

**Algorithmic Game Theory**  
**Prof. Palash Dey**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture - 44**  
**Gibbard-Satterthwaite Theorem**

Welcome. So, in the last lecture, we have studied some very important properties of social choice function. Namely, x efficiency or ex-post efficiency or it is also called Pareto optimality. Then, we talked about non-dictatorship social choice functions. Then, we talked about individual rationality. And under individual rationalities, we have studied 3 kinds of individual rationality; ex-post individual rational, interim individual rationality, and ex post individual rationality.

(Refer Slide Time: 01:00)

(1973) (1977)  
Gibbard-Satterthwaite Theorem

Lecture 9.4

For any player  $i \in [n]$  and any type  $\theta_i \in \Theta_i$ , we get a partial order  $R_i$  on the set  $X$  of outcomes.

$$x, y \in X, \quad x R_i y \stackrel{\text{def}}{=} u_i(x, \theta_i) \geq u_i(y, \theta_i)$$

$R_i$  is called the rational preference relation of player  $i$  when its type is  $\theta_i$ .

So, in today's class, we will characterize the social choice functions that are implementable in dominant strategy equilibrium and that famous theorem is called Gibbard-Satterthwaite Theorem. Discovered by both of them individually, Gibbard in 1973 and Satterthwaite in 1977.

So, what is the setup? So, we need to use some different notation to state the theorem. But once, but the theorem is really simple it is just the statement for to state the theorem we need some more notation. So, we first observe that for any player  $i \in [n]$  and any

type of that player,  $\theta_i \in \Theta_i$ , we get a relation, a partial order on the set of outcomes, set  $X$  of outcomes.

What is it? And let us call this partial order relation to be  $R_i$ . So, if I take two outcomes  $x, y \in X$ ;  $x R_i y$ , we define it, this is this definition. We say that  $x R_i y$ , if utility of player  $i$  in the outcome  $x$ , when its type is  $\theta_i$ . So, here is the first assumption of Gibbard-Satterthwaite theorem.

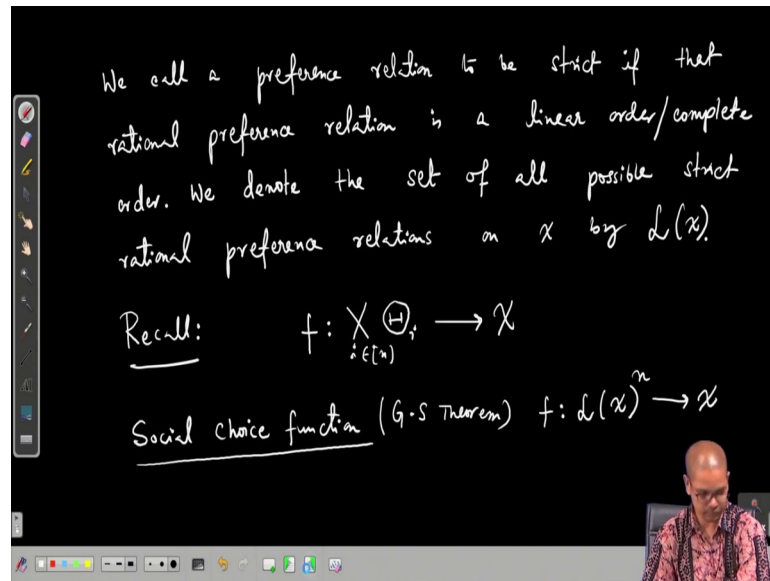
Although, these assumptions we will only make the theorem stronger, because it is a impossibility result. We will show that you know even under the under these assumptions; that means, under this very restricted setting also, very few things are implementable in dominant strategy equilibrium.

So, whatever assumption we put, it is only going to make the theorem stronger not weaker, because if something is impossible under a very restrictive scenario, under a very special kind of situations, then from the general situation also it is not possible so, in that sense.

So, the first assumption is that the utility of player  $i$ , depends only on the outcome  $x$  and the type of that particular player not on the type of other players. So, it is a function of outcome and  $\theta_i$ . It does not depend on  $\theta_{-i}$ . This is greater than equal to  $u_i(y, \theta_i)$ .

So, this  $R_i$  truly speaking is a function of not only player  $i$ , but of, but it depends on the type of player  $i$   $\theta_i$  also. So,  $R_i(\theta_i)$  you can see. So, it basically says that, at in terms of from player  $i$ 's perspective when its type is  $\theta_i$ , which outcome it likes more than other outcome and so on. That is the thing. This  $R_i(\theta_i)$  is called the preference the rational preference relation of player  $i$ , when its type is  $\theta_i$ , ok.

