Chapter 5

Arrow's Theorem and the Gibbard-Satterthwaite Theorem

In many voting systems, each voter must produce a ranked preference order of all candidates mentioned, and no ties are allowed. Such systems are called *ordinal*. However some voting systems, called *cardinal*, allow the voters to evaluate candidates separately, and a voter could say two candidates were equal. For the moment we shall concentrate on ordinal systems; cardinal systems will be studied in Chap. 7.

Given the ranked preferences of all voters, it is desirable to produce a list that represents the ranked preferences of the electorate. In other words, we would like to construct a list of all candidates in which, if A precedes B, then the electorate prefers A to B. (For convenience we write $A > B$ to mean A precedes B in the preference list.) We shall call a system that purports to produce such a list a *rank order voting system*. One example is Condorcet's extended method. Another is plurality voting: $A > B$ means A received more first-place votes than B, and the electorate ranks the candidates in order by the number of first-place votes received.

Not all ordinal voting methods are rank order voting systems; some simply find a winner and then the process stops. However, given any ordinal voting system, we can modify it so that it produces a ranking of any pair of candidates. For example, one could find the winner of an election, candidate A say. Then modify the preference profile by deleting A and recalculate the winner; say it is B. Delete B and continue. This will produce an ordered listing of all the candidates, which will be transitive.

But is it accurate?

In 1950, Kenneth J. Arrow [1] addressed this question, and proved that no rank order voting system for three or more candidates can convert the set of ranked preferences into a preference ranking for the whole electorate while meeting some very reasonable-sounding criteria. Arrow published his results in the book [2] in 1951, and a slightly improved version in the second edition in 1963 [3]. He subsequently received the Nobel Prize in Economics for his work in 1972.

W.D. Wallis, *The Mathematics of Elections and Voting*,
DOI 10.1007/978-3-319-09810-4_5

A quarter of a century after Arrow's first version was published, A. Gibbard [16] and M. A. Satterthwaite [25] independently proved a similar but stronger theorem that has become known as the *Gibbard-Satterthwaite Theorem*. Simply put, it essentially says that a dictatorship is the only voting system for three or more candidates that cannot be manipulated. We shall discuss their result in Sect. 5.5, below. For further discussion of strategy-proof voting, see [6].

5.1 Arrow's Criteria

Arrow originally stated five conditions that a fair system should satisfy. He subsequently modified them, resulting in three criteria: *No Dictators (ND)*, *Independence of Irrelevant Alternatives (IIA)* and *Pareto Efficiency(PE)*. We shall examine each of these.

No Dictators

This is the requirement that no single voter should have the power to determine the outcome. In a situation where one person (a "dictator") determines the result of an election, for example where the chairman of a company has the final say about all of its activities, voting would be a waste of time, so this condition is obvious. An election would only be held in those cases where the dictator declines to vote.

Say all voter rankings in a profile P remain fixed except for the ranking of one voter X, and however X relatively orders two candidates A and B, the electorate will order them in the same way. We shall say X is called *pivotal* for candidates A and B in the profile P. A voter is called *extremely pivotal* for candidate A in a profile if the voter is pivotal for all pairs involving A.

A voter who is pivotal for candidates A and B is also called a *pair dictator for* A, B in the profile P. Voter X is a *local dictator* for P if X is a pair dictator in P for all pairs. A *dictator* is a voter who is a local dictator for all profiles.

Independence of Irrelevant Alternatives

Say the election determines that the electorate as a whole prefers A to B, and suppose some electors change their preference lists. If no voter changes the relative positions of A and B—all those who initially ranked A ahead of B still do so in the final vote, and similarly those who preferred B to A continue to do so—then the system should continue to say that A is preferred to B.

Pareto Efficiency, or Unanimity

If every voter prefers A to B, then the system cannot say that the electorate prefers B to A.

Vilfredo Pareto (1848–1923) was an Italian engineer, sociologist, economist, political scientist and philosopher. An allocation of funds is called *Pareto efficient* if there is no other allocation in which some other individual is better off and no individual is worse off. The application of this term to voting systems is a little odd, but is now well-established.

