Comp 553: Game Theory

Jacob Reznikov

January 19, 2022

Abstract

My course notes for Game Theory

1 Classical Game Theory

1.1 Games

First we start with what we mean by a game.

Definition 1.1.1. *A **game** is any interaction between two or more agents with non-identical objectives.*

We next need a way to model the objectives of agents in a game.

Definition 1.1.2. *Let O be the set of possible outcomes of a game. The **preference list** of agent i is a relation $\succ_i$ over the set O of the game with the following properties:*

- ***Completeness**. For any two distinct states $a, b \in O$ we either have*

$$a \succ_i b \oplus b \succ_i a$$

- ***Transitivity**. For any states $a, b, c \in O$ we have*

$$a \succ_i b \wedge b \succ_i c \implies a \succ_i c$$

The set of all preference lists over outcomes O is denoted $\Pi(O)$.

It is difficult to study a pure relation and so instead we like to work with so called 'Utility Functions'.

Definition 1.1.3. *The **Utility Function** of agent i is a function $u_i : O \to \mathbb{R}$ that corresponds to the preference list $\succ_i$ by*

$$u_i(a) > u_i(b) \iff a \succ_i b, \forall a, b \in O$$

Remark 1.1.4. It is important to note that we cannot always assume the existence of a utility function.

The central assumption of game theory is that an agent seeks to maximize its utility.

Definition 1.1.5. *An agent that seeks to maximize its utility is called* ***rational****.*

Remark 1.1.6. The utility function of one agent can incorporate other agent and/or their goals.

1.2 Game Formulation

We now want to consider how we represent games mathematically.

Definition 1.2.1. *The* ***Normal Form*** *of a 2-player game is two payoff matrices A and B where for the player a choice i and player b choice j player a receives A_{ij} utility and player b receives B_{ij} utility.*

Example 1.2.2. The **chicken game** is defined as follows, both players have strategy 1 - eagle, and strategy 2 - chicken. The payoff matrices are then

$$A = \begin{pmatrix} 0 & 5 \\ 1 & 4 \end{pmatrix} \qquad B = \begin{pmatrix} 0 & 1 \\ 5 & 4 \end{pmatrix}$$

Here if both play eagle both players get 0. If one player plays eagle and the other players chicken the one who played eagle gets 5 while the other gets 1. If both players play chicken they both get 4.

It is also useful to combine the two matrices into one payoff table, as can be seen below.

	E	C
E	0,0	5,1
C	1,5	4,4

Definition 1.2.3. *A* ***best response*** *of agent i is their best outcome given some fixed configuration of the other agents choices.*

We say that an outcome is a ***Nash Equilibria*** *if it is the best response of every agent given the configuration of choices of the other agents that led to this outcome.*

Example 1.2.4. In the example above we have two Nash Equilibria which we have highlighted in yellow below, note that the blue value is the highest in its row and the red is the highest in its column in this table for both of the outcomes. This makes these outcomes Nash Equilibria.

	E	C
E	0,0	5,1
C	1,5	4,4

Because Nash Equilibria are mutual best responses, no rational agent will deviate from an Equilibria and so they can be considered stable points of a game. We can thus reframe questions about the nature of the game into questions about the properties of its Nash Equilibria.

The simplest questions we can ask then are what is the interpretation of multiple Nash Equilibria and of no Nash Equilibria.

Example 1.2.5. The simplest example of no Nash Equilibria is the game of rock paper scissors which has the payoff table

	R	P	S
R	0,0	-1,1	1,-1
P	1,-1	0,0	-1,1
S	-1,1	1,-1	0,0

On an intuitive level, the reason why no Nash Equilibria exists is because the concept of a best response is tied to having *knowledge* of the other players moves, which in this case allows for a self reference paradox if a best strategy existed. In order to circumvent this then we can introduce *uncertainty*. To do this we use probability theory:

1.3 Mixed Strategies

Definition 1.3.1. *A **Mixed Strategy** of agent i is a probability distribution p over the outcomes O. In the 2 agent case the value of a mixed strategy with distribution p given that the other agent's distribution is q is defined as*

$$p^T(Aq)$$

where A is its payoff matrix and p and q are written as vectors in $L^2(O)$. In the 3 or more agent case this can be generalized through tensor contractions.

