_i^e(v_i^*)v_i^* + t_i(v_i^*) - r_i v_i^* \geq 0$. Using Lemma 16, we know that $q_i^e(v_i^*) = G_(v_i^*) = r_i$. Hence, the above is equivalent to requiring $t_i(v_i^*) \geq 0$. ■

The analysis in this section is not much different from the optimal mechanism analysis of incentive and IR constraints. The only difference is that with property rights, the type that gets the minimum payoff in a mechanism is not necessarily the lowest type. The lowest payoff type depends on the property right structure.

4.3 Dissolving a partnership

The main theorem is a characterization of partnerships that can be dissoved. This theorem is due to Cramton et al. (1987).

THEOREM **14** *A partnership $\{r_i\}_{i\in N}$ can be dissolved efficiently if and only if*

$$n\int_0^{\beta}(1-F(x))xg(x)dx \geq \sum_{i=1}^{n}\int_0^{v_i^*} xg(x)dx, \tag{4.2}$$

where $G(v_i^*) = r_i$.

Proof: **Necessity.** Suppose there is an efficient, BIC, and IR feasible mechanism $\{Q_i^e, T_i\}_{i\in N}$. Then, we know that

$$\sum_{i\in N}\int_0^{\beta} t_i(v_i)f(v_i)dv_i = \sum_{i\in N}\int_v T_i(v)f_1(v_1)\dots f_n(v_n)dv = \int_v \big(\sum_{i\in N} T_i(v)\big)f_1(v_1)\dots f_n(v_n)dv = 0,$$

where the last equality follows from budget-balance. By Lemma 15 for every $i \in N$ and every $v_i \in V$, we have

$$\begin{aligned} t_i(v_i) &= t_i(v_i^*) - v_iG(v_i) + v_i^*G(v_i^*) + \int_{v_i^*}^{v_i} G(x)dx \\ &= t_i(v_i^*) - \int_{v_i^*}^{v_i} xg(x)dx \\ &\geq \int_{v_i}^{v_i^*} xg(x)dx, \end{aligned}$$

where the inequality follows from Lemma 17. Hence, $\int_{v_i}^{v_i^*} xg(x)dx$ is the difference in interim payment of agent i with type v_i and v_i^*. Since $t_i(v_i^*) \geq 0$. This difference in interim payment is also a lower bound on the interim payment at v_i. The necessity part uses this insight along with the budget-balance condition. Essentially, it computes the ex-ante value of this lower bound. That is, the sum of expected interim payments of all agents (which is zero due to budget-balance) must be greater than or equal to the sum of expected value of these lower bounds. Hence, we get

$$\begin{aligned} 0 &= \sum_{i\in N}\int_0^{\beta} t_i(v_i)f(v_i)dv_i \\ &\geq \sum_{i\in N}\Big[\int_0^{\beta}\big(\int_{v_i}^{v_i^*} xg(x)dx\big)f(v_i)dv_i\Big] \\ &= \sum_{i\in N}\Big[\int_0^{\beta}\big(\int_0^{v_i^*} xg(x)dx\big)f(v_i)dv_i - \int_0^{\beta}\big(\int_0^{v_i} xg(x)dx\big)f(v_i)dv_i\Big] \\ &= \sum_{i\in N}\Big[\int_0^{v_i^*} xg(x)dx - \int_0^{\beta}(1-F(x)xg(x)dx\Big] \end{aligned}$$

This gives us the required necessary condition.

Sufficiency. Suppose Inequality (4.2) holds. We will show that partnership $\{r_i\}_{i\in N}$ can be dissolved efficiently. Define for every $i \in N$ and every $v_i \in V$,

$$W(v_i) := \int_0^\beta (1-(F(x))xg(x)dx - \int_0^{v_i} xg(x)dx. \tag{4.3}$$

Notice that Inequality (4.2) is $\sum_{i\in N} W(v_i^*) \geq 0$. Define the following constants for each agent $i \in N$:

$$c_i = \frac{1}{n}\sum_{j\in N} W(v_j^*) - W(v_i^*).$$

Note that $\sum_{i\in N} c_i = 0$. Now, we define the transfer functions for our efficient mechanism. For every $i \in N$ and for every type profile $v \in V^n$, let

$$T_i(v) := \Big[c_i + W(v_i) - \frac{1}{n-1}\sum_{j\in N\setminus\{i\}} W(v_j)\Big]$$

Since $\sum_{i\in N} c_i = 0$, we get $\sum_{i\in N} T_i(v) = 0$ for all $v \in V^n$. Also, notice that

$$\int_0^\beta W(v_j)f(v_j)dv_j = \int_0^\beta (1-(F(x))xg(x)dx - \int_0^\beta \big(\int_0^{v_j} xg(x)dx\big)f(v_j)dv_j = 0$$

Hence, we can compute the interim payments of agents for this transfer rule. Fix agent $i \in N$. Then, for every $v_i \in V$, we see that

$$t_i(v_i) = c_i + W(v_i) - \frac{1}{n-1}\sum_{j\in N\setminus\{i\}} \int_0^\beta W(v_j)f(v_j)dv_j = c_i + W(v_i)$$

Further, we notice that

$$\begin{aligned} t_i(v_i) - t_i(v_i^*) &= W(v_i) - W(v_i^*) \\ &= \int_{v_i^*}^{v_i} xg(x)dx \\ &= v_iG(v_i) - v_i^*G(v_i^*) - \int_{v_i^*}^{v_i} G(x)dx, \end{aligned}$$

where the last equality follows by doing integration by parts. By Lemma 15, this mechanism is BIC (since interim allocation probability in efficient allocation share of type x is given by $G(x)$). Finally, by Lemma 17, we only need to check if $t_i(v_i^*) \geq 0$. To do so, note that

$$t_i(v_i^*) = c_i + W(v_i^*) = \frac{1}{n}\sum_{j\in N} W(v_j^*) \geq 0,$$

where the last inequality follows from Inequality (4.2). ■

A crucial observation from the proof is budget-balanced can be relaxed. Call a mechanism $(Q_i^e, T_i)_{i\in N}$ **feasible** if $\sum_{i\in N} T_i(v) \leq 0$. Notice that the necessity part of Theorem 14 works even if we replace budget-balance by the weaker condition feasibility. Hence, a corollary of this observation with Theorem 14 is that if a partnership can be dissolved using a BIC, efficient, feasible, and interim IR mechanism, then it can be efficiently dissolved (i.e., using a BIC, efficient, budget-balanced, and interim IR mechanism).

We will refer to the mechanism mentioned in the sufficiency part of the proof of Theorem 14 as **CGK** mechanism - due to Cramton et al. (1987). In their paper, Cramton et al. (1987) propose simple mechanisms. These simple mechanisms are easy to use - for instance, one of their mechanisms is that every agent submits a bid; highest bidder wins; and the winner's bid is equally distributed between all the agents. They show that such mechanisms dissolve a large subset of dissolvable partnerships.

4.3.1 Corollaries of Theorem 14

We can derive some easy corollaries from Theorem 14. The first is on one seller-many buyer partnerships. These are partnerships, where is some agent $s \in N$ such that $r_s = 1$ and $r_i = 0$ for all $i \neq s$. You can interpret s as a seller (who owns the object) and other agents as buyers. A special case of this is when $n = 2$, i.e., besides the seller, there is exactly one buyer. This setting is often referred to as the bilateral trading model. Though the following result is true for any $n \geq 2$ in the one seller-many buyer model, it was first shown in the bilateral trading model by Myerson and Satterthwaite (1983).

THEOREM **15** *One seller-many buyer partnerships cannot be dissolved efficiently.*

Proof: Let $r_s = 1$ for some $s \in N$ and $r_i = 0$ for all $i \neq s$. Then $v_s^* = \beta$ and $v_i^* = 0$ for all $i \neq s$. Now, that

$$\begin{aligned}\sum_{i\in N} W(v_i^*) &= W(\beta) + (n-1)W(0) \\ &= (n-1)\int_0^\beta (1-F(x))xg(x)dx - \int_0^\beta F(x)xg(x)dx \\ &= (n-1)\int_0^\beta xg(x)dx - n\int_0^\beta F(x)xg(x)dx.\end{aligned}$$

Using the fact that $g(x) = (n-1)[F(x)]^{n-2}f(x)$, we simplify as:

$$\begin{aligned}\frac{1}{n-1}\sum_{i\in N} W(v_i^*) &= (n-1)\int_0^\beta x[F(x)]^{n-2}f(x)dx - n\int_0^\beta x[F(x)]^{n-1}f(x)dx \\ &= \big[x[F(x)]^{n-1}\big]_0^\beta - \int_0^\beta [F(x)]^{n-1}dx - \big[x[F(x)]^n\big]_0^\beta + \int_0^\beta [F(x)]^n dx \\ &= \int_0^\beta [F(x)]^{n-1}\big(F(x)-1\big)dx] \\ &< 0.\end{aligned}$$

By Theorem 14, we are done. ■

Theorem 15 is a remarkable impossibility result. It says that many trading interactions have no hope of efficiency. It points to a clear explanation of this impossibility - extreme form of property rights structure. On the other hand, symmetric property rights structure allows partnerships to be dissolved efficiently.

THEOREM **16** *The set of partnerships that can be dissolved efficiently is a non-empty convex set centered around the equal partnership, and equal partnership can always be dissolved efficiently.*

Proof: Take any two partnerships $\{r_i\}_{i\in N}$ and $\{r_i'\}_{i\in N}$ which can be dissolved efficiently. Let M and M' be the respective mechanisms. Let $\{r_i''\}_{i\in N}$ be another partnership such that $r_i'' = \lambda r_i + (1-\lambda)r_i'$ for each $i \in N$, where $\lambda \in (0,1)$. Then, define the mechansim M'' as follows. The allocation rule is M'' is the efficient one and the transfer rule is $\{T_i''\}_{i\in N}$. For every valuation profile v and for every $i \in N$, let $T_i''(v) = \lambda T_i(v) + (1-\lambda)T_i'(v)$, where T and T' are the transfer rules in M and M' respectively. Since T and T' are budget-balanced, T'' is also budget-balanced. Also, since M and M' are BIC, M'' is also BIC. By construction, for every $i \in N$, the interim payment of these mechanisms are related as:

$$t_i''(v_i) = \lambda t_i(v_i) + (1-\lambda)t_i'(v_i) \;\forall\; v_i.$$

Hence, we have for every $i \in N$ and for every v_i,

$$\begin{aligned}v_i q_i^e(v_i) - t_i''(v_i) &= \lambda\big(v_i q_i^e(v_i) + t_i(v_i)\big) + (1-\lambda)\big(v_i q_i^e(v_i) + t_i'(v_i)\big) \\ &\geq \lambda r_i v_i + (1-\lambda)r_i' v_i \\ &= r_i'' v_i,\end{aligned}$$

which is the desired IIR constraint. Hence, $\{r_i''\}_{i\in N}$ can be dissolved efficiently. So, the set of partnerships that can be dissolved efficiently forms a convex set.

Now, consider the equal partnership $r_i = \frac{1}{n}$ for all $i \in N$. Let $G(v^*) = [F(v^*)]^{n-1} = \frac{1}{n}$. Then, we need to show that $W(v^*) \geq 0$, and by Theorem 14, we will be done. To see why it is the case, note the following.

$$\begin{aligned} W(v^*) &= \int_{v^*}^{\beta} xg(x)dx - \int_0^{\beta} xF(x)g(x)dx \\ &= (n-1)\int_{v^*}^{\beta} x[F(x)]^{n-2}f(x)dx - (n-1)\int_0^{\beta} x[F(x)]^{n-1}f(x)dx \end{aligned}$$

Hence, we get

$$\begin{aligned} \frac{1}{n-1}W(v^*) &= \int_{v^*}^{\beta} x[F(x)]^{n-2}f(x)dx - \int_0^{\beta} x[F(x)]^{n-1}f(x)dx \\ &= \frac{1}{n-1}\big[\beta - (v^*)[F(v^*)]^{n-1} - \int_{v^*}^{\beta}[F(x)]^{n-1}dx\big] \\ &- \frac{1}{n}\big[\beta - \int_0^{\beta}[F(x)]^n dx\big] \\ &= \frac{1}{n}\int_0^{\beta}[F(x)]^n dx + \frac{1}{n(n-1)}\big[\beta - v^*\big] - \frac{1}{(n-1)}\int_{v^*}^{\beta}[F(x)]^{n-1}dx \\ &= \frac{1}{n}\int_0^{v^*}[F(x)]^n dx + \frac{1}{n(n-1)}\big[\beta - v^*\big] - \frac{1}{n(n-1)}\int_{v^*}^{\beta}\big[n[F(x)]^{n-1} - (n-1)[F(x)]^n\big]dx \end{aligned}$$

Now, consider the function $\phi(x) := nF(x)^{n-1} - (n-1)F(x)^n$ for all $x \in [v^*, \beta]$. Note that $\phi(v^*) = 1 - \frac{n-1}{n}F(v^*) < 1$ and $\phi(\beta) = 1$. Further, $\phi'(x) = n(n-1)[F(x)]^{n-2}f(x) - n(n-1)[F(x)]^{n-1}f(x) = n(n-1)f(x)[F(x)]^{n-2}(1-F(x)) > 0$. Hence, ϕ is a strictly increasing function. So, $\phi(x) \leq 1$ for all $x \in [v^*, \beta]$. This means,

$$\frac{1}{n-1}W(v^*) \geq \frac{1}{n}\int_0^{v^*}[F(x)]^n dx \geq 0,$$

as desired. ■

We now consider the case where values of agents are distributed uniformly in $[0, 1]$. Then,

$G(x) = x^{n-1}$ and $g(x) = (n-1)x^{n-2}$. In that case, for each z, we have

$$\begin{aligned} W(z) &= \int_z^1 xg(x)dx - \int_0^1 xF(x)g(x)dx \\ &= \int_z^1 (n-1)x^{n-1}dx - (n-1)\int_0^1 x^n dx \\ &= \frac{n-1}{n}\left[1-z^n\right] - \frac{(n-1)}{n+1} \\ &= \frac{n-1}{n(n+1)} - \frac{n-1}{n}z^n \end{aligned}$$

Hence, for any partnership structure $\{r_i\}_{i\in N}$, Theorem 14 implies that it can be dissolved if and only if

$$\sum_{i\in N} W(v_i^*) = \frac{(n-1)}{n+1} - \frac{(n-1)}{n}\sum_{i\in N}(v_i^*)^n \geq 0$$

This is equivalently stated as

$$\sum_{i\in N}(v_i^*)^n \leq \frac{n}{n+1}. \tag{4.4}$$

One corollary is that that the one buyer-many sellers model can be dissolved efficiently for the uniform distribution.

LEMMA **18** *Suppose* $r_1 = 0, r_2 = \ldots = r_n = \frac{1}{n-1}$. *If values are distributed uniformly and* $n \geq 3$, *then this partnership can be dissolved efficiently.*

Proof: Let z be such that $G(z) = z^{n-1} = \frac{1}{n-1}$. Then using Inequality (4.4), we need to show that $(n-1)z^n = z \leq \frac{n}{n+1}$. Since G is strictly increasing, this is equivalent to showing $G(z) \leq G(\frac{n}{n+1})$ or $\frac{1}{n-1} \leq \left[1-\frac{1}{n+1}\right]^{n-1}$. But note that $(1-\frac{1}{n+1})^{n-1} \geq 1 - \frac{n-1}{n+1} = \frac{2}{n+1} \geq \frac{1}{n-1}$ if $n \geq 3$. ■

For three agents and uniform distribution, Figure 4.1 draws the simplex of partnerships and identifies those (in red color) that can be dissolved efficiently.

4.4 Dominant strategy redistribution

Even though the set of partnerships that can be dissolved is quite large, the mechansim required to dissolve them requires precise knowledge of the priors. That prompts the natural

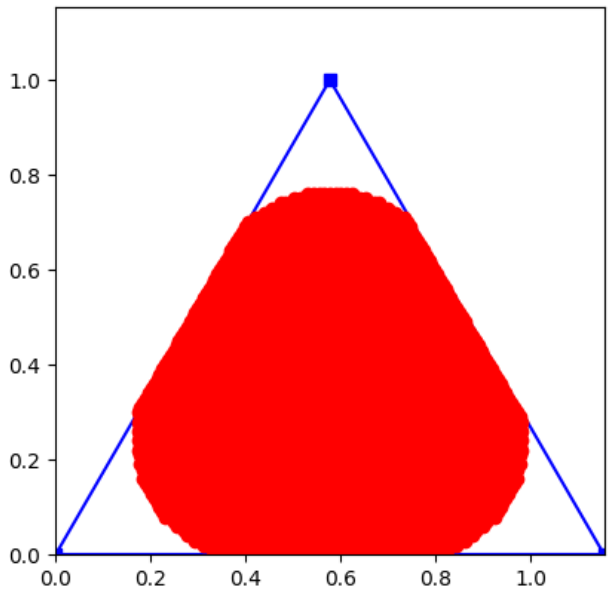

Figure 4.1: Dissolvable partnerships under uniform distribution

question whether there is a dominant strategy incentive compatible (DSIC) mechanism that can do the job. [1] We investigate this issue further here.

Suppose there are just two agents: $N = \{b, s\}$ (bilateral trading model with a buyer and a seller). Suppose values are distributed in $[0, \beta]$. The exact nature of distribution of values does not matter. Consider any DSIC, efficient, and budget-balanced mechanism - notice no mention of individual rationality constraint. Suppose $\mathcal{U}_b : [0, \beta]^2 \to \mathbb{R}$ and $\mathcal{U}_s : [0, \beta]^2 \to \mathbb{R}$ are the net utility functions of the two agents from this mechanism. Efficiency means the allocation rule Q^e satisfies $Q^e_b(v_b, v_s) + Q^e_s(v_b, v_s) = 1$ and

$$Q^e_b(v_b, v_s) = \begin{cases} 1 & \text{if } v_b > v_s \\ 0 & \text{if } v_b < v_s. \end{cases}$$

Further budget-balance gives us

$$\mathcal{U}_b(v_b, v_s) + \mathcal{U}_s(v_b, v_s) = \begin{cases} v_b & \text{if } v_b > v_s \\ v_s & \text{otherwise.} \end{cases}$$

Then, *payoff equivalence* formula implies that for any type profile (v_b, v_s) with $v_b > v_s > 0$, we have

$$\begin{aligned} \mathcal{U}_b(v_b, v_s) &= \mathcal{U}_b(0, v_s) + \int_0^{v_b} Q^e_b(x, v_s)dx \\ &= \mathcal{U}_b(0, v_s) + (v_b - v_s) \\ &= v_s - \mathcal{U}_s(0, v_s) + (v_b - v_s) \qquad \text{(Budget-balance and efficiency gives } \mathcal{U}_s(0, v_s) + \mathcal{U}_b(0, v_s) = v_s) \\ &= v_b - \mathcal{U}_s(0, 0) - v_s \qquad \text{(Using payoff equivalence again)} \\ &= (v_b - v_s) - \mathcal{U}_s(0, 0). \end{aligned}$$

Identical argument gives:

$$\begin{aligned} \mathcal{U}_s(v_b, v_s) &= \mathcal{U}_s(v_b, 0) + \int_0^{v_s} Q^e_s(v_b, x)dx \\ &= \mathcal{U}_s(v_b, 0) = v_b - \mathcal{U}_b(v_b, 0) = v_b - \mathcal{U}_b(0, 0) - v_b \\ &= -\mathcal{U}_b(0, 0) \end{aligned}$$

[1] Another drawback of the CGK mechanisms is that IR constraint is satisfied at the interim but not ex-post.

Hence, we have

$$\mathcal{U}_b(v_b, v_s) + \mathcal{U}_s(v_b, v_s) = (v_b - v_s) - \big(\mathcal{U}_b(0,0) + \mathcal{U}_s(0,0)\big)$$

By budget-balance and efficiency, LHS is equal to v_b and RHS is equal to $v_b - v_s$, a contradiction.

This argument generalizes to n agents (albeit complex). Notice that we did not use anything about the property rights of the two agents - r_b and r_s can be anything. The generalized version of this result is called the Green-Laffont impossibility result.

THEOREM **17** *There is no DSIC, budget-balanced, and efficient mechanism.*

This result was proved in very general models in Green and Laffont (1979) and Laffont and Maskin (1980). Hence, in this model, going from DSIC to BIC allows to overcome the impossibilites partially. However, there are models of interest where it is possible to have DSIC, budget-balanced, and efficient mechanisms.

Like every impossibility result, the Green-Laffont impossibility has inspired many researchers. Typical line of attack is to relax the assumptions in the model. I outline some ideas below.

- **Relax efficiency.** The first approach is to relax efficiency and find the "optimal" (appropriately defined to account for sum of utilities of agents). So, we look for some form of optimality under DSIC and budget-balancedness constraint. This turns out to be extremely difficult if we do an ex-ante expected welfare maximization. A famous mechanism by Green and Laffont proposes an asymptotically optimal mechanism. In that mechanism an agent called the *residual claimant* is picked and a Vickrey auction is done among remaining agents. The residual agent is given the payment in the Vickrey auction. This mechanism is DSIC and budget-balanced. If we pick the residual agent uniformly at random, then this guarantees that the highest valued agent wins the object with probability $(1 - \frac{1}{n})$ at each profile. So, for large n, we get close to an efficient allocation.

  Surprisingly, we can do better than this. Long et al. (2017) show that there are DSIC and budget-balanced mechanisms where we can allocate the object to the highest valued agent with probability $1 - H(n)$, where H vanishes to zero at an exponential rate with

n. In these mechanisms, about half the agents are given the agent and out of them all except the highest valued agent gets the object with equal but (vanishingly) small probability.

- **Relax budget-balance by burning money.** The other approach is to relax budget-balance. That is we look to maximize welfare (utilities) of agents under DSIC and efficiency. This automatically means we search within the class of Groves mechanisms - they are the unique class of DSIC and efficient mechanisms. Vickrey auction can be easily improved. Consider the following mechanism due to Cavallo (2006). A Vickrey auction is conducted and its revenue is redistributed smartly to maintain DSIC. In particular, at a valuation profile v with $v_1 \geq v_2 \geq \ldots \geq v_n$, we payment v_2 of winner is taken, and agents 1 and 2 are given back $\frac{v_3}{n}$ and others are given $\frac{v_2}{n}$. As a result, total money collected is:

$$v_2 - \frac{n-2}{n} v_2 - \frac{2}{n} v_3 = \frac{2}{n}(v_2 - v_3),$$

  which approaches zero for large n. In other words, the amount of *money burning* approaches zero for large n. Since this mechanism is efficient, we conclude that asymptotically, this mechanism redistributes all the surplus (v_1 here).

  Hence, there are classes of mechanisms in Groves class of mechanisms which can redistribute surplus better than the Vickrey auction. This idea has been extended to the limit in Moulin (2009); Guo and Conitzer (2009), where they identify Groves mechanisms that burn zero money at an exponential rate.

- **Relax efficiency by burning probabilities.** When we relaxed efficiency, we maintained the fact that we always allocate the object, although not necessarily to the highest valued agent. However, we can maintain the fact that the object *only* goes to the highest valued agent and search over the space of DSIC and budget-balanced mechanisms. Surprisingly, we can still achieve asymptotic results. This was shown in Mishra and Sharma (2018). They propose the following mechanism. In their mechanism, the highest valued agent is given the object with the following probability at valuation profile v with $v_1 \geq v_2 \geq \ldots \geq v_n$:

$$1 - \frac{2}{n} + \frac{2}{n}\frac{v_3}{v_2}$$

In fact, a Vickrey auction of this probability is done. Note that the revenue produced is v_2 times the above probability, which is

$$v_2 - \frac{2}{n}v_2 + \frac{2}{n}v_3 = (n-2)\frac{v_2}{n} + 2\frac{v_3}{n}.$$

Then, agents 1 and 2 are given back $\frac{v_3}{n}$ each and agents 3 to n are given back $\frac{v_2}{n}$ each. This maintains DSIC and budget-balance. It is not efficient because the highest valued agent is not given the entire object - some of the object is *wasted or burnt*. But as n tends to infinity, the probability that the highest valued agent gets the object approaches one.

- **Relax solution concept.** The final approach to circumvent the impossibility is to relax the solution concept to Bayesian equilibrium. We have already seen that the mechanism constructed in the proof of Theorem 14 satisfies Bayesian incentive compatibility, efficiency, and budget-balance for arbitrary partnership structure - it may fail interim individual rationality. Hence, at least for the single object model, the Green-Laffont impossibility fails if we weaken the solution concept to Bayesian equilibrium. The CGK mechanism is Bayesian incentive compatible, efficient, and budget-balanced. We show next that this result holds more generally.

4.5 The dAGV mechanism

We now show that the existence of a Bayesian incentive compatible, efficient, and budget-balanced mechanism can be guaranteed in very general settings - this covers single object case, multiple objects case, public goods case etc.

Let A be a finite set of alternatives and $v_i \in \mathbb{R}^{|A|}$ be the valuation vector of agent i. Let $\mathcal{V}_i$ be the type space of agent i - the set of possible valuation vectors. Let $\mathcal{V} \equiv \mathcal{V}_1 \times \ldots \times \mathcal{V}_n$. We will assume that types are drawn **independently**. An efficient mechanism is $(Q^e, T_i)_{i \in N}$ such that

$$Q^e(v) = \arg\max_{a \in A} \sum_{i \in N} v_i(a) \ \forall \ v \in \mathcal{V}.$$

A key construction is the map $r_i : \mathcal{V}_i \to \mathbb{R}$ for every $i \in N$. We define it as follows: for every $i \in N$,

$$r_i(v_i) = \mathbb{E}_{v_{-i}}\Big[\sum_{j\in N\setminus\{i\}} v_j(Q^e(v_i, v_{-i}))\Big] \qquad \forall\ v_i \in \mathcal{V}_i,$$

where $\mathbb{E}_{v_{-i}}$ is the expectation over valuations of other agents besides agent i. Without independence, this expression is a **conditional expectation**. As we will see, this will create problems since this is a term that needs to be calculated by the designer. For instance, if the true type is v_i and the agent reports v_i', then this will be conditioned on v_i' and not v_i. This is where independence helps because the expectation in $r_i(v_i)$ can be computed without conditioning on the true type v_i.

So, for agent $i \in N$, the expression $r_i(v_i)$ captures the **expected welfare of others** when her type is v_i - we will call this the **residual utility** of agent i at v_i. The idea is to use this expected welfare in a clever way to achieve BIC and budget-balance. Arrow (1979) and d'Aspremont and Gérard-Varet (1979) proposed the following remarkable mechanism which achieves this. Define the transfer rules $\{T_i^{dagv}\}_{i\in N}$ as follows: for every $i \in N$,

$$T_i^{dagv}(v) = r_i(v_i) - \frac{1}{n-1}\sum_{j\in N\setminus\{i\}} r_j(v_j) \qquad \forall\ v \in \mathcal{V}. \tag{4.5}$$

So, the payment of agent i is the difference between the average residual utility of other agents and her own residual utility. This is an interim analogue of the VCG idea - agents pay their **expected externality**. We will call the mechanism $(Q_i^e, T_i^{dagv})_{i\in N}$ the **dAGV mechanism**.

THEOREM **18** *The dAGV mechanism is efficient, budget-balanced, and Bayesian incentive compatible.*

Proof: Efficiency and budget-balancedness follows from the definition. To see BIC, fix

agent i and two types v_i, v_i'. Note the following.

$$\begin{aligned}
&\mathbb{E}_{v_{-i}}\Big[v_i(Q^e(v_i,v_{-i}))+T_i^{dagv}(v_i,v_{-i})\Big]=\mathbb{E}_{v_{-i}}\Big[v_i(Q^e(v_i,v_{-i}))+r_i(v_i)-\frac{1}{n-1}\sum_{j\in N\setminus\{i\}}r_j(v_j)\Big]\\
&=\mathbb{E}_{v_{-i}}\Big[v_i(Q^e(v_i,v_{-i}))-\frac{1}{n-1}\sum_{j\in N\setminus\{i\}}r_j(v_j)+r_i(v_i')-r_i(v_i')+r_i(v_i)\Big]\\
&=\mathbb{E}_{v_{-i}}\Big[v_i(Q^e(v_i,v_{-i}))+r_i(v_i')-\frac{1}{n-1}\sum_{j\in N\setminus\{i\}}r_j(v_j)\Big]\\
&+\mathbb{E}_{v_{-i}}\Big[\sum_{j\in N\setminus\{i\}}v_j(Q^e(v_i,v_{-i}))\Big]-\mathbb{E}_{v_{-i}}\Big[\sum_{j\in N\setminus\{i\}}v_j(Q^e(v_i',v_{-i}))\Big]\\
&=\mathbb{E}_{v_{-i}}\Big[\sum_{j\in N}v_j(Q^e(v_i,v_{-i}))+r_i(v_i')-\frac{1}{n-1}\sum_{j\in N\setminus\{i\}}r_j(v_j)\Big]-\mathbb{E}_{v_{-i}}\Big[\sum_{j\in N\setminus\{i\}}v_j(Q^e(v_i',v_{-i}))\Big]\\
&\geq\mathbb{E}_{v_{-i}}\Big[\sum_{j\in N}v_j(Q^e(v_i',v_{-i}))+r_i(v_i')-\frac{1}{n-1}\sum_{j\in N\setminus\{i\}}r_j(v_j)\Big]-\mathbb{E}_{v_{-i}}\Big[\sum_{j\in N\setminus\{i\}}v_j(Q^e(v_i',v_{-i}))\Big]\\
&=\mathbb{E}_{v_{-i}}\Big[v_i(Q^e(v_i',v_{-i}))+r_i(v_i')-\frac{1}{n-1}\sum_{j\in N\setminus\{i\}}r_j(v_j)\Big]\\
&=\mathbb{E}_{v_{-i}}\Big[v_i(Q^e(v_i',v_{-i}))+T_i^{dagv}(v_i',v_{-i})\Big],
\end{aligned}$$

where the inequality followed from efficiency. Thus, we satisfy the desired Bayesian incentive compatibility constraint. ■

As we discussed earlier, the independence plays a role in the BIC part of the proof of Theorem 18. Without independence, we will have to carry out conditional expectations and r_i will also be a conditional expectation. Without independence, it is sometimes possible to construct a BIC, efficient, and budget-balanced mechanism but not always (d'Aspremont et al., 2004).

The dAGV mechanism does not take into account any property rights structure. So, obviously, it will fail any form of individual rationality constraint. The objective of dAGV mechanism was to show that the Green-Laffont impossibility can be overturned by relaxing the solution concept from dominant strategies to Bayesian equilibrium.

Chapter 5

Multidimensional Mechanism Design

The analysis of optimal mechanism design and efficient budget-balanced mechanism design in previous chapters was possible because of the one-dimensional type space assumed. The problem of finding similar results when the type of each agent is multidimensional is a significantly challenging problem. However, some of the results that we discussed can still be generalized to the multidimensional environment. We discuss them next.

For simplicity of exposition, we assume that there is only one agent. In this case, the solution concept will not matter - dominant strategy and Bayesian reduce to the same thing. However, if you want to extend this result to a multiple agent framework, you need to add for all t_{-i} in the dominant strategy implementation and integrate out over T_{-i} in the Bayesian implementation.

The notation will be as before. Let A be some finite set of alternatives and $\mathcal{L}(A)$ be the set of all lotteries over A. There is a single agent. The type of the agent is $t \in \mathbb{R}^{|A|}$. Here, we will use $t(a)$ to denote the value of the agent for alternative a. The type space of the agent is some set $T \subseteq \mathbb{R}^{|A|}$. Some examples are useful to see the applicability of this setting is for both private good and public good allocation problems.

- **Multi-object auction with unit demand.** A seller is selling a set of objects to a buyer who can be assigned at most one object. The value for the buyer for each object is his type. The set of alternatives is the set of objects (and the alternative $\emptyset$ indicating not being assigned to any object).

- **Combinatorial auction.** This is the same model as the previous one but now the buyer

can buy multiple objects. Hence, the set of alternatives is the set of all subsets of objects. The value for each subset is the type of the agent.

- **Public project selection.** A planner needs to choose a project from multiple projects. The value of the agent for each project is his type.

Like in voting problems, it is expected that not all vectors in $\mathbb{R}^{|A|}$ are allowed to be types. Hence, the type space can be a strict subset of $\mathbb{R}^{|A|}$ with some restrictions. For instance, in the combinatorial auction problem, we may require that for any pair of objects a and b, at any type t, $t(\{a,b\}) = t(a) + t(b)$. This puts restrictions on how the type space looks. In the public project problem, type vector may be single peaked with respect to some exogenous ordering of the projects.

We will assume that all these restrictions are embedded in T. As in the analysis of the single object auction, we will first give a characterization of all incentive compatible mechanisms.

5.1 Incentive Compatible Mechanisms

A mechanism consists of an allocation rule $f : T \to \mathcal{L}(A)$ and a payment rule $p : T \to \mathbb{R}$. If type t is reported to the mechanism, then $f(t)$ is a probability distribution over alternatives at that type, where we denote by $f_a(t)$ the probability associated with alternative a. Hence, an agent with type s who reports type t to the mechanism (f, p) gets a net utility of

$$s \cdot f(t) - p(t),$$

where $s \cdot f(t) = \sum_{a \in A} s(a) f_a(t)$.

5.1.1 An illustration

Suppose there are two goods $\{a, b\}$ with a seller and one buyer. The seller decides to use a deterministic mechanism. So, it will either not sell any of the goods, or sell only good a or sell only good b or bundle both of them sell the bundles $\{a, b\}$. Normalize the value of the buyer to not getting the good to zero and assume that value of the bundle is $v_a + v_b$ if v_a is the value of good a alone and v_b is the value of good b alone. By incentive compatibility,

in any mechanism, the seller should announce the following prices: $p_\emptyset$ - this is the price of not getting any good; p_a - this is the price of getting good a; p_b - this is the price of getting good b; and p_{ab} - this is the price of getting the bundle $\{a, b\}$. For simplicity, let us assume that $p_\emptyset = 0$ - this can also be derived by imposing individual rationality and revenue maximization.

What more does incentive compatibility say? Take the mechanism (f, p), where p is defined as discussed above and f is the allocation rule. Suppose the type space is $V \equiv [0, 1]^2$; i.e., values of both the goods lie in $[0, 1]$. Partition the type space as follows

$$\begin{aligned} V_\emptyset &= \{v \in V : f(v) = \emptyset\} \\ V_a &= \{v \in V : f(v) = \{a\}\} \\ V_b &= \{v \in V : f(v) = \{b\}\} \\ V_{ab} &= \{v \in V : f(v) = \{a, b\}\} \end{aligned}$$

We wish to get a better understanding of this partitioning. To do so, pick $v \in V_{ab}$. There are three incentive constraints *from* v:

$$\begin{aligned} v_a + v_b - p_{ab} \geq 0 &\Leftrightarrow v_a + v_b \geq p_{ab} \\ v_a + v_b - p_{ab} \geq v_a - p_a &\Leftrightarrow v_b \geq p_{ab} - p_a \\ v_a + v_b - p_{ab} \geq v_b - p_b &\Leftrightarrow v_a \geq p_{ab} - p_b \end{aligned}$$

Similarly, pick $v \in V_a$. We have three incentive constraints.

$$\begin{aligned} v_a &\geq p_a \\ v_a - v_b &\geq p_a - p_b \\ v_b &\leq p_{ab} - p_a. \end{aligned}$$

Identical argument gives for any $v \in V_b$:

$$\begin{aligned} v_b &\geq p_b \\ v_a - v_b &\leq p_a - p_b \\ v_a &\leq p_{ab} - p_b. \end{aligned}$$

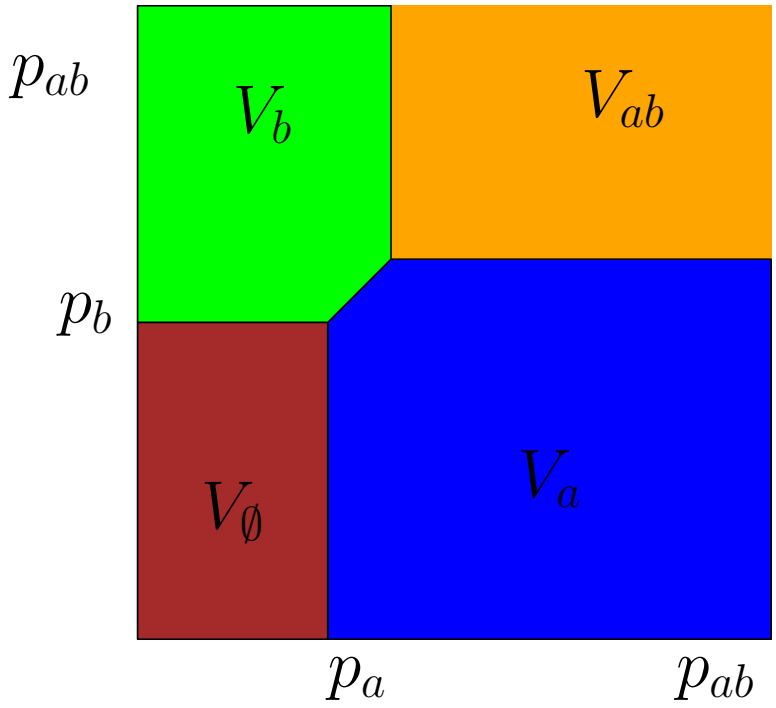

Figure 5.1: Partitioning the type space

Figure 5.1 shows the partition of this type space. As can be seen, it is not obvious what kind of monotonicity of the allocation rule we have.