(Refer Slide Time: 06:52)



For G-S theorem, we will make another assumption. And so, let us say, we call preference relation to be strict if that rational preference relation, the word rational sort of wants to point to the fact that it depends on the utility of player  $i$  because the rationality is encoded in terms of the utility of player  $i$ .

Rational preference, the preference relation if the is a linear order or complete order. What does it mean? It means that the player  $i$  is not indifferent between any two of the outcomes; and there exists a best outcome, there is a least outcome, there is a second best outcome. There is no tie in the order of this outcomes to player  $i$  when it is when its type is  $\theta_i$  and its order of  $X$ , ok.

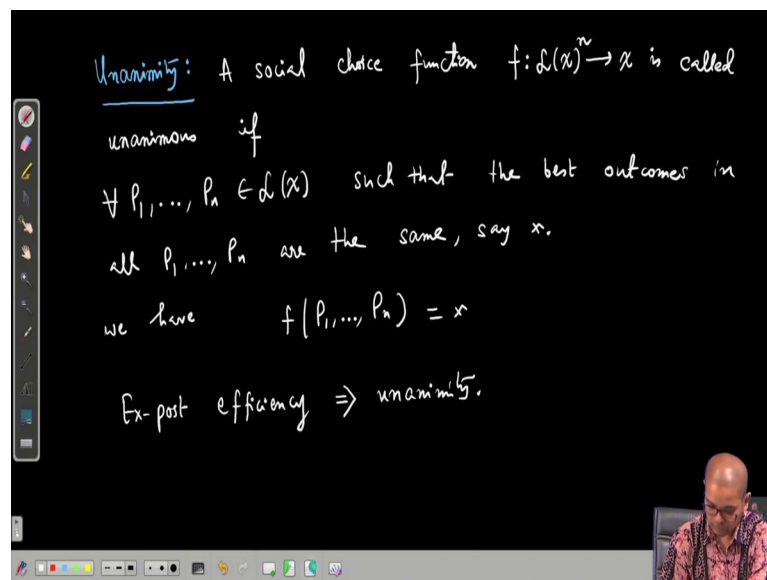
So, we denote the set of all set of all possible strict rational preference relation, we denote that, we denote the set of all possible strict rational preference relations on  $x$  by  $L(X)$ . Now, you see, recall that the social choice function was a function from the type profiles, product of type profiles  $(\theta_i)_{i \in [n]}$ .

Now, because now you focus on the type set of one particular players, say capital  $\theta_i$ , and if you focus on any particular type  $\theta_i$  of that  $\Theta_i$ , the rational, so thus the outcome if depends on  $\theta_i$  also and the rational preference relation depends on that  $\theta_i$ .

So, in the context of G-S theorem, we will assume that all preference relations are strict and this social choice function depends only on those strict preference orders. So, which is again an assumption. But again because let me repeat because Gibbard-Satterthwaite theorem is an impossibility theorem. It is a theorem about saying that much nothing much can be done. Then, all the assumptions the more the assumptions we have it will only going to make the theorem stronger.

If something is not, it is like if something is not possible, it is like if you cannot study with many helps then without that help also you cannot study. Something like that. So, social choice functions social choice function for Gibbard-Satterthwaite theorem. It is directly is a function of  $L(X)^n \rightarrow X$  ok. Very good. Now, we state few again few properties of social choice functions, when we view social choice functions as a function directly on the strict rational preference relations.

(Refer Slide Time: 12:55)



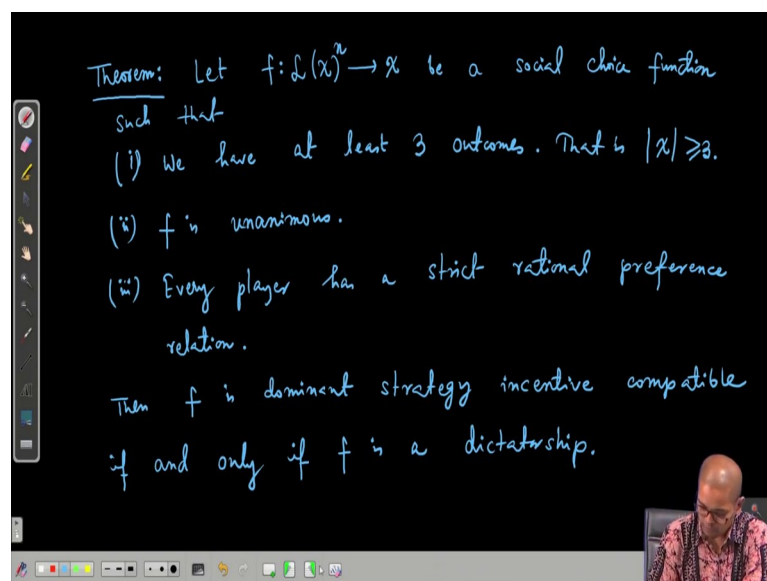
So, our first definition is what is called unanimity. Unanimity, a social choice function this particular property of unanimity makes sense only for social choice functions, which directly works on those strict rational preference relations. That means, social choice functions in the context of Gibbard-Satterthwaite theorem. For general social choice function we will see that this definition does not make any sense.