This probabilistic version of a strategy also carries with it the probabilistic version of a best response.

Definition 1.3.2. *A Mixed best response of agent i to a configuration of other agents mixed strategy q is defined as the probability distribution p_{max} maximizing its value, i.e.*

$$p_{max} := \text{argmax}_p\, p^T(Aq).$$

A Mixed Nash Equilibria (MNE) is a configuration of mixed strategies $\hat{p}$ and $\hat{q}$ such that

$$\hat{p} = \text{argmax}_p\, p^T(A\hat{q}), \quad \hat{q} = \text{argmax}_q(\hat{p}^T B)q$$

Example 1.3.3. In the rock paper scissors game there is only one unique Mixed Nash Equilibria and it is $\hat{p} = \hat{q} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$.

Proof. It is clear that $\hat{p}$ and $\hat{q}$ are best responses to each other as

$$\left(\begin{bmatrix} a \\ b \\ c \end{bmatrix}\right)^T \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{bmatrix} = 0$$

And by symmetry of the game they are both the best responses to each other.

For proof of uniqueness I am not sure about an exact proof but my intuition is that if the probability is not symmetric under the permutation (123) then applying that permutation makes one player win more and so that must be his best response but then the other player can apply (132) and win more than the other and so their initial probability could not have been the best response. □

In fact we have that in any game there is a Mixed Nash Equilibria.

Theorem 1.3.4 (Nash's Theorem). *Any game with finite players and finite outcomes has a MNE.*

It seems like mixed strategies solve most of our issues but we still want to be careful as they are not always good models of the real world. For example in an industrial location one will never flip a coin to decide whether they should fix a machine or not (hopefully). In such scenarios the justification for this model is to instead treat them as the limit of how an agent would play a game if it was played many times or how the average agent would play the game when drawn from a large population.

One needs to be very careful when applying models to games like these as it implies that the game being played is different then the exact one that the mathematics describes, i.e. game is repeated many times in a row.

Example 1.3.5. Consider the following game called the prisoner's dilemma, in this game each player can choose to deny (D) their involvement with a crime or confess (C). The payoffs are the years they get in prison are represented by the payoff table:

	D	C
D	-1,-1	-9,0
C	0,-9	-8,-8

Note that in this game no matter what the other player's strategy is you always gain by confessing rather than denying. Because of this the NE of this game is (C, C).

Definition 1.3.6. *Given an agent i and two of its strategies C and D we say that D (strictly) dominates C if i has a (strictly) bigger payoff with D than with C no matter what configuration of other strategies the other agents choose.*

Additionally a Nash Equilibrium where the strategy of every player is dominant is called a Nash Equilibrium in dominant strategies.

NE in dominant strategies are often much better at predicting real world behavior than pure NE.

However, note that the NE of the prisoner's dilemma is overall a bad outcome compared to the other outcomes of the game, we confirm this using another concept.

Definition 1.3.7. *An outcome $\hat{O}$ is called pareto optimal if there does not exist an outcome O where outcome O is at least as good as outcome $\hat{O}$ for every agent and strictly better for at least one agent.*

I.e. in order for an agent to improve its outcome from a pareto optimal one another agent must end up worse off.

Example 1.3.8. In the prisoners dilemma we can find 3 different pareto optimal outcomes shown here in green.

	D	C
D	-1,-1	-9,0
C	0,-9	-8,-8

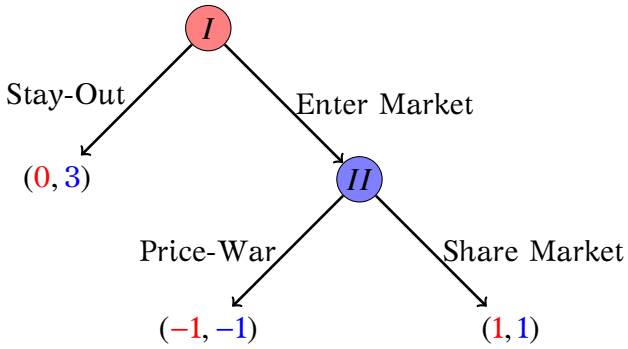
Note how the NE is the only non-pareto optimal outcome.