Other Criteria

Arrow's original Theorem as stated in [1] involved two other criteria, rather than Pareto efficiency. These were *monotonicity* and *non-imposition*. Arrow changed them to Pareto efficiency in 1963, in [3].

Monotonicity (also called *mono-raise*, see [36]) states that if one or more voters change their ranked preferences by putting one candidate higher, then the overall preference list should either be changed by ranking that candidate higher or else be unchanged; an individual cannot be made *less* popular overall by having one rating *improved*. Non-imposition means that every possible overall preference list should be achievable: if there are n candidates, then each of the $n!$ lists can be achieved.

It is not hard to show that Independence of Irrelevant Alternatives, monotonicity and non-imposition together imply Pareto Efficiency, and in fact non-imposition is not required, but Independence of Irrelevant Alternatives and Pareto Efficiency do not imply monotonicity. So the set of conditions is weaker, and the Theorem is stronger, in the 1963 version, which is now referred to as Arrow's Impossibility Theorem:

Theorem 2. *No rank order voting system for three or more candidates that has no dictator can satisfy both Independence of Irrelevant Alternatives and Pareto Efficiency.*

The proof is in the next section.

Sample Problem 5.1 *Prove that Independence of Irrelevant Alternatives and monotonicity together imply Pareto Efficiency.*

Solution. Suppose a voting system is not Pareto efficient, but it satisfies Independence of Irrelevant Alternatives and monotonicity, Select a preference profile in which every voter prefers A to B, but the system says the electorate prefers B to A. Change the profile by moving B up in every voter's list, until B is just above A. According to monotonicity, this cannot lower B's ranking overall, so B will still be preferred to A. Now change every ranking by moving A down to the position originally occupied by B; by Independence of Irrelevant Alternatives, B is still preferred to A. But the rankings are now exactly as they were originally, with A and B exchanged, so the end result is to exchange the positions of A and B in the final result, and A is now preferred to B—a contradiction.

Practice Exercise. Prove that Pareto efficiency implies non-imposition.

5.2 The Proof of Arrow's Theorem

Several proofs are available. What follows is based on one of three given by Geanakoplos in [15].

We begin with the following result, called the *Extremal Lemma.*

Lemma 1. *Assume the voting system satisfies Pareto Efficiency and Independence of Irrelevant Alternatives alternatives, and suppose there is one candidate, C, such that every voter either places C at the top of the preference list or at the bottom. Then either C is elected, or else the electorate prefers every candidate to C.*

Proof. Suppose not. Say there is a profile in which every voter either places C at the top or at the bottom, but the system places C somewhere in the middle; say $A > C > B$. By the Independence of Irrelevant Alternatives, this would continue to be true even if every voter changed their preferences to place B above A—this will not disturb the relative positions of A and C, or of B and C in any ranking. Every elector prefers B to A, so by Pareto efficiency, the electorate prefers B to A. But the electorate holds $A > C > B$, so by transitivity A is preferred to B—a contradiction. □

Proof of Arrow's Theorem. We assume that our voting system satisfies both Independence of Irrelevant Alternatives and Pareto Efficiency. We prove that it has a dictator.

Suppose there are n voters, $X_1, X_2, \dots, X_n$. (The ordering is completely arbitrary.) Consider a profile in which every voter has placed candidate C at the bottom of the ranking. We shall call this Profile P_0. Construct a series of profiles, Profile P_1, Profile $P_2, \dots$, Profile P_n, as follows. Profile P_1 is formed by changing X_1's preferences by moving C from the bottom to the top, leaving everything else unchanged. In general, Profile P_j is the same as profile Profile P_{j-1} except that X_j's preferences by moving C from the bottom to the top, with the rest of X_j's preference order unchanged.

