601.436/636 Algorithmic Game Theory **Lecturer:** Michael Dinitz
Topic: Voting Schemes and Social Choice **Date:** 4/30/20
Scribe: Michael Dinitz

25.1 Introduction

We're going to end this course by looking forward to November (or the June primary) by talking about voting schemes. For the vast majority of our discussions on mechanism design, we've been thinking of selling things, whether a single item, identical items, or more complicated combinatorial auctions. But, if you think about it, the "general mechanism design" setting that we defined two weeks ago just has "outcomes". And if we think of a "bid" as a "vote", then elections start to look like mechanism design. The concept of "utility" doesn't match up quite as well, but it's still reasonable and very interesting to think about elections from a mechanism design point of view.

25.2 Setup and Basics

The setup is pretty straightforward. There is a set A of candidates and there are n voters $[n]$. Let L be the set of all total orderings of A. Each $i \in [n]$ has a private ordering $\succ_i \in L$.

Even before we think about incentives, there's an obvious question: who should win the election? And if we ask for even more: is there a reasonable "group ordering" of A that aggregates all of the individual orderings?

If $|A| = 2$ then the answer to both of these questions is simple: there are only two possible orderings, so let's just take the majority vote (assume that n is odd). But what if $|A| \geq 3$?

25.2.1 Condorcet

One natural approach is the following. We say that x beats y in a pairwise election if a majority of the voters prefer x to y, i.e., $|\{i \in [n] : x \succ_i y\}| \geq n/2$.

Definition 25.2.1. *A candidate x is a* Condorcet winner *if x beats y in a pairwise election for all $y \neq x$.*

Condorcet winners are an obvious solution, since if x is the Condorcet winner then everyone can agree on why x won the election instead of their favorite candidate y. Unfortunately, Condorcet winners do not always exist. Consider the following example with three voters $1, 2, 3$ and three candidates a, b, c, with

$$a \succ_1 b \succ_1 c \qquad\qquad c \succ_2 a \succ_2 b \qquad\qquad b \succ_3 c \succ_3 a$$

Then no matter who wins, the majority would prefer someone else: if a wins then the majority would prefer c, if b wins then the majority would prefer a, and if c wins then the majority would prefer b. This is known as *Condorcet's Paradox*, even though it's not really a paradox.

Interestingly, even if a Condorcet winner *does* exist, obvious voting schemes fail. For example, many of you might have heard of "Single Transferable Vote", which is a voting scheme which is in use in real countries (Ireland, Australia, New Zealand, and possible others) and which some people are trying to get the US (or at least parts of it) to use. In single transferable vote, instead of just voting for their favorite, each voter submits their full ordering of the candidates. The mechanism then proceeds for $|A| - 1$ rounds, where in each round the candidate with the fewest first choice votes (out of the candidates who have not yet been eliminated) is eliminated. This seems much better than our usual voting scheme, since it lets me vote for a candidate who I like but who has no chance to win without any real penalty, since that candidate will be eliminated in the first round and then my vote will "count" for my second-choice until they're eliminated, when my vote will then go towards my third choice, etc. So minor parties in the US tend to be strongly in favor of STV.

But consider the following example with three voters $[3]$ and four candidates a, b, c, d:

$$b \succ_1 a \succ_1 c \succ_1 d \qquad c \succ_2 a \succ_2 d \succ_2 b \qquad d \succ_3 a \succ_3 b \succ_3 c$$

a is a Condorcet winner in this example, but it gets eliminated in the first round!

So it's not clear what to do when there is no Condorcet winner, and it's not clear how to design a mechanism that outputs the Condorcet winner if one does exist.

25.3 Arrow's Impossibility Theorem

Let's start with the first of these questions. And instead of thinking about exactly what our mechanism should do, let's think about what properties we would expect any mechanism to have.

Definition 25.3.1. *An* aggregation function *is a function $F : L^n \to L$.*

Definition 25.3.2. *An aggregation function F is a* dictatorship *if there exists an $i \in [n]$ such that $F(\succ_1, \dots, \succ_n) = \succ_i$ for all $(\succ_1, \dots, \succ_n) \in L^n$.*

What properties might we want from an aggregation function? The following is certainly very natural.