Characterization of Incentive Compatibility

As before, we associate with a mechanism $M \equiv (f, p)$, a net utility function $\mathcal{U}^M : T \to \mathbb{R}$, defined as

$$\mathcal{U}^M(t) := t \cdot f(t) - p(t) \quad \forall\, t \in T,$$

which is the truth-telling net utility from the mechanism.

DEFINITION **21** *A mechanism $M \equiv (f, p)$ is* **incentive compatible** *if for every $s, t \in T$, we have*

$$t \cdot f(t) - p(t) \geq t \cdot f(s) - p(s),$$

or equivalently,

$$\mathcal{U}^M(t) \geq \mathcal{U}^M(s) + (t - s) \cdot f(s).$$

An allocation rule f is **implementable** *if there exist a payment rule p such that (f, p) is incentive compatible.*

Our first step is to generalize the characterization of mechanisms in Theorem 5 to this environment. For this, we first need to define an appropriate notion of monotonicity of allocation rule in this type space. Since the type is multidimensional, it is not clear how this can be defined. But the following is a well-known form of monotonicity in multidimensional environment.

DEFINITION **22** *An allocation rule f is* **monotone** *if for every $s, t \in T$, we have*

$$(t - s) \cdot \big(f(t) - f(s)\big) \geq 0.$$

This condition is often referred to as the **2-cycle monotonicity** condition. We will discuss the reasons below. This is the correct extension of monotonicity in Theorem 5 to multi-dimensional environments. The following lemma proves that monotonicity is a necessary condition. The proof points out the similarity with the Myerson monotonicity.

LEMMA **19** *If (f, p) is incentive compatible, f is monotone.*

Proof: Suppose $M = (f, p)$ is incentive compatible. Consider types $s, t \in T$ and the pair of incentive constraints for these two types:

$$\mathcal{U}^M(t) \geq \mathcal{U}^M(s) + (t - s) \cdot f(s).$$
$$\mathcal{U}^M(s) \geq \mathcal{U}^M(t) + (s - t) \cdot f(t).$$

Adding these incentive constraints, we get $(t - s) \cdot \big[f(t) - f(s)\big] \geq 0$, which is the required monotonicity condition. ■

It is also instructive to see how monotonicity looks like when f is **deterministic**, i.e., $f : T \to A$. Then at every type t, $f(t)$ will be $0 - 1$ vector of dimension $|A|$. So the expression of monotonicity simplifies as follows:

$$(t - s) \cdot \big[f(t) - f(s)\big] = \big[t(f(t)) - t(f(s))\big] - \big[s(f(t)) - s(f(s))\big] \geq 0$$

This is a condition on differences of values of chosen alternatives.

We now prove an analogue of Theorem 5 in the multidimensional environment using monotonicity and a version of *payoff equivalence*.

THEOREM **19** *Suppose $T \subseteq \mathbb{R}^{|A|}$ is convex. A mechanism $M \equiv (f, p)$ is incentive compatible if and only if*

(a) f is monotone,

(b) for every $s, t \in T$,

$$\mathcal{U}^M(t) = \mathcal{U}^M(s) + \int_0^1 \psi^{s,t}(z)dz,$$

where $\psi^{s,t}(z) = (t-s) \cdot f(s + z(t-s))$ for all $z \in [0,1]$.

Proof: Suppose $M \equiv (f, p)$ is such that (a) and (b) hold. We will show that M is incentive compatible. Choose any $s, t \in T$.

STEP 1. We first show that for every $z, z' \in [0,1]$ with $z > z'$, we have $\psi^{s,t}(z) \geq \psi^{s,t}(z')$. Pick $z, z' \in [0,1]$ with $z > z'$. Since f is monotone, we have

$$\big[(s + z(t-s)) - (s + z'(t-s))\big] \cdot \big[f(s + z(t-s)) - f(s + z'(t-s))\big] \geq 0.$$

Simplifying, we get

$$(z - z')(t-s) \cdot \big[f(s + z(t-s)) - f(s + z'(t-s))\big] \geq 0.$$

But $z > z'$ implies $(t-s) \cdot \big[f(s + z(t-s)) - f(s + z'(t-s))\big] \geq 0$, which implies $\psi^{s,t}(z) - \psi^{s,t}(z') \geq 0$.

STEP 2. Now, we can write

$$\begin{aligned}
\mathcal{U}^M(t) - \mathcal{U}^M(s) - (t-s) \cdot f(s) &= \int_0^1 \psi^{s,t}(z)dz - (t-s) \cdot f(s) \\
&\geq \psi^{s,t}(0) - (t-s) \cdot f(s) \\
&= 0,
\end{aligned}$$

where the first equality follows from (b), the second inequality from Step 1 (non-decreasingness of $\psi^{s,t}$), and the last equality follows from the fact that $\psi^{s,t}(0) = (t-s) \cdot f(s)$. This shows that M is incentive compatible.

Now, for the other direction, we assume that $M \equiv (f, p)$ is incentive compatible. Monotonicity of (a) follows from Lemma 19. We show (b).

Consider any $s, t \in T$. We define for every $z \in [0, 1]$,

$$\phi(z) := \mathcal{U}^M(s + z(t - s)).$$

Incentive compatibility of M implies that for every $z, z' \in [0, 1]$, we have

$$\mathcal{U}^M(s + z(t - s)) \geq \mathcal{U}^M(s + z'(t - s)) + (z - z')(t - s) \cdot f(s + z'(t - s)).$$

This implies that for every $z, z' \in [0, 1]$, we have

$$\phi(z) \geq \phi(z') + (z - z')\psi^{s,t}(z'). \tag{5.1}$$

This also implies that ϕ is a convex function. To see this, pick $\bar{z}, \hat{z} \in [0, 1]$ and $\lambda \in (0, 1)$. Let $\tilde{z} = \lambda\bar{z} + (1 - \lambda)\hat{z}$. Then, using Inequality 5.1, we get

$$\phi(\bar{z}) \geq \phi(\tilde{z}) + (\bar{z} - \tilde{z})\psi^{s,t}(\tilde{z}).$$

Similarly, we have

$$\phi(\hat{z}) \geq \phi(\tilde{z}) + (\hat{z} - \tilde{z})(t - s) \cdot f(s + \tilde{z}(t - s)).$$

Multiplying the first inequality by λ and the second one by $(1 - \lambda)$ and summing them, we get

$$\lambda\phi(\bar{z}) + (1 - \lambda)\phi(\hat{z}) \geq \phi(\tilde{z}).$$

This show that ϕ is convex.

Hence, $\psi(z')$ is the subgradient of the convex function ϕ at z'. By Lemma 11, we get that for every $z' \in [0, 1]$,

$$\phi(z') = \phi(0) + \int_0^{z'} \psi^{s,t}(z)dz.$$

Hence,

$$\phi(1) = \phi(0) + \int_0^1 \psi^{s,t}(z)dz.$$

Substituting, we get

$$\mathcal{U}^M(t) = \mathcal{U}^M(s) + \int_0^1 \psi^{s,t}(z)dz.$$

■

Revenue/payoff equivalence. Theorem 19 immediately implies a payoff equivalence result. Consider two incentive compatible mechanisms $M = (f, p)$ and $M' = (f, p')$ using the same allocation rule f. Fix some type $t^0 \in T$. By Theorem 19, for every $t \in T$,

$$\mathcal{U}^M(t) - \mathcal{U}^M(t^0) = \mathcal{U}^{M'}(t) - \mathcal{U}^{M'}(t^0).$$

Hence, mechanisms M and M' assign different net utilities to the agent if and only if $\mathcal{U}^M(t^0)$ and $\mathcal{U}^{M'}(t^0)$ are different. In other words, if two incentive compatible mechanisms use the same allocation rule and assign the same net utility to the agent at *some* type, then they must assign the same net utility to the agent at *all* types. This is known as the **payoff equivalence** result.

One-dimensional problems. We remark that monotonicity reduces to "non-decreasingness" discussed in Theorem 5 for one-dimensional problem. We say a type space T is **one-dimensional** if there exists an alternative $a^* \in A$ such that $t(a) = 0$ for all $a \neq a^*$ and for all $t \in T$. In the single object auction setting a^* is the alternative where the agent wins the object. Note that if T is one-dimensional, then for any $s, t \in T$, $(t - s)$ is a vector whose components corresponding to any alternative $a \neq a^*$ are zero. Hence, for any $s, t \in T$, $(t - s)(f(t) - f(s))$, can be written as

$$\big(t(a^*) - s(a^*)\big)\big(f_{a^*}(t) - f_{a^*}(s)\big).$$

Monotonicity requires that the above term is non-negative. This is equivalent to saying that if $t(a^*) > s(a^*)$, then $f_{a^*}(t) \geq f_{a^*}(s)$.

In one-dimensional problem statement (b) in Theorem 19 also simplifies - compare it with the analogue statement in Theorem 5. Suppose the value for alternative a^* is lies in $[\ell, H]$. For any $x \in [\ell, H]$, we write the unique type t with $t(a^*) = x$ as t^x. Now, fix a mechanism $M \equiv (f, p)$. Then, statement (b) is equivalent to requiring that for any $x, y \in [\ell, H]$, we must have

$$\begin{aligned}\mathcal{U}^M(t^x) &= \mathcal{U}^M(t^y) + \int_0^1 (t^x - t^y) \cdot f(t^y + z(t^x - t^y))dz \\ &= \mathcal{U}^M(t^y) + \int_0^1 (x - y) f_{a^*}(t^y + z(t^x - t^y))dz\end{aligned}$$

Define $\phi(x') := f_{a^*}(t^{x'})$ for all $x' \in [\ell, H]$. So, the above equation reduced to

$$\begin{aligned}\mathcal{U}^M(t^x) &= \mathcal{U}^M(t^y) + \int_0^1 (x-y)\phi(y+z(x-y))dz \\ &= \mathcal{U}^M(t^y) + \int_y^x \phi(x')dx'\end{aligned}$$

Now, if we only require for every $x \in [\ell, H]$,

$$\mathcal{U}^M(t^x) = \mathcal{U}^M(t^\ell) + \int_\ell^x \phi(x')dx',$$

then this will imply that for any $x, y \in [\ell, H]$

$$\begin{aligned}\mathcal{U}^M(t^x) - \mathcal{U}^M(t^y) &= \int_\ell^x \phi(x')dx' - \int_\ell^y \phi(x')dx' \\ &= \int_y^x \phi(x')dx',\end{aligned}$$

as desired in (b). This explains the weaker analogue of (b) in the one-dimensional version in Theorem 5. However, in the multidimensional case we need the stronger version as stated in (b) of Theorem 19. In other words, when type space is multidimensional, requiring (b) in Theorem 19 to hold for every $t \in T$ with respect to some "base" type s_0 does not imply (b) to hold for every pair of types $s, t \in T$.

5.2 The Implementation Problem

We now turn to the implementation problem, i.e., identifying conditions on an allocation rule that characterizes implementability. Corollary 1 achieves this in the one-dimensional type space. It shows that non-decreasingness of an allocation rule characterizes implementability in the one-dimensional type space. Since monotonicity is the natural generalization (as we showed above) of non-decreasingness for multidimensional type space, a natural conjecture is then that monotonicity is equivalent to implementability. This conjecture is false. The reason for this is the same reason why (b) in Theorem 19 is stronger than the analogue statement in Theorem 5. In one-dimensional type space if an allocation rule is non-decreasing, then fixing the $\mathcal{U}^M$ value for the "lowest" type uniquely fixes the value of $\mathcal{U}^M$ for all other types using (b), and this automatically ensures the statement (b). However, in multidimensional

type space, fixing $\mathcal{U}^M$ for some "base" type and then using (b) to fix the value of $\mathcal{U}^M$ for all other types does not ensure (b) to hold for all pairs of types.

To extend Corollary 1, we need a stronger version of monotonicity. Consider an implementable allocation rule f and two types $s, t \in T$. Since f is implementable there exist a payment rule p such that the mechanism $M \equiv (f, p)$ is incentive compatible. Then,

$$t \cdot f(t) - p(t) \geq t \cdot f(s) - p(s)$$

$$s \cdot f(s) - p(s) \geq s \cdot f(t) - p(t).$$

Adding these two constraints, we get that $(t - s) \cdot \big(f(t) - f(s)\big) \geq 0$, i.e., f is monotone. We can do this exercise for a longer sequence of types too. For instance, take three types $s, t, x \in T$ and consider the incentive constraints

$$t \cdot f(t) - p(t) \geq t \cdot f(s) - p(s)$$

$$s \cdot f(s) - p(s) \geq s \cdot f(x) - p(x)$$

$$x \cdot f(x) - p(x) \geq x \cdot f(t) - p(t).$$

Again, adding these constraints will cancel the payment terms and we will be left with only a condition on allocation rules.

To define this longer sequence condition, we define some notation. Let $\ell^f(s, t) = t \cdot \big(f(t) - f(s)\big)$ for every $s, t \in T$. Note that incentive constraint from true type t to type s is: $p(t) - p(s) \leq \ell^f(s, t)$. A good way to interpret this is that we create a directed graph G^f with set of nodes T (possibly infinite nodes). For every pair of nodes $s, t \in T$, we put an edge from s to t and another from t to s. So, G^f is a complete directed graph. We call this the **type graph** of allocation rule. We assign a weight of $\ell^f(s, t)$ to the edge from s to t. Monotonicity requires that the for every $s, t \in T$, we must have $\ell^f(s, t) + \ell^f(t, s) \geq 0$, i.e., 2-cycles (cycles involving pairs of nodes) have non-negative length. The longer sequence condition requires cycles of arbitrary number of nodes must have non-negative length.

DEFINITION **23** *An allocation rule satisfies* K**-cycle monotonicity** *if for any finite sequence of types* $(s^1, \ldots, s^k)$*, with* $k \leq K$*, each belonging to* T*, we have*

$$\sum_{j=1}^{k} \ell^f(s^j, s^{j+1}) \geq 0,$$

where $s^{k+1} \equiv s^1$.

An allocation rule satisfies **cycle monotonicity** *if it is* K*-cycle monotone for all* K.

Using ideas explained above, it is routine to verify that every implementable allocation rule satisfies cycle monotonicity. The following theorem shows that the converse also holds - the theorem does not require any assumption on type spaces (Theorem 19 required the type space to be convex) or restriction on the number of alternatives.

THEOREM **20** *An allocation rule is implementable if and only if it is cyclically monotone.*

Proof: Suppose f is an implementable allocation rule. Consider a finite sequence of types $(t^1, t^2, \ldots, t^k)$ with $k \geq 2$. Since f is implementable, there exists a payment rule p such that

$$\begin{aligned} p(t^2) - p(t^1) &\leq \ell^f(t^1, t^2) \\ p(t^3) - p(t^2) &\leq \ell^f(t^2, t^3) \\ \ldots &\leq \ldots \\ \ldots &\leq \ldots \\ p(t^k) - p(t^{k-1}) &\leq \ell^f(t^{k-1}, t^k) \\ p(t^1) - p(t^k) &\leq \ell^f(t^k, t^1). \end{aligned}$$

Adding these inequalities, we obtain that $\ell^f(t^1, t^2)+\ell^f(t^2, t^3)+\ldots+\ell^f(t^{k-1}, t^k)+\ell^f(t^k, t^1) \geq 0$.

Now, suppose f satisfies cycle monotonicity. For any two types $s, t \in T$, let $P(s, t)$ denote the set of all (finite) paths from s to t in G^f. The set $P(s, t)$ is non-empty because the direct edge from s to t in G^f always exists. Also, note that a path $P \in P(s, t)$ may not contain distinct edges, i.e., it may contain a cycle. Define the **shortest path** length from s to t ($s \neq t$) as follows.

$$dist^f(s, t) = \inf_{P \in P(s,t)} \ell^f(P),$$

where $\ell^f(P)$ is the length of path P. Let $dist^f(s, s) = 0$ for all $s \in T$. First, we show that $dist^f(s, t)$ is finite. Consider any path $P \in P(s, t)$. If there are cycles in this path, then they have non-negative length. By removing those cycles from P, we get another path $Q \in P(s, t)$ with distinct vertices such that $\ell^f(P) \geq \ell^f(Q)$. But Q with the direct edge (t, s) is a cycle. Hence, $\ell^f(Q) + \ell^f(t, s) \geq 0$. Hence, $\ell^f(P) \geq \ell^f(Q) \geq -\ell^f(t, s)$. This implies that $dist^f(s, t) = \inf_{P \in P(s,t)} \ell^f(P) \geq -\ell^f(t, s)$. Since $\ell^f(t, s)$ is a real number, $dist^f(s, t)$ is finite.

Next, fix a type $r \in T$ and choose $s, t \in T$. Choose a path $P^1 \in P(r,s)$ and then the direct edge (s,t). Denote the path P^1 added to (s,t) as P^2 and note that $P^2 \in P(r,t)$. Hence, $\ell^f(P^1) + \ell^f(s,t) = \ell^f(P^2) \geq dist^f(r,t)$. Hence, $\ell^f(P^1) \geq distr^f(r,t) - \ell^f(s,t)$. Since P^1 was chosen arbitrarily, we conclude that

$$dist^f(r,s) \geq dist^f(r,t) - \ell^f(s,t). \tag{5.2}$$

Now, define the following payment rule: let $p(s) = dist^f(r,s)$ for all $s \in T$.

Take any $s, t \in T$. We have $p(t) - p(s) = dist^f(r,t) - dist^f(r,s) \leq \ell^f(s,t)$, where the inequality follows from Inequality 5.2. Hence, f is implementable. ■

Theorem 20 shows that if an allocation rule is cyclically monotone, then we can construct payments for implementing such an allocation rule by considering *shortest paths* in the underlying graph. For every allocation rule f, cycle monotonicity is best understood by its type graph G^f, whose set of vertices is the type space T and it is a complete directed graph, i.e., there is an edge from every vertex to every other vertex. The weight of an edge from type s to type t is $\ell^f(s,t) = t \cdot (f(t) - f(s))$. The cycle monotonicity condition is a no negative cycle condition of this graph. Moreover, the payments corresponding to a type graph having no negative cycles can be found by fixing *any* arbitrary type and fixing its payment at zero but fixing the type payment of other types as the shortest path in T^f from this type.

We give a simple example to illustrate the construction of payments. Consider an allocation rule, whose underlying type graph looks as in Figure 5.2. First, verify that all cycles in this type graph have non-negative length. If we fix type s and let $p(s) = 0$, then the payments of other two types can be found by taking shortest paths from s to these types. In Figure 5.2, this can be verified to be $p(t) = 1, p(x) = -1$. Theorem 20 shows that this payment rule implements the allocation rule corresponding to Figure 5.2.

There are other payment rules that also implement an allocation rule. Of course, adding a constant to this payment at all types will also generate another payment rule which will implement this allocation rule. But, you can verify that the following payment rule also implements a cyclically monotone allocation rule f. Fix a type s and set $p'(s) = 0$ and for every type t, let $p'(t) = -dist^f(t,s)$, i.e., take the negative of the shortest path from t to s. For the allocation rule in Figure 5.2, this generates $p'(s) = 0, p'(t) = 1, p'(x) = -1$.

There is an alternate way to think about incentive constraints. We say f is **deterministic**

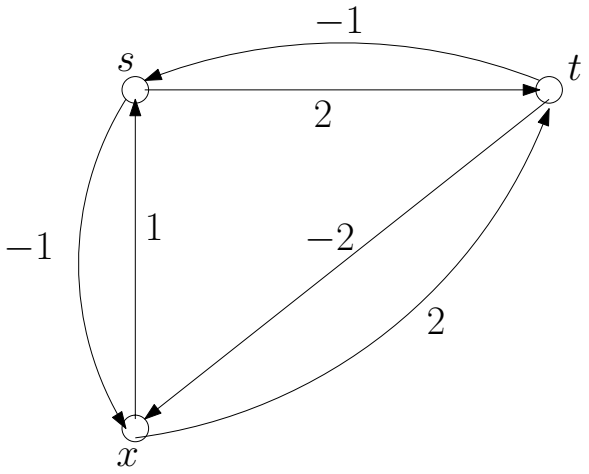

Figure 5.2: Cycle monotonicity

if for all $t \in T$ and for all $a \in A$, $f_a(t) \in \{0, 1\}$. For simplicity, let us consider deterministic rules. However, everything we discuss below can be extended to randomized allocation rules as follows: replace A by $\mathcal{L}(A)$ in the discussions below.

Given a type space T, we first partition T into various regions where each alternative is chosen. In particular, define for every $a \in A$,

$$T_a^f = \{t \in T : f(t) = a\}.$$

So $\{T_a^f\}_{a \in A}$ defines a partitioning of the type space. An example is shown in Figure 5.3. The boundaries of these partitions may belong to any of the alternatives which share the boundary. Futher, the partitioning will always *polyhedral* figures - this is because the incentive constraints are linear.

Now, incentive compatibility means that for every $a \in A$, the following holds: $p(t) = p(s)$ for all $s, t \in T_a$. Then, the implementability question can be slightly differently stated. We can say f is implementable if there exists a map $\pi : A \to \mathbb{R}$ such that for every $a \in A$ and for every $t \in T_a^f$, the following holds:

$$t(a) - \pi(a) \geq t(b) - \pi(b) \qquad \forall\ b \in A.$$

In other words, instead of finding the payment rule p, we can indirectly find it by constructing the π map.

Rewriting the above inequality, we get for all $a \in A$ and for all $t \in T_a^f$, the following must

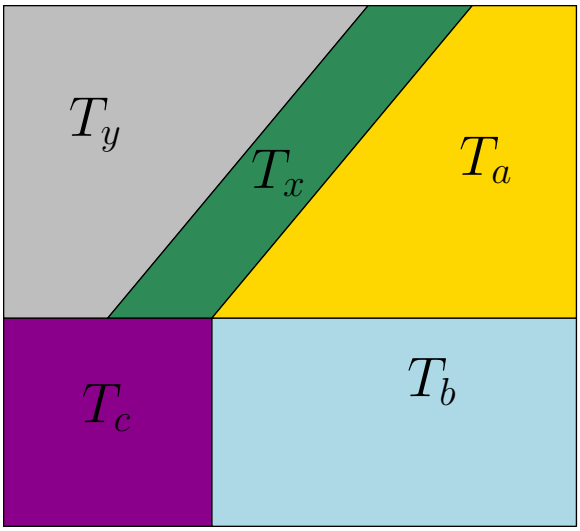

Figure 5.3: Partitioning the type space

hold:

$$\pi(a) - \pi(b) \leq t(a) - t(b) \qquad \forall\ b \in A.$$

The above inequality can be succinctly written as follows.

$$\pi(a) - \pi(b) \leq \inf_{t \in T_a^f} \big[t(a) - t(b)\big] \qquad \forall\ a, b \in A. \tag{5.3}$$

Now, we can attach a new graph to an allocation rule f. We call this the **allocation graph** and denote this as A^f. The set of nodes in this graph is A - for every alternative, we put a node. It is a complete directed graph. So, for every $a, b \in A$, there is an edge from a to b and an edge from b to a. The edge length of edge (a, b) is

$$d^f(a, b) := \inf_{t \in T_a^f} \big[t(a) - t(b)\big].$$

Just as we showed in Theorem 20, we can show that Inequality (5.3) has a solution if and only if the directed graph A^f has no cycles of negative length. Hence, we have shown the following.

THEOREM **21** *The following statements are equivalent.*

1. *f is an implementable allocation rule.*

2. *The type graph of G^f of f has no cycles of negative length.*

3. *The allocation graph A^f of f has no cycles of negative length.*

As noted cycle monotonicity is a significantly stronger condition than monotonicity. Monotonicity has been shown to imply cycle monotonicity if type space is convex and allocation rule is deterministic. We state this as a theorem below without giving a proof.

THEOREM **22** *Suppose T is convex and $f : T \to \mathcal{L}(A)$ is a deterministic allocation rule. Then, f is implementable if and only if it is monotone.*

The above result is not true if we consider randomized allocation rules. The following is an example taken from Bikh06.

EXAMPLE **1**

There are two units of a good and marginal values of the good to the buyer $(v_1, v_2) \in [0,1]^2$. Let $q = (q_1, q_2)$ be a random allocation rule which gives the probability of allocating the units. In particular, $q_k(v)$ with $k \in \{1,2\}$ is the probability with which *at least* k units are allocated at (v_1, v_2). It is easy to verify that 2-cycle monotonicity (or monotonicity) is equivalent to requiring for all $v, v' \in [0,1]^2$, we have $\big(q(v) - q(v')\big)(v - v') \geq 0$.

Now, suppose $q(v) = \frac{1}{3}Av$, where A is a 2×2 matrix with rows $(1,2)$ and $(0,1)$. It is easy to verify that q satisfies 2-cycle monotonicity. To refute implementability, we argue that q cannot be the subgradient of a convex function. Because if it is a subgradient of a convex function, the matrix of second-partials of such a convex function will be A, which is not possible since A is not symmetric.

5.3 REVENUE EQUIVALENCE

Consider an allocation rule f which is DSIC. Let p be a payment rule which makes f DSIC. Let $\alpha \in \mathbb{R}$. Define $q(t) = p(t) + \alpha$ for all $t \in T$. Since $q(t) - q(s) = p(t) - p(s) \leq \ell^f(s,t)$, we see that q is also a payment that makes f DSIC. Is it possible that all payments that make f DSIC can be obtained by adding a suitable constant $\alpha \in \mathbb{R}$ to p? This property of an allocation rule is called **revenue equivalence**. Not all allocation rules satisfy revenue

equivalence. Myerson (1981) showed that in the standard auction of single object (one-dimensional type space) every allocation rule satisfies revenue equivalence. The objective of this section is to identify allocation rules that satisfy revenue equivalence in more general settings.

DEFINITION **24** *An allocation rule f satisfies* **revenue equivalence** *if for any two payment rules p and $\hat{p}$ that make f DSIC, there exists a constant $\alpha \in \mathbb{R}$* [1] *such that*

$$p(t) = \hat{p}(t) + \alpha \qquad \forall\ t \in T. \tag{5.4}$$

The first characterization of revenue equivalence involves no assumptions on T or A or f.

THEOREM **23** *Suppose f is implementable. Then the following are equivalent.*

1. *The allocation rule f satisfies revenue equivalence.*

2. *For all $s, t \in T$, we have $dist^f(s,t) + dist^f(t,s) = 0$.*

Proof: We establish the equivalence of 1 and 2 first. Suppose f satisfies revenue equivalence. Consider any $s, t \in T$. Since f is DSIC, by Theorem 20, the following two payment rules makes f DSIC:

$$p^s(r) = dist^f(s,r) \qquad \forall r \in T$$
$$p^t(r) = dist^f(t,r) \qquad \forall r \in T.$$

Since revenue equivalence holds, $p^s(s) - p^t(s) = p^s(t) - p^t(t)$. But $p^s(s) = p^t(t) = 0$. Hence, $p^s(t) + p^t(s) = 0$, which implies that $dist^f(s,t) + dist^f(t,s) = 0$.

Now, suppose $dist^f(s,t) + dist^f(t,s) = 0$ for all $s, t \in T$. Consider any payment rule p that makes f DSIC. Take any path $P = (s, t_1, \ldots, t_k, t)$ from s to t. Now, $\ell^f(P) = \ell^f(s,t_1) + \ell^f(t_1,t_2) + \ldots + \ell^f(t_{k-1},t_k) + \ell^f(t_k,t) \geq [p(t_1) - p(s)] + [p(t_2) - p(t_1)] + \ldots + [p(t_k) - p(t_{k-1})] + [p(t) - p(t_k)] = p(t) - p(s)$. Hence, $p(t) - p(s) \leq \ell^f(P)$ for any path P from s to t. Hence, $p(t) - p(s) \leq dist^f(s,t)$. Similarly, $p(s) - p(t) \leq dist^f(t,s)$. Hence,

[1] In a model with more than one agent α can be a (agent-specific) mapping from type profile of other players to real numbers.

$0 = dist^f(s,t) + dist^f(t,s) \geq [p(s) - p(t)] + [p(t) - p(s)] = 0$. Hence, $p(s) - p(t) = dist^f(t,s)$, which is independent of $p(\cdot)$. Hence, revenue equivalence holds. ■

One can use this theorem to derive our familiar revenue equivalence theorem that we proved earlier - the second part of Theorem 19. However, the current theorem makes use of Theorem 23 to give a cleaner proof without resorting to particular form of payment.

THEOREM **24** *Suppose T is convex. If f is implementable, then f satisfies revenue equivalence.*

Proof: Pick any $s, t \in T$ and $\epsilon > 0$. Our first claim is that there exists points $r_1, \ldots, r_k$ in the convex hull of s and t such that

$$\ell^f(s, r_1) + \ell^f(r_1, r_2) + \ldots + \ell^f(r_k, t) + \ell^f(t, r_k) + \ell^f(r_k, r_{k-1}) + \ldots + \ell^f(r_1, s) < \epsilon.$$

To see this, let $\Delta = t - s$. Pick a positive integer k and define for every $j \in \{1, \ldots, k\}$,

$$r_j := s + \frac{j}{k+1}\Delta.$$

By convexity of T, $r_j \in T$. Denote $r_0 = s$ and $r_{k+1} = t$. Now, note that

$$\ell^f(r_j, r_{j+1}) + \ell^f(r_{j+1}, r_j) = (r_{j+1} - r_j)\Big(f(r_{j+1}) - f(r_j)\Big) = \frac{1}{k+1}\Delta \cdot \Big(f(r_{j+1}) - f(r_j)\Big).$$

Hence, we see that

$$\begin{aligned}\sum_{j=0}^{k} \ell^f(r_j, r_{j+1}) + \ell^f(r_{j+1}, r_j) &= \frac{1}{k+1}\sum_{j=0}^{k} \Delta \cdot \Big(f(r_{j+1}) - f(r_j)\Big) \\ &= \frac{1}{k+1}\Delta \cdot \Big(f(t) - f(s)\Big)\end{aligned}$$

Since $\Delta \cdot (f(t) - f(s))$ is a constant, the above expression can be made arbitrarily small by picking k large enough. This proves the claim.

Now, we show that $dist^f(s,t) + dist^f(t,s) = 0$. Since f is implementable, cycle monotonicity (Theorem 20) implies that $dist^f(s,t) + dist^f(t,s) \geq 0$. [2] Assume for contradiction $dist^f(s,t) + dist^f(t,s) = \epsilon > 0$. Then, by our claim in the first part, there is a path from

[2] If $dist^f(s,t) + dist^f(t,s) < 0$, there is some path from s to t and some path from t to s such that the sum of these path lengths is negative. This implies that there is some cycle whose length is negative, contradicting cycle monotonicity.

s to t and another from t to s such that their sum is strictly less that ϵ. This implies $dist^f(s,t) + dist^f(t,s) < \epsilon$, which is a contradiction. ■

With deterministic allocation rules, revenue equivalence can be shown to hold in any connected type space.

5.4 Optimal Multi-Object Auction

Our discussions so far have shown how many of the results for one-dimensional mechanism design can be extended when the type space is multidimensional. Though, it gives an expression for payment (via (b) of Theorem 19), this expression is not as easy to handle because the expectation over the type space is now complicated using a multidimensional joint distribution. As a result, the Myersonian technique that we employed for the one-dimensional type space does not yield any useful result. Finding an optimal multi-object auction for selling two objects to one buyer (the simplest possible setting) is very complicated. The nature of optimal multi-object auction is hard to describe.

We give below some complications involved in computing an optimal auction through some examples. All the examples are borrowed from Hart and Reny (2015).

Example **2**

There are two objects being sold by a seller to a single buyer. The buyer's type is her value for object 1, v_1, and her value for object 2, v_2. Her value for objects 1 and 2 together is $v_1 + v_2$ (this is called the additive values multi-object problem).

A mechanism in this setting can be described by a **menu** of outcomes - incentive compatibility implies that the buyer chooses the payoff-maximizing outcome from this menu. We describe one such menu below. We denote by (S, p) a generic element of the menu, where $S \subseteq \{1, 2\}$ is the bundle of goods and p is the price charged to S.

$$\mathcal{M} \equiv \Big\{ \big(\emptyset, 0\big), \big(\{1\}, 1\big), \big(\{2\}, 2\big), \big(\{1,2\}, 4\big) \Big\}.$$

Such a menu $\mathcal{M}$ can be associated with an incentive compatible mechanism by associating with each type its payoff maximizing outcome from the menu. Denote such a mechanism

(q, p). Notice that since our menu consists of $(\emptyset, 0)$, the buyer cannot do worse than getting zero payoff. Hence, this mechanism is also individually rational.

Now, consider a type $v \equiv (1, \frac{7}{3})$. Then, the optimal outcome in the menu $\mathcal{M}$ for this type is object $(\{2\}, 2)$. So, the revenue of the seller at this type is 2. Now, suppose we consider type $v' \equiv (2, \frac{8}{3})$. Then, the optimal outcome in the menu $\mathcal{M}$ for this type is $(\{1\}, 1)$. Hence, the revenue of the seller falls to 1. Note that $v'_1 > v_1$ and $v'_2 > v_2$ but revenue at v' falls below v.

The main intuition for why this non-monotonicity is happening is the following. First, the price of the bundle is high enough - higher than sum of prices of individual objects. Second, the price of object 1 is low but when the value of object 2 is high enough, the buyer chooses object 2. However, if we increase the value of object 1 higher than that of object 2, then the buyer switches to object 1 (cannot switch to the bundle because it has high price). But the seller gets a lower price.

The above example illustrates the point that there can be incentive compatible mechanisms where revenue is non-monotonic in values. Can this happen in an optimal mechanism? Indeed, the menu in an optimal mechanism can be quite different from the one shown in Example 2. To describe an optimal mechanism, we first describe the type space and show that in such a type space, the mechanism described above is optimal for a class of distributions.

EXAMPLE **3**

We continue with the same example: one buyer, one seller, two objects, and additive values. We describe a family of distributions F_α (parametrized by $\alpha \in [0, \frac{1}{12}]$) with finite support.

$$F_\alpha = \begin{cases} (1,1) & \text{with probability } \frac{1}{4} \\ (1,2) & \text{with probability } \frac{1}{4} - \alpha \\ (2,2) & \text{with probability } \alpha \\ (2,3) & \text{with probability } \frac{1}{2} \end{cases}$$

Notice that as α increases, the probability shifts from $(2,2)$ to $(1,2)$. Hence, $F_\alpha \succ_{FOSD} F_{\alpha'}$ for all $\alpha > \alpha'$. We prove the following claim.

CLAIM **1** *For all $0 \leq \alpha \leq \frac{1}{12}$, the optimal mechanism generates a revenue of $\frac{11}{4} - \alpha$ for distribution F_α.*

Proof: Consider the mechanism in Example 2 given by the menu:

$$\mathcal{M} \equiv \Big\{\big(\emptyset, 0\big), \big(\{1\}, 1\big), \big(\{2\}, 2\big), \big(\{1,2\}, 4\big)\Big\}.$$

In this mechanism type $(1,1)$ chooses $(\{1\},1)$; type $(1,2)$ chooses $(\{2\},2)$; type $(2,2)$ chooses $(\{1\},1)$; and type $(2,3)$ chooses $(\{1,2\},4)$. The expected revenue from this mechanism is

$$1 \times \frac{1}{4} + 2 \times \big(\frac{1}{4} - \alpha\big) + 1 \times \alpha + 4 \times \frac{1}{2} = \frac{11}{4} - \alpha.$$

We now show that any other incentive compatible (IC) and individually rational (IR) mechanism cannot generate more expected revenue. Pick an IC and IR mechanism (q,p), where $q_i(v)$ denotes the probability of getting object $i \in \{1,2\}$ at type v and $p(v)$ denotes the payment at type v. Since (q,p) is an IC and IR mechanism, the following constraints must hold.

$$\begin{aligned}
q_1(1,1) + q_2(1,1) - p(1,1) &\geq 0 && \text{Multiply } 1 \\
q_1(1,2) + 2q_2(1,2) - p(1,2) &\geq q_1(1,1) + 2q_2(1,1) - p(1,1) && \text{Multiply } \frac{1}{2} \\
2q_1(2,2) + 2q_2(2,2) - p(2,2) &\geq 2q_1(1,1) + 2q_2(1,1) - p(1,1) && \text{Multiply } 3\alpha \\
2q_1(2,3) + 3q_2(2,3) - p(2,3) &\geq 2q_1(1,1) + 3q_2(1,1) - p(1,1) && \text{Multiply } \frac{1}{4} - 3\alpha \\
2q_1(2,3) + 3q_2(2,3) - p(2,3) &\geq 2q_1(1,2) + 3q_2(1,2) - p(1,2) && \text{Multiply } \frac{1}{4} + \alpha \\
2q_1(2,3) + 3q_2(2,3) - p(2,3) &\geq 2q_1(2,2) + 3q_2(2,2) - p(2,2) && \text{Multiply } 2\alpha
\end{aligned}$$

We considered the IR constraint of the lowest type $(1,1)$, we considered the IC constraints of each type to the lowest type $(1,1)$, and finally, we considered the IC constraints from the highest type $(2,3)$ to every other type. In other words, we considered all **downward** IC constraints except the $(2,2) \to (1,2)$ IC constraint and the lowest type IR constraint. As it turns out, these are enough to show optimality of our mechanism.