From the Extremal Lemma we see that C is ranked either last or first in each of these profiles. By Pareto efficiency, C will be ranked last by the electorate under

P_1 and first—the winner—under P_n. Suppose P_k is the first profile in which C is the winner; C is ranked last under $P_1, P_2, \ldots, P_{k-1}$ and first under P_k. So X_j is extremely pivotal for C in profiles P_{k-1} and P_k.

We now show that X_j is a pair dictator in P_k for every pair not involving C. Suppose A and B are two other candidates. Select one of them, A say, and construct a new profile Q: X_j moves A above C, so that X_j ranks $A > C > B$, and let all other voters change their votes in any way provided that C stays in the same extremal position as it was in their vote in P_k. By Independence of Irrelevant Alternatives, the electorate will rank A above C (since all voters rank A and C in the same order as they did in profile P_{k-1}, where C was at the bottom overall), and it will rank C above B (since all voters rank A and C in the same order as they did in profile P_k, where C was at the top overall). So society ranks A above B whenever $X + j$ ranks A above B. But the same argument applies if B had been selected instead of A. So X_j is a pair dictator for every pair not involving C.

But the whole argument can be applied if some other candidate, D say, were used instead of C. (D could equal A, or B, or some other candidate.) Therefore there will be some voter, X_h say, who is a pair dictator for every pair not involving D. As $D \neq C$, X_h is a pair dictator for every pair involving C. In particular, we can assume D is neither A nor C. Then x_h is a pair dictator for A, C. If X_k and X_h are different, X_j cannot force a change in the electorate's relative ranking of A and C. But we have already seen that X_j can do just that, in the transformation from P_{j-1} to P_j. So X_k and X_h must be the same voter, and therefore be a dictator. So the voting system has a dictator. □

5.3 Systems that Do Not satisfy IIA

It is easy to see that the Borda count does not always satisfy the Independence of Irrelevant Alternatives, even in a very small case. Consider three candidates and eight voters, with the profile

3	5
A	C
B	B
C	A

Using a 3-2-1 Borda count, C wins with 18 points; B receives 16 and A 14. However, if the three voters who preferred A decided that B was a better candidate, the profile would be

3	5
B	C
A	B
C	A

and B would be the winner, with 19 points compared to C's 18 and A's 11. No voter changed their preference ordering of B and C, but the electorate's ordering of those two candidates has changed.

The Hare system also does not satisfy the Independence of Irrelevant Alternatives. For example, consider

2	4	2	3
C	A	B	C
A	B	C	B
B	C	A	A

Under the Hare system, no one meets the quota of 6, so B is eliminated; then C wins 7–4. But if just two voters changed their preference list from $C > B > A$ to $B > C > A$, the profile becomes

2	4	4	1
C	A	B	C
A	B	C	B
B	C	A	A

and C is eliminated on the first count. Then A wins the election. No voter changed their ordering of A and C.

Bucklin's method does not satisfy the Independence of Irrelevant Alternatives either. To see this, consider the preference profile

5	5	2	3	5
A	B	C	C	D
B	C	B	D	C
D	A	A	A	B
C	D	D	B	A

The quota is 11. No one meets the quota initially, but after second preferences are added, C has 15 votes, B has 12, and the others do not meet the quota. So C is elected.

However, suppose the five voters with first preference B were to exchange A and C in their preferences. The profile is now

5	5	2	3	5
A	B	C	C	D
B	A	B	D	C
D	C	A	A	B
C	D	D	B	A

Again no one meets the quota initially, but after second preferences are added, B has 12 votes; no one else meets the quota, and B is elected. No voters changed the relative rankings of B and C.

5.4 Systems that Do Not Satisfy Monotonicity

It is not hard to see that the Hare system does not satisfy monotonicity. For example, consider the profile

1	4	6	5
A	A	B	C
B	C	C	A
C	B	A	B

No candidate has met the quota; A and C each have one vote, so both are eliminated, and B wins the election. However, if one of A's supporters were to promote B to first place, resulting in the profile

1	4	6	5
B	A	B	C
A	C	C	A
C	B	A	B

then A would be eliminated, and C would beat B by 9–7. Examples that do not involve a tie would involve more than one voter having changed their preference list; see Exercise 4.

