Notes on Social Choice Theory

Social choice theory is the study of how decisions are made collectively. It examines the idea that, for a given society, the preferences of individuals can be directly aggregated to reflect a quintessential "social preference." An example of such a preference aggregation method is simple plurality rule, in which citizens vote for a candidate and the candidate winning the most votes is deemed society's most-preferred. In general, social choice theory does not aim to analyze or predict individual behavior, but rather to compare and evaluate different means of aggregating preferences.

1 Notation you should be comfortable with

Symbol	Definition	Example	Meaning
X	Set of all alternatives	$\|X\| \geq 3$	Assume 3 or more alternatives
w, x, y, z	Specific alternatives	$x \neq y$	Alternative x doesn't equal y
N	The set of all voters	$\|N\| \geq 2$	Assume 2 or more players (voters)
$\succ_i$	Strict individual preference relation	$x \succ_i y$	Player i strictly prefers x to y
$\succ$	Strict social preference relation	$x \succ y$	Society strictly prefers x to y
$\succeq$	Weak social preference relation	$x \succeq y$	Society weakly prefers x to y
$\sim$	Social indifference	$x \sim y$	Society indifferent between x and y
$\in$	Is an element of	$x \in X$	x is an element of the set X
$\forall$	For all	$\forall x \in X$	For all x in the set X
$\Rightarrow$	Implies	$A \Rightarrow B$	A implies B
$\exists$	There exists	$\exists x \in X$	There exists alternative x in set X

*Note: When there is no ambiguity, we will sometimes also use $\succ$ (without a subscript) to denote the strict individual preference relation.

2 Preliminary definitions to know

- **Binary relation**: A binary relation over the set X describes the relative merits of any two outcomes in X with respect to some criterion.

A simple example is letting $X = \{1, 2, 3, 4\}$, and choosing the binary relation $\geq$. We can compare any two elements of X with this relation. For example, choosing elements 1 and 2, we can say that $2 \geq 1$.

The following definitions can be used to describe a binary relation.

- **Complete**: A complete binary relation implies $\forall x, y \in X$, $x \succeq y$ or $y \succeq x$ or both.

This simply means that any two elements of X can be compared with each other, and either (1) one is strictly preferred to the other or (2) there's indifference between the two.

- **Transitive**: A transitive binary relation implies $\forall x, y, z \in X$, $x \succeq y$ and $y \succeq z \Rightarrow x \succeq z$.

For example, if I weakly prefer apples to oranges and oranges to pears, then I must weakly prefer apples to pears.

- **Strict**: A strict binary relation implies $\forall x, y \in X$ such that $x \neq y$, either $x \succ y$ or $y \succ x$.

This means that no indifference between different alternatives is allowed. Obviously I'm indifferent between apples and apples, but I must have a strict preference over the choice between apples and oranges.

*Note: Given the assumption of *strict*, all that *complete* adds is that $\forall x \in X$, $x \sim x$. In other words, any alternative is indifferent to itself.

- **Strict order**: Any binary relation that is complete, transitive and strict.

For example, the $\geq$ binary relation is a *strict order* over the set of integers. A number is only equal to itself. For any other pair of numbers, one is strictly greater than the other.

- **Cycle**: A binary relation admits a cycle if $\exists x, y, z \in X$ such that $x \succ y \succ z \succ x$.

In other words, if a binary relation violates transitivity, it must admit a cycle.

3 Voters

- A **profile**, p, is a listing of all voters and their preferences.

The following is an example of a profile:

Voter	Preferences
Player 1	$x \succ y \succ z$
Player 2	$z \succ y \succ x$
Player 3	$x \succ z \succ y$
Player 4	$x \succ y \succ z$
Player 5	$y \succ x \succ z$

In order to analyze the properties of different voting rules, we need to make certain assumptions about the behavior of voters. In particular, if voters are irrational, then the outcomes a voting procedure produces will also be irrational. Thus, we assume voters have *rational preferences*.

- Voters have **rational preferences**: This means that each voter's preference relation $\succ$ is a strict order (complete, transitive, and strict).

Essentially this means that no voter can order alternatives in a cycle. I can't prefer cats to dogs to ferrets to cats.

This is our *ONLY* assumption about voter preferences. Thus, we satisfy the condition of *unrestricted domain*.

- **Unrestricted domain** means that, other than voter preferences being a strict order, no other restrictions are made on voter preferences. Players can rank the alternatives however they want.

An example of a restriction on voter preferences would be requiring that no voter rank x as her favorite alternative, or requiring that every voter prefer y to z.

4 Voting rules and their properties

- A **social welfare function**, F, takes a profile of voter preferences, p, and assembles a societal ranking of the candidates. Unlike individual rankings, this societal ranking is not necessarily strict. Some alternatives may be socially indifferent to each other.