Add the constraints by multiplying with the quantities mentioned on the right. This gives us:

$$\begin{aligned}
&-\Big(\frac{3}{4} - 3\alpha\Big)q_2(1,1) - 2\alpha q_1(1,2) + \Big(\frac{1}{4} - 3\alpha\Big)q_2(1,2) + 2\alpha q_1(2,2) + q_1(2,3) + \frac{3}{2}q_2(2,3) \\
&\geq \frac{1}{4}p(1,1) + \Big(\frac{1}{4} - \alpha\Big)p(1,2) + \alpha p(2,2) + \frac{1}{2}p(2,3).
\end{aligned}$$

The RHS is exactly the expected revenue of the mechanism (q, p). The first two terms on LHS is negative since $\frac{3}{4} \geq 3\alpha$ (since $\alpha \leq \frac{1}{12}$). The rest of the terms are positive since $\alpha \leq \frac{1}{12}$. Hence, we can upper bound the LHS by $\frac{11}{4} - \alpha$, which is the expected revenue from our candidate mechansim with menu $\mathcal{M}$. Hence, this mechanism is optimal - further, if $\alpha < \frac{1}{12}$, then it is the unique optimal mechanism. ■

A straightforward corollary of Claim 1 is that with $\alpha, \alpha' \in [0, \frac{1}{12}]$ and $\alpha > \alpha'$, we have optimal revenue from F_α strictly lowere than the optimal revenue from $F_{\alpha'}$.

Our next example shows another property of optimal mechanisms which was not present in the single object case - the optimal mechanism may involve randomization. We continue with the same setting: one buyer, one seller, and two objects with additive valuations. Consider the following type space with associated probability distribution.

EXAMPLE **4**

$$F = \begin{cases} (1,0) & \text{with probability } \frac{1}{3} \\ (0,2) & \text{with probability } \frac{1}{3} \\ (3,3) & \text{with probability } \frac{1}{3} \end{cases}$$

The following randomized mechanism is shown to be the unique optimal mechanism in this example.

CLAIM **2** *The following is the unique optimal mechanism. Its menue consists of*

$$\Big(\emptyset, 0\Big), \Big(\frac{1}{2}\{1\}, \frac{1}{2}\Big), \Big(\{2\}, 2\Big), \Big(\{1,2\}, 5\Big),$$

where $\frac{1}{2}\{1\}$ *means object* 1 *is allocated with probability* $\frac{1}{2}$.

Proof: Let the outcomes at the three types be denoted by $(\alpha_1, \beta_1; \sigma_1)$ for type $(1,0)$; $(\alpha_2, \beta_2; \sigma_2)$ for type $(0,2)$; and $(\alpha_3, \beta_3; \sigma_3)$ for type $(3,3)$. Here, αs correspond to the probability of getting object 1, βs correspond to the probability of getting object 2, and σs are the payments. Since each type has equal probability, expected revenue is maximized by maximizing $\sigma_1 + \sigma_2 + \sigma_3$.

Now, consider the *relaxed problem*, where we consider the IR constraints of types $(1,0)$ and $(0,2)$, and consider the IC constraints from type $(3,3)$.

$$\alpha_1 - \sigma_1 \geq 0$$
$$2\beta_2 - \sigma_2 \geq 0$$
$$3\alpha_3 + 3\beta_3 - \sigma_3 \geq 3\alpha_1 + 3\beta_1 - \sigma_1$$
$$3\alpha_3 + 3\beta_3 - \sigma_3 \geq 3\alpha_2 + 3\beta_2 - \sigma_2.$$

Rewriting these inequalities:

$$\sigma_3 + 3\alpha_1 + 3\beta_1 - 3\alpha_3 - 3\beta_3 \leq \sigma_1 \leq \alpha_1$$
$$\sigma_3 + 3\alpha_2 + 3\beta_2 - 3\alpha_3 - 3\beta_3 \leq \sigma_2 \leq 2\beta_2.$$

So, to maximize $\sigma_1 + \sigma_2 + \sigma_3$, we set $\sigma_1 = \alpha_1$ and $\sigma_2 = 2\beta_2$ - so, IR constraints of low types bind. This simplifies the above inequalities to

$$\sigma_3 \leq 3\alpha_3 + 3\beta_3 - 2\alpha_1 - 3\beta_1$$
$$\sigma_3 \leq 3\alpha_3 + 3\beta_3 - 3\alpha_2 - \beta_2.$$

So, to maximize σ_3, we must take $\alpha_3 = \beta_3 = 1$ and $\beta_1 = \alpha_2 = 0$. Then, we should choose $\sigma_3 = \min(6-2\alpha_1, 6-\beta_2)$. So, the objective function becomes $\alpha_1+2\beta_2+\min(6-2\alpha_1, 6-\beta_2) = 6 + \min(2\beta_2 - \alpha_1, \beta_2 + \alpha_1)$. This means the objective function is increasing in β_2. So, we should set $\beta_2 = 1$. Then, the objective function becomes $6 + \min(2 - \alpha_1, 1 + \alpha_1)$. This is maximized when $2 - \alpha_1 = 1 + \alpha_1$ or $\alpha_1 = \frac{1}{2}$. This is exactly the mechanism we started out with. Hence, the given mechanism is optimal. ■

This example and Claim 2 establishes that the optimal mechanism may contain randomization. In particular, the set of IC and IR mechanisms form a convex set - convex combination of two IC and IR mechanisms produce an IC and IR mechanism. However, the *extreme points* of such IC and IR mechanisms may contain randomized mechanisms. This is not the case for single object problem - a formal proof of the fact that extreme points of single object problem is deterministic is left as an exercise. [3]

[3]Claim 2 seems to suggest that correlation is necessary to have randomization in the optimal mechanism menu. This is not the case. Hart and Reny (2015) have an example where values of each object is drawn independently from the same distribution and the optimal mechanism still involves randomization.

These examples highlight some important facts. First, the menu of the optimal mechanism may vary depending on the type space and distribution. Sometimes, the menu may contain randomization. The set of *relevant* IC and IR constraints also vary. This makes the optimal multi-object auction problem a notorious problem. A good starting point to understand these in detail is Manelli and Vincent (2007).

Chapter 6

Extensions

In this chapter, we present two extensions of the model we have been discussing so far. The first extension relaxes the quasilinearity assumption and investigates the single object allocation model. In particular, it shows a natural extension of Vickrey auction without quasilinearity. The second extension explores the implications of relaxing the private values assumption. We investigate a model of interdependent values and how to extend the Vickrey auction to such a model.

6.1 Classical Preferences

An important feature of mechanism design with transfers has been the quasilinearity assumption. Quasilinear preferences over transfers allows us to separate out transfer rule and allocation rule, say, by the revenue equivalence results. However, in many settings quasilinear preferences are not realistic. We start off by presenting a model of nonquasilinear preferences and some examples.

There is a single indivisible object. There are n agents interested in the object. The set of agents are denoted by $N := \{1, \ldots, n\}$. Monetary transfers are allowed. A typical consumption bundle of an agent i is (a, p), where $a \in \{0, 1\}$ indicates whether the agent is assigned ($a = 1$) the object or not ($a = 0$) and $p \in \mathbb{R}$ indicates his payment. Let $Z = \{0, 1\} \times \mathbb{R}$ be the set of all consumption bundles of any agent. Note that we do not deal with randomization in this model.

The **type** of an agent i is a map $u_i : Z \to \mathbb{R}$, where we normalize $u_i(0,0) = 0$. [1]

DEFINITION **25** *A type u_i is* **classical** *if it satisfies the following conditions:*

- **Money Monotonicity (M)**. *for every $p, p' \in \mathbb{R}$, $p > p'$ implies $u_i(a,p') > u_i(a,p)$ for all $a \in \{0,1\}$,*
- **Desirability (D)**. *for every $p \in \mathbb{R}$, $u_i(1,p) > u_i(0,p)$,*
- **Continuity (C)**. *for every $a \in \{0,1\}$, $u_i(a,p)$ is a continuous function of p.*
- **Finiteness (F)**. *for every $p \in \mathbb{R}$, there exists $d_1, d_2 \in \mathbb{R}_{++}$ such that $u_i(0,p) \geq u(1,p+d_1)$ and $u_i(0,p-d_2) \geq u(1,p)$.*

The set of all classical types is denoted by $\mathcal{U}^c$.

The conditions **M**, **D**, and **C** are reasonable and standard conditions. To understand finiteness, note that **D** implies that at every transfer, getting the object is strictly preferred to not getting it. Condition **F** ensures that there is no transfer amount, where getting the object is *infinitely* preferred to not getting it.

Note that the a type u_i is *quasilinear* if there exists a positive real number (valuation) v_i such that for all $a \in \{0,1\}$ and for all $p \in \mathbb{R}$, we have $u_i(a,p) = v_i \cdot a - p$. It is easy to verify that a quasilinear type is a classical type.

6.1.1 Type Spaces with Income Effects

An unpleasant feature of quasilinear type is the absence of any income effect. However, there are plenty of instances where income effects are plausible. We present two examples where income effects are apparent and quasilinearity is unrealistic.

- **Nonlinear cost of payments.** Suppose agent i has a budget b_i. If the payment exceeds b_i, then he takes a loan at an interest rate of r. Effectively, if agent i has to make a payment of p, his cost of payment is

$$c_i(p, b_i) = \begin{cases} p & \text{if } p \leq b_i \\ b_i + (p - b_i)(1+r) & \text{otherwise} \end{cases}$$

[1] We consider a cardinal model with utility only for convenience and the model can rewritten with ordinal preferences.

So, if agent i has a value v_i, his net utility (for winning the object) from a payment of p is

$$v_i - c_i(p, b_i),$$

where $c_i(p, b_i)$ is not linear in p and depends on another dimension of his private information - his budget b_i.

- **Surplus depending values.** In the previous example, the value of the object was fixed. In many scenarios, the value of a good itself depends on the amount of money you are left with. Consider the sale of a spectrum license. A company will be able to use the license more effectively if it is left with more money after it buys the license. So, the value of agent i is a map $v_i : \mathbb{R} \to \mathbb{R}$. In particular, if b_i is the initial endowment of money (budget) of agent i and he pays p, then $b_i - p$ is his surplus money. We assume that agent i has access to credit and allow this surplus to become negative. Hence, the net utility of agent i (for winning the object) with value function v_i, budget b_i, and payment p is

$$v_i(b_i - p) + b_i - p.$$

To formally define the notion of income effect, we will require some additional terminology. At every $p \in \mathbb{R}$, let $WP(u_i, p)$ denote the **willingness to pay** of agent i at type u_i, and it is defined as the solution to $u_i(1, p + x) = u_i(0, p)$.

FACT **1** *If u_i is a classical type, then for every $p \in \mathbb{R}$, $WP(u_i, p)$ is a unique positive real number.*

Proof: Consider a classical type u_i and $p \in \mathbb{R}$. By **D**, $u_i(1, p) > u_i(0, p)$. By **F**, there exists $d_1 \in \mathbb{R}_{++}$ such that $u_i(0, p) \geq u_i(1, p+d_1)$. Hence, we have, $u_i(1, p) > u_i(0, p) \geq u_i(1, p+d_1)$. By continuity of u_i, there must exist $p' \in (p, p + d_1]$ such that $u_i(1, p') = u_i(0, p)$. By **M**, such a p' is also unique. Hence, there is a unique $x \in \mathbb{R}_{++}$ such that $u_i(1, p + x) = u_i(0, p)$. ■

If a type u_i is quasilinear, then $WP(u_i, p) = WP(u_i, p')$ for all $p, p' \in \mathbb{R}$, and this willingness to pay is precisely the valuation for the object.

An easy interpretation of classical preferences is through indifference vectors. Consider Figure 6.1. It shows two parallel lines which represent the consumption bundles of the agent

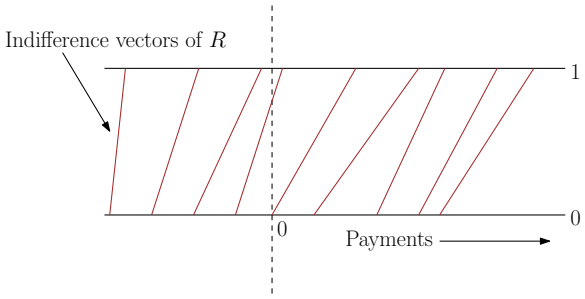

Figure 6.1: Classical preferences

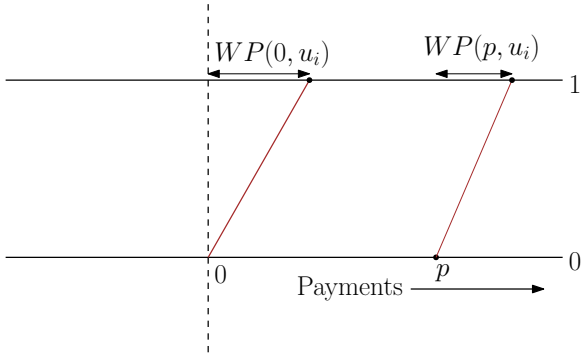

Figure 6.2: Classical preferences

- one line corresponds to not getting the object (0) and the other line corresponds to getting the object (1). As we go to the right on these lines, payment increases. Hence, on the same line, the agent is worse off as she goes to the right. On the other hand, we can also compare points across the two lines. The assumptions in classical preference implies that for every point on one line, there exists a point on the other line to which the agent is indifferent. This is what we refer to as an **indifference vector**. Figure 6.1 shows such collection of indifference vector for one preference. Indeed, a preference consists of infinite collection of such indifference vectors. An alternate way to think of a classical preference is through these indifference vectors. Because of desirability condition, the indifference vectors are slanted to right. For quasilinear preference, these indifference vectors are parallel.

The idea of WP can also be seen in such indifference figures. See Figure 6.2 for a representation of $WP(0, u_i)$ and $WP(p, u_i)$ for some utility function u_i, whose indifference vectors are shown.

Now, we define the notions of income effects. We define them using the notion of willing-

ness to pay. As transfer decreases (i.e, income increases), positive (negative) income effect requires that willingness to pay must increase (decrease).

DEFINITION **26** *A classical type u exhibits* **non-negative income effect** *if for all $p, p' \in \mathbb{R}$ with $p > p'$, we have $WP(u, p') \geq WP(u, p)$. A classical type u exhibits* **positive income effect** *if the above inequality is strict.*

A classical type u exhibits **non-postive income effect** *if for all $p, p' \in \mathbb{R}$ with $p > p'$, we have $WP(u, p') \leq WP(u, p)$. A classical type u exhibits* **negative income effect** *if the above inequality is strict.*

The set of all classical types with non-negative income effect, positive income effect, non-positive income effect, and negative income effect are denoted as $\mathcal{U}^+$, $\mathcal{U}^{++}$, $\mathcal{U}^-$, and $\mathcal{U}^{--}$ respectively.

Note that a quasilinear type exhibits both non-negative and non-positive income effect. Moreover, the set of all quasilinear types are precisely $\mathcal{U}^+ \cap \mathcal{U}^-$. We will denote the set of all quasilinear types as $\mathcal{U}^q$.

6.1.2 Mechanisms and Incentive Compatibility

A mechanism in this model consists of an allocation rule and a payment rule for every agent. Fix some type space $\mathcal{U} \subseteq \mathcal{U}^c$. An allocation rule is a map

$$f : \mathcal{U}^n \to \{0, 1\}^n.$$

At any type profile $\mathbf{u} \equiv (u_1, \ldots, u_n)$, we denote by $f_i(\mathbf{u}) \in \{0, 1\}$ the allocation of agent i - whether he has received the object or not. Of course, we assume that such an allocation rule is feasible, i.e., at every $\mathbf{u}$, we have $\sum_{i \in N} f_i(\mathbf{u}) \leq 1$.

A payment rule of agent $i \in N$ is a map $p_i : \mathcal{U}^n \to \mathbb{R}$. A mechanism is defined by $(f, p_1, \ldots, p_n)$. An alternative way to define a mechanism is to jointly write $(f, p_1, \ldots, p_n)$ as a map $F : \mathcal{U}^n \to Z^n$, where $Z = \{0, 1\} \times \mathbb{R}$ is the set of all consumption bundles of an agent.

DEFINITION **27** *A mechanism $(f, p_1, \ldots, p_n)$ is* **dominant strategy incentive compatible (DSIC)** *if for every $i \in N$, for every $u_{-i} \in \mathcal{U}^{n-1}$, and for every $u_i, u_i' \in \mathcal{U}$, we have*

$$u_i(f_i(\mathbf{u}), p_i(\mathbf{u})) \geq u_i(f_i(u_i', u_{-i}), p_i(u_i', u_{-i})).$$

Because of non-separability of payment terms from the utility function, it is difficult to see how the usual tricks of quasilinearity can be applied here. Still, we will construct DSIC mechanisms satisfying additional properties in this model. The first such axiom is efficiency.

Note that an outcome of an agent is a vector in Z consisting of an allocation decision and a payment decision. A typical outcome $z \in Z$ will be denoted as $z \equiv (a, t)$, where $a \in \{0, 1\}$ and $t \in \mathbb{R}$.

DEFINITION **28** *An outcome vector* $\mathbf{z} \equiv (z_1 \equiv (a_1, t_1), \ldots, z_n \equiv (a_n, t_n))$ *is* **Pareto efficient** *at* $\mathbf{u}$ *if there exists no outcome* $\mathbf{z}'$ *such that*

(A) $\sum_{i \in N} t'_i \geq \sum_{i \in N} t_i$

(B) $u_i(z'_i) \geq u_i(z_i)$ *for all* $i \in N$*,*

with strict inequality holding for one of the inequalities.

A mechanism $(f, p_1, \ldots, p_n)$ *is Pareto efficient if at every type profile* $\mathbf{u}$*, the outcome vector* $\{f_i(\mathbf{u}), p_i(\mathbf{u})\}_{i \in N}$ *is Pareto efficient at* $\mathbf{u}$*.*

Condition (A) in the definition is necessary. Without it, we can always subsidize every agent more than the given outcome vector, and Pareto improve it. The definition says that such improvements are only possible if the auctioneer does not have to spend extra money.

The second condition that we use is individual rationality.

DEFINITION **29** *A mechanism* $(f, p_1, \ldots, p_n)$ *is* **individually rational (IR)** *if at every type profile* $\mathbf{u}$*, we have* $u_i(f_i(\mathbf{u}, p_i(\mathbf{u})) \geq u_i(0, 0) = 0$ *for all* $i \in N$*.*

If an agent does not participate in the mechanism, he does not get the object and does not pay anything. The individual rationality assumption ensures that participation constraints are met.

The final axiom that we consider is no-subsidy.

DEFINITION **30** *A mechanism* $(f, p_1, \ldots, p_n)$ *satisfies* **no-subsidy** *if at every type profile* $\mathbf{u}$*, we have* $p_i(\mathbf{u}) \geq 0$ *for all* $i \in N$*.*

Though the no-subsidy axiom sounds natural in many auction environment, it may not be an attractive axiom in non-quasilinear models. For instance, we may want to subsidize low

utility agents to attract them to participate, and that may lead to increase in revenue. We present a mechanism which satisfies no-subsidy, and give examples of mechanisms that may violate no-subsidy.

6.1.3 Vickrey Auction with Income Effect

We will now extend the Vickrey auction to this setting and show that it is DSIC.

DEFINITION **31** *The* **Vickrey mechanism** *is a mechanism* $(f^*, p_1^*, \ldots, p_n^*)$ *such that every type profile* $\mathbf{u}$ *we have*

$$f_i^*(\mathbf{u}) = 1 \;\Rightarrow\; WP(u_i, 0) \geq WP(u_j, 0) \;\forall\; j \in N,$$

and for every $i \in N$*,*

$$p_i^*(\mathbf{u}) = \begin{cases} \max_{j \neq i} WP(u_j, 0) & \text{if } f_i(\mathbf{u}) = 1 \\ 0 & \text{otherwise} \end{cases}$$

Note that the Vickrey mechanism is a simple mechanism - the only information it needs from each agent is their willingness to pay at price zero. So, agents do not have to report their entire utility function to the designer.

THEOREM **25** *The Vickrey mechanism is DSIC, Pareto efficient, IR, and satisfies no-subsidy.*

Proof: It is clear that the Vickrey mechanism satisfies no-subsidy. To see that the Vickrey mechanism satisfies IR, note that an agent who does not win the object pays zero, and hence, his utility is zero. Suppose agent i wins the object at type profile $\mathbf{u}$. Then, his utility is $u_i(1, \max_{j \neq i} WP(u_j, 0))$. But $WP(u_i, 0) \geq \max_{j \neq i} WP(u_j, 0)$ means

$$u_i(1, \max_{j \neq i} WP(u_j, 0)) \geq u_i(1, WP(u_i, 0)) = u_i(0, 0).$$

Hence, IR is satisfied.

To show DSIC, if agent i wins the object, his payment remains the same as long as he wins the object. So, if he manipulates he needs to manipulate to a type where he does not get the object. But that gives him a payoff of zero. By IR, this is not a profitable manipulation. The other case is agent i does not win the object by being truthful, in which case he gets

zero payoff. By manipulating if he wins the object, he pays $\max_{j\neq i} WP(u_j, 0) \geq WP(u_i, 0)$. Hence, $u_i(1, \max_{j\neq i} WP(u_j, 0)) \leq u_i(1, WP(u_i, 0)) = u_i(0, 0) = 0$. Hence, this is not a profitable manipulation too.

Finally, we show that the Vickrey mechanism is Pareto efficient. Pick a type profile $\mathbf{u}$ and let $z_i \equiv (f_i(\mathbf{u}), p_i(\mathbf{u}))$ for all $i \in N$. Let $f_j(\mathbf{u}) = 1$ and $WP(u_k, 0) = \max_{i\neq j} WP(u_i, 0)$. Suppose there is another outcome z' that Pareto dominates z in the sense we defined. Denote $z'_i \equiv (a'_i, t'_i)$ for all i. Since sum of payments in z is $WP(u_k, 0)$. Hence,

$$\sum_{i\in N} t'_i \geq WP(u_k, 0) = \sum_{i\in N} t_i. \tag{6.1}$$

Suppose a' is such that agent $a'_{j'} = 1$. The case where $j = j'$ is easy, and left to the reader. Suppose $j \neq j'$. Pick any $i \notin \{j, j'\}$. Since $a'_i = 0 = f_i(\mathbf{u})$ and $u_i(z'_i) \geq u_i(z_i) = 0$, we get that

$$t'_i \leq 0. \tag{6.2}$$

Agent j also satisfies $u_j(0, t'_j) \geq u_j(1, WP(u_k, 0)) \geq u_j(1, WP(u_j, 0)) = u_j(0, 0)$. Hence, we have

$$t'_j \leq 0. \tag{6.3}$$

Finally, agent j' has $u_{j'}(1, t'_{j'}) \geq u_{j'}(0, 0) = u_{j'}(1, WP(u_{j'}, 0)) \geq u_{j'}(1, WP(u_k, 0))$. Hence, we have

$$t'_{j'} \leq WP(u_k, 0). \tag{6.4}$$

Adding all the Inequalities (6.2), (6.3), and (6.4), we get $\sum_i t'_i \leq WP(u_k, 0) = \sum_i t_i$. Using Inequality (6.1), we get $\sum_i t'_i = \sum_i t_i$. Hence, each of the Inequalities (6.2), (6.3), (6.4), and (6.1) are all equalities, contradicting the fact that z' Pareto dominates z. ■

The following theorem establishes some uniqueness of the Vickrey mechanism.

THEOREM **26** *Suppose $\mathcal{U} \subseteq \mathcal{U}^c$ satisfies the fact that $\mathcal{U}^q \subseteq \mathcal{U}$. The Vickrey mechanism is the unique mechanism which is DSIC, Pareto efficient, IR, and satisfies no-subsidy in the type space $\mathcal{U}$.*

Proof: We break the proof into steps.

STEP 1. Consider a profile of types $\mathbf{u}$ and an agent i. We show that if $f_i(\mathbf{u}) = 0$, then $p_i(\mathbf{u}) = 0$. By no-subsidy $p_i(\mathbf{u}) \geq 0$. Suppose $p_i(\mathbf{u}) > 0$. Then, $u_i(0, p_i(\mathbf{u})) < u_i(0,0) = 0$, a contradiction to IR. Hence, $p_i(\mathbf{u}) = 0$. Further, if $f_i(\mathbf{u}) = 1$, then $p_i(\mathbf{u}) \leq WP(u_i, 0)$. Suppose this is not true. Then, $u_i(1, p_i(\mathbf{u})) < u_i(1, WP(u_i, 0)) = u_i(0,0) = 0$, a contradiction to IR.

STEP 2. If agent i wins the object at a profile $\mathbf{u}$ and he changes his type to u_i' such that $WP(u_i', 0) > WP(u_i, 0)$, then $f_i(u_i', u_{-i}) = 1$. Assume for contradiction, $f_i(u_i', u_{-i}) = 0$. Then his utility is zero at u_i' (by Step 1). If he manipulates and reports u_i, then he gets the object and pays $p_i(u_i, u_{-i})$. By Step 1, $p_i(u_i, u_{-i}) \leq WP(u_i, 0) < WP(u_i', 0)$. Hence, $u_i'(1, p_i(u_i, u_{-i})) > u_i'(1, WP(u_i', 0)) = u_i'(0,0) = 0$. Hence, agent i can manipulate from u_i to u_i', a contradiction.

STEP 3. Suppose (u_i, u_{-i}) and (u_i', u_{-i}) are two type profiles such that $f_i(u_i, u_{-i}) = f_i(u_i', u_{-i}) = 1$. Then $p_i(u_i, u_{-i}) = p_i(u_i', u_{-i})$. This follows from DSIC - if $p_i(u_i, u_{-i}) < p_i(u_i', u_{-i})$, then we have $u_i'(1, p_i(u_i', u_{-i}) < u_i'(1, p_i(u_i, u_{-i})$, leading to a manipulation of agent i from u_i' to u_i, and a similar argument works if $p_i(u_i, u_{-i}) > p_i(u_i', u_{-i})$.

STEP 4. We show that if agent i is assigned the object at a profile $\mathbf{u}$, then $WP(u_i, 0) \geq WP(u_j, 0)$ for all $j \in N$. Assume for contradiction $WP(u_i, 0) < WP(u_j, 0) = \max_{k \in N} WP(u_k, 0)$ for some $j \in N$. Pick $\epsilon > 0$ but sufficiently close to zero. Since the type space contains $\mathcal{U}^q$, we can find a type $u_i' \in \mathcal{U}^q$ for agent i such that $WP(u_i', x) = WP(u_i, 0) + \epsilon$ for all x. By Step 2, $f_i(u_i', u_{-i}) = 1$. Denote $v_i' = WP(u_i', x)$ for all x. By Step 1, payment of agent i is less than or equal to v_i' and payment all other agents is zero. Now consider an outcome, where agent j wins the object and pays v_i' and agent i is paid an amount equal to $v_i' - p_i(u_i', u_{-i})$. The allocation of all other agents remain unchanged. Note all the agents except agent j are indifferent between the two outcomes. Agent j was getting 0 net utility earlier and now gets $u_j(1, v_i') > u_j(1, WP(u_j, 0)) = u_j(0,0) = 0$. Hence, agent j strictly improves. Finally, the sum of transfers in the new outcome is $p_i(u_i', u_{-i})$, which is the same as the old outcome. Hence, the outcome of the mechanism is not Pareto efficient at (u_i', u_{-i}).

This is a contradiction.

Also, at every type profile the object is assigned to some agent. This follows from Pareto efficiency. To see this, note that not assigning the object to anyone gives everyone zero utility. By assigning some agent i the object without any transfer gives him positive utility.

These two arguments imply that every type profile one of the agents in $\arg\max_{i \in N} WP(u_i, 0)$ is given the object.

STEP 5. Finally, we show that if at any type profile $\mathbf{u}$, we have $f_i(\mathbf{u}) = 1$, then $p_i(\mathbf{u}) = \max_{j \neq i} WP(u_j, 0)$. Let $x = \max_{j \neq i} WP(u_j, 0)$. We consider two cases.

CASE 1. Suppose $p_i(\mathbf{u}) > x$. Then, consider a type u_i' of agent i such that $p_i(\mathbf{u}) > WP(u_i', 0) > x$. Since $WP(u_i', 0) > x$, by Step 4, $f_i(u_i', u_{-i}) = 1$. By Step 3, $p_i(u_i, u_{-i}) = p_i(u_i', u_{-i})$. But then $WP(u_i', 0) < p_i(\mathbf{u}) = p_i(u_i', u_{-i})$ is a contradiction due to Step 2.

CASE 2. Suppose $p_i(\mathbf{u}) < x$. Then choose $u_i' \in \mathcal{U}$ such that $WP(u_i', 0) \in (p_i(\mathbf{u}), x)$. By Step 4, $f_i(u_i', u_{-i}) = 0$. By Step 2, $p_i(u_i', u_{-i}) = 0$. But DSIC implies that $u_i'(0, 0) = u_i'(1, WP(u_i', 0)) \geq u_i'(1, p_i(\mathbf{u}))$. Hence, $WP(u_i', 0) \leq p_i(\mathbf{u})$, which contradicts our assumption that $WP(u_i', 0) > p_i(\mathbf{u})$.

This completes the proof. ■

The conditions used in Theorem 26 are necessary. For instance, giving the object to some fixed agent for free is DSIC, IR, and satisfies no-subsidy but violates Pareto efficiency. On the other hand, modifying the Vickrey mechanism by charging a fixed amount to all the agents violates IR but satisfies DSIC, Pareto efficiency, and no-subsidy. Finally, modifying the Vickrey mechanism by subsidizing a fixed amount for not getting the object violates no-subsidy but satisfies DSIC, Pareto efficiency, and IR.

The type space assumptions in Theorem 26 can be relaxed - for instance, the proof goes through if $\mathcal{U} = \mathcal{U}^{++}$. Also, it is possible to get characterizations by relaxing some of the axioms in Theorem 26.

A small literature has developed on extending Theorem 26 into some specific mechanism design problems beyond single object auction. However, in general mechanism design prob-

lems, there is a lot that needs to be done when agents have non-quasilinear preferences. For instance, we have not discussed the implications of allowing for randomization. Allowing for all kinds of preferences over lotteries leads to incompatibility of Pareto efficiency, strategy-proofness, and individual rationality (more on this later).

6.2 Interdependent valuations

The models we studied so far assumed that each agent knows his preferences exactly. This may not be the case in many settings. Consider a simple setting of allocating a single object. The private values model assumed that each agent i has a value v_i and if he pays an amount p_i, his payoff is $v_i - p_i$. The value for the object for agent i may depend on the *private information* of all the agents. We give two examples to illustrate.

1. Perfect resale. Consider a model where once the object is allocated to agent i can be resold to the agent who values it the most. Further, this resale is perfect in the sense that agent i can charge the highest valued agent an amount equal to his value. In other words, the *ex-post* value of agent i is not his own value but the value of the highest valued agent. If we denote the value of each agent j as s_j, then the value of the object for any agent i is

$$v_i(s_1, \ldots, s_n) = \max_{j \in N} s_j,$$

where N is the set of n agents. Notice that in this model, every agent i has the same ex-post value for the object - this is an example of a *common value* model.

2. Signals as estimates. Consider a model where every agent only gets an estimate of the value of the object. For instance, if an oil field is auctioned, every firm conducts its own research to find out the estimated worth of the oil in the field. The *ex-post* value of the oil field is a function of the signals (estimates) of all the firms. This can still be a subjective assessment of valuation of each agent. For instance, suppose there are two firms. Firm 1 may decide to take the average of two firms' signals as his ex-post value, whereas firm 2 may decide to put more weight on his own signal. The literature assumes (mostly) that these subjective assessments of firm's ex-post values are known to the designer.

A model of interdependent valuations assumes that each agent $i \in N$ has a private **signal** $s_i \in \mathbb{R}_+$. Let S_i denote the set of all possible signals of agent i. The signal is the type of the agent. The valuation of agent i is given by a function

$$v_i : S_1 \times \ldots \times S_n \to \mathbb{R}_+.$$

Two standard assumptions that we will make: (1) v_i is **strictly increasing** in s_i and (2) v_i is (weakly) **increasing** in s_j for all $j \neq i$.

If the valuation functions of all the agents are the same, then we have a common values model.

DEFINITION **32** *The valuation functions* $(v_1, \ldots, v_n)$ *are* **common values** *if for all* $i, j \in N$ *and for all signal profiles* $s \equiv (s_1, \ldots, s_n)$*, we have* $v_i(s) = v_j(s)$.

Even in common values, we can think of many valuation functions $v(s_1, \ldots, s_n) = \max_j s_j$ is the resale model. Also, $v(s_1, \ldots, s_n) = \frac{1}{n}(s_1 + \ldots + s_n)$ is an average valuation model with common values. Consider the following valuation function:

$$v_i(s_1, \ldots, s_n) := s_i + \frac{1}{n-1} \sum_{j \neq i} s_j.$$

This is not a common values model since agent i puts more emphasis on his own signal and takes the average of others signals to compute valuations.

6.2.1 Mechanisms and Ex-post Incentive Compatibility

We now define mechanisms and a new notion of incentive compatibility. In general, a mechanism can be defined using message spaces, but the revelation principle continues to hold for the kind of equilibria we consider. Hence, we focus our attention on direct revelation mechanisms. For us, a mechanism is $(f, p_1, \ldots, p_n)$, where the allocation rule $f : S_1 \times \ldots \times S_n \to [0,1]^n$ and $p_i : S_1 \times \ldots \times S_n \to \mathbb{R}$ for each $i \in N$. We denote by $f_i(s)$ as the allocation probability of agent i at signal profile s.

One can start defining a notion of dominant strategy incentive compatibility in this setting. This will be something like the following. For every agent i and for every s'_{-i},

$$f_i(s_i, s'_{-i}) v_i(s_i, s_{-i}) - p_i(s_i, s'_{-i}) \geq f_i(s'_i, s_{-i}) v_i(s_i, s_{-i}) - p_i(s'_i, s'_{-i}) \ \forall \ s_i, s'_i \in S_i, \ \forall \ s_{-i} \in S_{-i}.$$

This allows agents other than i to report any signal s'_{-i} but the valuation of agent i will be determined by the true signals of others. There are two difficulties with this notion. First, it is *very* strong - it must hold for every $s_i, s'_i, s_{-i}, s'_{-i}$. Second, there is a conceptual problem with this notion. Agent i may *not* be able to evaluate his payoff even after everyone has reported their signals since N_{-i} may not report their true signals.

This leads us to the other popular solution concept: Bayesian incentive compatibility. For this, assume a common prior G of distributions of signal profiles. The conditional distribution of agent i having signal s_i about others' signals is $G_{-i}(\cdot|s_i)$, which we assume to admit a density function $g_{-i}(\cdot|s_i)$. So, we will say that $(f, p_1, \ldots, p_n)$ is **Bayesian incentive compatible** if for all $i \in N$

$$\int_{S_{-i}} \Big[f_i(s_i, s_{-i})v_i(s_i, s_{-i}) - p_i(s_i, s_{-i})\Big] g_{-i}(s_{-i}|s_i)ds_{-i}$$
$$\geq \int_{S_{-i}} \Big[f_i(s'_i, s_{-i})v_i(s_i, s_{-i}) - p_i(s'_i, s_{-i})\Big] g_{-i}(s_{-i}|s_i)ds_{-i} \qquad \forall\ s_i, s'_i \in S_i.$$

While this is a perfectly reasonable solution concept, it suffers from the usual prior-heavy criticisms. So, we adopt a stronger prior-free solution concept.

DEFINITION **33** *A mechanism* $(f, p_1, \ldots, p_n)$ *is* **ex-post incentive compatible (EPIC)** *if for all* $i \in N$

$$f_i(s_i, s_{-i})v_i(s_i, s_{-i}) - p_i(s_i, s_{-i}) \geq f_i(s'_i, s_{-i})v_i(s_i, s_{-i}) - p_i(s'_i, s_{-i}) \qquad \forall\ s_i, s'_i \in S_i,\ \forall\ s_{-i} \in S_{-i}.$$

Ex-post incentive compatibility says that if everyone else reports their true signals, then agent i is better off (ex-post) reporting his true signal. It can be seen that if a mechanism is EPIC, then it is Bayesian incentive compatible for all possible priors. So, EPIC is stronger than Bayesian incentive compatibility. Also, EPIC is much weaker than dominant strategy incentive compatibility since EPIC checks for *unilateral* deviations.