Runoff voting does not satisfy monotonicity. For example, with the preference profile

11	2	7	4	4
A	B	B	C	C
B	A	C	A	B
C	C	A	B	A

the runoff would be between A and B, and A would win by 15–13; if the two voters represented by the second column were to change their ballots by moving A to first place, the resulting profile

13	7	4	4
A	B	C	C
B	C	A	B
C	A	B	A

would see B eliminated, and C would win the runoff 15–13.

5.5 The Gibbard-Satterthwaite Theorem

As we said earlier, A. Gibbard [16] and M. A. Satterthwaite [25] independently proved the *Gibbard-Satterthwaite Theorem*, Theorem 3 below, which essentially says that a dictatorship is the only voting system for three or more candidates that cannot be manipulated.

Gibbard-Satterthwaite uses the ideas of No Dictators and Unanimity, although unanimity is expressed slightly differently: a system is *unanimous* if, whenever candidate A is ranked first by every voter, then A will win. Instead of IIE, it uses the following definition: a voting system is *strategyproof* (or *non-manipulable*) if a voter can never improve the chances of her favorite candidate by strategic voting; a voter will always obtain the best result by ranking the candidates according to her true preferences.

For the rest of this section we shall assume that there are at least three candidates, and that the voting system satisfies the Pareto condition, and is unanimous and strategyproof. Each voter is required to have a preference list of all candidates, with no ties allowed. We shall prove several lemmas, leading to the Gibbard-Satterthwaite Theorem, Theorem 3. Our proof is based on the work of Taylor [29, 30]; see also [31]. Two other relatively simple proofs are given by Benoît [7] and Ninjbat [20].

Say S is some set of voters and A and B are two candidates. We say "S can use A to block B" if, whenever every member of S ranks A above B then B will not be elected. We denote this by $A >_S B$. If there is even one preference profile in which every voter in S ranks A higher than B but every voter not in S ranks B higher than A, and B is not elected, then S can use A to block B. We state this as a Lemma:

Lemma 2. *To show that S can use $A >_S B$, it suffices to produce a (single) preference profile in which (i) every voter in S ranks A higher than B; (ii) every voter not in S ranks B higher than A; (iii) B is not elected.*

Lemma 3. *Say A, B and C are three candidates, and say S is a set of voters such that $A >_S B$. Select a partition $S = P \cup Q$ of S into two disjoint sets. Then either $A >_P C$ or $C >_Q B$.*

Proof. We use Lemma 2. It suffices to produce a preference profile in which either

(i) Everyone in P ranks A higher than C, every other voter ranks C over A, and C is not elected, so from Lemma 2 $A >_P C$; or

(ii) Everyone in Q ranks C higher than B, every other voter ranks B over C, and B is not elected, so from Lemma 2 $C >_Q B$.

We consider a profile of the form

P	Q	R
A	C	B
B	A	C
C	B	A

where all voters place any candidates other than A, B, C lower than those three, all voters in a set have the same preference list, and R is the set of all voters not in S. Clearly the Pareto condition implies that A, B or C will be elected; moreover, every member of S prefers A to B, and $A >_S B$, so B will not be elected. Either A or C will win.

Say A is elected. Every voter in P ranks A over C and every other voter ranks C over A. So $A >_P C$. Conversely, if C is elected, every voter in Q ranks C over B and every other voter ranks B over C. So $C >_Q B$. □

Lemma 4. *Say A and B are two candidates, and say S is a set of voters such that $A >_S B$. Then S can use A to block any third candidate, and S can use that same third candidate to block B.*

Proof. In Lemma 3, suppose Q is empty, and $P = S$. Say the third candidate is C. Either $A >_P C$ or $C >_Q B$. But an empty set of voters cannot block a candidate. So P—that is, S—can use A to block C. If instead we take P to be empty, we see that S can use C to block B. □

