Economics 200C, Supplementary Notes

1 General Comments on Cooperative Game Theory

In contrast to noncooperative game theory, cooperative game theory focuses on "normative" issues. "What is the right way to divide a surplus?" rather than "How will self interested actors distribute a surplus given certain rules?" Cooperative game theory typically assumes efficiency. Noncooperative game theory is filled with examples where outcomes are inefficient. The reason for the difference is that efficiency is clearly a good normative goal, but it may not always be achievable.

The basic formulation of a cooperative game is a player set (I will assume that there is a finite number n of players and denote the set of players by $N = \{1, \ldots, n\}$) and a "characteristic function" $v : 2^N \to \mathbb{R}$. For any subset S of N, $v(S)$ is the value available to the subset. I will assume that $v(\emptyset) = 0$. This formulation makes sense when utility is transferable. A more general formulation is that the characteristic function assigns to each subset S with cardinality s a subset $V(S)$ of $\mathbb{R}^s$, with the interpretation that if $x \in V(S)$ then it is possible for the coalition S to give agent $i \in S$ utility x_i. If v is the characteristic function of a transferable utility game, then $V(S) = \{x : \sum_{i \in S} v_i \leq v(S)\}$. I will (with the exception of two asides) limit attention to games with transferable utility.

For concreteness, imagine a public-good problem in which either you produce a project, which gives rise to utility u_i for agent i at the cost c or you do not, which leads to utility 0. Here $v(S) = \max\{0, \sum_{i \in S} u_i - c\}$. That is, a subset of agents can either decide not to produce the project or produce it. In the later case, the project generates $\sum_{i \in S} u_i$ for the group, but costs c.

Cooperative game theory does not worry about acquiring private information, so I do not include "θ" in the description.

Cooperative game theory looks for solutions to games. A solution is a function (or, sometimes, a coorrespondence) f from a game (that is, the set of characteristic functions) to a vector of payoffs for all players (an element of $\mathbb{R}^n$) with the property that $\sum_{i=1}^n f(v) \leq v(N)$. The interpretation is that $f_i(v)$ is the payoff that player i "should" get from the game with characteristic function v.

Just like noncooperative game theory has a seemingly endless supply of solution concepts (maxmin, dominant strategy, Nash, subgame perfect, ...), cooperative game theory has a long list of solutions (core, Shapley value, bargaining set, Nash bargaining solution, ...). Unless you become a cooperative game theorist, you need to know about at most three cooperative game theoretic solution concepts, core, Shapley value, and Nash bargaining solution).

The core is a set-valued solution concept defined for games with nontransferable utility (the definition specializes to games with transferable utility). An allocation x is in the core if $x \in V(N)$, for each $S \subset N$, there is no $y \in v(S)$ with $y_i > x_i$ for all $i \in S$. In words, a core allocation has the property that no subset of players can join together and create an allocation that is better for all members of the coalition. It features prominently in general equilibrium and identifies allocations that are individually rational (this "tests" the allocation against subsets S that contain only one element) and Pareto efficient (this "tests" the allocation against the coalition of all agents N).[1] In general equilibrium, the key result is the theorem that states that the core shrinks to the set of competitive equilibria in the limit for a class of economies. In general, the theory identifies games in which the core exists.

The Nash bargaining solution is a solution function defined on two player "bargaining games." These games have nontransferable utility. They are defined by a set of feasible utilities, which plays the role of

[1] Individual rationality and Pareto efficiency are necessary conditions for an allocation to be in the core. The conditions are not sufficient.

$v(N)$ and a disagreement point, which determines $v(\{i\})$ for $i = 1$ or 2. The Nash solution is a particular function. The key result is that the function is characterized by 4 assumptions. That is, there is one and only one solution function that satisfies 4 general properties. The function is easy enough to write down: If the disagreement outcome is (d_1, d_2), then the Nash bargaining solution maximizes $(x_1 - d_1)(x_2 - d_2)$ subject to $(x_1, x_2) \in V(N)$. (In this case $N = \{1, 2\}$.) The optimization problem has a utility solution provided that $V(N)$ is compact, convex and contains (d_1, d_2). The Nash bargaining solution specializes to games with transferable utility and generalizes to games with more than two players. The Nash bargaining solution appears a lot in applied literature. Authors postulate (without independent justification) that players solve bargaining problems using the Nash bargaining solution.