A social choice function, if from  $L(X)^n \rightarrow X$  is called unanimous, if for all strict preference relations  $P_1, \dots, P_n \in L(X)$ , such that the best outcome in all  $P_1, \dots, P_n$ , best outcomes in are the same, say  $x$ .

So, if I take  $n$  linear orders on the outcomes and the best outcome is same  $x$ , then for all such  $P_1, \dots, P_n$ , we have  $f(P_1, \dots, P_n)$  is equal to  $x$ . So, that means, if all players agree on the best possible outcome, then the social choice function must pick that filter pick that particular outcome.

You will see that the social choice with the property of ex-post efficiency, ex-post efficiency implies unanimity. In particular or equivalently, social choice function, if it is not unanimous then it is not ex-post efficient. These two are the same statement; although, a social choice function can be unanimous, but still not ex-post efficient. That is possible. So, what is what does Gibbard-Satterthwaite theorem say?

(Refer Slide Time: 16:48)



Theorem: Let  $f: L(X)^n \rightarrow X$  be a social choice function such that

- (i) We have at least 3 outcomes. That is  $|X| \geq 3$ .
- (ii)  $f$  is unanimous.
- (iii) Every player has a strict rational preference relation.

Then  $f$  is dominant strategy incentive compatible if and only if  $f$  is a dictatorship.

So, theorem, so let  $x$  be a social choice function, such that 3 properties should be satisfied. 1st property is we have at least 3 outcomes. That is cardinality  $x$  is greater than equal to 3. The 2nd one is, the social choice function  $f$  is unanimous; that means,  $f$  satisfies unanimity,  $f$  is unanimous. And the 3rd condition is every player has strict rational preference relation.

Then, the social choice function  $f$  or more formally the mechanism implementing  $f$  is dominant strategy incentive compatible if and only if,  $f$  is a dictatorship. So, with these 3 assumptions, it says that if you take a social choice function, and you want it to be dominant strategy incentive compatible; that means, the direct mechanism implementing it, is dominant strategy incentive compatible, then  $f$  must be a dictatorship function. And it is a if and only statement.

One direction is easy. So, if a  $f$  is a dictatorship function; that means, there exists a dictator, a player such that for all type profile, the outcome chosen by the, chosen by the social choice function is the best possible for that particular player. Then, it is it must be dominant strategy incentive compatible. Because the social choice function in some sense in plain English, it does not consider, it is not even looking at the preference of other players except the dictator.

So, for dictator there is no incentive to lie; it can report its true ranking, true type which is in this case true ordering of the outcomes or outcomes. And the social choice function will anyway pick the best one for the dictator. So, for dictator there is no incentive for lying, and for other players the social choice function just do not care. So, it does not matter unless the dictator changes his input his ranking and which he would not because the social choice function will always pick the best outcome of the dictator.

Then, there is no point of changing the or misreporting the ranking for other players. For other players not to; it does not have any incentive to misreport because the social choice function is simply going to ignore them. So, truth telling or revealing true ranking true type is the best strategy for dictator because social choice function will pick the top candidate according to that ranking. And it is the reporting true ranking is the best strategy for other players also because lying will not help. The social choice function simply dis disregards that.

And this theorem is very striking; it says that unless this social choice function is such a unfair dictatorship kind of function, it is not dominant strategy incentive compatible. But what does it mean? It means that it must be understood very carefully and clearly. A social choice function is dominant strategy incentive compatible, if you know at every play for all strategy profile, means truth telling is sort of the dominant strategy for all the players, irrespective of what other players does.