Lemma 5. *If there are candidates A and B such that $A >_S B$, then $B >_S A$.*

Proof. Select a third candidate C. From Lemma 4, $A >_S C$. Now apply Lemma 4 again, but with the roles of B and C reversed; Since $A >_S C$, it follows that $B >_S C$. Finally, use Lemma 4 again; since $B >_S C$, S can use B to block any candidate, and in particular $B >_S A$. □

We shall call a set S of voters a *dictating set* if $A >_S B$ for every pair of candidates A, B. A one-element set of voters is a dictating set if and only if the member is a dictator. In any unanimous system, the set of all electors is a dictating set: all that is required is for all the voters to put A first, and A will block any other candidate.

Lemma 6. *If there are candidates A and B such that $A >_S B$, then S is a dictating set.*

Proof. Suppose C and D are any two candidates. We show that $C >_S D$. Since C and D are any two candidates, this means that S is a dictating set.

First, suppose $D = A$. We know from Lemma 5 that $B >_S A$, so Lemma 4 says that S can use any third candidate to block A. Taking C as the third candidate, we have the result. If $D \neq A$, Lemma 4 says that S can use A to block any third candidate, D say. So, using Lemma 4 again, S can use any third candidate to block D; this time, take C as the third candidate. □

Lemma 7. *Say S is a dictating set of voters. Select a partition $S = P \cup Q$ of S into two disjoint sets. Then either P or Q is a dictating set.*

Proof. Say A, B and C are three candidates. Since S is a dictating set, $A >_S B$. From Lemma 3, either $A >_P C$ or $C >_Q B$. In the first case, Lemma 6 shows that P is a dictating set. In the second case, Lemma 6 shows that Q is a dictating set. □

Theorem 3. *Suppose there are at least three candidates, and the electoral system requires each voter to have a preference list of all candidates, with no ties allowed. If the system is unanimous and strategyproof, and satisfies the Pareto condition, then it is a dictatorship.*

Proof. Suppose there are n voters. Unanimity implies that the set of all voters is a dictating set. Partition the set of all voters into two non-empty sets, say P_1 and Q_1. It follows from Lemma 7 that either P_1 or Q_1 is a dictating set. Say it is P_1; this set will have at most $n - 1$ members. Applying Lemma 7 again, it follows that there is a proper subset of P_1 that is a dictating set, and that set will have at most $n - 2$ members. Continue in this way. After at most $n - 1$ iterations, we find that there is a dictating set with one element, that is, a dictator. □

Exercises 5

1. Prove that plurality voting satisfies monotonicity.

2. Prove that plurality voting satisfies Pareto Efficiency.

3. Prove that the runoff method does not satisfy Independence of Irrelevant Alternatives.

4. Here is the profile for 15 voters choosing between three candidates A, B and C. The Hare system is to be used.

4	5	3	3
A	C	B	B
B	A	C	A
C	B	A	C

Use this profile to show that the Hare system does not satisfy the requirement of monotonicity.

5. Prove that the runoff method satisfies Pareto Efficiency.

6. Prove that the Hare system satisfies Pareto Efficiency.

7. Prove that the Condorcet system satisfies No Dictators, Independence of Irrelevant Alternatives and Pareto Efficiency. Why does this not contradict Arrow's Theorem?

8. Use the preference profile

5	8	7
A	B	C
B	C	A
C	A	B

to prove that the Coombs rule does not satisfy Independence of Irrelevant Alternatives.

9. Use the preference profile

2	3	5
A	B	C
B	C	B
C	A	A

to prove that Bucklin's method does not satisfy Independence of Irrelevant Alternatives.

10. Does pairwise sequential voting satisfy Independence of Irrelevant Alternatives?

11. Recall that an electoral system satisfies *non-imposition* if, for every candidate, there exists a preference profile for which that candidate is the winner. Prove that if a system is Pareto efficient then it must also satisfy non-imposition, but that the converse is not true.

12. Prove that the Gibbard-Satterthwaite Theorem still holds if we replace the Pareto criterion by non-imposition.