Arrow's Theorem

1 Setup

There are n voters and a set A of candidates. Voter i casts a ranking $<_i$ of A, where $a <_i b$ means voter i ranks a worse than b (equivalently b better than a). A *social welfare function* F takes all the ballots $<= (<_1, \ldots, <_n)$ and outputs one ranking $<_F = F(<)$ of A.

Definition 1. F satisfies:

1. **Unanimity** if, whenever everyone casts the same ranking, F outputs that ranking.
2. **Independence of irrelevant alternatives (IIA)** if the output of a vs. b ($a >_F b$ or $a <_F b$) depends only on how each voter ranks a vs. b, and not on the rankings of any other candidates.

F is a **dictatorship** if there is a voter i^* such that, regardless of $<= (<_1, \ldots, <_n)$, we have $<_F = <_{i^*}$.

Theorem 2 (**Arrow's Theorem**)**.** *If F is a social welfare function where $|A| \geq 3$, and F satisfies unanimity and IIA, then F is a dictatorship.*

Throughout, F is a social welfare function on $|A| \geq 3$ candidates that satisfies unanimity and IIA.

2 Understanding IIA

Consider two different elections with two voters P_1, P_2 (each column lists a ballot from best to worst):

P_1	P_2	P_1	P_2
a	c	a	a
b	d	d	c
c	a	c	b
d	b	b	d

The two elections look very different, but neither person changed their ranking of a and b: both voters put a above b in both elections. So IIA says F must rank a vs. b the same way in both elections, whatever happens to c and d.

3 Pareto

Definition 3. F satisfies **Pareto** if, whenever all voters agree that $a <_i b$, then $a <_F b$.

Lemma 4. *If F satisfies both unanimity and IIA, then F satisfies Pareto.*

Proof. Suppose everyone agrees $a <_i b$ in the election $<$. Consider $<' = (<_1, <_1, \ldots, <_1)$, which casts the first person's vote for all people. By unanimity, $F(<') = <_1$, so in this election $a <'_F b$. Every voter ranks a below b in both $<$ and $<'$, so by IIA, $a <_F b$. □

Remark. For 2 candidates, plurality satisfies unanimity, IIA and Pareto without being a dictatorship. The theorem only applies once we move to 3 candidates.

4 Pairwise neutrality

The proof uses one more fact: F must treat every pair of candidates in the same way.

Lemma 5 (Pairwise neutrality)**.** *Let $< = (<_1, \ldots, <_n)$ and $<' = (<'_1, \ldots, <'_n)$ be two elections with $<_F = F(<)$ and $<'_F = F(<')$, and let $a \neq b$ and $c \neq d$ be candidates with $b \neq c$ such that*

$$\forall i: \quad a >_i b \iff c >'_i d.$$

Then $a >_F b \iff c >'_F d$.

In words: if the voters who put a above b in $<$ are exactly the voters who put c above d in $<'$, then when a beats b, c beats d.

Remark. In the lecture, this was written with the extra condition $a >_F b$. It is not needed.

5 Proof of Arrow's theorem

The proof has two steps: first we find a candidate for the dictator, then we show that this voter really is one.

5.1 Who is the dictator?

Fix any $a \neq b \in A$. For each $i = 0, 1, \ldots, n$, define an election $<^{(i)} = (<_1^{(i)}, \ldots, <_n^{(i)})$ such that

$$\text{for } j \leq i: \quad a >_j^{(i)} b,$$
$$\text{for } j > i: \quad b >_j^{(i)} a,$$

and all other candidates are put wherever. This defines n ballots $(n+1)$ times: $<^{(i)}$ is the election where the first i voters have switched from preferring b to preferring a. By IIA, the output of a vs. b in $<^{(i)}$ does not depend on where the other candidates were put.

- In $<^{(0)}$, everyone agrees $b >_j^{(0)} a$, so $b >_F^{(0)} a$ by Pareto.
- In $<^{(n)}$, everyone agrees $a >_j^{(n)} b$, so $a >_F^{(n)} b$ by Pareto.

So somewhere there is a switch between the winners: b is the winner at 0 and a is the winner at n (a discrete form of the intermediate value theorem). Let i^* be the first point where $a >_F^{(i^*)} b$. Then $i^* \geq 1$, and $b >_F^{(i^*-1)} a$.

Claim 6. *i^* is the dictator, i.e. regardless of $< = (<_1, \ldots, <_n)$, we have $<_F = <_{i^*}$.*

5.2 Proof of the claim

Take any election $<$ and any two candidates $c \neq d$ with $d <_{i^*} c$. We will show that $c >_F d$. Since $|A| \geq 3$, we can pick a third candidate e with $c \neq d \neq e$ and $c \neq e$.

Let's look at another election $<'$, equal to $<$ except that:

$$\begin{aligned} &\text{for } j < i^*: \quad \text{move } e \text{ to the top of } <'_j, \\ &\text{for } j > i^*: \quad \text{move } e \text{ to the bottom of } <'_j, \\ &\text{for } j = i^*: \quad \text{make } d <'_{i^*} e <'_{i^*} c. \end{aligned}$$

Only e was moved, so every voter ranks c vs. d in $<'$ exactly as in $<$.

The table below shows who wins each relevant comparison, voter by voter.

	in $<'$		in $<^{(i^*-1)}$	in $<^{(i^*)}$
voter j	c vs. e	e vs. d	a vs. b	a vs. b
$j < i^*$	$e >'_j c$	$e >'_j d$	$a >_j^{(i^*-1)} b$	$a >_j^{(i^*)} b$
$j = i^*$	$c >'_j e$	$e >'_j d$	$b >_j^{(i^*-1)} a$	$a >_j^{(i^*)} b$
$j > i^*$	$c >'_j e$	$d >'_j e$	$b >_j^{(i^*-1)} a$	$b >_j^{(i^*)} a$

Think about c vs. e. After and including i^*, $e <'_j c$; before i^*, e is on top. So we can see that, for every j,

$$c >'_j e \iff b >_j^{(i^*-1)} a.$$

By pairwise neutrality, $c >'_F e$, as $b >_F^{(i^*-1)} a$.

Think about e vs. d. Up to and including i^*, $d <'_j e$; after i^*, e is at the bottom. So, for every j,

$$e >'_j d \iff a >_j^{(i^*)} b.$$

By pairwise neutrality, $e >'_F d$, as $a >_F^{(i^*)} b$.

Combining. Thus, from both parts, $c >'_F e$ and $e >'_F d$, so $c >'_F d$ (we can use transitivity, since $<'_F$ is a ranking). As we only meddled with e, every voter ranks c vs. d the same way in $<$ and $<'$, so by IIA, $c >_F d$.

Since c and d were arbitrary with $d <_{i^*} c$, F agrees with $<_{i^*}$ on every pair of candidates, so $<_F = <_{i^*}$.

The election $<$ was also arbitrary, so F is a dictatorship with dictator i^*. □