6.2.2 Efficiency: Impossibility and Possibility

Our main concern here is with respect to efficiency (in an ex-post sense). We say a mechanism $(f, p_1, \ldots, p_n)$ is **efficient** if at every profile $s \equiv (s_1, \ldots, s_n)$, we have $\sum_{i \in N} f_i(s) = 1$ and

$$\Big[f_i(s) > 0\Big] \Rightarrow \Big[v_i(s) \geq v_j(s)\Big].$$

Notice that this does not necessarily mean that an efficient mechanism allocates the object to an agent who has the highest *signal*. We give a simple example to illustrate that efficiency and EPIC may be incompatible.

EXAMPLE. Suppose $N = \{1, 2\}$, $S_1 = S_2 = [0, 2]$. Finally, the valuation functions look as follows:

$$v_1(s_1, s_2) = s_1, \quad v_2(s_1, s_2) = s_2 + [s_1]^2.$$

Suppose there is an efficient mechanism (f, p_1, p_2) which is EPIC. Fix the signal of agent 2 at $s_2 = 0.1$. Choose $s_1 = 0.5, s_1' > 1$. By efficiency, note that object is always allocated and

$$f_1(s_1, s_2) = 1, f_1(s_1', s_2) = 0.$$

Now, EPIC constraints of agent 1 give:

$$\begin{aligned} s_1 - p_1(s_1, s_2) &\geq 0 - p_1(s_1', s_2) \\ 0 - p_1(s_1', s_2) &\geq s_1' - p_1(s_1, s_2). \end{aligned}$$

Adding them gives us $s_1 \geq s_1'$, which is a contradiction.

The problem with this example is that agent 1's signal influences valuation of agent 2 "more" on his own. This leads to a contradiction with the standard monotonicity - increasing signals makes agent 1 lose his object. We present a sufficient condition on valuations which ensures that efficiency and EPIC are compatible.

DEFINITION **34** *Collection of valuation functions* $(v_1, \ldots, v_n)$ *satisfy* **single crossing** *if for every* $i, j \in N$*, every* $s_{-i} \in S_{-i}$*, and every* $s_i > s_i'$*, we have*

$$v_i(s_i, s_{-i}) - v_i(s_i', s_{-i}) > v_j(s_i, s_{-i}) - v_j(s_i', s_{-i}).$$

Single crossing is a strict *supermodularity* property. It imposes some structure on the problem. This will become apparent when we show that a generalization of the Vickrey auction is EPIC with single crossing. For our purposes, we will assume that in an efficient allocation rule, ties in valuations of agents are broken deterministically - so, whenever there is a tie for the highest valuation, one of the highest valuation agent is given the object with probability one.

DEFINITION **35** *A mechanism* $(f^*, p_1^*, \ldots, p_n^*)$ *is the* **generalized Vickrey auction (GVA)** *if* f^* *is an efficient allocation rule and at every profile of signals* $s \equiv (s_1, \ldots, s_n)$*, for all* $i \in N$*,*

$$p_i^*(s) = \begin{cases} 0 & \text{if } f_i^*(s) = 0 \\ v_i(\kappa_i(s_{-i}), s_{-i}) & \text{if } f_i^*(s) = 1 \end{cases}$$

where

$$\kappa_i(s_{-i}) := \inf\{s_i' \in S_i : f_i^*(s_i', s_{-i}) = 1\}.$$

It is not clear that $\kappa_i(s_{-i})$ is a real number which lies in S_i in the above definition. A condition that ensures $\kappa_i(s_{-i}) \in S_i$ is a real number is the fact that S_i is a compact interval of the type $[s_i^\ell, s_i^h]$.

THEOREM **27** *Suppose for every agent* i*, the space of signals* S_i *is a compact subset of* $\mathbb{R}$*. Further, suppose that the collection of valuations* $(v_1, \ldots, v_n)$ *satisfy the single crossing property. Then, the GVA mechanism is efficient, EPIC, and ex-post individually rational.*

Proof: Note that by compactness of S_i, for every $i \in N$ and for every s_{-i}, the signal $\kappa_i(s_{-i})$ as defined above lies in S_i.

The GVA is efficient by definition. For ex-post IR, consider a profile $s \equiv (s_1, \ldots, s_n)$ and agent i. If $f_i^*(s) = 0$, then his payoff is zero. If $f_i^*(s) = 1$, then his payoff is

$$v_i(s) - v_i(\kappa_i(s_{-i}), s_{-i}) \geq 0,$$

where the inequality follows from the fact that $s_i \geq \kappa_i(s_{-i})$.

Now, for EPIC, consider agent i and signal profile s_{-i} of other agents. Choose $s_i, s_i' \in S_i$. If $f_i^*(s_i, s_{-i}) = f_i^*(s_i', s_{-i})$, then by construction $p_i^*(s_i, s_{-i}) = p_i^*(s_i', s_{-i})$, and there is nothing to show. Hence, we consider the case where $f_i^*(s_i, s_{-i}) \neq f_i^*(s_i', s_{-i})$. If $f_i^*(s_i, s_{-i}) = 1$, then by ex-post IR, his payoff from reporting true signal is non-negative. So, $f_i^*(s_i', s_{-i}) = 0$ implies that agent i cannot be better off by reporting s_i'.

Suppose $f_i^*(s_i, s_{-i}) = 0$. Then, his payoff from reporting the true signal is zero. If he reports s_i' such that $f^*(s_i', s_{-i}) = 1$, then his payoff from reporting s_i' is

$$v_i(s_i, s_{-i}) - v_i(\kappa_i(s_{-i}), s_{-i}).$$

We will show that $s_i \leq \kappa_i(s_{-i})$, and this will imply the above expression is non-positive, implying that the deviation is not profitable.

Assume for contradiction $s_i - \kappa_i(s_{-i}) = \epsilon > 0$. By definition of $\kappa_i(s_{-i})$, there exists $\epsilon' \geq 0$ but arbitrarily close to zero (in particular, it can be chosen such that $\epsilon' < \epsilon$) such that $f_i^*(\kappa_i(s_{-i}) + \epsilon', s_{-i}) = 1$. Notice that by construction, $s_i > \kappa_i(s_{-i}) + \epsilon'$. Hence, by single crossing

$$v_i(s_i, s_{-i}) - v_j(s_i, s_{-i}) > v_i(\kappa_i(s_{-i}) + \epsilon', s_{-i}) - v_j(\kappa_i(s_{-i}) + \epsilon', s_{-i}) \qquad \forall\ j \neq i. \tag{6.5}$$

Since $f_i^*(\kappa_i(s_{-i}) + \epsilon', s_{-i}) = 1$, we get

$$v_i(\kappa_i(s_{-i}) + \epsilon', s_{-i}) - v_j(\kappa_i(s_{-i}) + \epsilon', s_{-i}) \geq 0 \qquad \forall\ j \neq i. \tag{6.6}$$

Combining Inequalities (6.5) and (6.6), we get

$$v_i(s_i, s_{-i}) > v_j(s_i, s_{-i}) \qquad \forall\ j \neq i.$$

But this implies that $f_i^*(s_i, s_{-i}) = 1$, which is a contradiction. ■

It is also worthwhile to think that some sort of uniqueness of GVA can also be established. For simplicity, we assume that S_i is a compact interval for each $i \in N$ and v_i is continuous. Suppose we have an efficient, EPIC, ex-post individually rational mechanism $(f, p_1, \ldots, p_n)$ that satisfies the property that losing agents pay zero - note that the GVA mechanism satisfies this property. This pins down allocation by efficiency and payment of losing agents. We only need to show that the payment of losing agent coincides with the payment in the GVA mechanism. To see this, fix agent i and (s_i, s_{-i}) such that $f_i(s_i, s_{-i}) = 1$. Consider two possible cases.

Case 1. Suppose $p_i(s_i, s_{-i}) > v_i(\kappa_i(s_{-i}), s_{-i})$. By definition of $\kappa_i(s_{-i})$, there is a type s_i' which is arbitrarily close to $\kappa_i(s_{-i})$ such that $f_i(s_i', s_{-i}) = 1$. By EPIC, $p_i(s_i, s_{-i}) = p_i(s_i', s_{-i})$. By continuity, s_i' can be picked such that

$$p_i(s_i, s_{-i}) = p_i(s_i', s_{-i}) > v_i(s_i', s_{-i}).$$

But this contradicts ex-post individual rationality since $v_i(s_i', s_{-i}) - p_i(s_i', s_{-i}) < 0$.

Case 2. Suppose $p_i(s_i, s_{-i}) < v_i(\kappa_i(s_{-i}), s_{-i})$. Then, pick $s_i' < \kappa_i(s_{-i})$ but arbitrarily close to $\kappa_i(s_{-i})$. By definition $f_i(s_i', s_{-i}) = 0$ and by continuity, $v_i(s_i', s_{-i}) > p_i(s_i, s_{-i})$. But incentive compatibility requires that $v_i(s_i', s_{-i}) - p_i(s_i, s_{-i}) \leq 0$, where this inequality follows from

the fact that losing agent pays zero and i is a losing agent at (s_i', s_{-i}). This is a contradiction.

One way to summarize these discussions is that the GVA mechanism occupies the same role that the Vickrey auction occupies in the private values model. One criticism of the GVA mechanism is that it relies on the fact that the designer has complete knowledge of the v_i functions.

Chapter 7

The Strategic Voting Model

7.1 The Unrestricted Domain Problem

We now discuss a general model of voting and examine the consequence of incentive compatibility in this model. The model is very general and introduces us to the rich literature on strategic voting models where monetary transfers are excluded. The literature has origins in two seminal papers of Gibbard (1973) and Satterthwaite (1975).

Let A be a finite set of alternatives with $|A| = m$. Let N be a finite set of individuals or agents or voters with $|N| = n$. Every agent has a preference over the set of alternatives. Let P_i denote the preference of agent i, which is assumed to be a (strict) ordering of elements of A.

Given a preference ordering P_i of agent i, we say aP_ib if and only if a is strictly preferred to b under P_i. Further, the top ranked element of this ordering is denoted by $P_i(1)$, the second ranked element by $P_i(2)$, and so on. Let $\mathcal{P}$ be the set of all strict preference orderings over A. A profile of preference orderings (or simply a preference profile) is denoted as $P \equiv (P_1, \ldots, P_n)$. So, $\mathcal{P}^n$ is the set of all preference profiles. A **social choice function (SCF)** is a mapping $f : \mathcal{P}^n \to A$. Note that this definition of a social choice function implicitly assumes that all possible profiles of linear orderings are permissible. This is known as the **unrestricted domain** assumption in the strategic voting (social choice) literature. Later, we will study some interesting settings where the domain of the social choice function is restricted.

Every agent knows his own preference ordering (his type) but does not know the pref-

erence ordering of other agents, and the mechanism designer (planner) does not know the preference orderings of agents. This is a very common situation in many voting scenarios: electing a candidate among a set of candidates, selecting a project among a finite set of projects for a company, selecting a public facility location among a finite set of possible locations, etc. Monetary transfers are precluded in these settings. The objective of this section is to find out which social choice functions are implementable in dominant strategies in such strategic voting scenarios.

We first describe several desirable properties of an SCF. The first property is an efficiency property. We say an alternative $a \in A$ is **Pareto dominated** at a preference profile P if there exists an alternative $b \in A$ such that bP_ia for all $i \in N$. Efficiency requires that no Pareto dominated alternative must be chosen.

DEFINITION **36** *A social choice function f is* **efficient** [1] *if for every profile of preferences P and every $a \in A$, if a is Pareto dominated at P then $f(P) \neq a$.*

As we will see many SCFs will be efficient in this model (contrast this to our definition of Pareto efficiency with transfers and quasilinearity). The next property requires to respect unanimity.

DEFINITION **37** *A social choice function f is* **unanimous** *if for every preference profile $P \equiv (P_1, \ldots, P_n)$ with $P_1(1) = P_2(1) = \ldots = P_n(1) = a$ we have $f(P) = a$.*

Note that this version of unanimity is a stronger version than requiring that if the *preference ordering* of all agents is the same, then the top ranked alternative must be chosen. This definition requires only the top to be the same, but other alternatives can be ranked differently by different agents.

Next, we define the strategic property of a social choice function.

DEFINITION **38** *A social choice function f is* **manipulable by agent i at profile $P \equiv (P_i, P_{-i})$ by preference ordering** P_i' *if $f(P_i', P_{-i})P_if(P)$. A social choice function f is* **strategy-proof** *if it is not manipulable by any agent i at any profile P by any preference ordering P_i'.*

This notion of strategy-proofness is the dominant strategy requirement since no manipulation is possible for every agent for every possible profile of other agents.

Finally, we define a technical property on the social choice function.

[1] Such a social choice function is also called Pareto optimal or Pareto efficient or ex-post efficient.

DEFINITION **39** *A social choice function f is* **onto** *if for every $a \in A$ there exists a profile of preferences $P \in \mathcal{P}^n$ such that $f(P) = a$.*

7.1.1 Examples of Social Choice Functions

We give some examples of social choice functions.

- CONSTANT SCF. A social choice function f^c is a constant SCF if there is some alternative $a \in A$ such that for every preference profile P, we have $f^c(P) = a$. This SCF is strategy-proof but not unanimous.

- DICTATORSHIP SCF. A social choice function f^d is a **dictatorship** if there exists an agent i, called the dictator, such that for every preference profile P, we have $f^d(P) = P_i(1)$. Dictatorship is strategy-proof and onto. Moreover, as we will see later, they are also efficient and unanimous.

- PLURALITY SCF (WITH FIXED TIE-BREAKING). Plurality is a popular way of electing an alternative. Here, we present a version that takes care of tie-breaking carefully. For every preference profile P and every alternative $a \in A$, define the score of a in P as $s(a, P) = |\{i \in N : P_i(1) = a\}|$. Define $\tau(P) = \{a \in A : s(a, P) \geq s(b, P) \; \forall \; b \in A\}$ for every preference profile P, and note that $\tau(P)$ is non-empty. Let $\succ^T$ be a linear ordering over alternatives A that we will use to break ties. . A social choice function f^p is called a plurality SCF with tie-breaking according to $\succ^T$ if for every preference profile P, $f^p(P) = a$, where $a \in \tau(P)$ and $a \succ^T b$ for all $b \in \tau(P) \setminus \{a\}$.

  Though the plurality SCF is onto, it is not strategy-proof. To see this, consider an example with three agents $\{1, 2, 3\}$ and three alternatives $\{a, b, c\}$. Let $\succ^T$ be defined as: $a \succ^T b \succ^T c$. Consider two preference profiles shown in Table 7.1. We note first that $f(P) = a$ and $f(P') = b$. Since bP_3a, agent 3 can manipulate at P by P'_3.

- BORDA SCF (WITH FIXED TIE-BREAKING). The Borda SCF is a generalization of the Plurality voting SCF. The tie-breaking in this SCF is defined similar to Plurality SCF. Let $\succ^T$ be a linear ordering over alternatives A that we will use to break ties. Fix a preference profile P. For every alternative $a \in A$, the *rank* of a in P_i for agent i is given by $r(a, P_i) = k$, where $P_i(k) = a$. From this, the score of alternative a in preference

P_1	P_2	P_3	$P_1' = P_1$	$P_2' = P_2$	P_3'
a	b	c	a	b	b
b	c	b	b	c	a
c	a	a	c	a	c

Table 7.1: Plurality SCF is manipulable.

profile P is computed as $s(a, P) = \sum_{i \in N}[|A| - r(a, P_i)]$. Define for every preference profile P, $\tau(P) = \{a \in A : s(a, P) \geq s(b, P)\ \forall b \in A\}$. A social choice function f^b is called a Borda SCF with tie-breaking according to $\succ^T$ if for every preference profile P, $f^b(P) = a$ where $a \in \tau(P)$ and $a \succ^T b$ for all $b \in \tau(P) \setminus \{a\}$.

Like the Plurality SCF, the Borda SCF is onto but manipulable. To see this, consider an example with three agents $\{1, 2, 3\}$ and three alternatives $\{a, b, c\}$. Let $\succ^T$ be defined as: $c \succ^T b \succ^T a$. Consider two preference profiles shown in Table 7.2. We note first that $f(P) = b$ and $f(P') = c$. Since cP_1b, agent 1 can manipulate at P by P_1'.

P_1	P_2	P_3	P_1'	$P_2' = P_2$	$P_3' = P_3$
a	b	b	c	b	b
c	c	c	a	c	c
b	a	a	b	a	a

Table 7.2: Borda SCF is manipulable.

7.1.2 Implications of Properties

We now examine the implications of these properties. We start out with a simple characterization of strategy-proof social choice functions using the following monotonicity property. Such monotonicity properties are heart of every incentive problem - though the nature of monotonicity may differ from problem to problem.

For any alternative $a \in A$, let $B(a, P_i)$ be the set of alternatives below a in preference ordering P_i. Formally, $B(a, P_i) := \{b \in A : aP_ib\}$.

DEFINITION **40** *A social choice function f is* **monotone** *if for any two profiles P and P' with $B(f(P), P_i) \subseteq B(f(P), P_i')$ for all $i \in N$, we have $f(P) = f(P')$.*

Note that in the definition of monotonicity when we go from a preference profile P to P' with $f(P) = a$, whatever was below a in P for every agent continues to be below it in P' also, but other relations may change. For example, the following is a valid P and P' in the definition of monotonicity with $f(P) = a$ (see Table 7.3).

P_1	P_2	P_3	P'_1	P'_2	P'_3
a	b	c	a	a	a
b	a	a	b	c	c
c	c	b	c	b	b

Table 7.3: Two valid profiles for monotonicity

THEOREM **28** *A social choice function $f : \mathcal{P}^n \to A$ is strategy-proof if and only if it is monotone.*

Proof: Consider social choice function $f : \mathcal{P}^n \to A$ which is strategy-proof. Consider two preference profiles P and P' such that $f(P) = a$ and $B(a, P_i) \subseteq B(a, P'_i)$ for all $i \in N$. We define $(n-1)$ new preference profiles. Define preference profile P^1 as follows: $P^1_1 = P'_1$ and $P^1_i = P_i$ for all $i > 1$. Define preference profile P^k for $k \in \{1, \ldots, n-1\}$ as $P^k_i = P'_i$ if $i \leq k$ and $P^k_i = P_i$ if $i > k$. Set $P^0 = P$ and $P^n = P'$. Note that if we pick two preference profiles P^k and P^{k+1} for any $k \in \{0, \ldots, n-1\}$, then preference of all agents other than agent $(k+1)$ are same in P^k and P^{k+1}, and preference of agent $(k+1)$ is changing from P_{k+1} in P^k to P'_{k+1} in P^{k+1}.

We will show that $f(P^k) = a$ for all $k \in \{0, \ldots, n\}$. We know that $f(P^0) = f(P) = a$, and consider $k = 1$. Assume for contradiction $f(P^1) = b \neq a$. If $b \; P_1 \; a$, then agent 1 can manipulate at P^0 by $P'_1 = P^1_1$. If aP_1b, then aP'_1b, and agent 1 can manipulate at P^1 by $P_1 \equiv P^0_1$. This is a contradiction since f is strategy-proof.

We can repeat this argument by assuming that $f(P^q) = a$ for all $q \leq k < n$, and showing that $f(P^{k+1}) = a$. Assume for contradiction $f(P^{k+1}) = b \neq a$. If $bP_{k+1}a$, then agent $(k+1)$ can manipulate at P^k by $P^{k+1}_k \equiv P'_{k+1}$. If $aP_{k+1}b$ then $aP'_{k+1}b$. This means agent $(k+1)$ can manipulate at P^{k+1} by $P_k \equiv P^k_{k+1}$. This is a contradiction since f is strategy-proof.

Hence, by induction, $f(P^n) = f(P') = a$, and f is monotone.

Suppose $f : \mathcal{P}^n \to A$ is a monotone social choice function. Assume for contradiction that f is not strategy-proof. In particular, agent i can manipulate at preference profile P

by a preference ordering P_i'. Let $P' \equiv (P_i', P_{-i})$. Suppose $f(P) = a$ and $f(P') = b$, and by assumption bP_ia. Consider a preference profile $P'' \equiv (P_i'', P_{-i})$, where P_i'' is any preference ordering satisfying $P_i''(1) = b$ and $P_i''(2) = a$. By monotonicity, $f(P'') = f(P') = b$ and $f(P'') = f(P) = a$, which is a contradiction. ■

Theorem 28 is a strong result. The necessity of monotonicity is true in any domain - even if a subset of all possible preference profiles are permissible. Even for sufficiency, we just need a domain where we are able to rank any pair of alternatives first and second.

We now explore the implications of other properties.

LEMMA **20** *If an SCF f is monotone and onto then it is efficient.*

Proof: Consider $a, b \in A$ and a preference profile P such that aP_ib for all $i \in N$. Assume for contradiction $f(P) = b$. Since f is onto, there exists a preference profile P' such that $f(P') = a$. We construct another preference profile $P'' \equiv (P_1'', \ldots, P_n'')$ as follows. For all $i \in N$, let $P_i''(1) = a$, $P_i''(2) = b$, and $P_i''(j)$ for $j > 2$ can be set to anything. Since f is monotone, $f(P'') = f(P) = b$, and also, $f(P'') = f(P') = a$. This is a contradiction. ■

LEMMA **21** *If an SCF f is efficient then it is unanimous.*

Proof: Consider a preference profile $P \equiv (P_1, \ldots, P_n)$ with $P_1(1) = P_2(1) = \ldots = P_n(1) = a$. Consider any $b \neq a$. By definition, aP_ib for all $i \in N$. By efficiency, $f(P) \neq b$. Hence, $f(P) = a$. ■

LEMMA **22** *If a social choice function is unanimous then it is onto.*

Proof: Take any alternative $a \in A$ and a social choice function f. Consider a profile P such that $P_i(1) = a$ for all $i \in N$. Then $f(P) = a$ by unanimity. So, f is onto. ■

We can summarize these results in the following proposition.

PROPOSITION **4** *Suppose $f : \mathcal{P}^n \to A$ is a strategy-proof social choice function. Then, f is onto if and only if it is efficient if and only if it is unanimous.*

Proof: Suppose f is strategy-proof. By Theorem 28, it is monotone. Then, Lemmas 20, 21, and 22 establish the result. ■

7.1.3 The Gibbard-Satterthwaite Theorem

THEOREM **29 (Gibbard-Satterthwaite Theorem, Gibbard (1973); Satterthwaite (1975))** *Suppose* $|A| \geq 3$. *A social choice function* $f : \mathcal{P}^n \to A$ *is onto and strategy-proof if and only if it is a dictatorship.*

Before we discuss the proof, we make the following observations about the Gibbard-Satterthwaite (GS) theorem.

1. $|A| = 2$. The GS theorem fails when there are only two alternatives. An example of a non-dictatorial social choice function which is onto and strategy-proof is the plurality social choice function with a fixed tie-breaking. (The proof of this fact is an exercise.)

2. **Unrestricted domain.** The assumption that the type space of each agent consists of all possible strict orderings over A is critical in the GS theorem. The intuition about why the set of strategy-proof social choice functions become larger as we restrict the type space is very simple. In a smaller type space, agents have less opportunity to manipulate a social choice functions and, hence, it is easier for incentive constraints to hold. It is because of this reason, the GS theorem may fail in various *restricted domains.* We give a specific example in the next section.

3. **Indifference**. Suppose every agent has a preference ordering which is not necessarily anti-symmetric, i.e., there are ties between alternatives. Let $\mathcal{R}$ be the set of all preference orderings. Note that $\mathcal{P} \subsetneq \mathcal{R}$. Now, consider a domain $\mathcal{D} \subseteq \mathcal{R}$ such that $\mathcal{P} \subseteq \mathcal{D}$. Call such a domain **admissible**. A social choice function $f : \mathcal{D}^n \to A$ is **admissible** if $\mathcal{D}$ is admissible. In other words, if the domain of preference orderings include *all possible* linear orderings, then such a domain is admissible. The GS theorem is valid in admissible domains, i.e., if $|A| \geq 3$ and $f : \mathcal{D}^n \to A$ is admissible, onto, and strategy-proof, then it is a dictatorship. The proof follows from the observation that the proof of GS-Theorem only requires existence of certain strict preference orderings. So, as long as such preference orderings exist, the GS-Theorem proof goes through.

   However, dictatorship may not be strategy-proof when indifference is permitted. For instance, consider the dictatorship SCF f as follows. It always selects an alternative in agent 1's top - so, agent 1 is the dictator. However, if there are more than one

alternative in agent 1's top, then the following tie-breaking rule is followed. Let $\succ$ be a linear ordering over A. Consider a profile P such that $P_1(1)$ has more than one element [2]. Then, consider $P_2(1)$. If $P_2(1) \cap P_1(1)$ is non-empty, choose an element from $P_2(1) \cap P_1(1)$ using $\succ$, i.e., breaking ties according to $\succ$. Else, choose an alternative from $P_1(1)$ using $\succ$. As an example, suppose agent 1's top consists of $\{a, b, c\}$. Agent 2's top consists of b and c. The tie-breaking is done using $\succ$, and it has $b \succ c$. So, the outcome at this profile must be b. If agent 2's top did not have an element in $\{a, b, c\}$ and $a \succ b \succ c$, then the outcome will be a.

Such dictatorship SCFs are manipulable. To see this, consider a setting with three alternatives $\{a, b, c\}$ and two agents. Suppose we use the dictatorship of the previous example with $a \succ b \succ c$. Consider a profile where agent 1's top consists of b and c. But agent 2's top has a followed by c, and then followed by b at the bottom. Then, according to the SCF, b will be the outcome. Note that b is the worst alternative for agent 2. He can improve it by reporting c as his unique top since the outcome will now change to c.

7.1.4 Proof of the Gibbard-Satterthwaite Theorem

There are number of proofs available in the literature, including the proofs of Gibbard (1973) and Satterthwaite (1975). We follow the proof of Sen (2001), which is based on an induction argument on the number of agents. We first analyze the case when $n = 2$.

LEMMA **23** *Suppose $|A| \geq 3$ and $N = \{1, 2\}$. Suppose f is an onto and strategy-proof social choice function. Then for every preference profile P, $f(P) \in \{P_1(1), P_2(1)\}$.*

Proof: Fix a preference profile $P = (P_1, P_2)$. If $P_1(1) = P_2(1)$, the claim is due to unanimity

P_1	P_2	P_1	P_2'	P_1'	P_2'	P_1'	P_2
a	b	a	b	a	b	a	b
$\cdot$	$\cdot$	$\cdot$	a	b	a	b	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$

Table 7.4: Preference profiles required in proof of Lemma 23.

[2] Since we allow for indifference, $P_i(1)$ for any agent i is a subset of alternatives.

(Proposition 4). Else, let $P_1(1) = a$ and $P_2(1) = b$, where $a \neq b$. Assume for contradiction $f(P) = c \notin \{a, b\}$. We will use the preference profiles shown in Table 7.4.

Consider a preference ordering P_2' for agent 2 where $P_2'(1) = b$, $P_2'(2) = a$, and the remaining ordering can be anything. By efficiency, $f(P_1, P_2') \in \{a, b\}$. Further $f(P_1, P_2') \neq b$ since agent 2 can then manipulate at P by P_1. So, $f(P_1, P_2') = a$.

Now, consider a preference ordering P_1' for agent 1 where $P_1'(1) = a$, $P_1'(2) = b$, and the remaining ordering can be anything. Using an analogous argument, we can show that $f(P_1', P_2) = b$. Now, consider the preference profile (P_1', P_2'). By monotonicity (implied by strategy-proofness - Theorem 28) , $f(P_1', P_2') = f(P_1, P_2') = a$ and $f(P_1', P_2') = f(P_1', P_2) = b$. This is a contradiction. ■

LEMMA **24** *Suppose* $|A| \geq 3$ *and* $N = \{1, 2\}$. *Suppose* f *is onto and strategy-proof social choice function. Consider a profile* P *such that* $P_1(1) = a \neq b = P_2(1)$. *Consider a preference profile* $P' = (P_1', P_2')$ *with* $P_1'(1) = c$ *and* $P_2'(1) = d$. *If* $f(P) = a$, *then* $f(P') = c$ *and if* $f(P) = b$ *then* $f(P') = d$.

Proof: We can assume that $c \neq d$, since the claim is true due to unanimity when $c = d$. Suppose $f(P) = a$. We need to show that $f(P') = c$ - an analogous proof works if $f(P) = b$ (in which case, we need to show $f(P') = d$). We do the proof for different possible cases.

CASE 1: $c = a$, $d = b$. This case establishes a *tops-only* property. From Lemma 23, $f(P') \in \{a, b\}$. Assume for contradiction $f(P') = b$ (i.e., agent 2's top is chosen). Consider a preference profile $\hat{P} \equiv (\hat{P}_1, \hat{P}_2)$ such that $\hat{P}_1(1) = a, \hat{P}_1(2) = b$ and $\hat{P}_2(1) = b, \hat{P}_2(2) = a$ (See Table 7.5). By monotonicity, $f(\hat{P}) = f(P') = f(P)$, which is a contradiction.

P_1	P_2	P_1'	P_2'	$\hat{P}_1$	$\hat{P}_2$
a	b	a	b	a	b
.	.	.	.	b	a
.	.	.	.	.	.

Table 7.5: Preference profiles required in Case 1.

CASE 2: $c \neq a$, $d = b$. Consider any profile $\hat{P} = (\hat{P}_1, \hat{P}_2)$, where $\hat{P}_1(1) = c \neq a$, $\hat{P}_1(2) = a$,

and $\hat{P}_2(1) = b$ (See Table 7.6). By Lemma 23, $f(\hat{P}) \in \{b, c\}$. Suppose $f(\hat{P}) = b$. Then,

P_1	P_2	P_1'	P_2'	$\hat{P}_1$	$\hat{P}_2$
a	b	$c \neq a$	$d = b$	c	b
.	.	.	.	a	.
.	.	.	.	.	.

Table 7.6: Preference profiles required in Case 2.

agent 1 can manipulate by reporting any preference ordering where his top is a, and this will lead to a as the outcome (Case 1). Hence, $f(\hat{P}) = c = \hat{P}_1(1)$. Using Case 1, $f(P') = c$.

Case 3: $c \notin \{a, b\}$, $d \neq b$ [3]. Consider a preference profile $\hat{P}$ such that $\hat{P}_1(1) = c, \hat{P}_2(1) = b$ (See Table 7.7). Assume for contradiction $f(P') = d$. Then, by applying Case 2 from P' to $\hat{P}$, we get $f(\hat{P}) = b$. But applying Case 2 from P to $\hat{P}$, we get $f(\hat{P}) = c$. This is a contradiction.

P_1	P_2	P_1'	P_2'	$\hat{P}_1$	$\hat{P}_2$
a	b	$c \notin \{a, b\}$	$d \neq b$	c	b
.	.	.	.	.	.
.	.	.	.	.	.

Table 7.7: Preference profiles required in Case 3.

Case 4: $c = a$, $d \neq b$. By Lemma 23, $f(P') \in \{a, d\}$. Assume for contradiction $f(P') = d$. Consider a preference ordering $\hat{P}_2$ such that $\hat{P}_2(1) = b$ (See Table 7.8). By Case 2, from P' to $\hat{P}$, we get $f(\hat{P}) = b$. But applying Case 1 from P to $\hat{P}$, we get $f(\hat{P}) = a$, a contradiction.

Case 5: $c = b$, $d \neq a$. By Lemma 23, $f(P') \in \{b, d\}$. Assume for contradiction $f(P') = d$. Consider a preference ordering $\hat{P}_1$ such that $\hat{P}_1(1) = a$ and $\hat{P}_2$ such that $\hat{P}_2(1) = d$ (See Table 7.9). Applying Case 4 from P to $\hat{P}$, we get $f(\hat{P}) = a$. But applying Case 4 from P' to $\hat{P}$, we get $f(\hat{P}) = d$. This is a contradiction.

Case 6: $c = b$, $d = a$. Since there are at least three alternatives, consider $x \notin \{a, b\}$. Consider a preference ordering $\hat{P}_1$ such that $\hat{P}_1(1) = b$ and $\hat{P}_1(2) = x$ (See Table 7.10).

[3] This case actually covers two cases: one where $d = a$ and the other where $d \notin \{a, b\}$.

P_1	P_2	P'_1	P'_2	P'_1	$\hat{P}_2$
a	b	$c = a$	$d \neq b$	a	b
.	.	.	.	.	.
.	.	.	.	.	.

Table 7.8: Preference profiles required in Case 4.

P_1	P_2	P'_1	P'_2	$\hat{P}_1$	$\hat{P}_2$
a	b	$c = b$	$d \neq a$	a	d
.	.	.	.	.	.
.	.	.	.	.	.

Table 7.9: Preference profiles required in Case 5.

P_1	P_2	P'_1	P'_2	$\hat{P}_1$	P'_2	$\hat{P'}_1$	P'_2
a	b	$c = b$	$d = a$	b	a	x	a
.	.	.	.	x	.	.	.
.	.	.	.	.	.	.	.

Table 7.10: Preference profiles required in Case 6.

By Lemma 23, $f(\hat{P}_1, P'_2) \in \{b, a\}$. Assume for contradiction $f(\hat{P}_1, P'_2) = a$. Consider a preference ordering $\hat{P'}_1$ such that $\hat{P'}_1(1) = x$ (See Table 7.10). By Case 3, $f(\hat{P'}_1, P'_2) = x$. But $x\hat{P}_1 a$. Hence, agent 1 can manipulate $(\hat{P}_1, P'_2)$ by $\hat{P'}_1$. This is a contradiction. Hence, $f(\hat{P}_1, P'_2) = b$. By Case 1, $f(P') = b$. ■

PROPOSITION **5** *Suppose* $|A| \geq 3$ *and* $n = 2$. *A social choice function is onto and strategy-proof if and only if it is dictatorship.*

Proof: This follows directly from Lemmas 23 and 24 and unanimity (implied by onto and strategy-proofness - Proposition 4). ■

Once we have the theorem for $n = 2$ case, we can apply induction on the number of agents. In particular, we prove the following proposition.

PROPOSITION **6** *Let* $n \geq 3$. *Consider the following statements.*

(a) For all positive integer $k < n$, we have if $f : \mathcal{P}^k \to A$ is onto and strategy-proof, then f is dictatorial.

(b) If $f : \mathcal{P}^n \to A$ is onto and strategy-proof, then f is dictatorial.

Statement (a) implies statement (b).

Proof: Suppose statement (a) holds. Let $f : \mathcal{P}^n \to A$ be an onto and strategy-proof social choice function. We construct another social choice function $g : \mathcal{P}^{n-1} \to A$ from f by merging agents 1 and 2 as one agent. In particular, $g(P_1, P_3, P_4, \ldots, P_n) = f(P_1, P_1, P_3, P_4, \ldots, P_n)$ for all preference profiles $(P_1, P_3, P_4, \ldots, P_n)$. So agents 1 and 2 are "coalesced" in social choice function g, and will be referred to as agent 1 in SCF g.

We do the proof in two steps. In the first step, we show that g is onto and strategy-proof. We complete the proof in the second step, i.e., show that f is dictatorship.

STEP 1: It is clear that agents 3 through n cannot manipulate in g (if they can manipulate in g, they can also manipulate in f, which is a contradiction). Consider an arbitrary preference profile of $n-1$ agents $(P_1, P_3, P_4, \ldots, P_n)$. Suppose

$$f(P_1, P_1, P_3, P_4, \ldots, P_n) = g(P_1, P_3, P_4, \ldots, P_n) = a.$$

Consider any arbitrary preference ordering $\bar{P}_1$ of agent 1. Let

$$f(P_1, \bar{P}_1, P_3, P_4, \ldots, P_n) = b.$$

Let

$$f(\bar{P}_1, \bar{P}_1, P_3, P_4, \ldots, P_n) = g(\bar{P}_1, P_3, P_4, \ldots, P_n) = c.$$

If $a = c$, then agent 1 cannot manipulate g at $(P_1, P_3, P_4, \ldots, P_n)$ by $\bar{P}_1$. So, assume $a \neq c$. Suppose $a = b \neq c$. Then, agent 1 cannot manipulate f at $(P_1, \bar{P}_1, P_3, P_4, \ldots, P_n)$ by $\bar{P}_1$. So, $a = bP_1c$. Hence, agent 1 cannot manipulate g at $(P_1, P_3, P_4, \ldots, P_n)$ by $\bar{P}_1$. A similar logic works for the case when $b = c$.

Now, assume that a, b, and c are distinct. Since f is strategy-proof, agent 2 cannot manipulate f at $(P_1, P_1, P_3, P_4, \ldots, P_n)$ by $\bar{P}_1$. So, aP_1b. Similarly, agent 1 cannot manipulate

f at $(P_1, \bar{P}_1 P_3, P_4, \ldots, P_n)$ by $\bar{P}_1$. So, bP_1c. By transitivity, aP_1c. Hence, agent 1 cannot manipulate g at $(P_1, P_3, P_4, \ldots, P_n)$ by $\bar{P}_1$. This shows that g is strategy-proof.