Given the profile of the previous page

Voter	Preferences
Player 1	$x \succ y \succ z$
Player 2	$z \succ y \succ x$
Player 3	$x \succ z \succ y$
Player 4	$x \succ y \succ z$
Player 5	$y \succ x \succ z$

an example of a social welfare function is the *plurality vote*, where each player casts a single vote for his most-preferred candidate. Thus, x gets the votes of Players 1, 3, & 4, y gets the vote of Player 5 and z gets the vote of Player 2. So x gets a score of 3, and y and z get scores of 1 each. The social welfare function spits out the ranking:

$$x \succ y \sim z.$$

- A **social choice function** takes a profile of voter preferences and produces a winner.

This is more similar to the concept of an "electoral system" in practice. An example is the first-past-the-post system used in the United States, where the winner is the candidate that receives the

most votes. Using the profile above, this would label x as the winner. Note that you can create a social choice function from any social welfare function by simply picking out the top-ranked alternative.

Clearly we can imagine some social welfare functions that are better than others. No one wants a system that always enacts the favorite policy of a horrible person. Or a system that simply chooses a policy at random. Thus, there are certain normatively "nice" conditions that we might want our social welfare functions to satisfy.

4.1 Nice conditions

- **Pareto**: If $\forall i \in N \; x \succ_i y$, then F should rank x higher than y.

Pareto means that if *every* player strictly prefers x to y, then our social welfare function should not rank y higher than x.

- **Binary Independence / Independence of Irrelevant Alternatives / IIA**: All these terms mean the same thing. Suppose p_1 and p_2 are two profiles where each voter's ranking over the pair (x, y) is the same under both profiles. Then a social welfare function F satisfies IIA if it produces the same (xy) ranking when applied to each profile.

This concept is difficult, and is best explained with an example. Consider the following two preference profiles, p_1 and p_2:

Profile	Player	Preferences
p_1	Player 1	$x \succ y \succ z$
	Player 2	$y \succ z \succ x$
	Player 3	$z \succ x \succ y$

Profile	Player	Preferences
p_2	Player 1	$x \succ y \succ z$
	Player 2	$y \succ x \succ z$
	Player 3	$z \succ x \succ y$

Under both profiles, the players' rankings over the pair (y, z) remains the same: Players 1 &2 prefer y to z under both profiles, and Player 3 prefers z to y under both profiles. However, now consider the results we get with the social welfare function *Borda count*. Under Borda count, a

player's top-ranked alternative gets 2 points, their middle alternative gets 1 point, and their lowest-ranked alternative gets 0 points. Thus, under p_1 and p_2, Borda Count yields the following point scores and social ranking:

$$p_1 : x = 3, y = 3, z = 3 \text{ and } x \sim y \sim z$$

$$p_2 : x = 4, y = 3, z = 2 \text{ and } x \succ y \succ z.$$

This violates IIA because even though players' pairwise rankings of y and z are unchanged between the two profiles, under p_1 Borda count yields the ranking $y \sim z$ and under p_2 it yields the ranking $y \succ z$.

Important point: To prove a procedure violates IIA we must assume two profiles, p_1 and p_2, with the SAME number of voters and the SAME alternatives.

- **No Dictator**:A social welfare function satisfies *no dictator* if there is no particular voter such that the societal ranking of alternatives always agrees with that voter's ranking.

A silly example would be: Say we're trapped on a desert island and are deciding whether to pick bananas or coconuts. Suppose there's a person named Melissa such that, regardless of everyone else's preferences, we always pick Melissa's favorite. Melissa's favorite could be coconuts today and bananas tomorrow, and we'd pick coconuts today and bananas tomorrow. Then Melissa is a dictator.

- **Minimal Liberalism**: A social welfare function, F, satisfies minimal liberalism (ML) if there exist at least two voters who are each decisive over a specified pair of alternatives. *Decisive* means that the specified voter's ranking over an assigned pair equals the relative societal ranking.

Minimal liberalism models the rights of the individual; it implies that certain decisions are solely within our own spheres of influence (such as which shoes I chose to wear today). An example is: let there be three possible alternatives we can choose from. $X = \{$We eat tomatoes, We eat carrots, We eat celery $\}$. Suppose that I am very allergic to tomatoes and will die if I eat them. A

logical conclusion is that my preferences over tomatoes and all other veggies outweigh all others. But to be democratic, I'll let you decide between carrots and celery. This is an example of minimal liberalism being satisfied; I can completely determine the choice between the pairs of alternatives $\{$ We eat tomatoes, We eat carrots$\}$ and $\{$We tomatoes, We eat celery $\}$, and you can completely determine the choice between the two alternatives $\{$We eat carrots, We eat celery$\}$.

