(67686) Mathematical Foundations of AI June 18, 2008

Lecture 6

Lecturer: Ariel D. Procaccia *Scribe: Ezra Resnick & Ariel Imber*

1 Introduction: Social choice theory

Thus far in the course, we have dealt with topics in "hard-core" AI, that is, topics which have been in the main focus of the field since its inception. We now turn to more "modern" topics in AI, closer to the boundaries with economics and game theory.

One of the newer areas explored in AI in the past 15 years is *Multi-Agent Systems*, which analyzes interactions between multiple "agents," each with its own personal objectives. For example, we might model each router in the Internet as an agent, and consider how to get a packet forwarded from source to destination while each router prefers to do as little work as possible. Other examples: dividing processes between processors, stock exchange trading, auction bidding, and so on. Each agent is only looking out for its own best interests, but there may still be room for negotiation, cooperation, etc.

Ironically, the models developed by economists and game theorists often fail to predict real-world human behavior, since it has been shown that humans do not always behave rationally, in the sense that they do not always make the choice that is in their best "economic" interest.[1] However, the theory would apply perfectly to automated agents, "hard-wired" to care only about their own utility.

The first topic we will examine in this area is *social choice*. Social choice theory deals with *voting* scenarios: a set of individuals must select an outcome from a set of alternatives. Each individual ranks the possible alternatives, and a voting rule selects the winning alternative based on the voters' preferences.

2 The voting model

Let $N = \{1, 2, \ldots, n\}$ be a set of *individuals* (or *voters*) and let A be a set of *alternatives* (or *candidates*) such that $|A| = m$. A *preference* over A is a ranking of all the various candidates (i.e. a complete, transitive, asymmetric, binary relation). Let $\prec_i$ denote the preference of voter $i \in N$. (We will use $\preceq_i$ to denote the weak version of $\prec_i$, i.e. $a \preceq_i b$ if and only if $a \prec_i b$ or $a = b$.)

For example, assume $A = \{a, b, c\}$ and the preferences of voters 1 and 2 are given as follows:

1	2
b	a
a	b
c	c

In this case, we have $c \prec_1 a \prec_1 b$ and $c \prec_2 b \prec_2 a$.

[1] For example, consider a coin toss where the winner gets $1200 and the loser must pay $1000. Most people would not agree to play this game, even though the expected winnings are positive. This human tendency is called *loss aversion*.

Let $L(A)$ denote the set of all possible preferences over A (when there is only one set of candidates we will simply write L). A *preference profile* is an element in L^n (defining the preferences of all the voters). A preference profile $\prec = (\prec_1, \prec_2, \ldots, \prec_n)$ can also be denoted $(\prec_i, \prec_{-i})$ for $i \in N$ (meaning that voter i uses the preference $\prec_i$ and all other voters use the preferences defined in $\prec$).

A *voting rule* is a function $f : L^n \to A$ (selecting the winning candidate based on the given preference profile).

3 Voting rules

We will now examine some sample voting rules and their properties.

3.1 Plurality

In the *plurality* voting rule, each voter gives 1 point to the candidate she ranked first, and the winner is the candidate who receives the highest total number of points. In other words, the plurality rule selects the candidate who was ranked first by the most voters. (Technically, a method for breaking ties should also be specified. We will assume some such method exists and ignore this issue.) Plurality is the voting rule most often used in real-world elections, but note that it completely disregards all the information provided by the voter preferences except for the top ranking!

3.2 Borda

In the *Borda* voting rule,[2] each voter gives $m - 1$ points to the candidate she ranked first, $m - 2$ points to the candidate she ranked second, or in general $m - k$ points to the candidate she ranked k-th. The winner is the candidate who amasses the highest total number of points. This voting rule is used in the National Assembly of Slovenia, and is similar to that used in the Eurovision song contest.

3.3 Veto

The *veto* voting rule selects the candidate who was ranked last by the least number of voters.

3.4 Positional scoring voting rules

A *positional scoring* voting rule is defined by a non-negative vector $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$. Each voter gives α_j points to the candidate he ranked j-th, and the winner is the candidate who amasses the highest total number of points. The preceding three voting rules are all instances of positional scoring voting rules, where the plurality rule uses the vector $\alpha = (1, 0, 0, \ldots, 0)$, the Borda rule uses $\alpha = (m - 1, m - 2, \ldots, 0)$, and the veto rule uses $\alpha = (1, 1, \ldots, 1, 0)$.

[2] Named for the French mathematician, physicist, political scientist and sailor Jean-Charles de Borda (1733-1799).