It requires time to establish the connection between the core and competitive equilibria. The theory of core existence is deep and complicated (the core is empty is a lot of interesting situations). It is also sometimes tricky to define the core (for example, in markets with externalities).

The characterization of the Nash bargaining solution is elegant, but the basic assumptions that characterize the solution are not hard to understand and the characterization theorem is elementary (that is, doesn't require fancy mathematical techniques) and relatively short (this does not mean that it is simple).

You should have no trouble finding discussions of these ideas, but feel free to ask me if you want to know more.

2 Shapley Value

I discuss the Shapley Value in more detail because it is, in my mind, connected to the allocation problems we discussed at length.

The Shapley Value is defined to be:

$$f_i(v) = \frac{1}{n!} \sum_{\pi \in \Pi(N)} [v(P_i(\pi) \cup \{i\}) - v(P_i(\pi))]. \tag{1}$$

The definition does not make sense without some clarification of terminology and some explanations. First the terminology. $\Pi(N)$ is the set of permutations on the set N. (A permutation of a finite set is a 1-1 and onto mapping from the set to itself. More simply, it is just a rearrangement. There are $n!$ distinct permutations of a set with n elements.) If π is a permutation, $P_i(\pi) = \{j \in N : \pi(j) < \pi(i)\}$. Next the interpretation. Imagine that agents line up outside the room in some order (the permutation determines the order). I want to compute how much a particular agent i adds to surplus. As people enter the room, I keep a running total of the surplus. Suppose agent 1 enters first. Before he enters, I have a running total of $0 = v(\emptyset)$. When 1 enters, I increment the total to $v(\{1\})$ so what the first entrant adds is $v(\{1\}) - v(\emptyset)$. Furthermore, if 1 is the first in line, $\pi(1) < j$ for all $j \neq 1$ so $P_1(\pi \cup \{1\}) = \{1\}$. In general, for a fixed permutation, $v(P_i(\pi) \cup \{i\}) - v(P_i(\pi))$ is what i adds to the running total when she enters. That is, given π, $v(P_i(\pi) \cup \{i\}) - v(P_i(\pi))$ is the marginal contribution of agent i. The Shapley value therefore computes the expected marginal contribution of each agent under the assumption that all permutations are equally likely.

There is an equivalent way to write this.

$$f_i(v) = \frac{1}{n!} \sum_{S \subset N \setminus i} [s!(n - s - 1)!][v(S \cup \{i\}) - v(S)]. \tag{2}$$

(s is the cardinality of S and n is the cardinality of N.) Let me explain. $v(S \cup \{i\}) - v(S)$ is easy. It is the marginal contribution of i to a subset S that does not contain i. So the Shapley value gives agent i a

weighted average of her marginal contribution. What are the weights? There are lots of subsets S that do not contain i. $s!(n-s-1)!$ is the number of different ways in which i can "join" coalition S.

Why? The first s entrants must be the members of S. They can "enter" in any order. Hence there are $s!$ different ways in which they can enter. There is only one way in which i can enter as the $s+1$ person. The remaining $n-s-1$ agents can enter in any order. Hence there are $s!(n-s-1)!$ permutations in which i enters immediately after (exactly) the members of S enter. Since there are $n!$ permutations total, the weight in (2), $[s!(n-s-1)!]/n!$ is the fraction of permutations in which i follows S.

Coming up with this functional form is a big contribution, but Shapley actually characterized it. His famous (1953) result is that the value described above (he did not call it the Shapley value) satisfies four plausible assumptions and it is the only solution that does so.

Theorem 1. *The Shapley value is the only (single-valued) solution that satisfies efficiency, additivity, symmetry, and zero-payoff to dummies.*

Efficiency is the property that you distribute the entire surplus: $\sum_{i=1}^{n} f_i(v) = v(N)$. Given two characteristic functions, v and w, there is a new characteristic function $v+w$ defined as $(v+w)(S) = v(S)+w(S)$. The solution f is additive if $f(v+w) = f(v)+f(w)$. Symmetry is the property that two players who are identical should receive the same payoff. Formally, players i and j are symmetric if $v(S\cup\{i\}) = v(S\cup\{j\})$ for every $S \subset N$ that contains neither i nor j. The solution f is symmetric if $f_i(v) = f_j(v)$ for every game in which i and j are symmetric. Finally, i is a dummy in the game v if for all S that do not contain i, $v(S) = v(s\cup\{i\})$.