*Note: By relaxing IIA with an assumption like ML, we can look at interpersonal comparisons of welfare; ie. our rule can reflect the fact that an outcome may be much more important to one player than to another.

- **Gibbard Decisiveness**: For n people, define two social states (or alternatives) a and b, given by $a = (a_1, a_2, ..., a_n)$ and $b = (b_1, b_2, ..., b_n)$. a_i and b_i define what the i^{th} player does, and are thus in his *private domain*. A player is Gibbard decisive for "a over b" if and only if a and b differ only in that player's sphere of influence (ie. if $a_k = b_k \; \forall \; k \neq i$).

For example, if our choice of alternatives is $X = \{$(I wear red, you wear blue), (I wear blue, you wear blue), (I wear blue, you wear red)$\}$, then our procedure is Gibbard decisive if I can completely choose between the two alternatives$\{$(I wear red, you wear blue), (I wear blue, you wear blue)$\}$ and you can completely choose between the two alternatives $\{$(I wear blue, you wear blue),(I wear blue, you wear red)$\}$.

5 Theorems

Now we know about rational individual preferences and we know about various "nice" conditions a social welfare function can satisfy. What do all these conditions imply?

Arrow's Theorem: Consider any social welfare function F which always produces a transitive ordering over 3 or more alternatives. Assume at least two voters, and that every voter's preferences form a strict order and satisfy universal domain. Then if F satisfies Binary Independence and Pareto, F is a dictator.

Arrow's Theorem tells us that the only social welfare function satisfying transitivity, Pareto and Binary Independence is a dictatorship. There are many implications of this theorem, and one of the most important we've discussed is how most voting procedures produce rankings over the set of alternatives that conflict with the pairwise rankings of those alternatives. What does this mean?

- **Pairwise Ranking:** For the purposes of this class we'll think of the pairwise ranking of two alternatives as simply the majority's preference of one alternative over another. Thus, the pairwise ranking of x and y is

  $$x \succ y$$

  if x is majority-preferred to y,

  $$y \succ x$$

  if y is majority-preferred to x, and

  $$x \sim y$$

  if they're tied.

The following theorem is an immediate consequence of Arrow.

Theorem 2: Consider any non-dictatorial social welfare function for 3 or more alternatives that produces a transitive societal ranking of alternatives. Then there are instances (i.e. individual profiles) where the rankings of the procedure and of the pairwise vote are in conflict.

Theorem 2 implies that there is no transitive, non-dictatorial way of ranking *all* the alternatives that will always agree with the pairwise ranking of alternatives. If you think about, the proof of this is obvious. The pairwise ranking procedure is always Pareto and always satisfies Binary Independence (because only two alternatives are ever being compared at once). By Arrow we know that we can't have all four: Transitive, non-dictatorial, Pareto and Binary Independence. Thus, if a procedure satisfies the first two (Transitivity, non-dictator) it *can't* satisfy the second two (Pareto, Binary Independence).

We'll prove this theorem on the next page using a *proof by contradiction*.

- **Proof by Contradiction:** First make an assumption and derive a consequence of that assumption. If the consequence is false, then you've proved the assumption must be false.

Example: We're trying to prove that the # of female accountants is weakly less than the # of accountants. (If we assumed the opposite we would be committing a "conjunction fallacy"). To prove this by contradiction, assume the contrary: Assume

$$\#\text{ female accountants } > \#\text{ accountants}.$$

This implies that

$$\exists \text{ a female accountant who isn't an accountant.}$$

And this implies a contradiction. Thus we've proved that the # of female accountants is weakly less than the # of accountants.

Proof of Theorem 2: Assume the contrary. Assume that $\exists$ a non-dictatorial social welfare function F over 3 or more alternatives that produces a transitive societal ranking over alternatives, and *assume F always agrees with the pairwise ranking.*

Then we know the following:

1. F is non-dictatorial (we've assumed it)
2. F is transitive (we've assumed it)
3. F satisfies Pareto (because F has same ranking as the pairwise vote, and the pairwise vote satisfies Pareto)
4. F satisfies Binary Independence (because F has same ranking as the pairwise vote, and the pairwise vote satisfies Binary Independence)

We now immediately have a contradiction, because Arrow's Theorem tells us that F can't satisfy all four of those conditions. And thus, we've proved our result! F *can't* always produce the same rankings as the pairwise vote, because then Arrow's Theorem would be violated.

Sen's Theorem': Assume 3 or more alternatives and 2 or more voters with strict preferences and universal domain. If a social welfare function F satisfies Minimal Liberalism and Pareto, then there exist profiles where F cycles (is not transitive).