3.5 Plurality with runoff

Definition 1 *Candidate a beats candidate b in a* ***pairwise election*** *if a majority of the voters prefer a to b.*

The *plurality with runoff* voting rule selects a winner in two rounds. The two candidates who receive the highest scores using the plurality rule move on to the second round, where they compete in a pairwise election. (Note that the result of the second round can be calculated based on the original voter preferences, so there is no need for the voters to actually vote twice.) Two-round voting systems are widely used around the world for the election of legislative bodies and directly elected presidents (e.g. the French elections), although often the second round is held only if none of the candidates received an absolute majority of the votes in the first round.

3.6 Single transferable vote (STV)

In the *single transferable vote (STV)* voting rule, each voter has one vote, initially given to her highest ranked candidate. If no candidate receives a majority of first preference rankings, the candidate with the fewest number of votes is eliminated and that candidate's votes are redistributed to the voters' next preferences among the remaining candidates. This process is repeated until one candidate has a majority of votes among candidates not eliminated. This rule is designed to minimize "wasted" votes, and is used for certain elections in Ireland, Malta, Australia and New Zealand.

3.7 The Condorcet criterion

Definition 2 *The* ***Condorcet winner***[3] *for a given preference profile is the candidate who beats every other candidate in pairwise elections.*

Observation 3 *Not every preference profile has a Condorcet winner.*

For example, take the following preference profile:

1	2	3
a	c	b
b	a	c
c	b	a

In this case, each of the three candidates beats exactly one other candidate in pairwise elections, so there is no Condorcet winner. This preference profile is an example of *Condorcet's paradox*: the collective preferences are cyclic (a majority of the voters prefer candidate a to candidate b, a majority prefer b to c, and a majority prefer c to a).

Definition 4 *A voting rule is said to satisfy the* ***Condorcet criterion*** *if it chooses the Condorcet winner whenever one exists.*

[3] Named after the French mathematician and philosopher Marie Jean Antoine Nicolas de Caritat, the marquis de Condorcet (1743-1794).

The voting rules we have seen thus far do not satisfy the Condorcet criterion (see example below). We will now see two rules which do.

3.8 Copeland

The *Copeland* voting rule chooses the candidate who beats the highest number of other candidates in pairwise elections.

Claim 5 *The Copeland voting rule satisfies the Condorcet criterion.*

Proof Assume candidate a is the Condorcet winner for a given preference profile. This means (by definition) that he beats all $m-1$ other candidates in pairwise elections. Any candidate other than a beats at most $m-2$ other candidates in pairwise elections, since he definitely loses to a. Therefore, candidate a (the Condorcet winner) will be chosen by the Copeland rule. ■

3.9 Maximin

For $a, b \in A$, let $p(a,b)$ denote the number of voters who prefer a to b. Formally:

$$p(a,b) \overset{\text{def}}{=} |\{i \in N : b \prec_i a\}|$$

For every candidate a, we can consider the minimal value of $p(a,b)$ over all other candidates b (i.e. the worst pairwise election score a achieves against any of the other candidates). The *maximin* voting rule chooses the candidate for whom this value is maximal. Formally, the winner selected by the maximin rule is:

$$\arg\max_{a \in A} \min_{b \in A \setminus \{a\}} p(a,b)$$

Claim 6 *The maximin voting rule satisfies the Condorcet criterion.*

Proof Assume candidate a is the Condorcet winner for a given preference profile. This means (by definition) that a beats all other candidates in pairwise elections, so:

$$\forall c \in A \setminus \{a\} : p(a,c) > \frac{n}{2}$$

Therefore:

$$\min_{c \in A \setminus \{a\}} p(a,c) > \frac{n}{2}$$

On the other hand, for any candidate $b \in A \setminus \{a\}$ we know that $p(b,a) < \frac{n}{2}$, and so:

$$\min_{c \in A \setminus \{b\}} p(b,c) < \frac{n}{2}$$

Therefore:

$$\arg\max_{c \in A} \min_{d \in A \setminus \{c\}} p(c,d) = a$$

Meaning the maximin rule will indeed choose the Condorcet winner. ■

3.10 Example

We will use the preference profile below to demonstrate the use of the different voting rules. Interestingly, each of the five candidates is declared the winner by at least one of the voting rules we have seen!

33 voters	16 voters	3 voters	8 voters	18 voters	22 voters
a	b	c	c	d	e
b	d	d	e	e	c
c	c	b	b	c	b
d	e	a	d	b	d
e	a	e	a	a	a

- The plurality winner is a, since she is the candidate ranked first most often (33 votes).
- The Borda winner is b, since he has the highest Borda count ($33 \times 3 + 16 \times 4 + 3 \times 2 + 8 \times 2 + 18 \times 1 + 22 \times 2 = 247$).
- The Condorcet winner is c (she beats all other candidates in pairwise elections), so c will be chosen by the Copeland and maximin rules.
- Using STV, candidate c will be eliminated first, and her 11 votes will be redistributed between candidates d (3 votes) and e (8 votes). The next to be eliminated is candidate b, and his 16 votes pass to candidate d. Candidate e is eliminated next, and her 30 votes pass to d, allowing him to defeat a (67 to 33). So d is the STV winner.
- Using plurality with runoff, candidates a and e advance to the second round (with 33 and 22 votes respectively), where e goes on to defeat a in pairwise elections (64 to 36).