It is straightforward to show that if f is onto, then g is onto (follows from unanimity of f).

STEP 2: By our induction hypothesis, g is dictatorship. Suppose j is the dictator. There are two cases to consider.

CASE A: Suppose $j \in \{3, 4, \ldots, n\}$ is the dictator in g. We claim that j is also the dictator in f. Assume for contradiction that there is a preference profile $P \equiv (P_1, P_2, \ldots, P_n)$ such that

$$f(P) = b \text{ and } P_j(1) = a \neq b.$$

Since g is dictatorship, we get

$$f(P_1, P_1, P_3, P_4, \ldots, P_n) = g(P_1, P_3, P_4, \ldots, P_n) = a,$$
$$f(P_2, P_2, P_3, P_4, \ldots, P_n) = g(P_2, P_3, P_4, \ldots, P_n) = a.$$

We get bP_1a, since f is strategy-proof, and agent 1 cannot manipulate f at $(P_1, P_2, P_3, P_4, \ldots, P_n)$ by P_2. Similarly, agent 2 cannot manipulate at $(P_1, P_1, P_3, P_4, \ldots, P_n)$ by P_2. So, aP_1b. This is a contradiction.

CASE B: Suppose $j = 1$ is the dictator in g. In this case, we construct a family of 2-agent social choice function h as follows. Fix a profile P_{-12} of agents in $N \setminus \{1, 2\}$, and define for every preference profile (P_1, P_2), we define

$$h^{P_{-12}}(P_1, P_2) = f(P_1, P_2, \ldots, P_n).$$

Note that the social function Since agent 1 is the dictator in g, $h^{P_{-12}}$ is onto. Moreover, $h^{P_{-12}}$ is strategy-proof: if any of the agents can manipulate in $h^{P_{-12}}$, they can also manipulate in f. By our induction hypothesis, $h^{P_{-12}}$ is dictatorship. But $h^{P_{-12}}$ was defined for every $n-2$ agent profile $P_{-12} \equiv (P_3, P_4, \ldots, P_n)$. We show that the dictator does not change across two $n-2$ agent profiles.

Assume for contradiction that agent 1 is the dictator for profile $(P_3, P_4, \ldots, P_n)$ but agent 2 is the dictator for profile $(\bar{P}_3, \bar{P}_4, \ldots, \bar{P}_n)$. Now, progressively change the preference profile

$(P_3, P_4, \ldots, P_n)$ to $(\bar{P}_3, \bar{P}_4, \ldots, \bar{P}_n)$, where in each step, we change the preference of one agent j from P_j to $\bar{P}_j$. Then, there must exist a profile $(\bar{P}_3, \bar{P}_4, \bar{P}_{j-1}, P_j, P_{j+1}, \ldots, P_n)$ where agent 1 dictates and another profile $(\bar{P}_3, \bar{P}_4, \bar{P}_{j-1}, \bar{P}_j, P_{j+1}, \ldots, P_n)$ where agent 2 dictates with $3 \leq j \leq n$. Consider $a, b \in A$ such that aP_jb. Pick P_1 and P_2 such that $P_1(1) = b$ and $P_2(1) = a$ with $a \neq b$. By definition,

$$f(P_1, P_2, \bar{P}_3, \bar{P}_4, \bar{P}_{j-1}, P_j, P_{j+1}, \ldots, P_n) = P_1(1) = b,$$
$$f(P_1, P_2, \bar{P}_3, \bar{P}_4, \bar{P}_{j-1}, \bar{P}_j, P_{j+1}, \ldots, P_n) = P_2(1) = a.$$

This means agent j can manipulate in SCF f at $(P_1, P_2, \bar{P}_3, \bar{P}_4, \bar{P}_{j-1}, P_j, P_{j+1}, \ldots, P_n)$ by $\bar{P}_j$. This is a contradiction since f is strategy-proof. This shows that f is also a dictatorship.

This completes the proof of the proposition. ■

The proof of the Gibbard-Satterthwaite theorem follows from Propositions 5 and 6, and from the fact that the proof is trivial for $n = 1$.

Note that the induction step must start at $n = 2$, and not $n = 1$, since the induction argument going from k to $k + 1$ works for $k \geq 2$ only.

7.2 Single Peaked Domain of Preferences

While the Gibbard-Satterthwaite theorem is a negative result, there are important assumptions that drive the result: (a) at least three alternatives (b) domain contains all strict orderings of alterantives. While we have seen that having two alterantives allows us many strategy-proof mechanisms, we will now see the consequences of *restricting* the domain.

Why does domain restriction help? Incentive compatibility is a collection of constraints with variables being the mechanism. When we *restrict* the domain, the set of constraints become smaller. As a result, the feasible set (of mechanisms) enlarge – in the limiting case, when the domain contains just one preference ordering, then the designer knows everything and every mechanism is trivially strategy-proof.

Does restricting the set of admissible preferences (domain) make sense? In many problems, the designer can rule out the possibility of certain preferences. As an example, consider an election with several candidates. Candidates are *ordered* on a line so that candidate on

left is the most leftist, and candidates become more and more right wing as we move to right. Now, it is natural to assume that every voter has an ideal political position. As one moves away from his ideal political position, either to left or to right, his preference decreases. In this example, a designer may rule out many preference orderings.

To be more precise, let $\{a, b, c\}$ be three candidates, with a to extreme left, b in the center, and c to extreme right. Now, suppose a voter's ideal position is b. Then, he likes b over a and b over c, but can have any preference over a and c. On the other hand, suppose a voter likes a the most. Then, the only possible ordering is a better than b better than c. Hence, when a is on top, c cannot be better than b. This means that certain preference orderings cannot be in the domain. This is the single-peaked domain of preferences.

We now formally define the single-peaked preferences. First discussions of single-peaked preferences in economics go back to Black (1948), and its strategic investigation is due to Moulin (1980). Let $N = \{1, \ldots, n\}$ be the set of agents. Let A be a set of alternatives. We will assume A to be finite but with some modifications, most of the results generalize when $A = [0, 1]$. Consider a strict ordering $\succ$ on A. In the example above, $\succ$ corresponds to the ordering of the ideology of candidates in the election.

DEFINITION **41** *A preference ordering P_i of agent i is* **single peaked** *with respect to $\succ$ if*

- *for all $b, c \in A$ with $b \succ c \succ P_i(1)$ we have cP_ib, and*
- *for all $b, c \in A$ with $P_i(1) \succ b \succ c$ we have bP_ic.*

So, preferences away from the top ranked alternative or peak decreases, but no restriction is put on comparing alternatives when one of them is on the left to the peak, but the other one is on the right of the peak. We show some preference relations in Figure 7.1, and color the single-peaked ones in blue.

Since we fix the order $\succ$ throughout this section, we will just say single-peaked preferences instead of single-peaked with respect to $\succ$ and drop the reference to $\succ$. We illustrate the idea with four alternatives $A = \{a, b, c, d\}$. Let us assume that $a \succ b \succ c \succ d$. With respect to $\succ$, we give the permissible single peaked preferences in Table 7.11. There are sixteen more preference orderings that are not permissible here. For example, $bP_idP_iaP_ic$ is not permissible since c, d are on the same side of peak, and in that case c is nearer to b than d is to b. So, cP_id, which is not the case here.

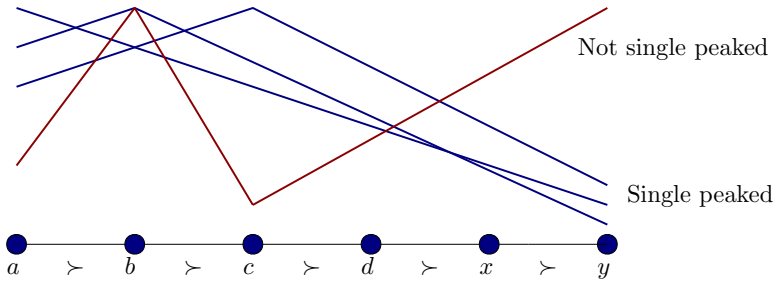

Figure 7.1: Examples of single-peaked preferences

a	b	b	b	c	c	c	d
b	a	c	c	d	b	b	c
c	c	d	a	b	a	d	b
d	d	a	d	a	d	a	a

Table 7.11: Single-peaked preferences

We now give some more examples of single-peaked preferences.

- An amount of public good (number of buses in the city) needs to be decided. Every agent has an optimal level of public good that needs to be consumed. The preferences decrease as the difference of the decided amount and optimal level increases.
- If we are locating a facility along a line, then agents can have single-peaked preferences. For every agent, there is an optimal location along a line where he wants the facility, and the preference decreases as the distance from the optimal location increases in one direction.
- Something as trivial as setting a time for a public seminar exhibit single-peaked preferences. Everyone has an ideal time slot, and as the difference from the ideal time slot increases, it is less preferred.

If $A = [0, 1]$, then $\succ$ corresponds to the natural ordering of $[0, 1]$. A preference P_i can be equivalently represented by a utility function u_i. Remember that a utility function $u_i : [0, 1] \to \mathbb{R}$ is **quasiconcave** if one of the three conditions hold:

1. u_i is increasing in $[0, 1]$
2. u_i is decreasing in $[0, 1]$

3. u_i is increasing till some point $x^* \in (0,1)$ and then decreases after x^*.

Hence, $\succ$ satisfying single-peakedness is *equivalent* to its utility representation being quasiconcave.

Let $\mathcal{S}$ be the set of all single-peaked preferences (with respect to $\succ$). A social choice function f is a mapping $f : \mathcal{S}^n \to A$. An SCF f is manipulable by i at (P_i, P_{-i}) if there exists another single-peaked preference $\hat{P}_i$ such that $f(\hat{P}_i, P_{-i}) P_i f(P_i, P_{-i})$. An SCF is strategy-proof if it is not manipulable by any agent at any preference profile.

7.2.1 Possibility Examples in Single-Peaked Domains

We start with an example to illustrate that many non-dictatorial social choice functions are strategy-proof in this setting. For any single-peaked preference ordering P_i, we let $P_i(1)$ to denote its peak. Now, consider the following SCF f: for every preference profile P, $f(P)$ is the minimal element with respect to $\succ$ among $\{P_1(1), P_2(1), \ldots, P_n(1)\}$.

Observe that f is not a dictatorship – at every profile, a different agent can have its peak to the left. Second, it is strategy-proof. To see this, note that the agent whose peak coincides with the chosen alternative has no incentive to deviate. If some other agent deviates, then the only way to change the outcome is to place his peak to the left of the chosen outcome. But that will lead to an outcome which is even more to the left of his peak, which he prefers less than the current outcome. Hence, no manipulation is possible.

One can generalize this further. Pick an integer $k \in \{1, \ldots, n\}$. In every preference profile, the SCF picks the k-th lowest peak. Formally, $f(P_1, \ldots, P_n)$ chooses among $\{P_1(1), \ldots, P_n(1)\}$ the k-th lowest alternative according to $\succ$. To understand why this SCF is not manipulable, note that those agents whose peak coincides with the k-th lowest peak have no incentive to manipulate. Consider an agent i, whose peak lies to the left of the k-th lowest peak. The only way he can change the outcome is to move to the right of the k-th lowest peak. In that case, an outcome which is even farther away from his peak will be chosen. According to single-peaked preferences, he prefers this less. A symmetric argument applies to the agents who are on to the right of k-th lowest peak.

7.2.2 Median voter social choice function

We now define the notion of a *median voter*. Consider an integer $k \geq 1$ and any sequence of points $B \equiv (x_1, \ldots, x_{2k-1})$ such that for all $j \in \{1, \ldots, 2k-1\}$, we have $x_j \in A$. Now $b \in B$ is the median if

$$|\{x \in B : x \succ b \text{ or } x = b\}| \geq k+1 \text{ and } |\{x \in B : b \succ x \text{ or } x = b\}| \geq k+1.$$

The median of a sequence of points B will be denoted as $med(B)$. Also, for any profile $(P_1, \ldots, P_n)$, we denote the sequence of peaks as $peak(P) \equiv (P_1(1), \ldots, P_n(1))$.

DEFINITION **42** *A social choice function $f : \mathcal{S}^n \to A$ is a* **median voter** *social choice function if there exists a collection of alternatives $B = (y_1, \ldots, y_{n-1})$ such that $f(P) = med(B, peak(P))$ for all preference profiles P. The alternatives in B are called the* **phantom peaks**.

Note that by adding $(n-1)$ phantom peaks, we have $(2n-1)$ peaks, and a median is well defined. We give an example to illustrate the ideas. Figure 7.2 shows the peaks of 4 agents. Then, we add 3 phantom peaks as shown at a, b, and d. The median voter SCF chooses the median of this set, which is shown to be c Figure 7.2.

$$f(P_1, P_2, P_3, P_4) = \text{median}(a, b, d, P_1(1), P_2(1), P_3(1), P_4(1)) = c$$

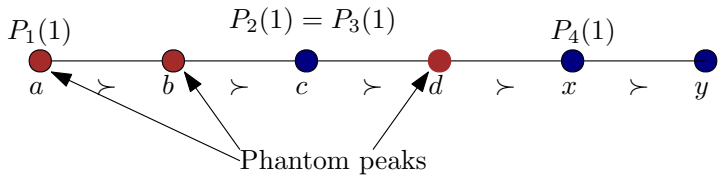

Figure 7.2: Phantom voters and the median voter

Of course, the median voter SCF is a class of SCFs. A median voter SCF must specify the peaks of the phantom voters (it cannot change across profiles). We can simulate the k-th lowest peak social choice function that we described earlier by placing the phantom peaks suitably. In particular, place $(n-k)$ phantom peaks at the lowest alternative according to $\succ$ and the remaining $(k-1)$ phantom peaks at the highest alternative according to $\succ$. It is clear that the median of this set lies at the kth lowest peak of agents.

PROPOSITION **7** (**Moulin (1980)**) *Every median voter social choice function is strategy-proof.*

Proof: Consider any profile of single-peaked preferences $P = (P_1, \ldots, P_n)$. Let f be a median voter SCF, and $f(P) = a$. Consider agent i. Agent i has no incentive to manipulate if $P_i(1) = a$. Suppose agent i's peak is to the left of a. The only way he can change the outcome is by changing the median, which he can only do by changing his peak to the right of a. But that will shift the median to the right of a which he does not prefer to a. So, he cannot manipulate. A symmetric argument applies if i's peak is to the right of a. ■

One may wonder if one introduces an arbitrary number of phantom voters whether the corresponding social choice function is still strategy-proof? We assume that whenever there are even number of agents (including the phantom voters), we pick the minimum of two medians. Along the lines of proof of Proposition 7, one can show that even this social choice function is strategy-proof.

Then, what is unique about the median voter social choice function where we take $n-1$ phantom voters. We discuss this next.

Median voter scfs are non-dictatorial. We saw in Lemma 23 in the proof of the Gibbard-Satterthwaite theorem that any strategy-proof and unanimous scf must select an alternative from the peaks of one of the agents. This *top-selection* property is not true for median voter scf. To see this, suppose there are two agents and one phantom peak. Then, there may be profiles where the peaks of the two agents are on the opposite side of the phantom peak. In that case, the scf chooses the phantom peak as the median.

7.2.3 Properties of Social Choice Functions

We first define some desirable properties of a social choice function. Most of these properties have already been discussed earlier for the Gibbard-Satterthwaite result.

DEFINITION **43** *A social choice function $f : \mathcal{S}^n \to A$ is* **onto** *if for every $a \in A$, there exists a profile $P \in \mathcal{S}^n$ such that $f(P) = a$.*

Onto rules out constant social choice functions.

DEFINITION **44** *A social choice function $f : \mathcal{S}^n \to A$ is* **unanimous** *if for every profile P with $P_1(1) = P_2(1) = \ldots = P_n(1) = a$ we have $f(P) = a$.*

DEFINITION **45** *A social choice function $f : \mathcal{S}^n \to A$ is* **efficient** *if for every profile of preferences P and every $b \in A$, if there exists $a \neq b$ such that aP_ib for all $i \in N$, then $f(P) \neq b$.*

Denote by $[a, b]$, the set of all alternatives which lie between a and b (including a and b) according to $\succ$.

LEMMA **25** *For every preference profile P, let p^{min} and p^{max} denote the smallest and largest peak (according to $\succ$) respectively in P. A social choice function $f : \mathcal{S}^n \to A$ is efficient if and only if for every profile P, $p^{max} \succeq f(P) \succeq p^{min}$, where $x \succeq y$ means $x \succ y$ or $x = y$.*

Proof: Suppose f is efficient. Fix a preference profile P. If $p^{min} \succ f(P)$, then choosing p^{min} is better for all agents. Similarly, if $f(P) \succ p^{max}$, then choosing p^{max} is better for all agents. Hence, by efficiency, $f(P) \in [p^{min}, p^{max}]$. For the converse, if $f(P) \in [p^{min}, p^{max}]$, then any alternative other than $f(P)$ will move it away from either p^{min} or p^{max}, making the corresponding agents worse off. Hence, f is efficient. ■

Median voting with arbitrary number of phantom voters may be inefficient (and may violate unanimity). Consider the median voting with $(3n - 1)$ phantom voters. Suppose we put all the phantoms at the minimal alterantive according $\succ$, and consider the preference profile where the peaks of the agents are at the maximum alterantive according to $\succ$. The outcome in this case is the minimal alternative according to $\succ$ since that is the median. But choosing agents' common peak make every agent better off.

DEFINITION **46** *A social choice function $f : \mathcal{S}^n \to A$ is* **monotone** *if for any two profiles P and P' with $f(P) = a$ and for all $b \neq a$, aP'_ib if aP_ib we have $f(P') = a$.*

Like in the unrestricted domain, strategy-proofness implies monotonicity.

LEMMA **26** *If a social choice function $f : \mathcal{S}^n \to A$ is strategy-proof, then it is monotone.*

Proof: The proof is exactly similar to the necessary part of Theorem 28. We take two preference profiles $P, P' \in \mathcal{S}^n$ such that $f(P) = a$ and aP'_ib if aP_ib for all $b \neq a$. As in the proof of Theorem 28, we can consider P and P' to be different in agent j's preference ordering *only* (else, we construct a series of preference profiles each different from the previous one by just one agent's preference). Assume for contradiction $f(P') = b \neq a$.

If bP_ja, then agent j can manipulate at P by P'. Hence, aP_jb. But that means aP'_jb. In such a case, agent j will manipulate at P' by P. This is a contradiction. ■

Like in the unrestricted domain, some of these properties are equivalent in the presence of strategy-proofness.

PROPOSITION **8** *Suppose $f : \mathcal{S}^n \to A$ is a strategy-proof social choice function. Then, f is onto if and only if it is unanimous if and only if it is efficient.*

Proof: Consider a strategy-proof social choice function $f : \mathcal{S}^n \to A$. We do the proof in three steps.

UNANIMITY IMPLIES ONTO. Fix an alternative $a \in A$. Consider a single peaked preference profile P where every agent has his peak at a. By unanimity, $f(P) = a$.

ONTO IMPLIES EFFICIENCY. Consider a preference profile P such that $f(P) = b$ but there exists a $a \neq b$ such that aP_ib for all $i \in N$. By single-peakedness, there is an alternative c which is a *neighbor* of b in $\succ$ and cP_ib for all $i \in N$. [4] Since f is onto, there exists a profile P' such that $f(P') = c$. Consider another preference profile P'' such that the peaks of every agent is c, but the second ranked alternative is b - such a preference is possible in a single-peaked domain. By Lemma 26, f is monotone. By monotonicity, we get $f(P'') = f(P') = c$ and $f(P'') = f(P) = b$. This is a contradiction.

EFFICIENCY IMPLIES UNANIMITY. In any profile, where peaks are the same, efficiency will imply that the peak is chosen. ■

We now define a new property which will be crucial for our main result in this section. For this, we need some definitions. A permutation of agents is denoted by a bijective mapping $\sigma : N \to N$. We apply a permutation σ to a profile P to construct another profile as follows: the preference ordering of agent i goes to agent $\sigma(i)$ in the new preference profile. We denote this new preference profile as P^σ.

[4] Two alternatives x and y are neighbors in $\succ$ if $x \succ y$ and there is no alternative z such that $x \succ z \succ y$ or $y \succ x$ and there is no alternative z such that $y \succ z \succ x$.

Table 7.12 shows a pair of profiles, one of which is obtained by permuting the other. We consider $N = \{1, 2, 3\}$ and σ as $\sigma(1) = 2, \sigma(2) = 3, \sigma(3) = 1$.

P_1	P_2	P_3	P_1^σ	P_2^σ	P_3^σ
a	b	b	b	a	b
b	a	c	c	b	a
c	c	a	a	c	c

Table 7.12: Example of permuted preferences

DEFINITION **47** *A social choice function $f : \mathcal{S}^n \to A$ is* **anonymous** *if for every profile P and every permutation σ such that $P^\sigma \in \mathcal{S}^n$, we have $f(P^\sigma) = f(P)$.*

Anonymous social choice functions require that the identity of agents are not important, and does not discriminate agents on that basis. An anonymous SCF counts the number of prefereces of each type at a profile and decides on the outcome. Dictatorial social choice functions are not anonymous (it favors the dictator). Any social choice function which *ignores* the preferences of some agent is not anonymous. Anonymity is a minimal form of fairness.

7.2.4 Characterization Result

We show now that the only strategy-proof social choice function which is onto and anonymous is the median voter.

THEOREM **30 (Moulin (1980))** *A social choice function $f : \mathcal{S}^n \to A$ is strategy-proof, unanimous, and anonymous if and only if it is the median voter social choice function.*

Moulin (1980) had an added condition called the *peaks-only* or *tops-only* property, which is not necessary - every strategy-proof and onto social choice function is tops-only (peaks-only) in the single peaked domain. Formally, *peaks-only* requires that only the peaks of agents matter for the social choice function.

DEFINITION **48** *A social choice function $f : \mathcal{S}^n \to A$ is* **peaks-only** *if at every pair of preference profiles $P, P' \in \mathcal{S}^n$ with $P_i(1) = P_i'(1)$ for all $i \in N$, we have $f(P) = f(P')$.*

Strategy-proofness and efficiency imply peaks-only property in the single-peaked domain, in the unrestricted domain, and in many other domains (Chatterji and Sen, 2011).

We give a proof which uses peaks-onlyness and we show that strategy-proofness and efficiency imply peaks-onlyness for the two-agents case.

Due to Proposition 8, we can replace unanimity with ontoness or efficiency in the statement of Theorem 30. We discuss the necessity of all the properties. First, a dictatorial social choice function is onto and strategy-proof. So, anonymity is crucial in the characterization. Second, choosing a fixed alternative at every preference profile is anonymous and strategy-proof, but it is not onto. Hence, all the conditions are necessary in the result. We now give the proof.

Proof: It is clear that the median voter social choice function is strategy-proof (Proposition 7), onto (all the peaks in one alternative will mean that is the median), and anonymous (it does not distinguish between agents). We now show the converse.

Suppose $f : \mathcal{S}^n \to A$ is a strategy-proof, onto, peaks-only, and anonymous social choice function. The following two preference orderings are of importance for the proof:

- P^0: this is the unique preference ordering where the peak of agent i is at the lowest alternative according to $\succ$.
- P^1: this is the unique preference ordering where the peak of agent i is at the highest alternative according to $\succ$.

We now do the proof in various steps.

Step 1. Finding the phantoms. For any $j \in \{1, \ldots, n-1\}$, define y_j as follows:

$$y_j = f(\underbrace{P^0, \ldots, P^0}_{(n-j) \text{ times}}, \underbrace{P^1, \ldots, P^1}_{j \text{ times}}).$$

So, y_j is the chosen alternative, when $(n-j)$ agents have their peak at the lowest alternative and the remaining j agents have their peak at the highest alternative. Notice that which of the j agents have their peaks at the highest alternative does not matter due to anonymity of f.

We show that $y_j = y_{j+1}$ or $y_{j+1} \succ y_j$ for any $j \in \{1, \ldots, n-1\}$. To see this consider two profiles

$$P = (\underbrace{P^0, \ldots, P^0}_{(n-j) \text{ times}}, \underbrace{P^1, \ldots, P^1}_{j \text{ times}})$$

$$P' = (\underbrace{P^0, \ldots, P^0}_{(n-j-1) \text{ times}}, \underbrace{P^1, \ldots, P^1}_{(j+1) \text{ times}}).$$

Only preference ordering of agent $k \equiv n - j$ is changing from P to P'. Note that $f(P) = y_j$ and $f(P') = y_{j+1}$. Since f is strategy-proof agent k cannot manipulate with true preference P^0: $y_j P^0 y_{j+1}$ or $y_j = y_{j+1}$. But the peak of agent k in P^0 is at the lowest alternative according to $\succ$. So, either $y_j = y_{j+1}$ or $y_{j+1} \succ y_j$.

Note that y_0 is the lowest alternative according to $\succ$ and y_n is the highest alternative according to $\succ$. By unanimity $f(P^0, \ldots, P^0) = y_0$ and $f(P^1, \ldots, P^1) = y_n$.

Step 2. Induction base. Consider a profile $P = (P_1, \ldots, P_n)$ such that $P_i \in \{P^0, P^1\}$ for each $i \in N$. We show that

$$f(P) = med(y_1, \ldots, y_{n-1}, P_1(1), \ldots, P_n(1)).$$

If $P_i(1) = P_j(1)$ for all $i, j \in N$, then the claim is true due to unanimity. Else, suppose $k > 1$ agents have preference P^1 and $(n - k) > 1$ agents have preference P^0. By definition, $f(P) = y_k$. By our earlier claim, $y_\ell \succ y_k$ or $y_\ell = y_k$ if $\ell > k$ and $y_k \succ y_\ell$ or $y_\ell = y_k$ if $\ell < k$. Hence, y_k is the median of $(y_1, \ldots, y_{n-1}, P_1(1), \ldots, P_n(1))$ as required.

Notation. From now on, at every preference profile, we write $med(P_1, \ldots, P_n)$ to denote $med((y_1, \ldots, y_{n-1}, P_1(1), \ldots, P_n(1)))$.

Step 3. Induction. Now, let K_P denote the number of agents who have either P^0 or P^1 as their preference. We complete the proof using an induction on K_P. The induction base case $K_P = n$ follows from Step 2. Fix a positive integer k such that $0 \leq k < n$ and assume that for all preference profiles P with $K_P > k$ we have $f(P) = med(P_1, \ldots, P_n)$. Now, pick a preference profile P with $K_P = k$. Suppose $f(P) = a$ and $med(P_1, \ldots, P_n) = b$. Assume for contradiction that $a \neq b$. Since $k < n$, there is an agent i such that $P_i \notin \{P^0, P^1\}$ and let $P_i(1) = c$. We consider various cases.

Case 1. $a = c$. In this case, if $b \succ a = c$, then moving peak P_i to P^0 does not change the median. Also, suppose $f(P^0, P_{-i}) = x$. Then, by strategy-proofness (P^0 not able to manipulate to P_i), we must have $xP^0a = c$ or $a = x$ or $a \succ x$. But that means $b \succ x$. As a result, we have $f(P^0, P_{-i}) \neq med(P^0, P_{-i})$. But (P^0, P_{-i}) is a profile where $K_P > k$, contradicting our induction hypothesis. Similarly, if $a = c \succ b$, we move peak P_i to P^1, which does not change the median. But by strategy-proofness $f(P^1, P_{-i}) = x \succ a$ or $x = a$. This again gives $x \succ b$, i.e., $f(P^1, P_{-1}) \neq med(P^1, P_{-i})$.

Case 2. Suppose $a \succ c$. Consider P_i' such that $P_i'(1) = P_i(1) = c$ but for every $y, y' \in A$ such that $c \succ y'$ and $y \succ c$ we have $y'P_i'y$ (left-side alternatives are preferred over right-side alternatives). By peak-onlyness, $f(P_i', P_{-i}) = a$. Further, $med(P_i', P_{-i}) = b$ (because peaks of agents did not change from (P_i, P_{-i}) to (P_i', P_{-i})). Then, consider a preference P_i'' of agent i where $P_i'' = P^0$. Figure 7.3 illstrates the layout of various alternatives.

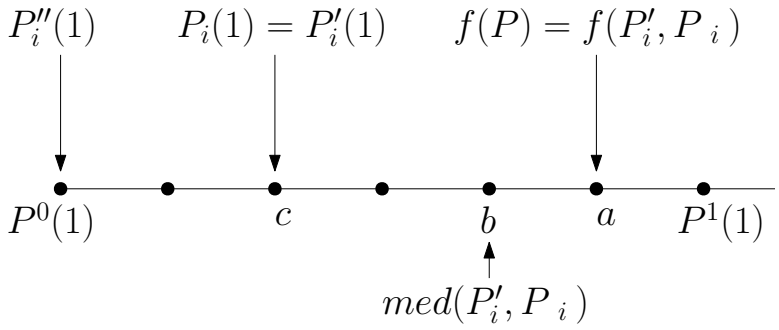

Figure 7.3: Illustration of Case 1

Suppose $f(P_i'', P_{-i}) = x$. We argue that $x = a$. By strategy-proofness (P_i'' not manipulating via P_i'), we have $x \not\succ a$. If x is between a and c, by single peakedness, $xP_i'a$, which means P_i' can manipulate to P_i''. If x is to the left of c, since a is to the right of c, by construction, $xP_i'c$ (left alternatives are better than right). Again, P_i' can manipulate to P_i''. Hence, $a = x$.

Now, suppose $b \succ c$. Then, by the property of median, $med(P_i'', P_{-i}) = b$. If $c \succ b$. Since $a \succ c$, we have $a \succ c \succ b$. As a result, by moving peak to the left from c to $P^0(1)$, we move the median to the left. Hence, $b \succ med(P_i'', P_{-i})$ or $b = med(P_i'', P_{-i})$. Since $f(P_i'', P_{-i}) = a \succ b \succeq med(P_i'', P_{-i})$, we have $f(P_i'', P_{-i}) \neq med(P_i'', P_{-i})$. But this is a contradiction to our induction hypothesis because $K_{(P_i'', P_{-i})} > k$.

CASE 3. Suppose $c \succ a$. This case is similar. Consider P_i' such that $P_i'(1) = P_i(1) = c$ but for every $y', y \in A$ such that $c \succ y'$ and $y \succ c$ we have $yP_i'y'$. By peak-onlyness, $f(P_i', P_{-i}) = a$. Further, the $med(P_i', P_{-i}) = b$. Then, consider a preference P_i'' of agent i where $P_i'' = P^1$. Figure 7.3 illstrates the layout of various alternatives.

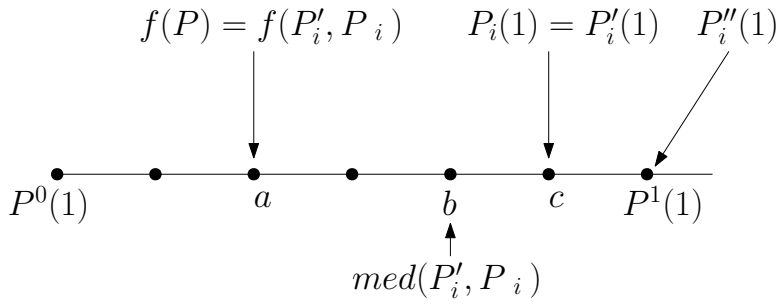

Figure 7.4: Illustration of Case 2

Suppose $f(P_i'', P_{-i}) = x$. We argue that $x = a$. By strategy-proofness (P_i'' not manipulating via P_i'), we have $a \not\succ x$. Also, by strategy-proofness (P_i' not manipulating to P_i''), we have $x \not\succ a$. Hence, $a = x$.

Now, suppose $c \succ b$. Then, by the property of median, $med(P_i'', P_{-i}) = b$. Next, suppose $b \succ c$ or $b = c$, then (i) $b \succ a$; and (b) $med(P_i'', P_{-i}) \succ b$ or $b = med(P_i'', P_{-i})$. In either case, $f(P_i'', P_{-i}) = a \neq med(P_i'', P_{-i})$. But this is a contradiction to our induction hypothesis because $K_{(P_i'', P_{-i})} > k$. ■

The peaks-only property assumed in the above proof (though natural) is implied by strategy-proofness and unanimity. It is quite cumbersome to prove generally. Below, we give a proof for two agents.

CLAIM **3** *Suppose $N = \{1, 2\}$ and f is a strategy-proof and efficient social choice function. Let P and P' be two profiles such that $P_i(1) = P_i'(1)$ for all $i \in N$. Then, $f(P) = f(P')$.*

Proof: Consider preference profiles P and P' such that $P_1(1) = P_1'(1) = a$ and $P_2(1) = P_2'(1) = b$. Consider the preference profile (P_1', P_2), and let $f(P) = x$ but $f(P_1', P_2) = y$. By strategy-proofness, xP_1y and $yP_1'x$. This implies, if x and y belong to the same side of a, then $x = y$. Then, the only other possibility is x and y belong to the different sides of a.

We will argue that this is not possible. Assume without loss of generality $y \succ a \succ x$. Suppose, without loss of generality, $a \succ b$. Then, by efficiency (Lemma 25) at profile P_1', P_2, we must have $y \in [b, a]$. This is a contradiction since $y \succ a$. Hence, it is not possible that x and y belong to the different sides of a. Thus, $x = y$ or $f(P_1, P_2) = f(P_1', P_2)$.

Now, we can replicate this argument by going from (P_1', P_2) to (P_1', P_2'). This will show that $f(P_1', P_2') = x = f(P_1, P_2)$. ■

The peaks of the phantom voters reflect the degree of compromise the social choice function has when agents have *extreme* preferences. If j agents have the highest alternative as the peak, and the remaining $n - j$ agents have the lowest alternative as the peak, then which alternative is chosen? A true median will choose the peak which has more agents, but the median voter social choice function may do something intermediate.

7.3 Randomized Social Choice Function

Randomization is a way of expanding the set of possible strategy-proof social choice function. Lotteries are also common in practice. So, it makes sense to study the effects of randomization on strategy-proofness.

As before let $A = \{a, b, c, \ldots\}$ be a finite set of alternatives with $|A| = m$ and $N = \{1, \ldots, n\}$ be the set of agents. Let $\mathcal{L}(A)$ denote the set of all probability distributions over A. We will refer to this set as the set of **lotteries** over A. A particular element $\lambda \in \mathcal{L}(A)$ is a probability distribution over A, and λ_a denotes the probability of alternative a. Of course $\lambda_a \geq 0$ for all $a \in A$ and $\sum_{a \in A} \lambda_a = 1$. As before, every agent i has a linear order over A, which is his preference ordering. A randomized social choice function picks a lottery over A at every profile of preference orderings. Hence, the set of outcomes is the set of all lotteries over A, i.e., $\mathcal{L}(A)$. Note that we have not defined a preference ordering over $\mathcal{L}(A)$. Hence, a crucial component of analyzing randomized social choice functions is

how should two lotteries $\lambda, \lambda' \in \mathcal{L}(A)$ be compared given a preference ordering over A?

We discuss below a very basic way of making such a comparison. Let $\mathcal{P}$ be the set of all linear orders over A. The domain of interest may be any subset $\mathcal{D} \subseteq \mathcal{P}$. A **randomized**

social choice function (RSCF) f is a mapping $f : \mathcal{D}^n \to \mathcal{L}(A)$. We let $f_a(P)$ to denote the probability of alternative a being chosen at profile P. To avoid confusion, we refer to $f : \mathcal{D}^n \to A$ as a **deterministic** social choice function (DSCF).

7.3.1 Defining Strategy-proof RSCF

There are several meaningful ways to define strategy-proofness in this setting. We follow one of the first-proposed approaches (by Gibbard). It requires that an RSCF be non-manipulable for every *utility representation* of linear orders when lotteries are evaluated using the **expected utility criteria**.

A utility function $u : A \to \mathbb{R}$ represents a preference ordering $P_i \in \mathcal{D}$ if for all $a, b \in A$, $u(a) > u(b)$ if and only if aP_ib. Given a utility representation u of P_i, the utility from a lottery $\lambda \in \mathcal{A}$ is computed using the expected utility criteria, and is given by

$$\sum_{a \in A} \lambda_a u(a).$$

Notice that this is a *domain restriction* - the utility of a lottery outcome is *restricted* to be the expected utility of the alternatives in its support. Hence, analysis of randomized social choice function is similar to analyzing restricted domains, and therefore, we hope to uncover more social choice functions than in the deterministic case.

Now, it is easy to define the notion of strategy-proofness. An RSCF is strategy-proof if for *every possible representation of orderings*, the expected utility of telling the truth is not less than the expected utility of lying.

Definition **49** *An RSCF $f : \mathcal{D}^n \to \mathcal{L}(A)$ is* **strategy-proof** *if for all $i \in N$, all $P_{-i} \in \mathcal{D}^{n-1}$, for all $P_i \in \mathcal{D}$, and for all utility functions $u : A \to \mathbb{R}$ representing P_i, we have*

$$\sum_{a \in A} u(a) f_a(P_i, P_{-i}) \geq \sum_{a \in A} u(a) f_a(P_i', P_{-i}) \qquad \forall\ P_i' \in \mathcal{D}.$$

For the strategy-proofness of DSCF, we did not require this utility representation. However, it is easy to verify that a DSCF is strategy-proof in the sense of Definition 38 if and only if it is strategy-proof in the sense of Definition 49. Also, the qualifier "for all utility functions" in the above definitions is extremely important. It underlines the fact that we are considering

ordinal social choice functions. If we were using "cardinal" social choice functions, then we will elicit utility functions from the agents instead of preference orderings.