These assumptions seem uncontroversial. Solutions that are inefficient are wasteful. Solutions that are not symmetric seem unfair. A solution that does not give 0 to a dummy seems unfair as well. The additivity property may be controversial, but when we form $v+w$ we assume that there are not interactions between the games, so it makes sense to assume that if we know the solution to both v and w we know the solution to the sum. I should mention that an implicit assumption is that the four assumptions hold for all games. It is conceivable that one might be willing to relax the assumption in some situations in order to have more freedom in another situation. Feel free to reflect on this.

For my purpose, I assert only that the assumptions are not crazy and it is a considerable accomplishment to know (a) that the assumptions are consistent (that is, you can impose the assumptions are still have a solution) and (b) the assumptions are restrictive (when you impose the assumptions only one solution can satisfy them). It is a bonus that the one way to satisfy the assumptions has an interpretation.

The proof of the theorem comes in two parts. The first part is showing that the Shapley value actually satisfies the four properties. This is a routine verification. Additivity and the dummy player properties are essentially automatic. Efficiency is clear (maybe!) from (2). Symmetry requires writing down the definition, but it is intuitively obvious. The second part of the theorem is showing that any solution that satisfies the axiom must be the Shapley value. The strategy of proof is to use linear-algebra logic. You examine the implications of symmetry on simple games of the form v_T, where $T \subset N$ and

$$v_T(S) = \begin{cases} 1 & \text{if } T \subset S \\ 0 & otherwise \end{cases}.$$

For these games, the dummy axiom requires that $f_i(v_T) = 0$ for $i \notin T$; symmetry requires that $f_i(v_T)$ is constant for $i \in T$ and efficiency requires $\sum_i f_i(v_T) = 1$. Hence $f(v_T)$ is uniquely determined for these games. Next you show that any game is the sum of games like v_T. Consequently, by additivity, there is (only one) formula for value. Because the Shapley value formula satisfies the assumptions, it must be the formula. The complete proof requires more than one paragraph, but this is the general idea.

3 Connection between Shapley Value and Implementation

I mentioned in class that I thought that there was a connection between the Shapley value and the "pivot" mechanism for eliciting private information in the case of transferable utility. The differences may be more salient than the similarities. One difference is that when we discuss implementation, individual's have private information about their valuations. The big concern is trying to figure out what these are (that is, providing proper incentives for truthful revelation). In the cooperative-game setting, we assume that we know the valuations. Another difference is that implementation theory did not worry about "budget balance." The Shapley value, because it is an efficient solution, must redistribute surplus. In the implementation problem we did not worry about what is fair, we just worried about extracting information. The Shapley value cares about (some sort of) fairness. So why do I think that they are related? Both procedures focus on marginal contributions. This is clear from the definition and discussion of the Shapley value. I hope that it is also clear from the class discussion is that the Groves mechanism also involves a marginal contribution. The differences are that for Shapley value, we average over all kinds of marginal contribution, while for Groves, we only ask what happens when the last person.