4 Manipulation

Consider an election using the Borda voting rule with the following preference profile:

1	2	3
b	b	a
a	a	b
c	c	c
d	d	d

Candidate b is the winner, beating candidate a 8 points to 7. However, if voter 3 lies about his preferences and ranks candidate b last (after a, c and d), b's score goes down to 6, and a (voter 3's favorite candidate) wins! Generally, if a voter knows the voting rule being used and the preferences of the other voters (or can guess them), he may be able to bring about a more preferable result for himself by reporting a preference different from his true preference.

This is called *manipulation*. We would like to have a voting rule which cannot be manipulated, meaning that no voter can ever profit from lying about her preferences. But is this possible?

Definition 7 *A voting rule f is* ***strategy-proof*** *if no (single) voter can ever benefit from lying about his preferences:*

$$\forall \prec \in L^n \;\; \forall i \in N \;\; \forall \prec'_i \in L : f(\prec'_i, \prec_{-i}) \preceq_i f(\prec)$$

Definition 8 *A voting rule is* ***manipulable*** *if it is not strategy-proof.*

Claim 9 *If there are exactly two candidates then the plurality voting rule is strategy-proof.*

Proof Denote $A = \{a, b\}$ and assume WLOG that candidate a was selected by the plurality voting rule (for some given preference profile $\prec$). Let $i \in N$ be one of the voters. If $b \prec_i a$, then voter i's favorite candidate has already won, and she certainly has nothing to gain by lying about her preferences. On the other hand, if $a \prec_i b$, then changing voter i's preference so that b is ranked below a only lowers b's score and cannot possibly cause b to win. In either case, voter i cannot benefit from lying about her preferences. ■

4.1 The Gibbard-Satterthwaite theorem

Definition 10 *A voting rule f is* ***dictatorial*** *if there is an individual (the dictator) whose most preferred candidate is always chosen by f:*

$$\exists i \in N \;\; \forall a \in A \;\; \forall \prec \in L^n : a \preceq_i f(\prec)$$

Definition 11 *A voting rule f is* ***onto*** *if it is possible for any of the candidates to win (given the right preference profile):*

$$\forall a \in A \;\; \exists \prec \in L^n : f(\prec) = a$$

Theorem 12 *(The Gibbard-Satterthwaite theorem) If there are at least three candidates, any voting rule that is strategy-proof and onto is dictatorial.*

Corollary 13 *If there are at least three candidates, any voting rule that is onto and non-dictatorial is manipulable.*

We will prove this theorem for the case where there are exactly two voters. First, we will prove two useful lemmas.

4.2 Two useful lemmas

The first lemma says that a strategy-proof voting rule's selected outcome remains constant for all changes to the preference profile such that candidates ranked below the winner before the change are also ranked below the winner after the change:

Lemma 14 *(monotonicity) Let f be a strategy-proof voting rule, and let $f(\prec) = a$ for some preference profile $\prec$ and $a \in A$. Then $f(\prec') = a$ for all preference profiles $\prec'$ such that:*

$$\forall i \in N \;\; \forall x \in A \setminus \{a\} : x \prec_i a \Rightarrow x \prec'_i a$$

Proof Starting from the preference profile $\prec$ we will change the voters' preferences one at a time to $\prec'$ showing that the winner remains constant at every step. We begin with the preference profile $(\prec'_1, \prec_{-1})$ and assume $f(\prec'_1, \prec_{-1}) = b$. Since f is strategy-proof, we know that $b \preceq_1 a$ (otherwise voter 1 would benefit from reporting his preference as $\prec'_1$ instead of $\prec_1$), so according to the lemma premise we get $b \preceq'_1 a$. If $b \prec'_1 a$ then voter 1 would benefit from reporting his preference as $\prec_1$ instead of $\prec'_1$, contradicting the assumption that f is strategy-proof. Therefore $a = b$, meaning that the winner has not changed due to voter 1's change in preference. In the same manner we show that the winner remains constant as each voter $i \in N$ changes his preference from $\prec_i$ to $\prec'_i$. We conclude that $f(\prec') = a$. ■

The second lemma says that the outcome of a strategy-proof and onto voting rule must be (weakly) Pareto optimal, meaning there is no candidate strictly preferred by all voters to the winning candidate:

Lemma 15 *(Pareto optimality) Let f be a strategy-proof voting rule which is onto, and let $a, b \in A, a \neq b$. If $\prec$ is a preference profile such that $\forall i \in N : b \prec_i a$ then $f(\prec) \neq b$.*

Proof Suppose that $f(\prec) = b$. Since f is onto, there exists a preference profile $\prec'$ such that $f(\prec') = a$. Let $\prec''$ be a preference profile where all voters rank candidate a first and candidate b second. This means that no voter ranked b lower in $\prec''$ than they did in $\prec$ (based on the assumption that all voters ranked a above b in $\prec$), so by monotonicity it follows that $f(\prec'') = f(\prec) = b$. On the other hand, no voter ranked a lower in $\prec''$ than they did in $\prec'$ (since a is always ranked first in $\prec''$), so by monotonicity it follows that $f(\prec'') = f(\prec') = a$, which is a contradiction (since $a \neq b$ and f is a function). Hence $f(\prec) \neq b$. ■

4.3 Proof of the Gibbard-Satterthwaite theorem

We will prove the Gibbard-Satterthwaite theorem under the simplifying assumption that $n = 2$.[4] First, let us consider an example which illustrates the idea used in the proof:

1	2
a	b
b	a
c	c

Assuming we use an onto, strategy-proof voting rule with the above preference profile, we would like to show that one of the voters must be a dictator. Notice that candidate c cannot win due to Pareto optimality. Assume that the winner selected by the voting rule is a. Now consider what happens if voter 2 changes his preference, and ranks c over a (leaving b ranked first). Pareto optimality still rules out c as the winner, and b cannot be the winner due to strategy-proofness (otherwise voter 2 benefits from changing his reported preference), so a must remain the winner. Since a wins when voter 1 ranks him first and voter 2 ranks him last, it follows from monotonicity that a will be the winner whenever he is ranked first by voter

[4] For the original proofs, see [1] and [2].

1, meaning voter 1 is a dictator for candidate a. Likewise, if we had initially assumed that candidate b was the winner, voter 2 would have been a dictator for candidate b. In the proof below, we will show that each candidate has a dictator, and that it must be the same dictator for all candidates.

Theorem 16 *If there are exactly two voters and at least three candidates, any voting rule f that is strategy-proof and onto is dictatorial.*

Proof Let $\prec$ be a preference profile and let $a, b \in A$ such that:

$$\forall x \in A \setminus \{a,b\} : (x \prec_1 b \prec_1 a) \wedge (x \prec_2 a \prec_2 b)$$

By Pareto optimality, $f(\prec) \in \{a,b\}$. Assume WLOG that $f(\prec) = a$. Now consider a preference $\prec'_2$ which satisfies:

$$\forall x \in A \setminus \{a,b\} : a \prec'_2 x \prec'_2 b$$

Due to Pareto optimality, $f(\prec_1, \prec'_2) \in \{a,b\}$. Due to strategy-proofness, $f(\prec_1, \prec'_2) \neq b$. And so $f(\prec_1, \prec'_2) = a$. Monotonicity now implies that f will select a as the winner for any preference profile where voter 1 ranks a first. So voter 1 is a dictator for candidate a. The analysis above can be repeated for all pairs of candidates $x, y \in A$, to show that either voter 1 is a dictator for candidate x or voter 2 is a dictator for candidate y.

For $i \in \{1,2\}$, let A_i denote the set of candidates for whom voter i is a dictator. Let $A_3 = A \setminus (A_1 \cup A_2)$. Note that $|A_3| \leq 1$ (otherwise we could repeat the analysis above for two candidates in A_3 and find that one of them must be in A_1 or A_2). Note that for two different candidates x, y, it is not possible that $x \in A_1$ and $y \in A_2$ (this would cause a contradiction if voter 1 ranks x first and voter 2 ranks y first). We know that $|A_3| \leq 1$ and $m \geq 3$ and therefore $|A_1 \cup A_2| \geq 2$, but we showed that two different candidates x and y cannot belong to A_1 and A_2 respectively, so $A_1 \cap A_2 = \emptyset$. It follows that $A_2 = \emptyset$ (we assumed $a \in A_1$). Finally, $A_3 = \emptyset$, since if we assume $c \in A_3$ then repeating the above analysis for c and a implies that either $c \in A_1$ or $a \in A_2$, which is a contradiction. We conclude that $A_1 = A$ and f is dictatorial with voter 1 as the dictator. (If we had initially assumed that $f(\prec) = b$, voter 2 would have been the dictator.) ■

References

[1] A. Gibbard. Manipulation of voting schemes: A general result. *Econometrica*, 41:587–601, 1973.

[2] M. Satterthwaite. Strategy-proofness and arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10:187–216, 1975.