Introduction to Multi-Agent Programming

13. Social Choice Theory

Arrow's Theorem

Games, Voting, Auctions, Mechanisms

Alexander Kleiner, Bernhard Nebel

Contents

- Introduction
- Game theory, social choice theory, voting, auctions, and mechanism design
- Social choice functions and social welfare functions
- Proving Arrow's result
- Incentive compatibility for social choice functions
- Gibbard-Satterthwaite's impossibility result
- Mechanism design
- Summary

Introduction

- We have learned about
  - Game Theory
  - Voting
  - Auctions
- What is the common theme behind all that?
- Groups of self-interested, rationale agents
  - who have to choose between options
  - in order to reach a common result
  - that (hopefully) brings maximal payoff to the individual

Social Choice Theory

- … is part of Game Theory
- How to aggregate individual preferences to a group preference or group decision (as in voting)
- Can also be viewed as a strategic game
- Each player
  - has private preferences
  - can declare preferences
  - based on the declared preferences, the group preference is computed
- It can make sense for a player to declare preferences strategically (i.e. to lie about one's preferences)
- Could this also occur in multi-agent/multi-robot teams?

Social Choice & Manipulation

- In Social Choice Theory, one designs and analyzes methods of how to aggregate (truly declared) individual preferences.
- If it makes sense for an agent to lie about her preferences (to manipulate the election), it becomes unclear what is aggregated.
- If most agents lie, the outcome may be in fact completely inefficient (violating the preferences of most group members)
- Is it possible to design social choice methods that avoid the problem of manipulation?

The Significance of Arrow's Result

- The result states that we cannot have "voting methods" (with at least 2 voters and at least 3 alternatives) that satisfy
  - Pareto efficiency (PE), i.e., $o \prec_i o'$ for all i then $o \prec o'$,
  - independence of irrelevant alternatives (IIA), i.e., the outcome $o \prec o'$ does not depend on how the agents order other alternatives
  - non-dictatorship, i.e., there is no single agent, such that the method always results in preferences identical to the one of the agent
- Since all voting methods have to satisfy PE and non-dictatorship, IIA will not be satisfied
- This gives us room to manipulate the outcome!
- This means that there is no way to come up with a "voting method" that cannot be manipulated (we will make that formal)

Auctions as Social Choice

- Auctions can be viewed as a way of aggregating the preferences of the bidders
- The social choice is about how the goods are allocated
- Should be as efficient as possible, i.e., the ones who values the goods the most should get the goods
- Can we get the bidders to state their true preferences?

- Yes, by introducing the notion of money they have to pay and by using the right mechanism, e.g. as implemented in some auction protocols, one can encourage the players to say the truth

Mechanism Design: Making Agents truthful

- Similar to auctions, we want that agents state their true preferences
- Mechanism design is the part of Game theory that is concerned about that
- Designing games by introducing payments so that telling the truth becomes the dominant strategy
- Based on the truly stated preferences one can then maximize the social welfare

Social Welfare Functions and Social Choice Functions

- Let L be the set of strict, linear orders (i.e. total, transitive and asymmetric binary relations) over the set of alternative A
- Elements of L are preferences of an agent i, denoted by $\prec_i$.
- $\pi = (\prec_1, \ldots, \prec_i, \ldots, \prec_n)$ is called the preference profile for agents $1, \ldots, n$
- A social welfare function F aggregates preferences to a group preference: $F: L^n \rightarrow L$
- A social choice functions f aggregates to a group choice: $f: L^n \rightarrow A$

Arrow's Conditions

- *Pareto efficiency* (or *unanimity*)
  - If for a profile $\pi = (\prec_1, \ldots, \prec_i, \ldots, \prec_n)$ such that $F(\pi) = \prec$ it is the case that $a \prec_i b$ for all *i,* then it shall hold that $a \prec b$
- *Independence of irrelevant alternatives (IIA)*
  - For all preferences $\prec_1, \ldots, \prec_i, \ldots, \prec_n, \prec, \prec'_1, \ldots, \prec'_i, \ldots, \prec'_n, \prec'$ with $F(\prec_1, \ldots, \prec_i, \ldots, \prec_n) = \prec$ and $F(\prec'_1, \ldots, \prec'_i, \ldots, \prec'_n) = \prec'$ s.t. $a \prec_i b$ *iff* $a \prec'_i b$, then it shall hold that $a \prec b$ *iff* $a \prec' b$
- *Non-dictatorship*
  - *It is not always the case that* $F(\prec_1, \ldots, \prec_i, \ldots, \prec_n) = \prec_i$ for a fixed player *i*