It is well known that the above formulation of strategy-proofness is equivalent to first-order stochastic dominance. To define this formally, let $B(a, P_i) = \{b \in A : b = a \text{ or } bP_ia\}$. We can define the strategy-proofness in the following equivalent way.

DEFINITION **50** *An RSCF $f : \mathcal{D}^n \to \mathcal{L}(A)$ is* **strategy-proof** *if for all $i \in N$, all $P_{-i} \in \mathcal{D}^{n-1}$, for all $P_i \in \mathcal{D}$, and for all $a \in A$, we have*

$$\sum_{b \in B(a,P_i)} f_b(P_i, P_{-i}) \geq \sum_{b \in B(a,P_i)} f_b(P_i', P_{-i}) \qquad \forall\ P_i' \in \mathcal{D}.$$

The necessity of this first-order stochastic dominance is easy to derive. Fix some $i \in N$, some P_{-i}, some P_i and some alternative $a \in A$. A particular u that represents P_i is of the following form: $u(b) \to 1$ for all $b \in B(a, P_i)$ and $u(b) \to 0$ for all $b \notin B(a, P_i)$. Then, strategy-proofness gives that for every P_i', we must have

$$\sum_{b \in A} u(b) f_b(P_i, P_{-i}) \geq \sum_{b \in A} u(b) f_b(P_i', P_{-i}).$$

Substituting for u, we get

$$\sum_{b \in B(a,P_i)} f_b(P_i, P_{-i}) \geq \sum_{b \in B(a,P_i)} f_b(P_i', P_{-i}).$$

It can also be shown that the first-order stochastic dominance condition is sufficient for strategy-proofness (see Chapter 6 in Mas-Collel-Whinston-Green).

To understand this definition a little better let us take an example with two agents $\{1, 2\}$ and three alternatives $\{a, b, c\}$. The preference of agent 2 is fixed at P_2 given by aP_2bP_2c. Let us consider two preference orderings of agent 1: $P_1 : bP_1cP_1a$ and $P_1' : cP_1aP_1b$. Denote $P = (P_1, P_2)$ and $P' = (P_1', P_2)$. Suppose $f_a(P) = 0.6$ and $f_b(P) = 0.1$ and $f_c(P) = 0.3$. First order stochastic dominance requires the following.

$$f_b(P) = 0.1 \geq f_b(P')$$
$$f_b(P) + f_c(P) = 0.4 \geq f_b(P') + f_c(P').$$

Summarizing, we consider randomization but ordinal social choice functions. Agents have preferences over alternatives and use that to evaluate lotteries. Our idea of truthfulness says

that the lottery given by the scf from truthtelling must first-order stochastically dominate *every* other lottery that this agent can get from lying. This notion of strategy-proofness is equivalent to preventing manipulation for *all* cardinalization of preferences when agents use expected utility to evaluate lotteries.

Of course, we can think of other notions of strategy-proofness. We discuss two such notions.

1. WEAK-STRATEGY-PROOF. In this notion, *manipulation* requires an agent has can get a lottery which FOSD dominates the truth-telling lottery. Contrast this to strategy-proofness, where truth-telling lottery was required to FOSD dominate every other lottery that the agent could get. More formally, fix agent i and fix the preferences of other agents at P_{-i}. We say that agent i **strongly manipulates** f at (P_i, P_{-i}) if there exists P_i' such that the lottery $f(P_i', P_{-i})$ first order stochastically dominates $f(P_i, P_{-i})$. Then, we can say that f is **weakly strategy-proof** if no agent can manipulate it at any profile.

2. LEX-STRATEGY-PROOF. Another method of defining strategy-proofness is *lexicographic*. Again, fix agent i and fix the preferences of other agents at P_{-i}. Take two preferences P_i, P_i' of agent i. Then, define a binary relation over every pair of lotteries using P_i in a lexicographic manner. It evaluates lotteries $f(P_i, P_{-i})$ and $f(P_i', P_{-i})$ in the following way: it first looks at $P_i(1)$ - the top ranked alternative in P_i, and compares the two lotteries; if they are the same, then it looks at $P_i(2)$, and so on. We can define *lex-strategy-proofness* easily now - $f(P_i, P_{-i})$ must be lexicographically better than $f(P_i', P_{-i})$, where the lexicographic comparison is done using P_i.

You are encouraged to reconsider the scfs discussed earlier (e.g., scoring rules) and see if they satisfy these notions of strategy-proofness.

7.3.2 Randomization over DSCFs

A natural way to construct an RSCF is to take a collection of DSCFs and randomize over them. We show a general result on strategy-proofness of RSCFs which can be expressed as a convex combination of other strategy-proof RSCFs.

PROPOSITION **9** *Let $f^1, f^2, \ldots, f^k$ be a set of k strategy-proof RSCFs, all defined on the domain $\mathcal{D}^n$. Let $f : \mathcal{D}^n \to \mathcal{L}(A)$ be defined as: for all $P \in \mathcal{D}^n$ and for all $a \in A$, $f_a(P) = \sum_{j=1}^k \lambda_j f_a^j(P)$, where $\lambda_j \in [0,1]$ for all $j \in \{1, \ldots, k\}$ and $\sum_{j=1}^k \lambda_j = 1$. Then, f is strategy-proof.*

Proof: Fix an agent i and a profile P_{-i}. For some preference P_i consider a utility representation $u : A \to \mathbb{R}$. Then, for any P_i',

$$\begin{aligned}\sum_{a\in A} u(a) f_a(P) &= \sum_{a\in A} u(a) \sum_{j=1}^k \lambda_j f_a^j(P) = \sum_{j=1}^k \lambda_j \sum_{a\in A} u(a) f_a^j(P) \\ &\geq \sum_{j=1}^k \lambda_j \sum_{a\in A} u(a) f_a^j(P_i', P_{-i}) = \sum_{a\in A} u(a) \sum_{j=1}^k \lambda_j f_a^j(P_i', P_{-i}) \\ &= \sum_{a\in A} u(a) f_a(P_i', P_{-i}).\end{aligned}$$

■

Another way to interpret Proposition 9 is that the set of strategy-proof RSCFs form a convex set. Since a DSCF cannot be written as convex combination of other social choice functions, a strategy-proof DSCF forms an *extreme point* of the set of strategy-proof RSCFs. Knowing the deterministic strategy-proof social choice functions automatically gives you a class of strategy-proof RSCFs.

7.3.3 The Counterpart of Gibbard-Satterthwaite Theorem

To understand the implication of randomization, we go back to the complete domain model in the Gibbard-Satterthwaite theorem. First, we define the notion of unanimity that we will use in this model. [5] The notion of unanimity we use is the exact version of unanimity we used in the deterministic social choice functions.

DEFINITION **51** *An RSCF $f : \mathcal{P}^n \to \mathcal{L}(A)$ satisfies* **unanimity** *if for all $i \in N$, all $P \in \mathcal{P}^n$ such that $P_1(1) = P_2(1) = \ldots = P_n(1) = a$, we have $f_a(P) = 1$.*

[5] In the deterministic model, there was an equivalence between unanimity, efficiency, and ontoness under strategy-proofness - this is no longer true in the model with randomization.

As in the deterministic SCF case, we can see that the constant social choice function is not unanimous. But there is even a bigger class of RSCFs which are strategy-proof but not unanimous.

DEFINITION **52** *An RSCF f is a* **unilateral** *if there exists an agent i and $\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_{|A|}$ with $\alpha_j \in [0,1]$ and $\sum_{j=1}^{|A|} \alpha_j = 1$ such that for all P we have $f_{P_i(j)} = \alpha_j$ for all $j \in \{1, \ldots, |A|\}$.*

In a unilateral RSCF, there is a **weak dictator** i such that top ranked alternative of i gets probability α_1, second ranked alternative of i gets probability α_2, and so on. Notice that every unilateral is strategy-proof, but not unanimous.

We now define another broad class of RSCFs which are strategy-proof and unanimous.

DEFINITION **53** *An RSCF $f : \mathcal{P}^n \to \mathcal{L}(A)$ is a* **random dictatorship** *if there exists weights $\beta_1, \ldots, \beta_n \in [0,1]$ with $\sum_{i \in N} \beta_i = 1$ such that for all $P \in \mathcal{P}^n$,*

$$f_a(P) = \sum_{i \in N : P_i(1) = a} \beta_i.$$

If a particular agent i has $\beta_i = 1$, then such a random dictatorship is the usual dictatorship. A random dictatorship can be thought to be a randomization over deterministic dictatorships, where β_i reflects the probability with which agent i is a dictator. For example, if $N = \{1,2,3\}$ and $A = \{a,b,c\}$ and $\beta_1 = \frac{1}{2}$, $\beta_2 = \beta_3 = \frac{1}{4}$, then at a profile P where $P_1(1) = a, P_2(1) = a, P_3(1) = c$, the output of this random dictatorship will be $f_a(P) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$ and $f_c(P) = \frac{1}{4}$.

Random dictatorship can be thought of as a convex combination of dictatorships, where β_i is the probability with which agent i is the dictator. Since dictatorship is strategy-proof, one can show that random dictatorship is also strategy-proof. As a corollary of Proposition 9, we get the following.

COROLLARY **5** *Every random dictatorship is strategy-proof.*

Proof: A random dictatorship is a convex combination of dictatorships. Hence, it is strategy-proof by Proposition 9. ■

We are now ready to state the counterpart of the Gibbard-Satterthwaite theorem for RSCFs. This was proved by Gibbard.

THEOREM **31** *Suppose $|A| \geq 3$. An RSCF is unanimous and strategy-proof if and only if it is a random dictatorship.*

The proof of this theorem is more involved than the Gibbard-Satterthwaite theorem. We only do the case with two agents.

Proof: We have already shown that a random dictatorship is strategy-proof (Corollary 5). It is also unanimous - if all agents have the same alternative as top ranked, βs will sum to 1 for that alternative. We now prove that any RSCF which is unanimous and strategy-proof must be a random dictatorship for $n = 2$ case. We do the proof by showing two claims. Let f be a strategy-proof and unanimous RSCF.

CLAIM **4** *Let $P \in \mathcal{P}^2$ be a preference profile such that $P_1(1) \neq P_2(1)$. If $f_a(P) > 0$ then $a \in \{P_1(1), P_2(1)\}$.*

Proof: Consider a preference profile P such that $P_1(1) = a \neq b = P_2(1)$. Let $f_a(P) = \alpha$ and $f_b(P) = \beta$. Consider a preference ordering P_1' such that $P_1'(1) = P_1(1) = a$ and $P_1'(2) = P_2(1) = b$. Similarly, consider a preference ordering P_2' such that $P_2'(1) = P_2(1) = b$ and $P_2'(2) = P_1(1) = a$.

Strategy-proofness implies that $f_a(P_1', P_2) = \alpha$. Also, by unanimity the outcome at (P_2, P_2) is b. So, strategy-proofness implies that $f_a(P_1', P_2) + f_b(P_1', P_2) \geq f_a(P_2, P_2) + f_b(P_2, P_2) = 1$. Hence, $f_a(P_1', P_2) + f_b(P_1', P_2) = 1$.

Using a symmetric argument, we can conclude that $f_b(P_1, P_2') = \beta$ and $f_a(P_1, P_2') + f_b(P_1, P_2') = 1$.

Strategy-proofness implies that $f_b(P_1', P_2') = f_b(P_1', P_2) = 1 - \alpha$. and $f_a(P_1', P_2') = f_a(P_1, P_2') = 1 - \beta$. But $f_a(P_1', P_2') + f_b(P_1', P_2') \leq 1$ implies that $\alpha + \beta \geq 1$ and $f_a(P) + f_b(P) \leq 1$ implies $\alpha + \beta \leq 1$. Hence, $\alpha + \beta = 1$. ■

CLAIM **5** *Let $P, \bar{P} \in \mathcal{P}^2$ be such that $P_1(1) = a \neq b = P_2(1)$ and $\bar{P}_1(1) = c \neq d = \bar{P}_2(1)$. Then $f_a(P) = f_c(\bar{P})$ and $f_b(P) = f_d(\bar{P})$.*

Proof: We consider various cases.

CASE 1: $c = a$ and $d = b$. Strategy-proofness implies that $f_a(P_1, P_2) = f_a(\bar{P}_1, P_2)$. By Claim 4, $f_a(P_1, P_2) + f_b(P_1, P_2) = f_a(\bar{P}_1, P_2) + f_b(\bar{P}_1, P_2) = 1$. Hence, $f_b(P_1, P_2) = f_b(\bar{P}_1, P_2)$. Repeating this argument for agent 2 while going from $(\bar{P}_1, P_2)$ to $(\bar{P}_1, \bar{P}_2)$, we get that $f_a(\bar{P}) = f_a(P)$ and $f_b(\bar{P}) = f_b(P)$.

CASE 2: $c = a$ or $d = b$. Suppose $c = a$. Consider a preference profile $(P_1, \hat{P}_2)$ such that $\hat{P}_2(1) = d \notin \{a, b\}$ and $\hat{P}_2(2) = b$. Assume without loss of generality that $P_2(1) = b$ and $P_2(2) = d$. Then, strategy-proofness implies that $f_b(P_1, \hat{P}_2) + f_d(P_1, \hat{P}_2) = f_b(P) + f_d(P)$. By Claim 4, $f_b(P_1, \hat{P}_2) = f_d(P) = 0$. Hence, $f_b(P) = f_d(P_1, \hat{P}_2)$. This further implies that $f_a(P) = f_a(P_1, \hat{P}_2)$. By Case 1, $f_a(P) = f_a(\bar{P})$ and $f_b(P) = f_d(\bar{P})$. An analogous proof works if $d = b$.

CASE 3: $c = b$ and $d \notin \{a, b\}$. Let $\hat{P} = (P_1, \bar{P}_2)$. By Case 2, $f_a(P) = f_a(\hat{P})$ and $f_b(P) = f_d(\hat{P})$. Again, applying Case 2, we get $f_a(P) = f_a(\hat{P}) = f_b(\bar{P})$ and $f_b(P) = f_d(\hat{P}) = f_d(\bar{P})$.

CASE 4: $c \notin \{a, b\}$ and $d = a$. A symmetric argument to Case 3 can be made.

CASE 5: $c = b$ and $d = a$. Since there are at least three alternatives there is a $x \notin \{a, b\}$. We construct a profile $\hat{P} = (\hat{P}_1, \bar{P}_2)$ such that $\hat{P}_1(1) = x$. By Case 4, $f_x(\hat{P}) = f_a(P)$ and $f_b(P) = f_a(\hat{P})$. Now, applying Case 2, we can conclude that $f_x(\hat{P}) = f_b(\bar{P})$ and $f_a(\hat{P}) = f_a(\bar{P})$.

CASE 6: $c \notin \{a, b\}$ and $d \notin \{a, b\}$. Consider a profile $\hat{P} = (\hat{P}_1, P_2)$ such that $\hat{P}_1(1) = c$. By Case 2, $f_c(\hat{P}) = f_a(P)$ and $f_b(\hat{P}) = f_b(P)$. Applying Case 2 again, we get $f_c(\bar{P}) = f_c(\hat{P}) = f_a(P)$ and $f_d(\bar{P}) = f_b(\hat{P}) = f_b(P)$. ■

Claims 4 and 5 establishes that f is a RSCF. ■

As we have seen a unilateral SCF is not unanimous but strategy-proof. Hence, unanimity is a crucial assumption in Theorem 31. This shows that randomization does not expand the set of strategy-proof SCFs in a satisfactory way. One may argue that a random dictatorship is a *reasonable* SCF, but scoring rules may not be representable by a random dictatorship.

Chapter 8

Matching Theory

Matching theory refers to theory of a rich class of models where there are two "sides" and one side is matched to the other. In the graph theory literature, a matching is formulated more abstractly.[1] Over the years, matching theory has been one of the prominent success stories of economics. It has been applied extensively in practice: in school admissions; in allocating houses and dorm rooms; in assigning doctors to internships; in exchanging organs (liver and kidney) among patients. Sönmez and Ünver (2011) is an excellent starting point to learn about these applications.

We study the theory of two kinds of matching: (a) one-sided matching and (b) two-sided matching. In one-sided matching, there are agents on one side and objects on the other side, and the objective is to match agents to objects. In this model, agents have preferences over objects but objects do not have any preferences, and hence, the name one-sided matching. We will refer to this model as the *object assignment model*. The other model is the two-sided matching, where agents (workers) on one side are matched to the agents (firms) on the other side. This model is often referred to as the *marriage market model*.

[1] Consider a graph with a set of vertices and (undirected) edges between the vertices. A matching is a collection of edges that have no common end points. The matching theory we study here (and most of economic theory) concerns with a particular kind of graph, called the *bipartite* graph, where the set of vertices can be partitioned into two groups and edges run from one group to the other.

8.1 Object Assignment Model

In this section, we consider the object allocation model. There is a finite set of objects $M = \{a_1, \ldots, a_m\}$ and a finite set of agents $N = \{1, \ldots, n\}$. We assume that $m \geq n$. The objects can be houses, jobs, projects, positions, candidates or students etc. Each agent has a linear order over the set of objects, i.e., a complete, transitive, and anti-symmetric binary relation. In this model, this ordering represents the preferences of the agents, and this is their private information (type). The preference ordering of agent i will be denoted as $\succ_i$. A profile of preferences will be denoted as $\succ \equiv (\succ_1, \ldots, \succ_n)$. The set of all preference orderings over M will be denoted as $\mathcal{M}$. The top element amongst a set of objects $S \subseteq M$ according to ordering $\succ_i$ is denoted as $\succ_i (1, S)$, and the k-th ranked object by $\succ_i (k, S)$.

The main departure of this model is that agents do not have direct preference over alternatives. We need to extract their preference over alternatives from their preference over objects. What are the alternatives? An alternative is a *feasible matching*, i.e., an injective mapping from N to M. The set of alternatives will be denoted as A, and this is the set of all injective mappings from N to M. For a given alternative $a \in A$, if $a(i) = j \in M$, then we say that agent i is assigned object j (in a).

Consider two alternatives a and b. Suppose agent 1 is is assigned the same object in both a and b (this is possible if there are at least three objects). Then, it is reasonable to assume that agent 1 will **always** be indifferent between a and b. Hence, for any preference ordering of agent 1, aP_1b and bP_1a are not *permissible*. This restriction implies that the domain of preference orderings over alternatives is not the unrestricted domain, which was the case in the GS theorem. Because of this reason, we cannot apply the GS theorem. Indeed, we will show that non-dictatorial social choice functions are strategy-proof in these settings.

A social choice function (direct mechanism) f is a mapping $f : \mathcal{M}^n \to A$. We denote by $f(\succ)$ the matching produced at a preference profile $\succ$. We denote by $f_i(\succ)$ the object assigned to agent i at a preference profile $\succ$.

8.1.1 The fixed priority mechanism

A prominent mechanism in the object assignment model is the **fixed priority (serial dictatorship)** mechanism. We call this a mechanism but not a social choice function since it is *not* a direct revelation mechanism. A **priority** is a bijective mapping $\sigma : N \to N$, i.e., an

ordering over the set of agents. The fixed priority mechanism is defined inductively. Fix a preference profile $\succ$. We now construct a matching a as follows:

$$
\begin{aligned}
a(\sigma(1)) &=\succ_{\sigma(1)} (1, M)\\
a(\sigma(2)) &=\succ_{\sigma(2)} (1, M \setminus \{a(\sigma(1))\})\\
a(\sigma(3)) &=\succ_{\sigma(3)} (1, M \setminus \{a(\sigma(1)), a(\sigma(2))\})\\
&\cdots\cdots\\
a(\sigma(i)) &=\succ_{\sigma(i)} (1, M \setminus \{a(\sigma(1)), \ldots, a(\sigma(i-1))\}\\
&\cdots\cdots\\
a(\sigma(n)) &=\succ_{\sigma(n)} (1, M \setminus \{a(\sigma(1)), \ldots, a(\sigma(n-1))\}.
\end{aligned}
$$

Now, at preference profile $\succ$, the fixed priority mechanism (and the underlying SCF) assigns $f^{\sigma}(\succ) = a$.

Let us consider an example. The ordering over houses $\{a_1, a_2, \ldots, a_6\}$ of agents $\{1, 2, \ldots, 6\}$ is shown in Table 8.1. Fix a priority σ as follows: $\sigma(i) = i$ for all $i \in N$. According to this

$\succ_1$	$\succ_2$	$\succ_3$	$\succ_4$	$\succ_5$	$\succ_6$
a_3	a_3	a_1	a_2	a_2	a_1
a_1	a_2	a_4	a_1	a_1	a_3
a_2	a_1	a_3	a_5	a_6	a_2
a_4	a_5	a_2	a_4	a_4	a_4
a_5	a_4	a_6	a_3	a_5	a_6
a_6	a_6	a_5	a_6	a_3	a_5

Table 8.1: An example for housing model

priority, the fixed priority mechanism will let agent 1 choose his best object first, which is a_3. Next, agent 2 chooses his best object among remaining objects, which is a_2. Next, agent 3 gets his best object among remaining objects $\{a_1, a_4, a_5, a_6\}$, which is a_1. Next, agent 4 gets his object among remaining objects $\{a_4, a_5, a_6\}$, which is a_5. Next, agent 5 gets his best object among remaining objects $\{a_4, a_6\}$, which is a_6. So, agent 6 gets a_4.

Note that a fixed priority mechanism is a generalization of dictatorship. Hence, this mechanism is often referred to as the **serial dictatorship** mechanism. But unlike a dictator-

ship, here every agent can change the outcome by reporting different preferences. This is the reason that the term "dictatorship" is not entirely appropriate to use for this mechanism.

We show below (quite obvious) that a fixed priority mechanism is strategy-proof. Moreover, it is efficient in the following sense.

DEFINITION **54** *A social choice function f is* **efficient** *(in the object assignment model) if for every preference profile $\succ$, there exists no matching $a \neq f(\succ)$ such that $a(i) = f_i(\succ)$ or $a_i \succ_i f_i(\succ)$.*

PROPOSITION **10** *Every fixed priority social choice function (mechanism) is strategy-proof and efficient.*

A word of caution here about the notion of strategy-proofness for the fixed priority mechanism (and other mechanisms discussed in this section). The fixed priority mechanism is not a direct mechanism. However, using revelation principle, one can think of the associated direct mechanism - agents report their entire ordering, and the mechanism designer executes the fixed priority SCF on this ordering. Whenever we say that the fixed priority mechanism is strategy-proof, we mean that the underlying direct mechanism is strategy-proof.

Proof: Fix a priority σ, and consider f^σ- the associated fixed priority mechanism. The strategy of any agent i is any ordering over M. Suppose agent i wants to deviate. Fix the preferences reported (strategy) by other agents. When agent i is truthful, let M^{-i} be the set of objects allocated to agents who have higher priority than i (agent j has higher priority than agent i if and only if $\sigma(j) < \sigma(i)$). So, by being truthful, agent i gets $\succ_i (1, M \setminus M^{-i})$. When agent i deviates, any agent j who has a higher priority than agent i continues to get the same object that he was getting when agent i was truthful. So, agent i gets an object in $M \setminus M^{-i}$. Hence, deviation cannot be better.

To show efficiency, assume for contradiction that f^σ is not efficient. Consider a profile $\succ$ such that $f(\succ) = a$. Let $a' \neq a$ be another matching satisfying $a'(i) \succ_i a(i)$ or $a'(i) = a(i)$ for all $i \in N$. Then, consider the first agent j in the priority σ such that $a'(j) \succ_j a(j)$. Since agents before j in priority σ got the objects of matching a', object $a'(j)$ was still available to agent j. This is a contradiction since agent j chose $a(j)$ with $a'(j) \succ_j a(j)$. ■

Note that every fixed priority mechanism f^σ is a dictatorship. In the fixed priority mechanism f^σ corresponding to priority σ, agent $\sigma(1)$ is assigned his top object, and hence, his top alternative. So, $\sigma(1)$ is a dictator in f^σ.

One can construct social choice functions which are strategy-proof but not a fixed priority mechanism in this model. We show this by an example.

EXAMPLE **5**

Let $N = \{1, 2, 3\}$ and $M = \{a_1, a_2, a_3\}$. The social choice function we consider is f, and is *almost* a fixed priority SCF. Fix a priority σ as follows: $\sigma(i) = i$ for all $i \in N$. Another priority is σ': $\sigma'(1) = 2, \sigma'(2) = 3, \sigma'(3) = 1$. The SCF f generates the same outcome as f^σ whenever $\succ_2 (1, M) \neq a_1$. If $\succ_2 (1, M) = a_1$, then it generates the same outcome as $f^{\sigma'}$. To see that this is strategy-proof, it is clear that agents 1 and 3 cannot manipulate since they cannot change the priority. Agent 2 can change the priority. But, can he manipulate? If his top ranked house is a_1, he gets it, and he cannot manipulate. If his top ranked house is $\in \{a_2, a_3\}$, then he cannot manipulate without changing the priority. If he does change the priority, then he gets a_1. But being truthful, either he gets his top ranked house or second ranked house. So, he gets a house which is either a_1 or some house which he likes more than a_1. Hence, he cannot manipulate.

Notice that it is possible for agent 2 to change the priority in Example 5. This means it can change the allocation of other agents by reporting a different preference. Further, it can do so when her own allocation is unchanged. For instance, consider two preferences of agent 2 as follows:

$$a_2 \ \succ_2 \ a_1 \ \succ_2 \ a_3; \qquad a_1 \ \succ'_2 \ a_2 \ \succ'_2 \ a_3.$$

Suppose agent 1's preference is such that her top is a_2. When agent 2 reports $\succ_2$ as her preference, then priority is σ. So, agent 1 gets a_2 and agent 2 gets a_1. But if agent 2 reports $\succ'_2$ as her preference, the priority is σ'. So, agent 2 gets a_1 too. However, now agent 3 chooses next. If agent 3's top-ranked object is also a_2, then agent 1 can no longer get a_2.

So, agent 2 can change the outcome of the mechanism without changing her own matching. In this case, we call agent 2 *bossy*. The following axiom rules out bossiness. It was first introduced in mechanism design literature in Satterthwaite and Sonnenschein (1981).

DEFINITION **55** *A social choice function f is* **non-bossy** *if for every $i \in N$, every $\succ_{-i}$, and every $\succ_i, \succ_i'$ we have*

$$\Big[f_i(\succ_i, \succ_{-i}) = f_i(\succ_i', \succ_{-i})\Big] \Rightarrow \Big[f_j(\succ_i, \succ_{-i}) = f_j(\succ_i', \succ_{-i}) \; \forall \; j \in N\Big]$$

Though Example 5 is *not* non-bossy, every fixed priority mechanism is non-bossy. In fact, fixed priority mechanisms are characterized by strategy-proofness, non-bossiness, and *neutrality*. To define neutrality, let $\rho : M \to M$ be a permutation of objects and $\succ^\rho$ be the permutation of preference profile $\succ$ using ρ.

DEFINITION **56** *A social choice function f is* **neutral** *if for every permutation of objects ρ and every preference profile $\succ$*

$$f_i(\succ^\rho) = \rho(f_i(\succ)) \qquad \forall \; i \in N.$$

Svensson (1999) shows the following result.

THEOREM **32 (Svensson (1999))** *A social choice function is strategy-proof, non-bossy, and neutral if and only if it is a fixed priority mechanism.*

Proof: One direction is straightforward and omitted. For the other direction, let f be a strategy-proof, non-bossy, and neutral scf. We start off by establishing a version of monotonicity in this model.

CLAIM **6** *Suppose $\succ$ and $\succ'$ are two preference profiles with $f(\succ) = a$ and $\{b_k : a(i) \succ_i b_k\} \subseteq \{b_k : a(i) \succ_i' b_k\}$ for all $i \in N$. Then, $f(\succ') = a$.*

Proof: Without loss of generality, we assume that $\succ$ and $\succ'$ differ only in the preference of agent i – if they differ in more than one agent's preference, then we can repeat the argument below iteratively to conclude. So, $\succ' \equiv (\succ_i', \succ_{-i})$. Let $f(\succ') = b$. By strategy-proofness, $b(i) = a(i)$ or $a(i) \succ_i b(i)$. If $a(i) \succ_i b(i)$, then by assumption, $a(i) \succ_i' b(i)$. But this contradicts strategy-proofness since agent i with true preference $\succ_i'$ can manipulate by reporting $\succ_i$. Hence, $a(i) = b(i)$. By non-bossiness, $a = b$. ■

Notice that Claim 6 did not require neutrality in the proof. So, strategy-proofness and non-bossiness imply monotonicity.[2] The second step is to show what happens at profiles

[2]Recall that in strategic voting models, strategy-proofness alone implies monotonicity.

where everyone has identical preferences.[3]

CLAIM **7** *Let $\succ$ be a preference profile such that $\succ_1 = \ldots = \succ_n$. Then, the matching $f(\succ)$ is efficient, i.e., there does not exist $b \neq f(\succ)$ such that $b(i) = f_i(\succ)$ or $b(i) \succ_i f_i(\succ)$ for all $i \in N$.*

Proof: Let the common preference of agents in preference profile $\succ$ be:

$$a_1 \succ_i a_2 \succ_i \ldots \succ_i a_n \succ_i \ldots \succ_i a_m.$$

Since $m \geq n$, efficiency at such a profile means objects in $\{a_1, \ldots, a_n\}$ must be allocated. Assume for contradiction this is not the case. Then, there exists objects a_j, a_{j+1} such that a_j is not allocated in matching $f(\succ)$ but a_{j+1} is allocated in matching $f(\succ)$. Consider another common ranking preference profile, where we permute the role of objects a_j and a_{j+1}, i.e, now object a_{j+1} is preferred to a_j by everyone (and other objects are ranked the same way). Denote the new preference profile as $\succ'$. Notice that $\succ'$ is obtained from $\succ$ by permuting a_j and a_{j+1}. Table 8.2 shows the two profiles $\succ$ and $\succ'$.

$\succ$				$\succ'$			
a_1	a_1	$\ldots$	a_1	a_1	a_1	$\ldots$	a_1
a_2	a_2	$\ldots$	a_2	a_1	a_1	$\ldots$	a_1
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
a_j	a_j	$\ldots$	a_j	a_{j+1}	a_{j+1}	$\ldots$	a_{j+1}
a_{j+1}	a_{j+1}	$\ldots$	a_{j+1}	a_j	a_j	$\ldots$	a_j
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
a_m	a_m	$\ldots$	a_m	a_m	a_m	$\ldots$	a_m

Table 8.2: Common ranking preference profile

[3]A notable difference between voting models and private good (matching) models are predictions in preference profiles where everyone has identical preference. In voting models, this indicates agreement, and unanimity immediately gives us the answer. However, in matching models, this is where the most disagreement occurs – if everyone likes the same object, who should get it?

Suppose $f_i(\succ) = a_{j+1}$. By neutrality,

$$f_k(\succ') = \begin{cases} f_k(\succ) & \text{if } k \neq i \\ a_j & \text{if } k = i. \end{cases}$$

But by Claim 6, $f(\succ') = f(\succ)$, and hence, $f_i(\succ') = a_{j+1}$. This is a contradiction. ■

We can now conclude the proof of the theorem. We first construct a permutation σ of agents. Consider any preference profile $\succ$ where preferences of agents are identical as in Claim 7. Without loss of generality, for every $i \in N$, let

$$a_1 \succ_i a_2 \succ_i \ \dots \ \succ_i a_n \succ_i \ \dots \ \succ_i a_m.$$

By Claim 7, agents in N get objects in $\{a_1, \dots, a_n\}$. For each $i \in N$, let $\sigma(i)$ be the agent such that $f_{\sigma(i)}(\succ) = a_i$. Notice that by neutrality, at any preference profile where agents have the same preference, the outcome coincides with the fixed priority mechanism with respect to σ.

Now, pick an arbitrary preference profile $\bar{\succ}$. Let

$$\bar{a}_1 := \bar{\succ}_{\sigma(1)}(1, M); \bar{a}_2 := \bar{\succ}_{\sigma(2)}(1, M \setminus \{\bar{a}_1\}); \dots; \bar{a}_k := \bar{\succ}_{\sigma(k)}(1, M \setminus \{\bar{a}_1, \dots, \bar{a}_{k-1}\}); \dots;$$
$$\bar{a}_n := \bar{\succ}_{\sigma(n)}(1, M \setminus \{\bar{a}_1, \dots, \bar{a}_{n-1}\})$$

By construction, $\bar{a}_1$ is the highest ranked object of $\sigma(1)$ in $\bar{\succ}_1$. For agent $\sigma(2)$, object $\bar{a}_2$ is better than $\{\bar{a}_3, \dots, \bar{a}_m\}$. Similarly, for agent $\sigma(k)$, object $\bar{a}_k$ is better than $\{\bar{a}_{k+1}, \dots, \bar{a}_m\}$; and so on. Now, consider the common preference profile $\bar{\succ}'$, where all the agents have the common preference:

$$\bar{a}_1 \ \bar{\succ}'_i \ \bar{a}_2 \ \bar{\succ}'_i \ \dots \ \bar{\succ}'_i \ \bar{a}_m.$$

Notice that for any k, agent $\sigma(k)$ prefers object $\bar{a}_k$ to objects in $\{\bar{a}_{k+1}, \dots, \bar{a}_m\}$ at preference profile $\bar{\succ}'$. Hence, applying monotonicity of Claim 6, we conclude that $f(\bar{\succ}) = f(\bar{\succ}')$. But by construction $f(\bar{\succ}')$ is the matching produced by the fixed priority mechanism according to σ. Hence, we are done. ■

Theorem 32 shows what additional axioms are needed to get an analogue of a dictatorship characterization in the object assignment model. Of course, unlike the characterizations

in the voting model, we did not use Pareto efficiency, but we used neutrality and non-bossiness. How large are the class of strategy-proof and Pareto efficient mechanisms? The class of strategy-proof, non-bossy, and Pareto efficient mechanisms are quite rich – these are characterized by *trading cycle* (Pycia and Ünver, 2017). Pápai (2000) characterizes a smaller family of mechanisms, which she calls *hierarchical exchange rules*, by strategy-proofness, non-bossy, Pareto efficiency, and an additional property called *reallocation-proofness*. Both classes of rules contain fixed priority mechanisms as a special case. In the next section, we discuss a prominent subclass of hierarchical exchange rules.

8.1.2 Top Trading Cycle Mechanism with Fixed Endowments

The **top trading cycle mechanism (TTC) with fixed endowment** is a class of general mechanisms which are strategy-proof and Pareto efficient. We will study them in detail here. We assume here $m = n$ for simplicity. In the next subsection, we show how to relax this assumption. To explain the mechanism, we start with an example. Consider the preference profile shown in Table 8.3.

$\succ_1$	$\succ_2$	$\succ_3$	$\succ_4$	$\succ_5$
a_4	a_1	a_5	a_2	a_4
a_3	$[a_2]$	a_1	a_5	a_2
$[a_1]$	a_4	a_2	a_1	$[a_5]$
a_2	a_5	a_4	$[a_4]$	a_3
a_5	a_3	$[a_3]$	a_3	a_1

Table 8.3: A preference profile

The mechanism is defined by an *endowment*, which is fixed across all profiles. An endowment is another matching a^*. One can think of the mechanism to work as an *algorithm* at each preference profile. This algorithm starts with matching a^* and makes a sequence of Pareto improvements to reach a Pareto efficient matching. For the example in Table 8.3, we take a^* as (shown in square brackets in Table 8.3):

$$a^*(i) = a_i \qquad \forall\ i \in \{1, 2, 3, 4, 5\}.$$

Notice that a^* is Pareto inefficient: agents 4 and 5 can exchange their endowments to get a matching that they prefer. The main idea of the algorithm is to repeatedly perform such exchanges or *trades*.

Construct a directed graph with five vertices, one for each agent. Put a directed edge from vertex of agent i to vertex of agent j if the top ranked object of agent i is endowed with agent j. For the preference profile of Table 8.3, the directed graph looks as in Figure 8.1.