1.4 Extensive Form

There is also another way to formulate a game that can often be more intuitive than the normal form.

Definition 1.4.1. *The extensive form of a game is a tree where each internal node represents a choice that an agent can make and each leaf is an outcome.*

Example 1.4.2. The game represented by the following extensive form is called Price-Wars.



Note that in extensive form the strategy of a player is a fixed choice at *every* choice node in the tree, even if it is not reached by the choices of other players.

This game has two NE, one is (Stay-Out, Price-War) and the other is (Enter Market, Share Market). However, these are practically very different Equilibria, note that in the first Equilibrium player II is using the threat of a Price-War to keep player I out of the market

but if player I were to enter the market despite the threat it is in player II's best interest to not follow up on the threat and instead share the market. On the other hand the second Equilibrium does not have a similar issue. They key difference is that the subtree of II's choice is a not in Equilibrium in the first NE but it is in the second one.

Definition 1.4.3. *A NE of a game represented in extensive form is called* ***subgame perfect equilibrium*** *(SPE) if it is also an equilibrium of every of every subtree.*

In a similar way as dominant equilibrium a subgame perfect equilibrium is more credible than a regular equilibrium.

2 Rationality

2.1 Rational Agents

We now dive more formally into the properties of a rational Agent.

Definition 2.1.1. *A rational agent is an agent i with preference $\succ_i$ is an agent that when he has to make a choice over some subset $S \subseteq O$ he chooses the outcome $O^* \in S$ where*

$$O^* \succ_i O \quad \forall O \in S$$

A rational agent can always evaluate which options are available to it, know how to evaluate them based on his preference relation, and will always choose the optimal outcome regardless of description.

It may seem like this is way too strong of an assumption to apply to the real world but in reality most non rational agents can be re-imagined as rational agents with unintuitive utility functions.

Example 2.1.2. Let agent I be a student with a weekly budget of $\$B$ which spends the same fractions $f_1, \dots, f_n$ on goods $x_1, \dots, x_n$ regardless of their prices $p_1, \dots, p_n$. It might seem like I is an irrational agent but he could just be a rational agent with the Cobb-Douglas utility function

$$U(x_1, \dots, x_n) = x_1^{f_1} \cdot x_2^{f_2} \dots x_n^{f_n}$$

In general, however, this solution can often be problematic because of problems like overfitting (we can fit any set choices with a detailed enough utility function) and the fact that human interactions very often do not involve goals.

We also have an equivalent definition for a rational agent that is often more useful

Definition 2.1.3. *Let i be an agent with a choice function f where given a subset $S \subseteq O$ of options he chooses $f(S)$. We say that f satisfies* ***independence of irrelevant alternatives (IIA)*** *if*

$$\forall S' \subseteq S \subseteq O, f(S) \in S' \implies f(S') = f(S)$$

Theorem 2.1.4. *Rationality Theorem An agent is rational if and only if its choice function satisfies IIA.*

Proof. It is a clear that if its choice function is based on a preference relation $\succ$ then its choice function satisfies IIA.

On the other hand assume that its choice function satisfies IIA. In that case we define

$$a \succ b \iff f(\{a, b\}) = a$$

with that in mind we can check the requirements for a preference relation.

- Completeness is trivial by definition.
- For transitivity we assume $a \succ b$ and $b \succ c$ then

  $$f(\{a, b, c\}) = c \implies f(\{b, c\}) = c \text{ and } f(\{a, b, c\}) = b \implies f(\{a, b\}) = b$$

  both of these are contradictions and so $f(\{a, b, c\}) = a$ and so by IIA $f(\{a, c\}) = a$ giving us $a \succ c$.

□

Example 2.1.5. An example of an IIA decision process is what we call a satisficer.

The process is described by the following algorithm:

- Select a set $S \subseteq O$ that we declare to be satisfactory.
- Fix an ordering $\{O_1, O_2, \ldots, O_m\}$ of O.
- Given a subset S' of options order S' according to the above ordering and find the first element that is in $S \cap S'$, select that element.