Arrow's Impossibility Result

Theorem: If a social welfare function over *at least three alternatives* with *at least two players* satisfies Pareto efficiency and independence of irrelevant alternatives, then it cannot satisfy non-dictatorship.

- In other words, PE, IIA, and non-dictatorship cannot hold at the same time
- Proof uses only elementary arguments over preference profiles that are constructed

Pairwise Neutrality

Lemma: Let *F* be a social welfare function satisfying *IIA* and *PE* with $F(\prec_1, \ldots, \prec_n) = \prec$ and $F(\prec'_1, \ldots, \prec'_n) = \prec'$ such that $a \prec_i b$ *iff* $c \prec'_i d$. Then it holds that $a \prec b$ *iff* $c \prec d$.

- If we rename *a* and *b* to *c* and *d, respectively,* then IIA should hold regardless.

Proof of Pairwise Neutrality

1. Consider two profiles ($\preceq_i$) and ($\preceq'_i$) over at least *a, b, c, d* *s.t.* $F(\preceq_i) = \preceq$ and $F(\preceq'_i) = \preceq'$ and assume w.l.g. $a \prec b$ (otherwise exchange *a* and *b*) and $b \neq c$ (otherwise exchange *a* and *c* exchange *b* and *d*).
2. Construct a new profile ($\preceq''_i$) with $F(\preceq''_i) = \preceq''$ such that
   - $c \prec''_i a$ and $b \prec''_i d$ for all *i*
   - the order between a and b is from $\preceq_i$
   - the order between c and d is from $\preceq'_i$
3. Because of PE, we have $c \prec'' a$ *and* $b \prec'' d$.
4. Because of IIA, we have $a \prec'' b$
5. With transitivity, it follows that $c \prec'' d$.
6. Again, with IIA, we have $c \prec' d$.
7. Other direction analog

qed

Proof of Arrow's Theorem (1)

1. Consider *n* players and two alternatives *a, b* with *a≠b.*
2. Construct a series of profiles $\pi^i=(\prec_j)$ *s.t.* in the *i*th profile exactly the first i players prefer *b* over *a*, i.e. $a \prec_j b$ *iff* $j \leq i$

	π^0	...	...	π^n
$1:$	$b \prec_1 a$	...	...	$a \prec_1 b$
$\vdots$	$\vdots$	$\ddots$	$\ddots$	$\vdots$
$i^*-1:$	$b \prec_{i^*-1} a$	...	...	$a \prec_{i^*-1} b$
$i^*:$	$b \prec_{i^*} a$	...	...	$a \prec_{i^*} b$
$\vdots$	$\vdots$	$\ddots$	$\ddots$	$\vdots$
$n:$	$b \prec_n a$	...	...	$a \prec_n b$
$F:$	$b \prec^0 a$	...	...	$a \prec^n b$

3. At some point i^*, one has to change from $b \prec^j a$ to $a \prec^j b$
4. We will show that i^* is a dictator!

Proof of Arrow's Theorem (2)

- For dictatorship of i^*, we have to show that for all possible preference profiles, it is always the case that for two alternatives c and d, $c \prec_{i^*} d$ implies that $c \prec d$ *for* $F(\prec_1, \ldots, \prec_{i^*}, \ldots, \prec_n) = \prec$.
- Take such a profile, consider a third element $e \notin \{c,d\}$ and construct a new profile ($\prec'_i$) such that
  - for $j < i^*$:
    - $e \prec'_j c \prec'_j d$ if $c \prec_j d$
    - $e \prec'_j d \prec'_j c$ if $d \prec_j c$
  - *for* $j = i^*$
    - $c \prec'_j e \prec'_j d$ if $c \prec_j d$ (which we have!)
    - $d \prec'_j e \prec'_j c$ if $d \prec_j c$
  - for $i^* < j$:
    - $c \prec'_j d \prec'_j e$ if $c \prec_j d$
    - $d \prec'_j c \prec'_j e$ if $d \prec_j c$