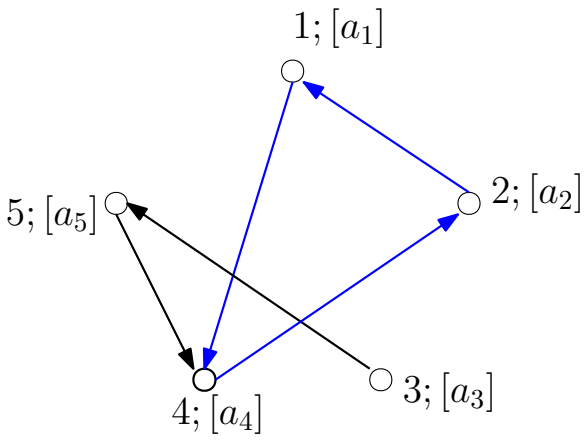

Figure 8.1: Top trading cycle trading graph

One notices from the *cycle* in the directed graph (blue edges) that agents 1,2, and 4 can trade their endowments and get their top ranked objects. In fact, any such directed graph will have at least one cycle (a cycle may involve only one vertex and an edge from that vertex to itself) and trading along these cycles makes the agents in the cycle improve from their endowments. In the Figure 8.1, we make agents $1, 2, 4$ trade their endowments along the cycle: this results in a matching where $a(1) = a_4, a(2) = a_1, a(4) = a_2$. After this trade, agents $1, 2, 4$ leave with objects $\{a_1, a_2, a_4\}$ and we are only left with agents 3 and 5 and their endowments $\{a_3, a_5\}$. This is a crucial *greedy* feature of the algorithm: at any iteration, agents who trade leave with their assigned objects after trade and we are left with the remaining agents and their endowments. Then, we apply the same step to this smaller problem.

In this case, agents 3 and 5 point to their top ranked objects from $\{a_3, a_5\}$. This results in a directed graph as shown in Figure 8.2. Here, we see that there is only one cycle involving agent 5. Hence, we set $a(5) = a_5$. Then, only agent 3 is left with her endowment a_3, and so,

we set $a(3) = a_3$. So, the final matching of the algorithm is

$$a(1) = a_4, a(2) = a_1, a(3) = a_3, a(4) = a_2, a(5) = a_5.$$

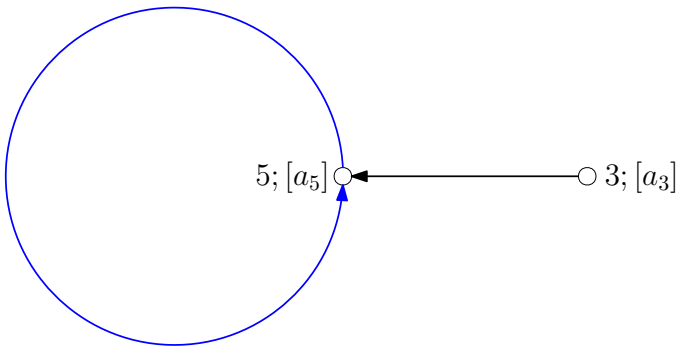

Figure 8.2: Top trading cycle trading graph

We now formally describe the top trading cycle (TTC) mechanism. At every preference profile $\succ$, the mechanism runs the following algorithm, which is referred to as the *top trading cycle (TTC)* algorithm. The TTC algorithm is due to Shapley and Scarf (1974), but they acknowledge (See Section 6 of the paper) that the algorithm was suggested to them by David Gale.

Fix an endowment of agents a^*. The mechanism maintains the remaining set of objects M^k and remaining set of agent N^k in every Step k of the mechanism.

- Step 1: Set $M^1 = M$ and $N^1 = N$. Construct a directed graph G^1 with nodes N^1. There is a directed edge from node (agent) $i \in N^1$ to agent $j \in N^1$ if and only if $\succ_i (1, M^1) = a^*(j)$.

  Allocate objects along every cycle of graph G^1. Formally, if $(i^1, i^2, \ldots, i^p, i^1)$ is a cycle in G^1 then set $a(i^1) = a^*(i^2), a(i^2) = a^*(i^3), \ldots, a(i^{p-1}) = a^*(i^p), a(i^p) = a^*(i^1)$. Let $\widehat{N}^1$ be the set of agents allocated in such cycles in G^1, and $\widehat{M}^1$ be the set of objects assigned in a to N^1.

  Set $N^2 = N^1 \setminus \widehat{N}^1$ and $M^2 = M^1 \setminus \widehat{M}^1$.

- Step k: Construct a directed graph G^k with nodes N^k. There is a directed edge from node (agent) $i \in N^k$ to agent $j \in N^k$ if and only if $\succ_i (1, M^k) = a^*(j)$.

Allocate objects along every cycle of graph G^k. Formally, if $(i^1, i^2, \ldots, i^p, i^1)$ is a cycle in G^k then set $a(i^1) = a^*(i^2), a(i^2) = a^*(i^3), \ldots, a(i^{p-1}) = a^*(i^p), a(i^p) = a^*(i^1)$. Let $\widehat{N}^k$ be the set of agents allocated in such cycles in G^k, and $\widehat{M}^k$ be the set of objects assigned in a to N^k.

Set $N^{k+1} = N^k \setminus \widehat{N}^k$ and $M^{k+1} = M^k \setminus \widehat{M}^k$. If N^{k+1} is empty, STOP, and a is the final matching chosen. Else, repeat.

Note that each TTC mechanism is defined by an endowment matching. We show below that each TTC mechanism is strategy-proof and efficient.

PROPOSITION **11** *A TTC with fixed endowment mechanism is strategy-proof and efficient.*

Proof: Consider agent i who wants to deviate. Suppose agent i is getting assigned in Step k of the TTC mechanism if he is truthful. Hence, agent k gets the best object among M^k (set of objects remaining in Step k).

Consider any Step j, where $j \leq k$. We argue that agent i cannot gain by misreporting her preference such that her outgoing edge changes in Step j. This of course will prove that any manipulation can never given an object in $M \setminus M^k$, thus completing the proof of strategy-proofness.

We say agent i' *chases* agent i in Step j if there exists a sequence of edges $(i', i^1), (i^1, i^2), \ldots, (i^k, i)$ in the directed graph in Step j. Note that when agent i is truthful, i is available in Step k. Hence, if i^k points to i in Step j (i.e., the endowment of agent i is the best object of agent i^k in M^j), it continues to point to i in Step k. So, i^k is available in Step k. Repeating this argument along the path $(i', i^1), (i^1, i^2), \ldots, (i^k, i)$, we conclude that i' is available in Step k. Hence, i does not strictly prefer the endowment of agent i' to its matching when she is truthful. As a consequence of this argument, we can conclude that agent i has no incentive to manipulate so that she is matched to the endowment of agent i', where i' chases her in Step j. But agent i cannot be assigned to the endowment of agent i'' if agent i'' is not chasing her in Step j. Hence, agent i cannot gain by misreporting her preference such that her outgoing edge changes in Step j.

Now, we prove efficiency. Let a be a matching produced by the TTC mechanism for preference profile $\succ$. Assume for contradiction that this matching is not efficient, i.e., there exists a different matching a' such that $a'(i) \succ_i a(i)$ or $a'(i) = a(i)$ for all $i \in N$. Consider

the first step of the TTC mechanism where some agent i gets $a(i) \neq a'(i)$. Since all the agents get the same object in a and a' before this step, object $a'(i)$ is available in this step, and since $a'(i) \succ_i a(i)$, agent i cannot have an edge from i to the "owner" of $a(i)$ in this step. This means that agent i cannot be assigned to $a(i)$. This gives a contradiction. ■

Note that a TTC mechanism need not be a dictatorship. To see this, suppose there are three agents and three houses. Fix an endowment a^* as $a^*(i) = a_i$ for all $i \in \{1, 2, 3\}$. Let us examine the TTC mechanism corresponding to a^*. Consider the profile $(\succ_1, \succ_2, \succ_3)$ such that $\succ_i (1, N) = a_1$ for all $i \in \{1, 2, 3\}$, i.e., every agent has object a_1 as his top ranked object. Clearly, only agent 1 gets one of this top ranked alternatives (matchings) in this profile according to this TTC mechanism. Now, consider the profile $(\succ'_1, \succ'_2, \succ'_3)$ such that $\succ'_i (1, N) = a_2$ for all $i \in \{1, 2, 3\}$, i.e., every agent has object a_2 as his top ranked object. Then, only agent 2 gets one of his top ranked alternatives (matchings) according to this TTC mechanism. Hence, this TTC mechanism is not a dictatorship.

Further, TTC mechanism violates *neutrality*, the axiom used to characterize fixed priority mechanisms in Theorem 32. A characterization of TTC with fixed endowment is still illusive.

8.1.3 Stable House Allocation with Existing Tenants

We consider a variant of the house allocation problem. In this model, each agent already has a house that he owns - if an agent i owns house j then he is called the tenant of j. This is the model studied in Shapley and Scarf (1974). Immediately, one sees that the TTC mechanism can be applied in this setting with initial endowment given by the house-tenant relationship. This is, as we have shown, strategy-proof and efficient (Proposition 11).

We address another concern here, that of *stability*. In this model, agents own resources that are allocated. So, it is natural to impose some sort of stability condition on the mechanism. Otherwise, a group of agents can break away and trade their houses amongst themselves.

Consider the example in Table 8.1. Let the existing tenants of the houses be given by matching a^*: $a^*(1) = a_1, a^*(2) = a_3, a^*(3) = a_2, a^*(4) = a_4, a^*(5) = a_5, a^*(6) = a_6$. Consider a matching a as follows: $a(i) = a_i$ for all $i \in N$. Now consider the coalition of agents $\{3, 4\}$. In the matching a, we have $a(3) = a_3$ and $a(4) = a_4$. But agents 3 and 4 can reallocate the houses they own among themselves in a manner to get a better matching for themselves.

In particular, agent 3 can get a_4 (house owned by agent 4) and agent 4 can get a_2 (house owned by agent 3. Note that $a_4 \succ_3 a_3$ and $a_2 \succ_4 a_4$. Hence, both the agents are better off trading among themselves. So, they can potentially *block* matching a. We formalize this idea of blocking below.

Let a^* denote the matching reflecting the initial endowment of agents. We will use the notation a^S for every $S \subseteq N$, to denote a matching of agents in S to the houses owned by agents in S. Whenever we write a matching a without any superscript we mean a matching of all agents. Formally, a coalition (group of agents) $S \subseteq N$ can **block** a matching a at a preference profile $\succ$ if there exists a matching a^S such that $a^S(i) \succ_i a(i)$ or $a^S(i) = a(i)$ for all $i \in S$ with $a^S(j) \succ_j a(j)$ for some $j \in S$. A matching a is in the **core** at a preference profile $\succ$ if no coalition of agents can block a at $\succ$. A social choice function f is **stable** if for all preference profile $\succ$, $f(\succ)$ is in the core at preference profile $\succ$. Note that stability implies efficiency - efficiency requires that the grand coalition cannot block.

We will now analyze if the TTC mechanism is stable. Note that when we say a TTC mechanism, we mean the TTC mechanism where the initial endowment is the endowment given by the house-tenant relationship.

THEOREM **33** *The TTC mechanism is stable. Moreover, there is a unique core matching for every preference profile.*

Proof: Assume for contradiction that the TTC mechanism is not stable. Then, there exists a preference profile $\succ$, where the matching a produced by the TTC mechanism at $\succ$ is not in the core. Let coalition S block this matching a at $\succ$. This means there exists another matching a^S such that $a^S(i) \succ_i a(i)$ or $a^S(i) = a(i)$ for all $i \in S$, with equality not holding for all $i \in S$. Let $T = \{i \in S : a^S(i) \succ_i a(i)\}$. Assume for contradiction $T \neq \emptyset$.

To remind notation, we denote $\widehat{N}^k$ to be the set of agents allocated houses in Step k of the TTC mechanism, and $\widehat{M}^k$ be the set of these houses. Clearly, agents in $S \cap \widehat{N}^1$ are getting their respective top ranked houses. So, $(S \cap \widehat{N}^1) \subseteq (S \setminus T)$. Define $S^k = S \cap \widehat{N}^k$ for each stage k of the TTC mechanism. We now complete the proof using induction. Suppose $(S^1 \cup \ldots \cup S^{k-1}) \subseteq (S \setminus T)$ for some stage k. We show that $S^k \subseteq (S \setminus T)$. Now, agents in $S \cap \widehat{N}^k$ are getting their respective top ranked houses amongst houses in $M \setminus (\widehat{M}^1 \cup \ldots \cup \hat{M}^k)$. Given that agents in $(S^1 \cup \ldots \cup S^{k-1})$ get the same set of houses in a^S and a, any agent in S^k cannot be getting a better house in a^S than his house in a. Hence, again $S^k \subseteq (S \setminus T)$.

By induction, $S \subseteq (S \setminus T)$ or $T = \emptyset$, which is a contradiction.

Finally, we show that the core matching returned by the TTC mechanism is the unique one. Suppose the core matching returned by the TTC mechanism is a, and let a' be another core matching for preference profile $\succ$. Note that (a) in every Step k of the TTC mechanism agents in $\widehat{N}^k$ get allocated to houses owned by agents in $\widehat{N}^k$, and (b) agents in $\widehat{N}^1$ get their top ranked houses. Hence, if $a(i) \neq a'(i)$ for any $i \in \widehat{N}^1$, then agents in $\widehat{N}^1$ will block a'. So, $a(i) = a'(i)$ for all $i \in \widehat{N}^1$.

Now, we use induction. Suppose, $a(i) = a'(i)$ for all $i \in \widehat{N}^1 \cup \ldots \cup \widehat{N}^{k-1}$. We will argue that $a(i) = a'(i)$ for all $i \in \widehat{N}^k$. Agents in $\widehat{N}^k$ get their highest ranked house from $M \setminus \widehat{M}^1 \cup \ldots \cup \widehat{M}^{k-1}$. So, given that agents in $\widehat{N}^1 \cup \ldots \cup \widehat{N}^{k-1}$ get the same houses in a and a', if some agent $i \in \widehat{N}^k$ get different houses in a and a', then it must be $a(i) \succ_i a'(i)$. This means, agents in $\widehat{N}^k$ will block a'. This contradicts the fact that a' is a core matching.

This shows that $a = a'$, a contradiction. ■

The TTC mechanism with existing tenants has another nice property. Call a mechanism f **individually rational** if at every profile $\succ$, the matching $f(\succ) \equiv a$ satisfies $a(i) \succ_i a^*(i)$ or $a(i) = a^*(i)$ for all $i \in N$, where a^* is the matching given by the initial endowment or existing tenants.

Clearly, the TTC mechanism is individually rational. To see this, consider a profile $\succ$ and let $f(\succ) = a$. Note that the TTC mechanism has this property that if the house owned by an agent i is matched in Step k, then agent i is matched to a house in Step k too. If $a(i) \neq a^*(i)$ for some i, then agent i must be part of a trading cycle where he is pointing to a house better than $a^*(i)$. Hence, $a(i) \succ_i a^*(i)$.

This also follows from the fact that the TTC mechanism is stable and stability implies individual rationality - individual rationality means no coalition of single agent can block.

In the model of house allocation with existing tenants, the TTC mechanism satisfies three compelling properties along with stability - it is strategy-proof, efficient, and individually rational. Remarkably, these three properties characterize the TTC mechanism in the existing tenant model. We skip the proof.

THEOREM **34 (Ma (1994))** *A mechanism is strategy-proof, efficient, and individually rational if and only if it is the TTC mechanism.*

Proof: Since the TTC mechanism is strategy-proof, efficient, and individually rational, it

is enough to show that there is a unique mechanism satisfying these properties. The proof we give is due to Sethuraman (2016).

Let f be a strategy-proof, efficient, and individually rational mechanism and $f \neq f^{TTC}$, where f^{TTC} is the TTC mechanism. NOTATIONS. At any profile of preferences $\succ$, define

$$B_i(\succ_i) = \{a_k : a_k \succ_i a^*(i)\} \cup \{a^*(i)\} \qquad \forall\ i \in N$$

These are the set of houses agent i prefers to her endowment $a^*(i)$ and $a^*(i)$ itself.

Since $f \neq f^{TTC}$, there is some profile of preferences $\succ$ such that $f(\succ) \neq f^{TTC}(\succ)$. Choose $\succ$ such that $\sum_{i \in N} |B_i(\succ_i)|$ is minimized. Let $N^* := \{i \in N : f_i(\succ) \neq f_i^{TTC}(\succ)\}$. The proof works in two steps.

STEP 1. We show that $|B_i(\succ_i)| = 2$ for all $i \in N^*$. Pick any $i \in N^*$. First, $B_i(\succ_i) = \{a^*(i)\}$ is impossible because by individual rationality, $f_i(\succ) = f_i^{TTC}(\succ) = a^*(i)$, a contradiction to the definition of N^*. Hence, assume for contradiction $|B_i(\succ_i)| > 2$. There are two cases to consider.

CASE 1. Suppose $f_i(\succ) \ \succ_i \ f_i^{TTC}(\succ)$. Hence, by individual rationality, $f_i(\succ) \neq a^*(i)$.

Consider another preference $\succ_i'$ such that the top two objects $\succ_i'$ is $\{f_i(\succ), a^*(i)\}$, and the ranking of these two objects is the same as in $\succ_i$ (i.e., $f_i(\succ) \ \succ_i \ a^*(i)$). Suppose $f_i(\succ) = a_k, f_i^{TTC}(\succ) = a_t$ and $f_i(\succ_i', \succ_{-i}) = a_{k'}, f_i^{TTC}(\succ_i', \succ_{-i}) = a_{t'}$. By construction, top-ranked object in $\succ_i'$ is a_k followed by $a^*(i)$. By individual rationality $\{a_{k'}, a_{t'}\} \subseteq \{a_k, a^*(i)\}$. By strategy-proofness of f,

$$a_{k'} \ \succ_i' \ a_k \text{ or } a_{k'} = a_k.$$

If $a_{k'} \neq a_k$, then $a_{k'} = a^*(i)$. Hence, agent i can manipulate to $\succ_i$ at preference $\succ_i'$ to get $a_k \ \succ_i' \ a^*(i)$, which contradicts strategy-proofness. So, we have $a_{k'} = a_k$.[4] We argue that $a_{t'} = a^*(i)$. By our assumption (that $f_i(\succ) \ \succ_i \ f_i^{TTC}(\succ)$), we have $a_k \ \succ_i \ a_t$. By individual rationality, $a_{t'} \in \{a_k, a^*(i)\}$. Assume for contradiction $a_{t'} = a_k$. Then, agent i can manipulate in TTC when her preference is $\succ_i$ by reporting $\succ_i'$, a contradiction to strtaegy-proofness. Hence, $a_{t'} = a^*(i)$. As a result, $f_i(\succ_i', \succ_{-i}) \neq f_i^{TTC}(\succ_i', \succ_{-i})$. But $|B_i(\succ_i')| = 2$

[4]This argument is similar to the proof of Claim 6.

and $|B_i(\succ_i)| > 2$, contradicting minimality of $\succ$.

Case 2. Suppose $f_i^{TTC}(\succ) \ \succ_i \ f_i(\succ)$. Hence, by individual rationality, $f_i^{TTC}(\succ) \neq a^*(i)$. Consider another preference $\succ_i'$ such that the top two objects $\succ_i'$ is $\{f_i^{TTC}(\succ), a^*(i)\}$, and the ranking of these two objects is the same as in $\succ_i$ (i.e., $f_i^{TTC}(\succ) \ \succ_i \ a^*(i)$). The rest of the proof is identical to Case 1 by changing the position of f_i^{TTC} and f_i.

Step 2. Let $\bar{N}$ be the set of agents who strictly prefer their matched objects in f to f^{TTC} (according to $\succ$). Similarly, let $\tilde{N}$ be the set of agents who strictly prefer their matched objects in f^{TTC} to f (according to $\succ$). Since f and f^{TTC} are individually rational, for each $i \in \bar{N}$,

$$\succ_i (1, M) = f_i(\succ) \ \succ_i \ f_i^{TTC}(\succ) = a^*(i) = \succ_i (2, M).$$

Similarly, for each $i \in \tilde{N}$,

$$\succ_i (1, M) = f_i^{TTC}(\succ) \ \succ_i \ f_i(\succ) = a^*(i) = \succ_i (2, M).$$

Now, pick any $i \in \bar{N}$. We argue that $a^*(i)$ must be assigned to an agent in $\bar{N}$ in f. We know that i is assigned $a^*(i)$ in TTC. Suppose j is assigned to $a^*(i)$ in f. If $j \notin \bar{N} \cup \tilde{N}$, then j must be assigned $a^*(i)$ in TTC too, which is not the case. Hence, if $j \notin \bar{N}$, then $j \in \tilde{N}$. But this implies that $f_j(\succ) = a^*(j) \neq a^*(i)$, a contradiction.

This means that agents in $\bar{N}$ are assigned objects in $\cup_{i \in \bar{N}} a^*(i)$ in the mechanism f. A similar argument shows that agents in $\tilde{N}$ are assigned objects owned by agents in $\tilde{N}$ in TTC.

Clearly $\bar{N} \cup \tilde{N} \neq \emptyset$. Hence, $\bar{N}$ is non-empty or $\tilde{N}$ is non-empty. If $\bar{N}$ is non-empty, we observe the following. Agents in $\bar{N}$ prefer their matched object in f to their matched object in the TTC mechanism. Hence, the following matching Pareto dominates $f^{TTC}(\succ)$.

$$a(i) = \begin{cases} f_i(\succ) & \text{if } i \in \bar{N} \\ f_i^{TTC}(\succ) & \text{otherwise.} \end{cases}$$

Notice that a is a feasible matching because

- agents in $\bar{N}$ are assigned objects in $\cup_{i \in \bar{N}} a^*(i)$ in the mechanism f (we showed this)
- each agent $i \in \bar{N}$ is assigned $a^*(i)$ in f^{TTC} (by our construction)

- hence, agents in $N \setminus \bar{N}$ are assigned objects in $\cup_{i \in N \setminus \bar{N}} a^*(i)$ in f^{TTC}

This contradicts Pareto efficiency of TTC. Similarly, if $\tilde{N}$ is non-empty, then we can reallocated objects in $\tilde{N}$ according to the match in f^{TTC} to Pareto dominate the matching in f, contradicting Pareto efficiency of f. ■

Note that the serial dictatorship with a fixed priority is strategy-proof and efficient but not individually rational. The "status-quo mechanism" where everyone is assigned the houses they own is strategy-proof and individually rational but not efficient. So, the properties of individual rationality and efficiency are crucial for the characterization of Theorem 34.

8.1.4 Generalized TTC Mechanisms

In this section, we generalize the TTC mechanisms in a natural way so that one extreme covers the TTC mechanism we discussed and the other extreme covers the fixed priority mechanism. We can now handle the case where the number of objects is not equal to the number of agents. We now define **fixed priority TTC (FPTTC)** mechanisms. In a FPTTC mechanism, each house a_j is endowed with a priority $\sigma_j : N \to N$ over agents. This generates a profile of priorities $\sigma \equiv (\sigma_1, \ldots, \sigma_n)$. Every FPTTC mechanism is defined by a profile of priorities σ. We denote the FPTTC mechanism corresponding to a priority profile σ as f^σ.

The FPTTC mechanism then goes in stages, with each stage executing a TTC mechanism but the endowments in each stage changing with the fixed priority profile σ.

We first illustrate the idea with the example in Table 8.4.

$\succ_1$	$\succ_2$	$\succ_3$	$\succ_4$
a_3	a_2	a_2	a_1
a_2	a_3	a_4	a_4
a_1	a_4	a_3	a_3
a_4	a_1	a_1	a_2

Table 8.4: An example for housing model

Consider two priorities σ_1 and σ_2, where $\sigma_1(i) = i$ for all $i \in N$ and σ_2 is defined as $\sigma_2(1) = 2, \sigma_2(2) = 1, \sigma_2(3) = 4, \sigma_2(4) = 3$. Suppose houses a_1 and a_2 are assigned priority σ_1 but houses a_3 and a_4 are assigned priority σ_2.

In stage 1, the endowments are derived from the priorities of houses. Since houses a_1 and a_2 have agent 1 as top in their priority σ_1, agent 1 is endowed with these houses. Similarly, agent 2 is endowed houses a_3 and a_4 by priority σ_2. Now, the TTC phase of stage 1 begins. By the preferences of agents, each agent points to agent 1, except agent 1, who points to agent 2 (agent 2 is endowed house a_3, which is agent 1's top ranked house). So, trade takes place between agents 1 and 2. This is shown in Figure 8.3 - the endowments of agents are shown in square brackets. The edges also reflect which object it is pointing to.

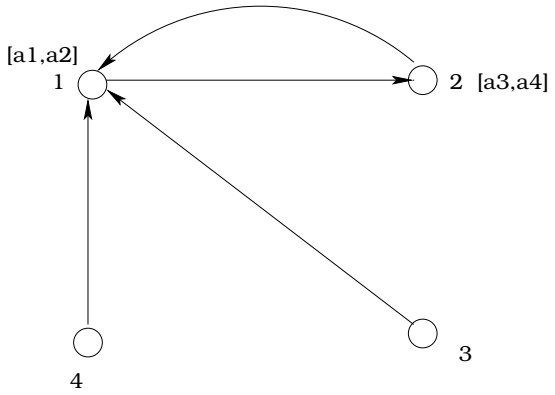

Figure 8.3: Cycle in stage 1 of the FPTTC mechanism

In the next stage, only agents 3 and 4 remain. Also, only houses a_1 and a_4 remain. We look at the priority of σ_1 of house a_1. Of the remaining agents, agent 3 is the top. Then, for priority σ_2 of house a_4, the top agent among remaining agent is agent 4. So, the new endowment is agent 3 gets a_1 and agent 4 gets a_4. We run the TTC phase now. Agent 3 points to agent 4 and agent 4 points to agent 3. So, they trade, and the FPTTC mechanism gives the following matching $\bar{a}$: $\bar{a}(1) = a_3, \bar{a}(2) = a_2, \bar{a}(3) = a_4, \bar{a}(3) = a_1$. This is shown in Figure 8.4.

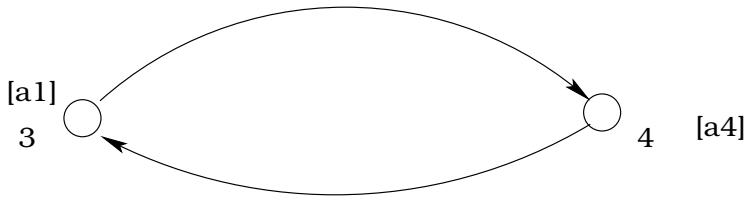

Figure 8.4: Cycle in stage 2 of the FPTTC mechanism

If all the houses have the same fixed priority, then we recover the fixed priority mechanism.

To see this, notice that because of identical priority of houses, all the houses are endowed to the same agent in every stage of the FPTTC mechanism. As a result, at stage i, the ith agent in the priority gets his top-ranked house. Hence, we recover the fixed priority (serial dictatorship) mechanism.

On the other extreme, if all the houses have priorities such that the top ranked agents in the priorities are distinct (i.e., for any two houses a_j, a_k with priorities σ_j and σ_k, we have $\sigma_j(1) \neq \sigma_k(1)$), then the endowments of the agents do not change over stages if the number of houses is equal to the number of agents. If there are more houses than number of agents, the endowment of each agent increases (in terms of set inclusion) across stages. So, we recover the traditional TTC mechanism for the case of equal number of agents and houses.

The following proposition can now be proved using steps similar to Proposition 11.

PROPOSITION **12** *For every priority profile σ, the FPTTC mechanism f^σ is strategy-proof and efficient.*

8.2 The Two-sided Matching Model

The house allocation model is a model of one-sided matching - only agents (one side of the market) had preference over the houses. In many situations, the matching market can be partitioned into two sides, and an agent on one side will have preference over agents on the other side. For instance, consider the scenario where students are matched to schools. It is plausible that not only students have preferences over the schools but schools also have a preferences over students. Other applications of two-sided matching include job applicants matched to firms, doctoral students matched to faculty etc.

Let M be a set of men and W be a set of women. For simplicity, we will assume that $|M| = |W|$ - but this is not required to derive the results. Every man $m \in M$ has a *strict* preference ordering $\succ_m$ over the set of women W. So, for $x, y \in W$, $x \succ_m y$ will imply that m ranks x over y. A matching is a bijective mapping $\mu : M \to W$, i.e., every man is assigned to a unique woman. If μ is a matching, then $\mu(m)$ denotes the woman matched to man m and $\mu^{-1}(w)$ denotes the man matched to woman w. This model is often called the "marriage market" model or "two-sided matching" model. We first discuss the stability aspects of this model, and then discuss the strategic aspects.

8.2.1 Stable Matchings in Marriage Market

As in the house allocation model with existing tenants, the resources to be allocated to agents in the marriage market model are owned by agents themselves. Hence, stability becomes an important criteria for designing any mechanism.

We consider an example with three men and three women. Let $M = \{m_1, m_2, m_3\}$ and $W = \{w_1, w_2, w_3\}$. Their preferences are shown in Table 8.5.

$\succ_{m_1}$	$\succ_{m_2}$	$\succ_{m_3}$	$\succ_{w_1}$	$\succ_{w_2}$	$\succ_{w_3}$
w_2	w_1	w_1	m_1	m_3	m_1
w_1	w_3	w_2	m_3	m_1	m_3
w_3	w_2	w_3	m_2	m_2	m_2

Table 8.5: Preference orderings of men and women

Consider the following matching μ: $\mu(m_1) = w_1, \mu(m_2) = w_2, \mu(m_3) = w_3$. This matching is *unstable* in the following sense. The pair $(m_1, \mu(m_2) = w_2)$ will *block* this matching (ex post) since m_1 likes w_2 over $\mu(m_1) = w_1$ and w_2 likes m_1 over $\mu^{-1}(w_2) = m_2$. So, (m_1, w_2) will break away, and form a new pair. This motivates the following definition of stability.

DEFINITION **57** *A matching μ is* **pairwise unstable** *at preference profile $(\succ)$ if there exists $m, m' \in M$ such that (a) $\mu(m') \succ_m \mu(m)$ and (b) $m \succ_{\mu(m')} m'$. The pair $(m, \mu(m'))$ is called a* **blocking pair** *of μ at $(\succ)$. If a matching μ has no blocking pairs at a preference profile $\succ$, then it is called a* **pairwise stable matching** *at $\succ$.*

The following matching μ' is a pairwise stable matching at $\succ$: $\mu'(m_1) = w_1, \mu'(m_2) = w_3, \mu'(m_3) = w_2$ for the example in Table 8.5. The question is: Does a pairwise stable matching always exist? The answer to this question is remarkably yes, as we will show next.

One can imagine a stronger requirement of stability, where groups of agents block instead of just pairwise blocking. We say that a coalition $S \subseteq (M \cup W)$ **blocks** a matching μ at a profile $\succ$ if there exists another matching μ' such that (i) for all $m \in M \cap S$, $\mu'(m) \in W \cap S$ and for all $w \in W \cap S$, $\mu'^{-1}(w) \in M \cap S$, and (ii) for all $m \in M \cap S$, $\mu'(m) \succ_m \mu(m)$ and for all $w \in M \cap S$, $\mu'^{-1}(w) \succ_w \mu^{-1}(w)$. We say a matching μ is in core at a profile $\succ$ if no coalition can block μ at $\succ$. The following theorem suggests that this notion of stability is equivalent to the pairwise notion of stability we have initially defined.

THEOREM **35** *A matching is pairwise stable at a profile if and only if it belongs to the core at that profile.*

Proof: Consider a matching μ which is pairwise stable at $\succ$. Assume for contradiction that μ is not in the core at $\succ$. Then, there must exist $S \subseteq (M \cup W)$ and a matching $\hat{\mu}$ such that for all $m \in M \cap S$ and for all $w \in W \cap S$ with $\hat{\mu}(m), \hat{\mu}^{-1}(w) \in S$ we have $\hat{\mu}(m) \succ_m \mu(m)$ and $\hat{\mu}^{-1}(w) \succ_w \mu^{-1}(w)$. This means for some $m \in S$ we have $\hat{\mu}(m) \in W \cap S$. Let $\hat{\mu}(m) = w$. We know $w \succ_m \mu(m)$. Then, we have $m \succ_w \mu^{-1}(w)$. Hence, (m, w) is a blocking pair at $\succ$ for μ. This implies that μ is not pairwise stable, which is a contradiction.

The other direction of the proof is trivial. ■

For this reason, we will say a matching is **stable** at a preference profile if it is pairwise stable at that preference profile. We will also drop that qualified "at a preference profile" at some places where the preference profile in question is clear from the context.

8.2.2 Deferred Acceptance Algorithm

In this section, we show that a stable matching always exists in the marriage market model. The fact that a stable matching always exists is proved by constructing an algorithm to find such a matching (this algorithm is due to David Gale and Lloyd Shapley, and also called the Gale-Shapley algorithm). There are two versions of this algorithm. In one version men propose to women and women either accept or reject the proposal. In another version, women propose to men and men either accept or reject the proposal. We describe the men-proposal version.

- S1. First, every man proposes to his top ranked woman.
- S2. Then, every woman who has at least one proposal keeps (tentatively) the top man amongst these proposals and rejects the rest.
- S3. Then, every man who was rejected in the last round, proposes to the top woman amongst those women who have not rejected him in earlier rounds.
- S4. Then, every woman who has at least two proposals, including any proposal tentatively kept from earlier rounds, keeps (tentatively) the top man amongst these proposals and rejects the rest. The process is then repeated from Step S3 till each woman

has a proposal, at which point, the tentative proposal accepted by a woman becomes permanent.

Since each woman is allowed to keep only one proposal in every round, no woman will be assigned to more than one man. Since a man can propose only one woman at a time, no man will be assigned to more than one woman. Since the number of men and women are the same, this algorithm will terminate at a matching. Also, the algorithm will terminate finitely since in every round, the set of women a man can propose does not increase, and strictly decreases for at least one man.

We illustrate the algorithm for the example in Table 8.5. A proposal from $m \in M$ to $w \in W$ will be denoted by $m \to w$.

- In the first round, every man proposes to his best woman. So, $m_1 \to w_2, m_2 \to w_1, m_3 \to w_1$.
- Hence, w_1 has two proposals: $\{m_2, m_3\}$. Since $m_3 \succ_{w_1} m_2$, w_1 rejects m_2 and keeps m_3.
- Now, m_2 is left to choose from $\{w_2, w_3\}$. Since $w_3 \succ_{m_2} w_2$, m_2 now proposes to w_3.
- Now, every woman has exactly one proposal. So the algorithm stops with the matching μ_m given by $\mu_m(m_1) = w_2, \mu_m(m_2) = w_3, \mu_m(m_3) = w_1$.

It can be verified that μ_m is a stable matching. Also, note that μ_m is a different stable matching than the stable matching μ' which we discussed earlier. Hence, there can be more than one stable matching.

One can also state a women proposing version of the deferred acceptance algorithm. Let us run the women proposing version for the example in Table 8.5. As before, a proposal from $w \in W$ to $m \in M$ will be denoted by $w \to m$.

- In the first round, every woman proposes to her top man. So, $w_1 \to m_1, w_2 \to m_3, w_3 \to m_1$.
- So, m_1 has two proposals: $\{w_1, w_3\}$. We note that $w_1 \succ_{m_1} w_3$. Hence, m_1 rejects w_3 and keeps w_1.

- Now, w_3 is left to choose from $\{m_2, m_3\}$. Since $m_3 \succ_{w_3} m_2$, w_3 proposes to m_3.
- This implies that m_3 has two proposals: $\{w_2, w_3\}$. Since $w_2 \succ_{m_3} w_3$, m_3 rejects w_3 and keeps w_2.
- Now, w_3 is left to choose only m_2. So, the algorithm terminates with the matching μ_w given by $\mu_w(m_1) = w_1, \mu_w(m_2) = w_3, \mu_w(m_3) = w_2$.

Note that μ_w is a stable matching and $\mu_m \neq \mu_w$.

8.2.3 Stability and Optimality of Deferred Acceptance Algorithm

THEOREM **36** *At every preference profile, the Deferred Acceptance Algorithm terminates at a stable matching for that profile.*

Proof: Consider the Deferred Acceptance Algorithm where men propose (a similar proof works if women propose) for a preference profile $\succ$. Let μ be the final matching of the algorithm. Assume for contradiction that μ is not a stable matching. This implies that there exists a pair $m \in M$ and $w \in W$ such that (m, w) is a blocking pair. By definition $\mu(m) \neq w$ and $w \succ_m \mu(m)$. This means that w rejected m earlier in the algorithm (else m would have proposed to w at the end of the algorithm). But a woman rejects a man only if she gets a better proposal, and her proposals improve in every round. This implies that w must be assigned to a better man than m, i.e., $\mu^{-1}(w) \succ_w m$. This contradicts the fact that (m, w) is a blocking pair. ■

The men-proposing and the women-proposing versions of the Deferred Acceptance Algorithm may output different stable matchings. Is there a reason to prefer one of the stable matchings over the other? Put differently, should we use the men-proposing version of the algorithm or the women-proposing version?

To answer this question, we start with some notations. A matching μ is **men-optimal stable** matching if μ is stable and for every other stable matching μ' we have $\mu(m) \succ_m \mu'(m)$ or $\mu(m) = \mu'(m)$ for all man $m \in M$. Similarly, a matching μ is **women-optimal stable** matching if μ is stable and for every other stable matching μ' we have $\mu^{-1}(w) \succ_w \mu'^{-1}(w)$ or $\mu^{-1}(w) = \mu'^{-1}(w)$ for all woman $w \in W$.