Definition 25.3.3. *F satisfies* unanimity *if the following holds: if $a \succ_i b$ for all $i \in [n]$, then $a \succ_{F(\succ_1, \dots, \succ_n)} b$*

The next condition is also pretty intuitive, but is a little harder to understand. Intuitively, it says that the ordering between two candidates a and b in the aggregated function should only depend on the voters' preference between a and b. In other words, if we're only looking at deciding between a and b, what the voters think of c shouldn't matter. Formally, this is the following.

Definition 25.3.4. *F satisfies* independence of irrelevant alternatives *(IIA) if the following holds. Let $(\succ_1, \dots, \succ_n)$ and $(\succ'_1, \dots, \succ'_n)$ be two preference profiles such that $\succ_i$ and $\succ'_i$ have the same ordering between a and b for all $i \in [n]$. Then a and b have the same ordering in $F(\succ_1, \dots, \succ_n)$ and $F(\succ'_1, \dots, \succ'_n)$.*

The following is the main result that we're going to prove (or at least sketch a proof of) today.

Theorem 25.3.5 (Arrow's Impossibility Theorem)**.** *When $|A| \geq 3$,* every *aggregation function that satisfies unanimity and IIA is a dictatorship.*

We'll spend the rest of this section sketching a proof of this theorem. To keep things simple, we'll assume that $A = \{a, b, c\}$, i.e., that $|A| = 3$. This proof can be extended to the general case, but this special case is significantly simpler. Let F satisfy unanimity and IIA. We want to show that F is a dictatorship.

We'll start by finding a "pivotal voter" for b over a. To define this, we create $n+1$ different profiles $P^0, \dots, P^n$. In profile P^i (with $0 \leq i \leq n$) the first i voters rank b above a (we don't care where they rank c, since by IIA the ranking of c cannot affect the relative ranking of a and b), and the remaining $n - i$ voters rank a above b. Then unanimity implies that $F(P^0)$ rank a above b, while $F(P^n)$ rank b above a. Thus there must exist some $k \in [n]$ such that b is above a in $F(P^k)$ but a is above b in $F(P^{k-1})$. Call this the *pivotal voter* k_{ba} for b above a.

Now we claim that voter k_{ba} is actually a partial dictator: it is a dictator for b over c. More formally, we have the following lemma.

Lemma 25.3.6. *Let $P' = (\succ_1, \dots, \succ_n)$ be some preference profile. If $b \succ_{k_{ba}} c$ then $b \succ_{F(P')} c$.*

Proof Sketch. Consider the profile $P = (\succ_1, \dots, \succ_n)$ where all voters $i \in \{1, \dots, k_{ba}-1\}$ have $b \succ_i c \succ_i a$ and all voters $i \in \{k_{ba}, \dots, n\}$ have $a \succ_i b \succ_i c$. Then this is a valid choice for profile $P^{k_{ba}-1}$, since the first $k_{ba} - 1$ voters prefer b to a and the rest prefer a to b. This implies that $F(P)$ is exactly $a \succ_{F(P)} b \succ_{F(P)} c$, where the first is by the definition of k_{ba} and the second is by unanimity.

Now fix some set of voters $S \subseteq [n] \setminus \{k_{ba}\}$. Consider a strategy profile $P^S = (\succ_1^S, \dots, \succ_n^S)$ where $b \succ_{k_{ba}}^S a \succ_{k_{ba}}^S c$, every voter in $[n] \setminus (S \cup \{k_{ba}\})$ has the same preference as in P, and every voter in S has the same preference as in P *except* that b and c are swapped. Note that this does not change the relative ordering of a and b in any of these preferences (other than k_{ba}) or the relative ordering of a and c.

Then P^S is a valid choice for profile $P^{k_{ba}}$, so we know that $b \succ_{F(P^S)} a$ by the definition of k_{ba}. Moreover, since in P^S the relative ordering of a and c is the same in every preference as in P, we know from IIA that $F(P^S)$ must order them the same way as $F(P)$, and so $a \succ_{F(P^S)} c$. Thus $b \succ_{F(P^S)} a \succ_{F(P^S)} c$.