It is a simple procedure to prove that this process satisfies IIA.

2.2 Non-Rational Agents

Satisificers provide a great distinction between how real world humans act and how rational agents act. For non-rational agents like humans the way the search is performed is very important, we often can change how good our decision making is by reframing our search.

These then lead to the 3 different approaches we have to model these type of agents

2.2.1 Optimization under Constraints

We can try and model the fact that agents are constrained by computation-time and other resources and so cannot do arbitrarily large searches. We do this by saying that a constrained agent stops searching when its best guess for *expected future gains* is less than its best guess for *expected future costs*.

The problem with this method is that these types of guesses are often even more complicated to compute than the full search space.

2.3 Heuristics and Biases

The next model we can try is that of a heuristic. A heuristic is a very quick algorithm to quickly make decisions instead of doing the full computations. These Heuristics often violate the basic laws of probability and logic. This then means that humans are demonstrably not rational and therefore make bad decisions.

Example 2.3.1. Anchoring Humans usually rely on the first piece of information they see to anchor the rest of their reasoning. This is why stores use one bad deal to anchor the human mind to make all other deals seem better.

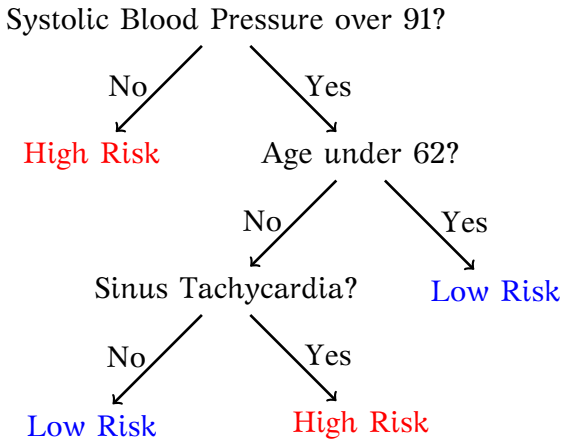
Example 2.3.2. Law of Small Numbers Humans assume that random samples will act similar to their limiting average. This is apparent in many examples

- Gambler's Fallacy: A series of losses increase the chance of a win next.
- Pattern Spotting: Irrational confidence in small early trends.
- Clustering: It is unlikely that random clusters occur in data.
- Medical Trials: Humans overestimate certainty based on data.

However, just because heuristics can lead to bad decisions doesn't mean they have to. A lot of economists argue that heuristic's speed, simplicity and robustness make them extremely useful decision making tools.

Example 2.3.3. Catching a ball When catching a ball we could go through the long process of calculating the trajectory of the ball and computing where we need to stand to catch it, or we can use a simple heuristic like moving to keep it at the same angle compared to us and the ground.

Example 2.3.4. Heart Attacks



This quick decision tree outperforms many more complex algorithms when it comes to this task.

3 Social Choice Theory

3.1 Voting

we now change our focus to societal decision making. Specifically, how society should make decisions bade on the individual preferences of its members. We can model hit by an election of n agent voters and k outcome candidates ($O = [k]$). Each voter i has a preference $\succ_i$ over all candidates.

Given such a setup we can either try to find the 'Winner' candidate out of the group or a total ordering over all candidates.

Definition 3.1.1. *A single choice function is a function*

$$f : (\Pi(O))^n \to O$$

modeling the selection of a single candidate from n preference lists.

Definition 3.1.2. *A social welfare function is a function*

$$f : (\Pi(O))^n \to \Pi(O)$$

modelling the selection of a total ordering of candidates from n preference lists.

There are many systems that are used in practice.

Example 3.1.3. Plurality Plurality is a single choice function that counts the number of voters with candidate j as their top candidate and chooses the one with the highest such number.

Example 3.1.4. Borda Count Borda assigns each position in the list a point value (i.e. 1st place gets 5 points, 2nd place 2 points, etc.) and sums up the points for each candidate over all voters preference lists. The candidate with the most points wins.

Example 3.1.5. Majority Majority is not really a single choice function nor a social welfare function as it only compares candidates pairwise. A candidate a wins over candidate b if the majority of voters prefer a to b. This cannot be made into a social welfare function because it is not necessarily transitive.