Proof of Arrow's Theorem (3)

- From the construction, we get the following profile that resembles the profiles π^{i^*-1} and π^{i^*}

	π^{i^*-1}	$(\prec'_i)_{i=1,\ldots,n}$	π^{i^*}	$(\prec'_i)_{i=1,\ldots,n}$
$1:$	$a \prec_1 b$	$e \prec_1 c$	$a \prec_1 b$	$e \prec_1 d$
$i^*-1:$	$a \prec_{i^*-1} b$	$e \prec_{i^*-1} c$	$a \prec_{i^*-1} b$	$e \prec_{i^*-1} d$
$i^*:$	$b \prec_{i^*} a$	$c \prec_{i^*} e$	$a \prec_{i^*} b$	$e \prec_{i^*} d$
$n:$	$b \prec_n a$	$c \prec_n e$	$b \prec_n a$	$d \prec_n e$
$F:$	$b \prec^{i^*-1} a$		$a \prec^{i^*} b$	

1. By pairwise neutrality (consider a and b in π^{i^*-1}), we must have $c \prec' e$
2. Similarly (consider a and b in π^{i^*}), we get $e \prec' d$.
3. By transitivity, we get $c \prec' d$.
4. Hence, by IIA, it follows that $c \prec d$.

qed

Does Arrow's Result also Effect Elections?

- Usually, in elections we just have one candidate to elect (and not an ordered list)
- Arrow's result only applies to social welfare functions, i.e., methods to aggregate preferences into a group preference ordering.
- Maybe, we can elect just one candidate without being effected?

Manipulations

- A social choice function f is said to be manipulable if for some preferences, $\prec_1, \ldots, \prec_i, \ldots, \prec_n, \prec'_i$, it holds that
  - a $\prec_i$ *b*
  - *a* = $f(\prec_1, \ldots, \prec_i, \ldots, \prec_n)$, *and*
  - *b* = $f(\prec_1, \ldots, \prec'_i, \ldots, \prec_n)$
  - i.e., "cheating" is profitable
- A social choice function is incentive compatible if it is not manipulable.
- Is it possible to design incentive compatible social choice function?

Dictatorship

- Defined similar to social welfare functions
- For a given social choice function *f*, a player *i* is called dictator if for all preferences $\prec_1, \ldots, \prec_i, \ldots, \prec_n$,
  - $f(\prec_1, \ldots, \prec_i, \ldots, \prec_n) = a$,
  - where a is the unique element, *s.t.* for all $b \neq a$: $b \prec_i a$
- f satisfies non-dictatorship if it does not contain a dictator.

Gibbard-Satterthwaite's Impossibility Result

Theorem: If *f* is a surjective ("onto") and incentive compatible social choice function with more than than two alternatives, then *f* does not satisfy non-dictatorship.

Proof idea:

1. A social choice function can be extended to a social welfare function using the results of pairwise comparisons of the social choice function (if the pair were the top two candidates, who would win?)
2. If the social choice function is surjective, incentive compatible and does not contain a dictator, the constructed social welfare function will satisfy PE, IIA, and non-dictatorship.

What can we do?

- We could devise voting protocols, which are difficult to manipulate (i.e., it is NP-hard to compute an effective manipulation)
- We could have a more fine-grained specification of the preferences (e.g., by giving utility values)
- We could try to concentrate on some forms of preferences (e.g., so-called single-peaked), for which we can find incentive compatible social choice functions
- We could extend the game by requiring payments for outcomes (as in auctions) ➜ mechanism design

Summary

- Social Choice Theory is about aggregation of preferences
  - social welfare functions
  - social choice functions
- Arrow's and Gibbard-Satterthwaite's impossibility results show that there are no mechanisms that are provably immune against strategic manipulations
- There are some cures against it,
  - NP-hard voting protocols
  - specialized preferences
  - mechanism design