Example: Suppose 4 voters and 4 alternatives, so $X = \{A, B, C, D\}$. Since ML holds, assume Player 1 decides between the pair $\{A, B\}$, and Player 2 decides between the pair $\{C, D\}$. Preferences look like this (note Players 2, 3 and 4 all have the same ranking):

$$\text{Player 1} : D \succ A \succ B \succ C$$

and

$$\text{Players 2, 3, and 4} : B \succ C \succ D \succ A.$$

Then we get

Voter	$\{A, B\}$	$\{B, C\}$	$\{C, D\}$	$\{A, D\}$
1	$A \succ B$	$B \succ C$		$D \succ A$
2		$B \succ C$	$C \succ D$	$D \succ A$
3 and 4		$B \succ C$		$D \succ A$
Outcome	$A \succ B$	$B \succ C$	$C \succ D$	$D \succ A$

The "outcomes" row shows that a societal preference cycle has occurred. All we needed for this result was the assumption of Pareto, and the assumption that Player 1 makes society's choice between A and B and Player 2 makes society's choice between C and D. Note that preferences were unanimous over the $\{B, C\}$ and $\{A, D\}$ choices.

Gibbard's Theorem: With 2 or more people, there does not exist a social choice function that is transitive and satisfies Gibbard Decisiveness.

The reason for this is because, although agents decide their own actions, the societal state each agent wants to occur depends on the actions of others. Be careful to remember that Gibbard Decisiveness is only with respect to *actions*, not *payoffs*. In being able to choose whether to wear a red or blue shirt I may not affect the *choices* available anyone else, but I may affect their *payoffs*. I'll provide the classic example of the **Prisoner's Dilemma**.

The story behind the Prisoner's Dilemma is that two criminals are arrested. The police don't have enough evidence to convict both, so they are put in separate interrogation rooms. Under questioning, the prisoners have to choice to either Cooperate (with each other, and not give the police any information), or Defect (and rat out their accomplice). The following table gives the payoff each player receives dependent on his action and the action of his accomplice.

Prisoner's Dilemma		Player 1	
		Cooperate	Defect
Player 2	Cooperate	5,5	10,0
	Defect	0,10	1,1

Player 1's payoffs appear first in each cell. Thus, if Player 1 cooperates and Player 2 defects,

Player 1 receives a payoff of zero and Player 2 receives a payoff of 10. You can see that regardless of the action of the other player, it is always in the best interest of a player to defect (if the other cooperates you get 10 from defecting and 5 from cooperating, if the other defects you get 1 from defecting and 0 from cooperating).

Think of the alternatives (or states of the world) as consisting of an action by each player. Thus, $X = \{(C, C), (C, D), (D, C), (D, D)\}$. Using the concept of Gibbard Decisiveness, it is clear that Player 1 is Gibbard decisive over the choice between the alternatives$\{(C, D), (D, D)\}$ and between the alternatives $\{(C, C), (D, C)\}$; i.e. given that Player 2 is going to either cooperate or defect, Player 1 can decide on his own action. Player 2 is also Gibbard decisive over his corresponding pairs of alternatives. Using the table above, we can see how the players rank the alternatives:

$$\text{Player 1: } (D, C) \succ (C, C) \succ (D, D) \succ (C, D)$$

and

$$\text{Player 2: } (C, D) \succ (C, C) \succ (D, D) \succ (D, C).$$

To see how a majority-preference cycle occurs in this example, we'll construct the following table:

Player	(D,C),(C,C)	(D,D),(C,D)	(D,C),(D,D)	(C,C),(C,D)	(CC),(DD)
Player 1	$(D, C) \succ (C, C)$	$(D, D) \succ (C, D)$			$(C, C) \succ (D, D)$
Player 2			$(D, D) \succ (D, C)$	$(C, D) \succ (C, C)$	$(C, C) \succ (D, D)$
Outcome	$(D, C) \succ (C, C)$	$(D, D) \succ (C, D)$	$(D, D) \succ (D, C)$	$(C, D) \succ (C, C)$	$(C, C) \succ (D, D)$

All we have assumed in this example in order to get outcomes is the Pareto condition and Gibbard Decisiveness. Looking at the "outcomes" row, we can see that two majority preference cycles have occurred:

$$(D, C) \succ (C, C) \succ (D, D) \succ (D, C)$$

and

$$(C, D) \succ (C, C) \succ (D, D) \succ (C, D).$$

6 A last example

Condorcet's Paradox: A classic example of a majority-preference cycle. Everyone should be very familiar with this example.

Players	Preferences
Player 1	$x \succ y \succ z$
Player 2	$z \succ x \succ y$
Player 3	$y \succ z \succ x$

Consider pairwise majority rule, where an alternative x defeats another alternative y if a majority prefers x to y. Players 1&2 prefer x to y, Players 1 & 3 prefer y to z, and Players 2 & 3 prefer z to x. We get the following social preference relation:

$$x \succ y \succ z \succ x.$$