Here we see our first problem. In plurality it is possible to benefit by lying about your preference list.

Example 3.1.6. Let there be 100 voters with preference list counts

$$a \succ c \succ b : 49$$
$$b \succ a \succ c : 46$$
$$c \succ b \succ a : 5$$

Here a wins, but the voters with lists $c \succ b \succ a$ can change lie to appear to be $b \succ c \succ a$, in that case b would win. But all those voters prefer b to a so they are incentivized to lie to get their preferred outcome.

3.2 Properties of Voting Systems

We can formalize this idea as follows.

Definition 3.2.1. *A social choice function f is strategically manipulable if for some set of preferences $\{\succ_1, \ldots, \succ_n\}$ there is a $\succ_i'$ such that*

$$f(\succ_1, \ldots, \succ_{i-1}, \succ_i, \succ_{i+1}, \ldots, \succ_n) = a$$
$$f(\succ_1, \ldots, \succ_{i-1}, \succ_i', \succ_{i+1}, \ldots, \succ_n) = b$$
$$b \succ_i a$$

*In this situation voter i has an incentive to lie. A social choice function f is called **incentive compatible** if it is not manipulable.*

Example 3.2.2. Dictatorship A dictatorship social choice function f^D that always outputs the top preference of voted D.

Similarly a dictatorship welfare function f^D always outputs the entire preference list of voted D.

It turns out that a social choice dictatorship is incentive compatible, and furthermore is the only possible incentive compatible choice function.

Theorem 3.2.3. *Gibbard-Satterthwaite Theorem Any incentive compatible social choice function is a dictatorship.*

We also have properties of a social welfare function.

Definition 3.2.4. *A social welfare function f is called **unanimous** if*

$$a \succ_i b \forall i \in [n] \implies a \; f(\succ_1, \ldots, \succ_n) \; b$$

Definition 3.2.5. *A social welfare function f is **independent of irrelevant alternatives** or **consistent** if for any two sets of preference lists $\{\succ_1, \succ_2, \ldots, \succ_n\}$ and $\{\succ_1', \succ_2', \ldots, \succ_n'\}$ where*

$$a \succ_i b \iff a \succ_i' b \quad \forall i$$

then

$$a \; f(\succ_1, \ldots, \succ_n) \; b \iff a \; f(\succ_1', \ldots, \succ_n') \; b$$

Example 3.2.6. Again a dictatorship welfare function satisfies both of the above criterion and is in fact the only such function.

3.3 Arrow's Impossibility Theorem

Theorem 3.3.1. *Arrow's Impossibility Theorem (1950) Any social welfare function f that satisfies unanimity and IIA is a dictatorship.*

Proof. This proof has 3 parts.

- First we show that unanimity and IIA together imply that the social welfare function doesn't distinguish between any pair of candidates.
- Next we show that for some instance of voter preference lists there is a voted D which is locally influential in determining the relative position of some specific a and b.
- Next we combine these two to get that D has complete influence over the relative position of every pair of candidates in all circumstances and so this system is a dictatorship.

Lemma 3.3.2. *f is pairwise neutral, i.e.*

$$(a \succ_i b \iff x \succ_i y \quad \forall i) \implies (a \; f(\succ_1, \ldots, \succ_n) \; b \iff x \; f(\succ_1, \ldots, \succ_n) \; y)$$

Proof. Lets call $\succ_f = f(\succ_1, \ldots, \succ_n)$ and assume that $a \succ_i b \iff x \succ_i y \quad \forall i$ and WLOG assume that $a \succ_f b$ and we want to show that $x \succ_f y$.

First we will simplify our scenario, we do this by defining new preference lists $\succ'_i$ where we move x immediately above a and y immediately below b. This does not change the relative positionof a, b and so by IIA $\succ'_f = f(\succ'_1, \ldots, \succ'_n)$ ranks a, b in the same way as $\succ_f$.