Let me try to make this concrete. Consider the allocation problem. The decision is which agent should receive an item. The value of the item is θ_i to agent i. Assume that all $\theta_i > 0$. When we studied the implementation problem, the goal was to give the item to the agent who had the highest valuation. The problem was that we needed to find some way for the agents to announce their valuations honestly. We did this by creating a tax system. Let $\theta^* = \max_j \theta_j$ and $\hat{\theta}$ be the second highest valuation. Given an announcement $\theta = (\theta_1, \ldots, \theta_n)$ the transfer rule is $t_i(\theta) = \sum_{j \neq i} v_j(k^*(\theta), \theta_j) - h_i(\theta_{-i})$. Consider the first term. Either $\theta_i = \theta^*$, in which case $k^*(\theta) = i$ and $\sum_{j \neq i} v_j(k^*(\theta), \theta_j) = 0$ (this is the case where i gets the item and none of the other agents receives value) or i doesn't have the highest valuation ($\theta_i < \theta^*$), in which case the sum is equal to θ^*. In the "pivot mechanism" (or Clarke mechanism), we further set $h_i(\theta_{-i}) = \sum_{j \neq i} v_j(k^*_{-i}(\theta_i), \theta_j) = \max_{j \neq i} \theta_i$. This is equal to either θ^* (if θ_i is the largest) or $\hat{\theta}$ (otherwise). Consequently the total transfer is zero when θ_i is not equal to θ^* and $\hat{\theta}$ when $\theta_i = \theta^*$. That is, just like a Vickery auction, the winner pays the second highest valuation and no one else pays anything. It rewards the highest valuer by allowing her to keep the marginal surplus (difference between her valuation and the next highest valuer). How does this connect to Shapley? In this setting we are trying to divide up a total surplus equal to the maximum valuation, θ^*. The Shapley value allocated the surplus according to the "average marginal contribution" in particular it gives the high valuer more than the difference between the two highest valuations, because the value is an average of marginal contributions and the high valuer's marginal contribution is higher than $\theta^* - \hat{\theta}$ when she enters prior to the second highest valuation. Stated differently, the Groves mechanism gives i $v(S \cup \{i\}) - v(S)$ for subsets S that include the second highest valuer, while the Shapley value averages over all S, including some in which $v(S \cup \{i\}) - v(S)$ is larger than $\theta^* - \hat{\theta}$.

What about balance? The Shapley value would distribute the entire surplus to the agents. That is, the sum of the utilities would be equal to θ^*. I don't know how much each agent will receive (this depends on the values and must be computed by the formula), but I know that the agent who values the item most will receive at least $\theta^* - \hat{\theta}$ (because no matter which permutation of agents you use, the high valuer adds this much total surplus). The pivot mechanism does not distribute all of the surplus – it taxes the high valuer $\hat{\theta}$ and gives everyone else 0. Hence the mechanism distributes a total surplus equal to $\theta^* - \hat{\theta}$ and gives it all to the high valuer. The mechanism would still work if it redistributed the surplus, for example by giving a rebate of $\hat{\theta}/n$ to each agent. (This is one possibility. What is important is that agent i's rebate is independent of agent i's reported type.)

My message is that both the Groves mechanism and the Shapley value distribute surplus following a rule that uses "marginal contribution," but that Groves imagines that every agent is the final contributor, while Shapley takes an average. Because Shapley takes an average, it can allocate the entire surplus. Because Groves looks at the final contribution, it can preserve incentives.

Final comment: As an example, imagine that there are three agents and their valuations are 1, 2, and 3 respectively. The Shapley value allocates the total surplus 3 as 1/6 to Agent 1; 5/6 to Agent 2; and 11/6 to Agent 3. (This follows from the Shapley value formula.) The pivot mechanism says that the third agent must pay 2 (and gives utility 0 to the first two agents and $3-2$ to the third agent). Again, these two distributions are not fully comparable because Shapley distributes all of the surplus but the pivot mechanism does not.

4 Shapley with Non-transferable utility

It is possible to define the value in games without transferable utility. This is a technical and specialized literature. Two key results: The non-transferable value agrees with the Nash bargaining solution in cases where the Nash bargaining solution is defined. It is possible to prove "value equivalence theorems" that are analogs to core equivalence theorems (loosely, the value of a large game is equal to the utility generated from a competitive equilibrium).[2]

5 Arrow's Impossibility Theorem

This result is in many ways the first result in modern microeconomic theory. No core micro sequence is complete without it. Sorry.

Arrow posed the following question about preference aggregation. Given a society describe as a finite collection of individuals, each with complete, transitive preferences over a finite set of alternatives, is there a "reasonable" way to aggregate these preferences to obtain a complete, transitive social preference relationship over the same finite set of alternatives.