Which of the voters actually prefer b to c in P^S? Only k_{ba} and the voters not in S. So even if an arbitrary set S of other voters prefer c to b, we'll end up with $F(P^S)$ preferring b to c. And now we can use the fact that even though P was constructed in a very special way, by IIA all that matters is the relative ordering between b and c. □

Now let's use this lemma. First, clearly there's nothing special about b and a. For example, we could also define the pivotal voter k_{bc} for b over c, and then Lemma 25.3.6 would imply that k_{bc} is a partial dictator for b over a. We're going to reason about all six of the pivotal voters $k_{ba}, k_{ab}, k_{ac}, k_{ca}, k_{bc}, k_{cb}$.

Since k_{ba} is a dictator for b over c it must come no earlier than the pivotal voter for b over c, since the pivotal voter was defined to be the *first* voter where flipping it (and all of its predecessors) to prefer b over c instead of c over b would actually result in preferring b over c. Thus $k_{bc} \leq k_{ba}$. But since k_{ba} is a dictator for b over c, it must come *before* the pivotal voter for c over b, since by the definition of the dictator as long as k_{ba} prefers b to c so will the aggregate. Thus

$$k_{bc} \leq k_{ba} \leq k_{cb}$$

Now we can do the same argument and just switch the names of b and c to get

$$k_{cb} \leq k_{ca} \leq k_{bc}.$$

Thus all these are actually equal:

$$k_{bc} = k_{ba} = k_{cb} = k_{ca}$$

Repeating this with the other pairs implies that all pivotal voters are actually the same voter! Thus all of the partial dictators are the same voter, so that voter is actually a full dictator.

25.4 Gibbard-Satterthwaite

Arrow's impossibility theorem is the cornerstone of social choice theory, and is often interpreted as something along the lines of "Every voting method that only uses ordinal preferences is bad", since any such voting method must be either a dictatorship or violate unanimity or IAA. This is a reasonable interpretation, but it's certainly not the only possible one. For example, we might not really believe in IAA (wikipedia has some examples where it's "obvious" that IAA should be violated). Or we might not care about outputting an actual aggregation, since we only care about the one winner.

Unfortunately, it turns out that we can use Arrow's theorem to give an extremely bad result in the game-theoretic setting. Arrow's theorem says nothing about incentives or incentive compatibility, and it turns out we can basically replace the "less believable" part of Arrow's theorem with an incentive compatibility requirement.

We're going to think about social choice functions rather than aggregations:

Definition 25.4.1. *A* social choice function *is a function $f : L^n \to A$.*

So we're going to think of "voting" as submitting a "bid" which is your full ordering, but now we're just outputting a single winner rather than an aggregated preference order. Incentive compatibility means that for every player, they prefer the candidate who wins if they tell the truth to the candidate who wins if they lie.

Theorem 25.4.2 (Gibbard-Satterthwaite)**.** *Let f be an incentive compatible social choice function which is surjective with $|A| \geq 3$. Then f is a dictatorship.*

The surjective requirement feels like a minor assumption, since if we had a social choice function which was not surjective then there would be a candidate who would not be able to win even if they were everyone's first choice. So assuming f is surjective is not much of an assumption.

We're not going to prove this, but you can find it in NRTV 9.2.4. At a very high level, the proof goes as follows. Let f be an incentive compatible and surjective social choice function. Since we want to use Arrow's theorem, we're going to want to somehow extend f to an aggregation function. To do this, we'll make the following definition.

Definition 25.4.3. *Let $S \subseteq A$ and let $\succ \in L$. Then we define $\succ^S \in L$ to the be ordering where:*

- *If $a, b \in S$ or $a, b \notin S$ then $a \succ^S b$ if and only if $a \succ b$*
- *If $a \in S$ and $b \notin S$ then $a \succ^S b$.*

This will let us extend f to an aggregation function F by setting $F(\succ_1, \ldots, \succ_n) =\succ$ where $a \succ b$ if and only if $f(\succ_1^{\{a,b\}}, \ldots, \succ_n^{\{a,b\}}) = a$. In other words, to determine if $a \succ b$ we move a and b to the top two choices in each ordering but keep them in the same relative order, and then see who wins. It's not entirely obvious, but it is possible to prove that this is indeed a well-defined aggregation function, i.e., $\succ \in L$.

The key lemma is the following.

Lemma 25.4.4. *If f is not a dictatorship, then F satisfies unanimity and IIA and is not a dictatorship.*

This lemma immediately implies the theorem, since Arrow's theorem implies that no such F exists. Thus f must be a dictatorship.