Now since $x \succ'_i a$ for all i we also have $x \succ'_f a$ by unanimity and similarly $b \succ'_f y$ by unanimity. But then we have

$$x \succ'_f a \succ'_f b \succ'_f y$$

and so by transitivity $x \succ'_f y$. But then x and y do not move relative positions in $\succ'_i$ compared to $\succ_i$ and so again by IIA we have

$$x \succ'_f y \implies x \succ_f y$$

giving us what we want. □

Lemma 3.3.3. *There is a pivotal voter D for some pair a, b of candidates.*

Proof. Let $\succ_i$ be arbitrary such that $a \succ_i b$ for all i. Then let $\succ'_i$ be the modified preference lists such that a and b are swapped and so $b \succ'_i a$ for all i. By unanimity we have

$$a \; f(\succ_1, \ldots, \succ_n) \; b \text{ and } b \; f(\succ'_1, \ldots, \succ'_n) \; a$$

So now we know the set $\{i : b \; f(\succ_1, \ldots, \succ_{i-1}, \succ'_i, \ldots, \succ'_n) \; a\}$ is neither empty nor the whole set. Thus there is a minimal element D such that

$$b \; f(\succ_1, \ldots, \succ_{D-1}, \succ'_D, \succ'_{D+1}, \ldots, \succ'_n) \; a \text{ and } a \; f(\succ_1, \ldots, \succ_{D-1}, \succ_D, \succ'_{D+1}, \ldots, \succ'_n) \; b$$

Now by pairwise neutrality we know that if for any a, b we have a situation where $a \succ_i b$ for $i \leq D$ and $b \succ_i a$ for all $i > D$ then $a\ f(\succ_1, \ldots, \succ_n)\ b$. Similarly if we ever have a situation where $a \succ_i b$ for all $i < D$ and $b \succ_i \geq a$ for all $i \geq D$ then $b\ f(\succ_1, \ldots, \succ_n)\ a$. □

Now we can bring it all together. Let D be as above and assume that there are at least 3 candidates x, y, z. Now we want to show that D's preference of x, y decides $f(\succ_1, \ldots, \succ_n)$'s preference and so lets assume $x \succ_D y$. Now by IIA the position of candidate z in the preference lists cannot change the outcome of the vote and so we can freely assume whatever we want about it. Thus let us assume that voters $\{1, \ldots, D-1\}$ rank z as their first choice and voters $\{D+1, \ldots, n\}$ rank z as their last choice and the dictator D ranks it somewhere between x and y so that $x \succ_D z \succ_D y$.

Now we see that clearly we have $z \succ_i y$ for all $i \leq D$ and $y \succ_i z$ for $i > D$ and so by our previous lemma $z\ f(\succ_1, \ldots, \succ_n)\ y$. But also $z \succ_i x$ for all $i < D$ and $x \succ_i z$ for all $i \geq D$ and so again by the previous lemma $x\ f(\succ_1, \ldots, \succ_n)\ z$. We can then apply transitivity to get $x\ f(\succ_1, \ldots, \succ_n)\ y$.

But note that we assumed *nothing* of the preferences of the other voters. Thus their votes did not matter and so D is in fact the dictator for every pair x, y of candidates and thus f is a dictatorship system. □

3.4 Gibbard-Satterthwaite Theorem

Now we will prove 3.2.3 by using Arrow's Theorem.

We will prove this by showing that it is a special case of Arrow's Theorem. The proof will have 5 steps.

- First we show that if f is incentive compatible then it is monotone.
- Then we build a social welfare function g based on f.
- We then show that g is unanimous and consistent.
- Then we use Arrow's Theorem that g is a dictatorship.
- This will imply that f is also a dictatorship.

Incentive compatibility is a hard thing to study and so first we will reformulate it into simpler terms.

Definition 3.4.1. *A social choice function f is called* ***monotone*** *if for some preference lists $\{\succ_j\}_{j \in [n]}$ and $\succ_i'$ we have*

$$(f(\succ_1, \ldots, \succ_i, \ldots, \succ_n) = x \wedge f(\succ_1, \ldots, \succ_i', \ldots, \succ_n) = y) \implies (x \succ_i y \wedge y \succ_i' x)$$

Lemma 3.4.2. *A social choice function is incentive compatible if and only if it is monotone.*

Proof. Assume that f is monotone and not incentive compatible. It is then manipulable and so there exist preference lists $\{\succ_j\}_{j\in[n]}$ and $\succ_i'$ such that

$$f(\succ_1,\ldots,\succ_{i-1},\succ_i,\succ_{i+1},\ldots,\succ_n) = x$$
$$f(\succ_1,\ldots,\succ_{i-1},\succ_i',\succ_{i+1},\ldots,\succ_n) = y$$
$$y \succ_i x$$

but then monotonicity means that $x \succ_i y$ and so we have a contradiction and thus monotonicity implies compatibility.