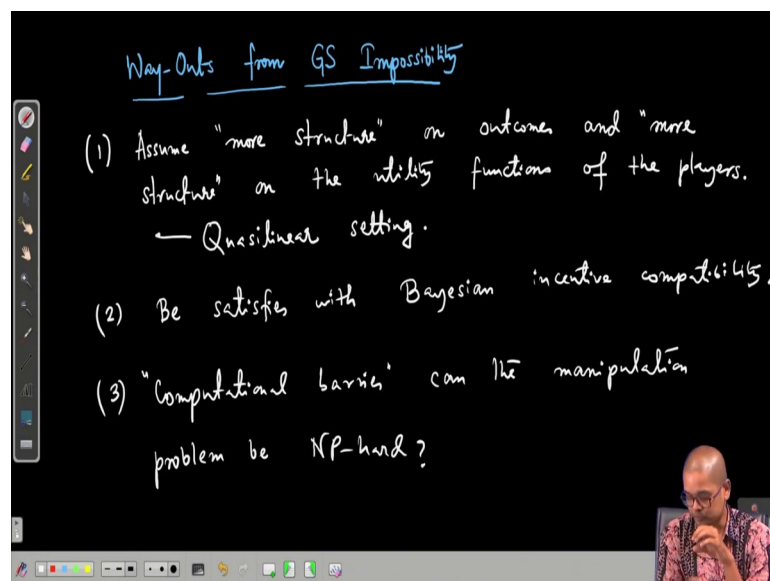
So, it just says that there exist a type profile, where there may exist certain player for which you know under certain circumstances truth telling or reporting true ranking may not be the best of his or her interest. It does not say that for all type profile, for all players, it is always better to misreport the preference. It does not say that.

As social choice function being not demonstrating incentive compatible, does not mean that for all profile all players will benefit by line. No, it is not that. It just says that there exist at least one preference profile or strategy profile, and at least one player for whom it is better to misreport his or her preference and get better outcome. And make the social choice function  $f$  choose the better outcome, if  $f$  is not dictatorship.

Then, it will not be dominant strategy incentive compatible. And this is what it means. There are multiple proofs for this Gibbard-Satterthwaite theorem. And it ranges, it varies from different complexities the proof that I have in my lecture notes is simple, completely accessible, but little long.

And in the interest of time, we will not go into the proof. It is not very much relevant for the course also, but if you are interested you can look at the lecture notes and also you can Google it out and find many proofs of Gibbard-Satterthwaite theorem. So, we are not going into that. What we are interested in is what is next.

(Refer Slide Time: 25:21)



Way-Outs from GS Impossibility

- (1) Assume "more structure" on outcomes and "more structure" on the utility functions of the players.  
— Quasilinear setting.
- (2) Be satisfied with Bayesian incentive compatibility.
- (3) "Computational barriers" can the manipulation problem be NP-hard?

The image shows a blackboard with handwritten text in blue and white. The title is 'Way-Outs from GS Impossibility'. There are three numbered points. A small video inset in the bottom right corner shows a man speaking.

So, way-out from way-outs from G-S impossibility. The 1st and most popular and successful way-out is called, assume “more structured” on outcomes. Outcomes are not just arbitrary any element from a set of all possible outcomes, no. There are there is certain kind of structure on the outcomes. So, we assume more structures on outcomes, and more structure on the utility functions of the players.

These structures, what structures or outcomes? What structures and utility functions? This is the topic of our next lectures discussion. And this is called quasi linear setting. This called quasi linear setting. And this is among the most successful approaches to escape the Gibbard-Satterthwaite impossibility.

The 2nd approach is we can let go the requirement of being dominant strategy incentive compatible, and be satisfied with Bayesian incentive compatibility. It is not so successful and very little or not so much can be done in this direction. The 3rd direction is what is called; you can say the computational barrier, ok. So, social choice function is not dominant strategy incentive compatible. It means that there exist a type profile and there exists a player for whom it is better to misreport or lie about his true type.

But like what? So, this particular job is called say manipulation. Manipulating a social choice function, where I am not a player is manipulating a social choice function, if he is he or she is not reporting her true type, his or her true type and is misreporting something.

These are this called the manipulation problem. And what if; can the manipulation problem be NP hard? And the idea is that although theoretically there exist a type profile, where there exist a player for whom there exists another type which it if it reports that type, then its outcome is, then it will be more happy, its utility will increase, but how to find that outcome for that particular player?

In reality, all players are computationally limited. So, if that computational task is very hard, it is if it is intractable, then players will not be able to find. And there is also much kind of much research, and this is also I would say sort of okish kind of successful, not as successful as quasi linear setting. So, yes, there are very and various other approaches. So, we will see, we will start exploring some of this approaches from the next class. So, we will stop here, ok.



Thank you.