Note that by definition, a men-optimal stable matching is unique - if there are two men optimal stable matchings μ, μ', then they must differ by at least one man's match and this man must be worse in one of the matchings. Similarly, there is a unique women-optimal stable matching.

THEOREM **37** *The men-proposing version of the Deferred Acceptance Algorithm terminates at the unique men-optimal stable matching and the women-proposing version of the Deferred Acceptance Algorithm terminates at the unique women-optimal stable matching.*

Proof: We do the proof for men-proposing version of the algorithm. The proof is similar for the women-proposing version. Let $\hat{\mu}$ be the stable matching obtained at the end of the men-proposing Deferred Acceptance Algorithm. Assume for contradiction that $\hat{\mu}$ is not men-optimal. Then, there exists a stable matching μ such that for some $m \in M$, $\mu(m) \succ_m \hat{\mu}(m)$. Let $M' = \{m \in M : \mu(m) \succ_m \hat{\mu}(m)\}$. Hence, $M' \neq \emptyset$.

Now, for every $m \in M'$, since $\mu(m) \succ_m \hat{\mu}(m)$, we know that m is rejected by $\mu(m)$ in some round of the algorithm. Denote the round in which $m \in M'$ is rejected by $\mu(m)$ by t_m. Choose $m' \in \arg\min_{m \in M'} t_m$, i.e., choose a man m' who is the first to be rejected by $\mu(m')$ among all men in M'. Since $\mu(m')$ rejects m', she must have got a better proposal from some other man m'', i.e.,

$$m'' \succ_{\mu(m')} m'. \tag{8.1}$$

Now, consider $\mu(m')$ and $\mu(m'')$. If $m'' \notin M'$, then $\hat{\mu}(m'') = \mu(m'')$ or $\hat{\mu}(m'') \succ_{m''} \mu(m'')$. Since m'' is eventually assigned to $\hat{\mu}(m'')$, it must be the last woman that m'' must have proposed in DAA. The fact that m'' proposed to $\mu(m')$ earlier means $\mu(m') \succ_{m''} \hat{\mu}(m'')$. Using, $\hat{\mu}(m'') = \mu(m'')$ or $\hat{\mu}(m'') \succ_{m''} \mu(m'')$, we get

$$\mu(m') \succ_{m''} \mu(m'').$$

If $m'' \in M'$, then, since $t_{m''} > t_{m'}$ and the fact that m'' proposed to $\mu(m')$ in round $t_{m'}$ implies that m'' has not been rejected by $\mu(m'')$ till round $t_{m'}$. This means, again, m'' proposed to $\mu(m')$ before proposing to $\mu(m'')$. Hence, as in the earlier case, we get

$$\mu(m') \succ_{m''} \mu(m''). \tag{8.2}$$

By Equations 8.1 and 8.2, $(m'', \mu(m'))$ forms a blocking pair. Hence, μ is not stable. This is a contradiction. ■

We will denote the unique men-optimal stable matching as μ^M and the unique women-optimal stable matching as μ^W. The natural question is then whether there exists a stable matching that is optimal for both men and women. The answer is no. The example in Table 8.5 has two stable matchings, one is optimal for men but not for women and one is optimal for women but not for men.

We explore the structure of stable matchings a bit further.

THEOREM **38** *Let μ and μ' be a pair of stable matchings. Then, $\mu(m) \succ_m \mu'(m)$ or $\mu(m) = \mu'(m)$ for all $m \in M$ if and only if $\mu'^{-1}(w) \succ_w \mu^{-1}(w)$ or $\mu'^{-1}(w) = \mu^{-1}(w)$ for all $w \in W$.*

Proof: Let μ, μ' be two stable matchings with $\mu(m) \succ_m \mu'(m)$ or $\mu(m) = \mu'(m)$ for all $m \in M$. Assume for contradiction that $\mu^{-1}(w) \succ_w \mu'^{-1}(w)$ for some w. Suppose $m \equiv \mu^{-1}(w)$. By definition, $w \succ_m \mu'(m)$. Hence, (m, w) is a blocking pair of μ', contradicting the fact that μ' is stable.

A similar proof can be given to show the other direction. ■

Theorem 38 and Theorem 37 suggest that the men-optimal stable matching μ^M is the worst stable matching for women and the women-optimal stable matching μ^W is the worst stable matching for men. Indeed, we can now define a preference relation $\lhd$ on the set of stable matchings. For any pair of stable matchings μ, μ', we say $\mu \lhd \mu'$ if for every $m \in M$, $\mu(m) = \mu'(m)$ or $\mu(m) \succ_m \mu'(m)$ (or equivalently, $\mu'^{-1}(w) = \mu^{-1}(w)$ or $\mu'^{-1}(w) \succ_w \mu^{-1}(w)$ for every w). Note that $\lhd$ is not a complete relation. Similarly, we can define the relation $\rhd$ over the set of stable matchings by considering the preferences of women. In particular, for any pair of stable matchings μ, μ', we say $\mu \rhd \mu'$ if for every $\mu^{-1}(w) = \mu'^{-1}(w)$ or $\mu^{-1}(w) \succ_w \mu'^{-1}(w)$ for every w. The following is an immediate corollary to Theorems 37 and 38.

THEOREM **39** *For any stable matching μ, the following holds:*

$$\mu^M \lhd \mu \lhd \mu^W$$
$$\mu^W \rhd \mu \rhd \mu^M.$$

	m	m'
μ	w	$\mu(m')$
μ'	$\mu'(m)$	w
μ''	w	w

Table 8.6: μ'' is a matching

There is more to the structure of stable matchings. For any pair of stable matchings μ, μ', we construct another matching $\mu'' \equiv (\mu \vee^m \mu')$ as follows: for every $m \in M$, we define $\mu''(m) = \mu(m)$ if $\mu(m) \succ_m \mu'(m)$ or $\mu(m) = \mu'(m)$, and $\mu''(m) = \mu'(m)$ if $\mu'(m) \succ_m \mu(m)$. We write this equivalently as for all $m \in M$,

$$(\mu \vee^m \mu')(m) = \max_{\succ_m}(\mu(m), \mu'(m)).$$

It is not clear if μ'' is a matching. The following theorem shows that μ'' is a stable matching. Similarly, we can define the matching $\mu \vee^w \mu'$ as follows: for every $w \in W$, we define

$$(\mu \vee^w \mu')^{-1}(w) := \max_{\succ_w}(\mu^{-1}(w), \mu'^{-1}(w)).$$

The following theorem shows that these are stable matchings.

THEOREM **40** *For every pair of stable matchings μ, μ', $(\mu \vee^m \mu')$ and $(\mu \vee^w \mu')$ are also stable matchings.*

Proof: Take a pair of stable matchings μ, μ' and let $\mu'' \equiv (\mu \vee^m \mu')$. Assume for contradiction that μ'' is not a matching. Then, there must exist $m, m' \in M$ such that $\mu''(m) = \mu''(m')$.

Then, it must be that (see Table 8.6) there is some $w \in W$ with $w := \mu(m) = \mu'(m')$ and $w \succ_m \mu'(m)$ and $w \succ_{m'} \mu(m')$ - there is also a similar case where the role of μ and μ' is reversed. Since μ' is stable, then $m' \succ_w m$. But then (m', w) form a blocking pair of μ, contradicting the fact that μ is stable.

Next, we show that μ'' is a stable matching. Assume for contradiction (m, w) is a blocking pair of μ''. Suppose m is matched to w_1 and w is matched to m_1 in μ''. Hence, since (m, w) is a blocking pair

$$m \succ_w m_1, \quad \text{and} \quad w \succ_m w_1$$

	m	m_1
μ	w_1	$w?$
μ'	w_2	$w?$
μ''	w_1	w

Table 8.7: μ'' is a stable matching

Now, suppose $w_1 = \mu(m)$ and $w_2 = \mu'(m)$ - it is possible that $w_1 = w_2$. See Table 8.7.

By definition of μ'', $w_1 \succ_m w_2$ or $w_1 = w_2$. Hence, we conclude that

$$w \succ_m w_1 \quad \text{and} \quad w \succ_m w_2.$$

We now argue that w is not matched to m_1 in μ. Suppose w is matched to m_1 in μ. Note that $\mu(m) = w_1$. Since $m \succ_w m_1$ and $w \succ_m w_1$, (m, w) form a blocking pair for μ.

Similarly, w is not matched to m_1 in μ'. Suppose w is matched to m_1 in μ'. Note that $\mu'(m) = w_2$. Since $m \succ_w m_1$ and $w \succ_m w_2$, (m, w) form a blocking pair for μ'.

Since w is not matched to m_1 in both μ and μ', w and m_1 cannot be matched with each other in μ''. This is a contradiction.

A similar proof shows that $(\mu \vee^w \mu')$ is a stable matching. ■

For every pair of stable matchings μ and μ', we can also define the minimum matchings as $(\mu \wedge^m \mu')$ and $(\mu \wedge^w \mu')$, where $(\mu \wedge^m \mu')(m) = \min_{\succ_m}(\mu(m), \mu'(m))$ for all m and $(\mu \wedge^w \mu')^{-1}(w) = \min_{\succ_w}(\mu^{-1}(w), \mu'^{-1}(w))$ for all w. You are encouraged to think (a) if these are stable matchings and (b) their relationship with the matchings $(\mu \vee^m \mu')$ and $(\mu \vee^w \mu')$.

An example clarifies the results. Consider the preferences of four men and four women in Table 8.8. It can be verified that there are four stable matchings in this example:

$$\mu_1 : (m_1, w_1), (m_2, w_2), (m_3, w_4), (m_4, w_3)$$
$$\mu_2 : (m_1, w_2), (m_2, w_1), (m_3, w_4), (m_4, w_3)$$
$$\mu_3 : (m_1, w_1), (m_2, w_2), (m_3, w_3), (m_4, w_4)$$
$$\mu_4 : (m_1, w_2), (m_2, w_1), (m_3, w_3), (m_4, w_4)$$

Now, it is easy to verify that $\mu_2 \vee^m \mu_3 = \mu_1$ and $\mu_2 \vee^w \mu_3 = \mu_4$. Indeed μ_1 is the men-optimal stable matching and μ_4 is the women-optimal stable matching.

$\succ_{m_1}$	$\succ_{m_2}$	$\succ_{m_3}$	$\succ_{m_4}$	$\succ_{w_1}$	$\succ_{w_2}$	$\succ_{w_3}$	$\succ_{w_4}$
w_1	w_2	w_4	w_3	m_2	m_1	m_3	m_4
w_2	w_1	w_3	w_4	m_1	m_2	m_4	m_3
w_3	w_3	w_1	w_2	m_3	m_3	m_2	m_1
w_4	w_4	w_2	w_1	m_4	m_4	m_1	m_2

Table 8.8: Preference orderings of men and women

8.2.4 Unequal number of men and women

In the model in the previous section, we assumed that $|M| = |W|$. This was only for simplicity. As far as stability is concerned, none of the results change if $|M| \neq |W|$. To see this, assume $|M| > |W|$. Then, introduce a set of *dummy* women D such that $|D| = |M| - |W|$. For every dummy woman, $d \in D$, endow her with some arbitrary preference $\succ_d$ over M. For every $m \in M$, extend his preference $\succ_m$ over $W \cup D$ such that for every $w \in W$ and every $d \in D$, we have $w \succ_m d$. Call this the *extended marriage market*. The extended marriage market has the same number of men and women. The DAA produces a stable matching of the extended marriage market. Further, the men-proposing (women-proposing) DAA produces men-optimal (women-optimal) stable matching.

A matching in the extended marriage market has the following interpretation: if a man is matched to a dummy woman, then he is not matched to any woman. We argue that the matching μ produced in the extended marriage market is a stable matching in the original market. To see, assume for contradiction some (m, w) blocks this matching. So, $m \succ_w \mu^{-1}(w)$ and $w \succ_m \mu(m)$. If m is unmatched in μ, then it is matched to some dummy woman d in extended marriage market. Since w is preferred to being unmatched, it is also preferred to being the dummy woman d. Hence, (m, w) also block in the extended marriage market. If m is matched in μ, then it is matched to a non-dummy woman in the extended market. This means that (m, w) also block in the extended marriage market. So, in both cases, we have a contradiction since μ is a stable matching of the extended marriage market. A similar analysis shows that the men-optimal and women-optimal stable matchings in the extended marriage market corresponds to the men-optimal and women-optimal stable matchings in the original market.

The following result shows what happens if $|M| \neq |W|$. In this case, we will assume that

every man (woman) prefers to be matched to some woman (man) than to remain unmatched.

THEOREM **41** *The set of matched agents is the same at every stable matching.*

Proof: Let μ^M be the men-optimal stable matching and μ be any arbitrary stable matching. Let $\widetilde{M}$ and $\widetilde{W}$ be the set of men and women matched in μ^M respectively. Similarly, let $\widehat{M}$ and $\widehat{W}$ be the set of men and women matched in μ. By definition

$$|\widetilde{M}| = |\widetilde{W}| \quad |\widehat{M}| = |\widehat{W}|$$

Since μ^M is men-optimal stable matching, any man matched in μ must be matched in μ^M. So, $\widehat{M} \subseteq \widetilde{M}$. Hence, we get

$$|\widetilde{W}| = |\widetilde{M}| \geq |\widehat{M}| = |\widehat{W}| \tag{8.3}$$

By Theorem 38, every woman prefers μ to μ^M. As a result, any woman matched in μ^M is also matched in μ. Hence, $\widetilde{W} \subseteq \widehat{W}$. Hence, $|\widehat{W}| \geq |\widetilde{W}|$. Combining with (8.3), we get

$$|\widetilde{W}| = |\widetilde{M}| = |\widehat{M}| = |\widehat{W}|$$

As a result, $\widehat{M} = \widetilde{M}$ and $\widehat{W} = \widehat{M}$. ■

A result similar in spirit to Theorem 41 is the following.

THEOREM **42** *Suppose w is ranked last by every man at a preference profile. Then w is matched to the same man in every stable matching. Similarly, suppose m ranked last by every woman at a preference profile. Then m is matched to the same woman in every stable matching.*

Proof: We do the proof for man, and a similar proof works for woman. Suppose m is ranked last by all women at a preference profile. We will argue that at every stable matching the match of m is the same as that in the men-optimal stable matching μ^M. By Theorem 41, if m is unmatched in μ^M, it is unmatched in all stable matchings, and we are done. Suppose m is matched to w_1 in an arbitrary stable matching μ. Assume for contradiction that m is not matched to w_1 in μ^M. Then, w_1 is matched to some $m_1 \neq m$ in μ^M. Further, suppose m_1 is matched to w_1' in μ. The matchings are shown in Table 8.9.

μ	μ^M
(m, w_1)	
(m_1, w_1')	(m_1, w_1)

Table 8.9: Stable matchings μ and μ^M

Since (m_1, w_1) do not block μ and $m_1 \succ_{w_1} m$ (since m is last ranked), it must be that

$$w_1' \succ_{m_1} w_1$$

This implies m_1 strictly prefers his match in μ to his match in μ^M, a contradiction to men-optimality of μ^M. ■

Theorems 41 and 42 are famously referred to as the *rural hospital theorem*. An important application of matching is to match doctors to hospitals. Often, hospitals in rural areas are the worst choice for doctors. Theorem 41 and 42 says that if we choose a stable matching, then these hospitals see the same matching in all stable matchings.

8.2.5 Strategic Issues in Deferred Acceptance Algorithm

We next turn to strategic properties of the Deferred Acceptance Algorithm (DAA). We first consider the men-proposing version. We define the notion of strategyproofness informally here. Strategyproofness is with respect to the direct revelation mechanism. The DAA is strategy-proof if reporting a non-truthful preference ordering does not result in a better outcome for an agent for any reported preferences of other agents.

We first show that the men-proposing version of the Deferred Acceptance Algorithm is not strategyproof for women (i.e., women can manipulate). Let us return to the example in Table 8.5. We know if everyone is truthful, then the matching is: $\mu(m_1) = w_2, \mu(m_2) = w_3, \mu(m_3) = w_1$. We will show that w_1 can get a better outcome by not being truthful. We show the steps here.

- In the first round, every man proposes to his best woman. So, $m_1 \to w_2, m_2 \to w_1, m_3 \to w_1$.

- Next, w_2 only has one proposal (from m_1) and she accepts it. But w_1 has two proposals: $\{m_2, m_3\}$. If she is truthful, she should accept m_3. We will see what happens if she is not truthful. So, she accepts m_2.

- Now, m_3 has two choices: $\{w_2, w_3\}$. He likes w_2 over w_3. So, he proposes to w_2.

- Now, w_2 has two proposals: $\{m_1, m_3\}$. Since she likes m_3 over m_1, she accepts m_3.

- Now, m_1 has a choice between w_1 and w_3. Since he likes w_1 over w_3, he proposes to w_1.

- Now, w_1 has two proposal: $\{m_1, m_2\}$. Since she prefers m_1 over m_2 she accepts m_1.

- So, m_2 is only left with $\{w_2, w_3\}$. Since he likes w_3 over w_2 he proposes to w_3, which she accepts. So, the final matching $\hat{\mu}$ is given by $\hat{\mu}(m_1) = w_1, \hat{\mu}(m_2) = w_3, \hat{\mu}(m_3) = w_2$.

Hence, w_1 gets m_1 in $\hat{\mu}$ but was getting m_3 earlier. The fact that $m_1 \succ_{w_1} m_3$ shows that not being truthful helps w_1. However, the same result does not hold for men. Similarly, the women-proposing DAA is not strategy-proof for men.

THEOREM **43** *The men-proposing version of the Deferred Acceptance Algorithm is strategyproof for men. The women-proposing version of the Deferred Acceptance Algorithm is strategyproof for women.*

Proof: Suppose there is a profile $\pi = (\succ_{m_1}, \ldots, \succ_{m_n}, \succ_{w_1}, \ldots, \succ_{w_n})$ such that man m_1 can misreport his preference to be $\succ_*$, and obtain a better matching. Let this preference profile be π'. Let μ be the stable matching obtained by the men-proposing deferred acceptance algorithm when applied to π. Let ν be the stable matching obtained by the men-proposing algorithm when applied to π'. We show that if $\nu(m_1) \succ_{m_1} \mu(m_1)$, then ν is not stable at π', which is a contradiction.

Let $R = \{m : \nu(m) \succ_m \mu(m)\}$. Since $m_1 \in R$, R is not empty. We show that $\{w : \nu^{-1}(w) \in R\} = \{w : \mu^{-1}(w) \in R\}$ - in other words, the set of women matched to men in R is the same in μ and ν. Take any w such that $\nu^{-1}(w) \in R$, we will show that $\mu^{-1}(w) \in R$, and this will establish the claim. If $\mu^{-1}(w) = m_1$, then we are done by definition. Else, let $w = \nu(m)$ and $m' = \mu^{-1}(w)$ with $m' \neq m_1$. Since $w \succ_m \mu(m)$, stability of μ at π implies

that $m' \succ_w m$. Stability of ν at π' implies that $\nu(m') \succ_{m'} w$. Therefore, $m' \in R$. Let $S = \{w : \nu^{-1}(w) \in R\} = \{w : \mu^{-1}(w) \in R\}$. Note that for any $w \in S$, $\mu^{-1}(w) \neq \nu^{-1}(w)$ (if this is true, then $m := \mu^{-1}(w) = \nu^{-1}(w)$, and since $m \in R$ this gives $w \succ_m w$).

By definition $\nu(m) \succ_m \mu(m)$ for all $m \in R$. By stability of μ (at π), we then have $\mu^{-1}(w) \succ_w \nu^{-1}(w)$ for all $w \in S$. Now, pick any $w \in S$. By definition, $w \succ_{\nu^{-1}(w)} \mu(\nu^{-1}(w))$. This implies that during the execution of the men-proposing deferred acceptance algorithm at π, $\nu^{-1}(w) \in R$ must have proposed to w which she had rejected.

Let $m \in R$ be the last man in R to make a proposal and get accepted during the execution of the men-proposing deferred acceptance algorithm at π. Suppose this proposal is made to $w \equiv \mu(m) \in S$. As argued, w rejected $\nu^{-1}(w)$ earlier. This means that when m proposed to w, she had some tentative matching in DAA, say to m', which she rejected. By definition, m' cannot be in R - because m' must have proposed after m. This means that $m' \neq \nu^{-1}(w)$, and since m' was accepted by w after rejecting $\nu^{-1}(w)$,

$$m' \succ_w \nu^{-1}(w).$$

Since $m' \notin R$, $\mu(m') \succ_{m'} \nu(m')$ or $\mu(m') = \nu(m')$. Also, since m' proposed to w before proposing to $\mu(m')$, $w \succ_{m'} \mu(m')$. This shows that

$$w \succ_{m'} \nu(m').$$

This shows that (m', w) form a blocking pair for ν at π'. ■

Does this mean that no mechanism can be both stable and be strategyproof to all agents? The answer is yes. We give a proof when agents do not necessarily rank themselves at the bottom. Hence, now, a man m ranks all the women in W and himself. Similarly, a woman w ranks all the men in M and herself.

THEOREM **44** *No mechanism which gives a stable matching can be strategy-proof for both men and women.*

Proof: Consider the following preference profile of two men and two women.

$$\succ_{m_1}: w_1 \ \succ_{m_1} \ w_2 \ \succ_{m_1} \ m_1$$
$$\succ_{m_2}: w_2 \ \succ_{m_2} \ w_1 \ \succ_{m_2} \ m_2$$
$$\succ_{w_1}: m_2 \ \succ_{w_1} \ m_1 \ \succ_{w_1} \ w_1$$
$$\succ_{w_2}: m_1 \ \succ_{w_2} \ m_2 \ \succ_{w_2} \ w_2$$

At this profile, there are two stable matchings: $(m_1, w_1), (m_2, w_2)$ and $(m_1, w_2), (m_2, w_1)$. If there is a stable mechanism, it must choose one of them.

Suppose it chooses the men-optimal one: $(m_1, w_1), (m_2, w_2)$. Suppose w_1 misrepresents her preference to be

$$\succ'_{w_1}: m_2 \ \succ_{w_1} \ w_1 \ \succ_{w_1} \ m_1$$

In the new preference profile $(\succ_{m_1}, \succ_{m_2}, \succ'_{w_1}, \succ_{w_2})$, the unique stable matching is $(m_1, w_2), (m_2, w_1)$. Hence, the mechanism must select it. But then, w_1 is better off by manipulation.

Suppose it chooses the women-optimal one: $(m_1, w_2), (m_2, w_1)$. Suppose m_1 misrepresents her preference to be

$$\succ'_{m_1}: w_1 \ \succ_{m_1} \ m_1 \ \succ_{m_1} \ w_2$$

Then, there is a unique stable mechanism $(m_1, w_1), (m_2, w_2)$ at this preference profile. This means m_1 is better off by manipulating. ■

However, one can trivially construct strategy-proof mechanisms for both men and women. Consider a mechanism which ignores all men (or women) orderings. Then, it can run a fixed priority mechanism for men (or women) or a TTC mechanism with fixed endowments for men (or women) to get a strategy-proof mechanism.

8.2.6 College Admission Problem

The deferred acceptance algorithm can be suitably modified to handle the generalization of college admission problems. In a college admission problem, there is a set of students S. There is a set of colleges C. Each student s has a preference $\succ_s$ over set of colleges. We assume these preferences are strict orderings.

Each college c has a quota $q_c \geq 1$, which is the maximum number of students it can take. One approach is to model preferences over subsets of students for each college. We assume that this preference over subsets is "responsive" to preference over singleton students in the following way. We allow for the fact that a college may find a particular student "unacceptable" – so, not admitting a student may be preferred over admitting a particular student. For a matching μ, we denote by $\mu(s)$, the college assigned to student s. By $\mu(c)$, we denote the *set of* students assigned to college c. We will assume that $\mu(c)$ only contains acceptable students of college c (this is a weak form of stability condition).

DEFINITION **58** *Preference P_c of college c over subsets of students satisfy* **responsiveness** *if*

1. *for all $T \subseteq S$ with $|T| < q_c$ and for all $s \in S \setminus T$,*

$$(T \cup \{s\}) \; P_c \; T \text{ if and only if } \{s\} \; P_c \; \emptyset$$

2. *for all $T \subseteq S$ with $|T| = q_c$ and for all $s \in S$ and $s' \notin S \setminus T$,*

$$(T \setminus \{s'\} \cup \{s\}) \; P_c \; T \text{ if and only if } \{s\} \; P_c \; \{s'\}$$

In other words, choice of taking an extra student depends on whether that student is acceptable. Further, choice between two students depend on the ranking of those two students only.

The deferred acceptance algorithm can be modified in a straightforward way in these settings. The student proposing version works as follows. Each student proposes to its favorite remaining acceptable school. A college c evaluates the set of proposals it has, and accepts the top subset of students from the proposals. The procedure is repeated as was described earlier. One can extend the stability, student-optimal stability, and strategy-proofness results of previous section to this setting in a straightforward way. Here, we give a formal definition of stability.

DEFINITION **59** *A student, college pair (s, c)* **blocks** *matching μ if $c \succ_s \mu(s)$ and*

$$\begin{aligned} \{s\} \; P_c \; \{s'\} \quad & \text{for some } s' \in \mu(c) \quad \text{if } |\mu(c)| = q_c \\ \{s\} \; P_c \; \emptyset \quad & \text{if } |\mu(c)| < q_c \end{aligned}$$

A matching is **stable** *if there are no blocking pairs.*

THEOREM **45** *The student-proposing DAA produces a stable matching.*

Proof: Let S^t and S^{t+1} be the subset of students accepted by a college c in round t and $(t+1)$ with $|S^t| = |S^{t+1}| = q_c$. We first prove the following. If $s \notin S^t$ and for every $s' \in S^t$, $\{s'\}\ P_c\ \{s\}$, then for every $s'' \in S^{t+1}$, we have $\{s''\}\ P_c\ \{s\}$. To see this, pick $s \notin S^t$ such that for every $s' \in S^t$, $\{s'\}\ P_c\ \{s\}$. Then, pick $s'' \in S^{t+1}$. If $s'' \in S^t$, we are done. Else, $s'' \notin S^t$. Then, there must exist a student $\hat{s} \in S^t$ which is not there in S^{t+1} (because $|S^t| = |S^{t+1}| = q_c$). Hence, $S^{t+1}\ P_c\ S^{t+1} \setminus \{s''\} \cup \{\hat{s}\}$. By responsiveness, $\{s''\}\ P_c\ \{\hat{s}\}$. But, we know that $\{\hat{s}\}\ P_c\ \{s\}$ for every $s \notin S^t$. By transitivity of P_c, we get $\{s''\}\ P_c\ \{s\}$ for every $s \notin S^t$.

Next, we argue that if at any round, the subset of accepted students S^t by college c is such that $|S^t| = q_c$, then in every future round $\hat{t} > t$, we have $|S^{\hat{t}}| = q_c$. Suppose not. Then, there is some round t' such that $|S^{t'}| = q_c$ but $|S^{t'+1}| < q_c$. But the students in $S^{t'}$ are acceptable and available in $t'+1$. Since $|S^{t'+1}| < q_c$, there must exist some acceptable student $s \in S^t$ which is not in $S^{t'+1}$. By responsivness, $S^{t'+1} \cup \{s\}\ P_c\ S^{t'+1}$, a contradiction to the fact that college c chose the best set of students in $S^{t'+1}$.

Now, assume for contradiction that the DAA did not produce a stable matching. Then, there exists a pair (s, c) that can block the DAA matching μ. Then, student s must have proposed to c and got rejected. Let S_c be the students tentatively accepted by c at that time.

If $|S_c| < q_c$, it must be that c is unacceptable to s (else, by responsive preferences, $S_c \cup \{s\}$ is preferred to S_c). But this cannot be because (s, c) blocks μ. Hence, $|S_c| = q_c$. In that case, by our earlier arguments, the set of students accepted in any future rounds must have size q_c. Further, for every $\hat{s}$ accepted in future round, we must have $\{\hat{s}\}\ P_c\ \{s\}$. If $\hat{S}_c$ be the set of students matched to college c at the end of DAA, by responsive preferences, $\hat{S}_c\ P_c\ \hat{S}_c \setminus \{\hat{s}\} \cup \{s\}$ for each $\hat{s} \in \hat{S}_c$. Hence, (s, c) cannot block, a contradiction. ■

Similarly, one can prove that the college-proposing DAA produces a stable matching and the lattice properties continue to hold in the college admission problem.

8.3 Two-sided matching with priorities

Now, consider a matching model, where agents have preferences over objects and objects have *priorities* over agents. So, if N is the set of agents, then each $i \in N$ has a strict preference ordering $\succ_i$ over the set of objects in M. Similarly, each object $a_k \in M$ has a *priority* (a strict ranking) of the set of agents. An object may be a school, which can admit more than one student (agent). In that case, we can assume *responsive priorities*.

This formulation of two-sided matching is slightly different from the formulation we had earlier. Here, the priorities of the objects are *not* preferences – they are assumed to be common knowledge and given exogenously. For instance, this can be some rules imposed by regulators on how schools can have priorities over students. This allows us to use two mechanisms in this model.

- **The agent-proposing DAA**. This is strategy-proof for agents. It is stable with respect to preferences of agents and priorities of objects. Indeed, it gives the agent-optimal stable matching.
- **The fixed priority TTC**. This is strategy-proof for agents. It is Pareto efficient with respect to preferences of agents.

There is a potential tradeoff between stability and Pareto efficiency as the following example illustrates. Suppose there are three agents $N = \{1, 2, 3\}$ and three objects $M = \{a_1, a_2, a_3\}$. The preferences and priorities are shown in Table 8.10. In this problem, if we run agent-proposing DAA or object-proposing DAA we get the same matching: $\mu(1) = a_1, \mu(2) = a_2, \mu(3) = a_3$. Hence, this is the unique stable matching. But this is not Pareto efficient for agents: the matching μ' given by $\mu'(1) = a_2, \mu'(2) = a_1, \mu'(3) = a_3$ dominates matching μ with respect to preferences of agents. Indeed, μ' is the matching obtained when we run the FPTTC mechanism with priority profile σ. Notice that μ' can be blocked by $(3, a_1)$. This shows that the FPTTC may not be stable and DAA may not be Pareto efficient.

σ_{a_1}	σ_{a_2}	σ_{a_3}	$\succ_1$	$\succ_2$	$\succ_3$
1	2	2	a_2	a_1	a_1
3	1	1	a_1	a_2	a_2
2	3	3	a_3	a_3	a_3

Table 8.10: Two-sided matching with priorities

Bibliography

Kenneth J Arrow. The property rights doctrine and demand revelation under incomplete information. In *Economics and human welfare*, pages 23–39. 1979.

Larry Ausubel and Paul Milgrom. Ascending proxy auctions. In Peter Cramton, Yoav Shoham, and Richard Steinberg, editors, *Combinatorial Auctions*, chapter 3, pages 266–290. MIT Press, Cambridge, MA, 2006.

Helmut Bester and Roland Strausz. Imperfect commitment and the revelation principle: the multi-agent case. *Economics Letters*, 69(2):165–171, 2000.

Helmut Bester and Roland Strausz. Contracting with imperfect commitment and the revelation principle: the single agent case. *Econometrica*, 69(4):1077–1098, 2001.

Duncan Black. On the rationale of group decision-making. *The Journal of Political Economy*, pages 23–34, 1948.

J. Bulow and P. Klemperer. Auctions versus negotiations. *American Economic Review*, 86: 180–194, 1996.

Ruggiero Cavallo. Optimal decision-making with minimal waste: Strategyproof redistribution of vcg payments. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, pages 882–889. ACM, 2006.

Shurojit Chatterji and Arunava Sen. Tops-only domains. *Economic Theory*, 46(2):255–282, 2011.

Edward H Clarke. Multipart pricing of public goods. *Public choice*, 11(1):17–33, 1971.

Peter Cramton, Robert Gibbons, and Paul Klemperer. Dissolving a partnership efficiently. *Econometrica: Journal of the Econometric Society*, pages 615–632, 1987.

Jacques Cremer and Richard P McLean. Full extraction of the surplus in bayesian and dominant strategy auctions. *Econometrica: Journal of the Econometric Society*, pages 1247–1257, 1988.

Claude d'Aspremont and Louis-André Gérard-Varet. Incentives and incomplete information. *Journal of Public economics*, 11(1):25–45, 1979.

Claude d'Aspremont, Jacques Crémer, and Louis-André Gérard-Varet. Balanced bayesian mechanisms. *Journal of Economic Theory*, 115(2):385–396, 2004.

Geoffroy De Clippel. Behavioral implementation. *American Economic Review*, 104(10): 2975–3002, 2014.

Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *The American economic review*, 97(1):242–259, 2007.

Allan Gibbard. Manipulation of voting schemes: a general result. *Econometrica: journal of the Econometric Society*, pages 587–601, 1973.

Jerry R Green and Jean-Jacques Laffont. *Incentives in public decision making*. North-Holland, 1979.

Theodore Groves. Incentives in teams. *Econometrica: Journal of the Econometric Society*, pages 617–631, 1973.

Mingyu Guo and Vincent Conitzer. Worst-case optimal redistribution of vcg payments in multi-unit auctions. *Games and Economic Behavior*, 67(1):69–98, 2009.

John C Harsanyi. Games with incomplete information played by ?bayesian? players part i. the basic model. *Management Science*, 14(3):159–182, 1968a.

John C Harsanyi. Games with incomplete information played by ?bayesian? players part ii. bayesian equilibrium points. *Management Science*, 14(5):320–334, 1968b.

John C Harsanyi. Games with incomplete information played by'bayesian'players, part iii. the basic probability distribution of the game. *Management Science*, 14(7):486–502, 1968c.

Sergiu Hart and Philip J Reny. Maximal revenue with multiple goods: Nonmonotonicity and other observations. *Theoretical Economics*, 10(3):893–922, 2015.

Leonid Hurwicz. *Optimality and informational efficiency in resource allocation processes*. Stanford University Press, 1960.

Jean-Jacques Laffont and Eric Maskin. A differential approach to dominant strategy mechanisms. *Econometrica*, pages 1507–1520, 1980.

Yan Long, Debasis Mishra, and Tridib Sharma. Balanced ranking mechanisms. *Games and Economic Behavior*, 105:9–39, 2017.

Jinpeng Ma. Strategy-proofness and the strict core in a market with indivisibilities. *International Journal of Game Theory*, 23(1):75–83, 1994.

Alejandro M Manelli and Daniel R Vincent. Multidimensional mechanism design: Revenue maximization and the multiple-good monopoly. *Journal of Economic theory*, 137(1):153–185, 2007.

R Preston McAfee and Philip J Reny. Correlated information and mecanism design. *Econometrica: Journal of the Econometric Society*, pages 395–421, 1992.

Debasis Mishra and Tridib Sharma. A simple budget-balanced mechanism. *Social Choice and Welfare*, 50(1):147–170, 2018.

Hervé Moulin. On strategy-proofness and single peakedness. *Public Choice*, 35(4):437–455, 1980.

Hervé Moulin. Almost budget-balanced vcg mechanisms to assign multiple objects. *Journal of Economic theory*, 144(1):96–119, 2009.

Michael Mussa and Sherwin Rosen. Monopoly and product quality. *Journal of Economic theory*, 18(2):301–317, 1978.

Roger B Myerson. Incentive compatibility and the bargaining problem. *Econometrica*, pages 61–73, 1979.

Roger B Myerson. Optimal auction design. *Mathematics of operations research*, 6(1):58–73, 1981.

Roger B Myerson and Mark A Satterthwaite. Efficient mechanisms for bilateral trading. *Journal of economic theory*, 29(2):265–281, 1983.

Szilvia Pápai. Strategyproof assignment by hierarchical exchange. *Econometrica*, 68(6): 1403–1433, 2000.

Marek Pycia and M Utku Ünver. Incentive compatible allocation and exchange of discrete resources. *Theoretical Economics*, 12(1):287–329, 2017.

Rene Saran. Menu-dependent preferences and revelation principle. *Journal of Economic Theory*, 146(4):1712–1720, 2011.

Mark A Satterthwaite and Hugo Sonnenschein. Strategy-proof allocation mechanisms at differentiable points. *The Review of Economic Studies*, 48(4):587–597, 1981.

Mark Allen Satterthwaite. Strategy-proofness and arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of economic theory*, 10(2):187–217, 1975.

Arunava Sen. Another direct proof of the gibbard–satterthwaite theorem. *Economics Letters*, 70(3):381–385, 2001.

Jay Sethuraman. An alternative proof of a characterization of the ttc mechanism. *Operations Research Letters*, 44(1):107–108, 2016.

Lloyd Shapley and Herbert Scarf. On cores and indivisibility. *Journal of mathematical economics*, 1(1):23–37, 1974.

Tayfun Sönmez and M Utku Ünver. Matching, allocation, and exchange of discrete resources. In *Handbook of social Economics*, volume 1, pages 781–852. 2011.

Roland Strausz. Deterministic mechanisms and the revelation principle. *Economics Letters*, 79(3):333–337, 2003.

Lars-Gunnar Svensson. Strategy-proof allocation of indivisible goods. *Social Choice and Welfare*, 16(4):557–567, 1999.

William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 16(1):8–37, 1961.