On the other hand if f is incentive compatible but not monotone then we have some list such that

$$(f(\succ_1,\ldots,\succ_i,\ldots,\succ_n) = x) \wedge (f(\succ_1,\ldots,\succ_i',\ldots,\succ_n) = y) \wedge (y \succ_i x \vee x \succ_i' y)$$

Now assume that $y \succ_i x$, then this is exactly our definition of a manipulable social choice function and so this is a contradiction. Thus we must have $x \succ_i' y$ but then we can switch the roles of x, y and $\succ_i, \succ_i'$ then in that case we have

$$(f(\succ_1,\ldots,\succ_i',\ldots,\succ_n) = y) \wedge (f(\succ_1,\ldots,\succ_i,\ldots,\succ_n) = x) \wedge (x \succ_i' y)$$

this is again our definition of a strategically manipulable social choice function. This is again a contradiction and so our proof is done. □

Next we want to transform f into a social welfare function g. We do this using the Up-Operation.

Definition 3.4.3. *Let $S \subseteq [k]$ be a subset of candidates. Given a preference list $\succ_i$ over $[k]$ we define the new preference list $\succ_i^S$ to be the one created by sliding all the elements of S in to the top of the list while conserving their relative order inside S as given by $\succ_i$.*

Example 3.4.4. Let

$$\succ_i := 4\ 5\ 9\ 1\ 3\ 2\ 8\ 7\ 6$$

then

$$\succ_i^{\{3,7\}} = 3\ 7\ 4\ 5\ 9\ 1\ 2\ 8\ 6$$
$$\succ_i^{\{1,5,6\}} = 5\ 1\ 6\ 4\ 9\ 3\ 2\ 8\ 7$$
$$\succ_i^{\{9\}} = 9\ 4\ 5\ 1\ 3\ 2\ 8\ 7\ 6$$

Lemma 3.4.5. ***Up-Lemma*** *Take any set of preferences* $\{\succ_1, \ldots, \succ_n\}$ *and any subset* S *of candidates then if* f *is monotone and surjective then*

$$f(\succ_1^S, \ldots, \succ_n^S) \in S$$

Proof. Take any candidate $x \in S$ as f is surjective there is a set of preferences $\{\succ_{1,x}, \ldots, \succ_{n,x}\}$ such that

$$f(\succ_{1,x}, \ldots, \succ_{n,x}) = x$$

Now one by one for each voter we replace $\succ_{i,x}$ with $\succ_i^S$. Now assume that with any individual swap the vote changed from an element in S to an element not in S (note that we start in S so if this is not true we also end in S and so our lemma is proven). We then have

$$\left(f(\succ_1^S, \ldots, \succ_{i,x}, \ldots, \succ_{n,x}) = x\right) \wedge \left(f(\succ_1^S, \ldots, \succ_i^S, \ldots, \succ_{n,x}) = y\right) \wedge (x \in S) \wedge (y \notin S)$$

but note that monotonicity implies that $x \succ_{i,x} y$ and $y \succ_i^S x$. But this is impossible as since y is not in S it cannot be above any x in S in the preference list $\succ_i^S$. This is then a contradiction and so this proves our lemma. □

Now with this lemma in hand we can transform f into a social welfare function g. We define

$$g : (x\ g(\succ_1, \ldots, \succ_n)\ y) \iff \left(f(\succ_1^{\{x,y\}}, \ldots, \succ_n^{\{x,y\}}) = x\right)$$

Now we need to check g outputs a preference list.

Lemma 3.4.6. *g is a preference list*

Proof. Completeness is trivial by definition.