When there are two alternatives, the answer is yes and a good procedure is majority voting. Imagine that the choices are either A or B so that preferences either rank A first, B first, or say that they are indifferent. For simplicity, assume strict preferences, so that indifference does not arise. What we want is to take a group of individual rankings (some people prefer A to B and others B to A and decide whether society prefers A to B). Surely you'd like the rule to respect individual preferences in the sense that if everyone prefers A to B, then so does society. You may want it to be democratic in the sense that no one person can always determine the outcome. You can formalize this by saying that the group's preferences must not always agree with one agent's preferences. Otherwise, some agent is a dictator. These two conditions (respecting individual preferences and nondictatorial) are not compatible when society has only one person (but there it is natural to equate the group with the individual) and maybe when there are only people (essentially one side must win when the agents disagree, although the rules that says that "A wins when the two people disagree" does not create a dictator). When there are more than 2 agents, majority rule works well.

Arrow's monumental contribution is to think through the problem when there are more than two people and more than two options. He proposes a small number of axioms that a good preference aggregation procedure should satisfy and then shows that these axioms are incompatible.

The basic problem begins with a finite set of alternatives, X; a finite set of agents, $i = 1, \ldots, N$. For reasons suggested above, assume that N and the cardinality of X and both at least 3. Individuals have

[2]This statement is loose, so if you find it interesting, look for formal treatments.

preference orderings over X. That is, for each i, there is a relationship R_i defined on $X \times X$. We write "xR_ix'" and say "i (weakly) prefers x to x'." We assume that R_i is complete (either xR_iy or yR_ix for all $x, y \in X$); and transitive (if xR_iy, yR_iz, then xR_iz). The goal is to come up with a function F that maps $(R_1, \ldots, R_N)$ into complete, transitive relationship. $F(R_1, \ldots, R_N)$ is society's preferences when individual preferences are R_i, $i = 1, \ldots N$. Formally, denote the set of preference relations by $\mathcal{R}$. A social welfare function is a mapping F from $\mathcal{R}^n$ to $\mathcal{R}$.

There are lots of ways to aggregate preferences. One way is to let $F(R_1, \ldots, R_N) = R_i$. In this case, we say that Agent i is a dictator. Only her preferences matter. Another way is to imagine that ranks the options, with her favorite option receiving v_1 points, her second option $v_2 < v_1$ points, and so on. We aggregate preferences by adding up all of the points received by each option. The option with the most points is society's favorite and so on. Another option is to try to implement pairwise majority rule: society prefers A to B if and only if AR_iB for a majority of the i. The procedure does not work, due to the famous Condorcet paradox. (Three agents, three alternatives, Agent 1 strictly prefers A to B to C; Agent 2 strictly prefers B to C to A; Agent 3 strictly prefers C to A to B.) Using pairwise majority voting, A beats B (because of Agents 1 and 3); B beats C (Agents 1 and 2); and C beats A (Agents 2 and 3). So pairwise majority rule does not lead to transitive group preferences. A glib way to characterize Arrow's Theorem is to say that the Condorcet example is unavailable.

Theorem 2. *Suppose that the number of alternatives is at least three. Every social welfare function that satisfies Pareto efficiency and independence of irrelevant alternatives is dictatorial.*

The theorem has several explicit and several tacit assumptions. The first explicit assumption is that there are at least three alternatives. The result is false without this assumption as I have discussed.

The second explicit assumption is Pareto efficiency. This assumption is what I described earlier as "respecting individual preferences." Precisely, F satisfies Pareto efficiency if xR_iy for all i implies that $xF(R_1, \ldots R_N)y$. This property seems desirable.

The third explicit assumption is independence of irrelevant alternatives. This assumption compares different profiles of preferences. Suppose that $R = (R_1, \ldots, R_N)$ and $R' = (R'_1, \ldots, R'_N)$ are both elements of $\mathcal{R}^N$. Assume that these profiles treat the options $x, y \in X$ in the same way (xR_iy if and only if xR'_iy for all i). The independence of irrelevant alternatives assumption requires that $F(R)$ and $F(R')$ treat x and y in the same way. That is, $xF(R)y$ if and only if $x'F(R)y'$. This assumption is powerful but controversial. The argument in favor of the assumption is this. Going from R to R' the only changes made are "irrelevant" to the ranking of x and y. It could be that R_i ranks z higher than x and R'_i ranks z lower than x, but, by assumption R_i and R'_i express that same ranking of the pair (x, y). If this is true for all members of society, then the assumption posits it should also hold for society. Notice that the assumption does not say how society orders x and y, only that the ordering should not change if you make changes in the individual rankings of the other alternatives. The independence of irrelevant alternatives assumption does not hold in practice.[3] It generally fails in the weighted voting scheme described above (if you get more points for being the top choice, then your weight – and possibly the group ranking – will change if an irrelevant option is elevated to top choice). It is easy to think of examples in which the inclusion of an irrelevant candidate changes electoral outcomes. Abstractly, if people vote sincerely, it could be that a majority favor A to B, but that if you included a third candidate C(perhaps one "more extreme" than A), enough voters would prefer C to A to make B a plurality winner. Hence adding an "irrelevant" alternative (the existence of C does not change preferences between A and B), the electoral outcome may switch from A to B. This story is not perfect, but it does suggest that some commonly used ways to determine preference rankings of