For transitivity it suffices to prove there are no 3-cycles since the graph is complete and so any larger cycle can be cut in half reducing it down to the 3-cycle case by induction. Thus assume that there are candidates x, y, z such that

$$x\ g(\succ_1, \ldots, \succ_n)\ y \wedge y\ g(\succ_1, \ldots, \succ_n)\ z \wedge z\ g(\succ_1, \ldots, \succ_n)\ x$$

then let $S = \{x, y, z\}$ and then assume WLOG that

$$f(\succ_1^S, \ldots, \succ_n^S) = x$$

now we again swap each voter's preferences one by one from $\succ_i^S$ to $\succ_i^{\{x,z\}}$ at each individual swap there is no candidate that is valued higher than x with the new preference list that wasn't valued higher with the old preference list and so by monotonicity we also have

$$f(\succ_1^{\{x,z\}}, \ldots, \succ_n^{\{x,z\}}) = x$$

this then means that $x\ g(\succ_1, \ldots, \succ_n)\ z$ by definition and so we get a contradiction and thus g gives a transitive output. □

Now we need to confirm unanimity and consistency of g.

Lemma 3.4.7. *g is unanimous.*

Proof. Assume that $x \succ_i y, \forall i$ we then have that x is the top candidate in $\succ_i^{\{x,y\}}$ and so sliding it to the front does nothing. This then means that

$$(\succ_i^{\{x,y\}})^{\{x\}} = \succ_i^{\{x,y\}}$$

and so by 3.4.5 we have

$$f\left(\succ_1^{\{x,y\}}, \ldots, \succ_n^{\{x,y\}}\right) = f\left(\left(\succ_1^{\{x,y\}}\right)^{\{x\}}, \ldots, \left(\succ_n^{\{x,y\}}\right)^{\{x\}}\right) \in \{x\}$$

and so $f(\succ_1^{\{x,y\}}, \ldots, \succ_n^{\{x,y\}}) = x$ giving us that $x \; g(\succ_1, \ldots, \succ_n) \; y$. □

Lemma 3.4.8. *g is consistent.*

Proof. Assume that $x \succ_i y \iff x \succ_i' y, \forall i$. We need to show that

$$x \; g(\succ_1, \ldots, \succ_n) \; y \iff x \; g(\succ_1', \ldots, \succ_n') \; y$$

Now assume WLOG that $x \; g(\succ_1, \ldots, \succ_n) \; y$ and so $f(\succ_1^{\{x,y\}}, \ldots, \succ_n^{\{x,y\}}) = x$ we then replace each preference list one by one from $\succ_i^{\{x,y\}}$ to $(\succ_i')^{\{x,y\}}$. Now assume that at some swap the winner changes from x to y. Then note that for that deciding voter the order of x and y are the same at the top of the list by our assumption, then by monotonicity the winner cannot change from x to y as that would have to mean that y became higher on the deciding voter's preference list. Thus the winner never changes and so we get

$$f((\succ_1')^{\{x,y\}}, \ldots, (\succ_n')^{\{x,y\}}) = x \implies x \; g(\succ_1', \ldots, \succ_n') \; y$$

□

Now we can finally complete the proof.

Proof. Let f be an incentive compatible, we construct g as above and the previous lemmas show that g is unanimous and consistent.

Then we apply 3.3.1 to see that g must be a dictatorship. Now let D be the dictator and then let $\succ_D$ be some preference list of the dictator with their top candidate being some candidate y. Then assume that $f(\succ_1, \ldots, \succ_n) = x$, i.e. the winner is not who the dictator chose. Then again we replace $\succ_i$ by $\succ_i^{\{x,y\}}$ one voter at a time. If the winner ever changes from x to not x then by monotonicity the deciding voter must have change preference so that another candidate is above x. But note that x's position in the list only increased by shifting it to the top and so this cannot happen. Thus we have a contradiction and so

$$f(\succ_1^{\{x,y\}}, \ldots, \succ_n^{\{x,y\}}) = x \implies x \; g(\succ_1, \ldots, \succ_n) \; y \implies x \succ_D y$$

and yet the dictator's top choice is y so this is clearly a contradiction. This means that the winner is always the dictator's choice and so this f is a dictatorship. □