[3]In some sense, Arrow's theorem says that it cannot hold in practice.

groups will not satisfy the independence condition. Arrow's approach is normative. That is, he asserts that Independence of Irrelevant Alternatives is a good property and investigates whether it is possible to satisfy it. If you agree that the property is good, then the theorem is a big disappointment. The theorem is not a reason to conclude that independence is a bad property, but it is a reason to expect it to be violated.

There are two implicit assumptions in the theorem. The first implicit assumption is that we care about obtaining a social ranking. Perhaps you only want society to make a choice. This requires identifying the best option, but not necessarily ranking the others. That is, instead of looking for a map from $\mathcal{R}^R$ to $\mathcal{R}$, we could ask for the range to be X (so the aggregation process would take a profile of preferences and identify a winning option). If you think that elections select winning candidates, but do not order the losers (Mitt Romney won the 2012 Republican Presidential nomination, but there is a sense in which we don't know who finished second), then Arrow's framework is too demanding.

The second implicit assumption is that Arrow requires us to have a rule that aggregates all possible preference rankings. It could be that some orders are unlikely. There is a lot of work on the preference aggregation problem on restricted domains. The best known work assumes that one can order the outcomes in X and that each R_i is single peaked (that is if $x \succ y$ and xR_iy, then xR_izR_iy for any z such that $x \succ z \succ y$).[4] Arrow's theorem is false on this (and other) restricted domains.

The final element of the statement of the Theorem is the conclusion. F is dictatorial if there exists i such that xR_iy implies $xF(R)y$.

The reason the result is "Arrow's Impossibility Theorem" is because it demonstrates that it is impossible to aggregate preferences in a way that is efficient, consistent with independence of irrelevant alternatives, and non-dictatorial. Stated crudely, it is impossible to have a universal democratic way to aggregate preferences. The mathematical logic behind this assertion is unquestionable. The interpretation, of course, depends on whether you believe that the assumptions are appropriate. The two standard ways to try to avoid dictatorship are to relax the assumption of unrestricted domain (there are well known possibility results for single-peaked preferences as I mentioned above) or abandoning the independence of irrelevant alternatives assumption. Many common voting methods fail independence of irrelevant alternatives.

There are many proofs of the theorem. A particularly short and pretty proof appears in "Three Brief Proofs of Arrow's Theorem" by John Geanokoplos (in *Economic Theory* 2005). Arrow's original formulation has been refined (his axioms where somewhat different).

6 Connection between Arrow's Theorem and Implementation

Arrow's theorem asks whether is possible to find a find a social welfare function that satisfies some assumptions. The research stimulated similar approaches in other areas. Axiomatic decision theory makes assumptions about preferences and asks whether these assumptions are consistent with a particular representation of preferences. (The characterization of von Neumann Morgenstern utility and Savage's axioms are leading examples.) Axiomatic cooperative games theory and social choice theory ask whether there are choice functions that satisfy certain assumptions. The Nash Bargaining Solution and the Shapley value are examples of this approach.

Implementation theory asks whether it is possible to construct a game that has an equilibrium that implements a particular social welfare function. There are at least two differences between this topic and the approach in Arrow's theorem. First, usually you do not begin with an axiomatization of the function that you wish to implement. This difference may be cosmetic. Second, the implementation literature is explicitly noncooperative. You implement by constructing a game and looking for how people will play the

[4] Here $\succ$ is the ordering on X.

game (typically by invoking some non-cooperative solution concept). In the Arrow framework, there is no problem with incentives.