Strategic Polarization

Adam Kalai*
Carnegie Mellon University
akalai@cs.cmu.edu

Ehud Kalai†
Northwestern University
kalai@nwu.edu

June 7, 2000

Abstract

In joint decision making, similarly minded people may take opposite positions. Consider the example of a marriage in which one spouse gives generously to charity while the other donates nothing. Such "polarization" may misrepresent what is, in actuality, a small discrepancy in preferences. It may be that the donating spouse would like to see 10% of their combined income go to charity each year, while the apparently frugal spouse would like to see 8% donated. A simple game-theoretic analysis suggests that the spouses will end up donating 10% and 0%, respectively.

By generalizing this argument to a larger class of games, we provide strategic justification for polarization in many situations such as debates, shared living accommodations, and disciplining children. In some of these examples, an arbitrarily small disagreement in preferences leads to an arbitrarily large loss in utility for all participants. Such small disagreements may also destabilize what, from a game-theoretic point of view, is a very stable equilibrium.

*Supported by an IBM Distinguished Graduate Fellowship. Address: Dept. of Computer Science, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh PA 15213.

†Charles E. Morrison Professor of Decision Sciences, Supported in part by NSF grant #SES 9904931.

1 Introduction

In public debates, private organizations, shared living accommodations, and many other types of interaction, participants' behavior is seen to polarize. Even similarly minded moderate individuals will take opposite, extreme positions in such settings. To add to the frustration, such polarization can be quite costly: arbitrarily small disagreements in players' preferences may lead to arbitrarily large losses in utility for all participants.

This note presents a simple game-theoretic analysis that explains why selfishly-motivated participants polarize[1]. But like the prisoners' dilemma [1], polarization games are designed and played by seemingly rational individuals who suffer utility loss as a result. This suggests that, despite the apparent game-theoretic simplicity, polarization is also subtle and easily overlooked.

The remainder of this section includes two examples where players choose drastically opposing strategies. The first example gives a quick understanding of the phenomenon. The second example shows how costly and destabilizing the phenomenon may be. In that example, two players that are very happy with the outcome of the game under full agreement of preferences, become equally miserable when an ϵ disagreement enters their preferences. This is despite the fact that the outcome in the case of agreement is highly stable from a game-theoretic point of view.

Section 2 formally defines a polarized player as one who is playing a strategy on the boundary of her set of possible strategies. A polarization lemma shows that, for a large class of games which we call aggregation games, every player is either polarized or completely satisfied with the outcome. In particular, for situations in which the players disagree on the best outcome, all but at one player must be polarized. Sections 3 and 4 conclude with additional illustrations and discussion.

1.1 Example: How much to give to charity?

A newly married couple, Gene and Pat, each contribute from their joint bank account a certain amount of money to charity. Suppose, for a moment, that Pat would like to see 10% of their combined income donated to charity and

[1]Our use of the term "polarization" differs from other uses, such as "group polarization," which refers to a phenomenon in which all members of a group support the *same* decision more emphatically than they would as individuals.[3]

Gene would prefer 8%. One sensible scenario is that they would begin by donating 5% and 4% respectively, which gives a compromise of 9% donated. With time, however, we may expect Gene's donations to decrease to 0, in an attempt to reduce the total contribution, and Pat's donation to increase to 10%. Although this is not the only possible outcome, it is a natural one when both players act strategically to achieve their own goals.

More formally, modeled as a strategic game, we show that this polarized outcome is the only Nash Equilibrium. Say Gene and Pat donate amounts of g and p dollars respectively. This determines their total contribution, $t = g + p$. Suppose that they each have in mind a desired value for this total contribution. We denote their ideal total contributions by t_g and t_p, respectively. Further assume that they each would like the total to be as close as possible to their ideals, with their preferences described by utility functions $u_g(t) = |t - t_g|$ and $u_p(t) = |t - t_p|$.

Case 1. Perfect agreement, $t_g = t_p$. By each donating one half of the ideal amount, both people are completely satisfied.

Case 2. No agreement, say $t_g < t_p$. Various models of players' dynamics will lead to polarized behavior with $g = 0$ and $p = t_p$. For example, under Cournot best response adjustments[2], where players adjust their individual current period choices to be best response to observed opponents' last period choices, the following occurs. In the initial decision period, say they each donate one half of their ideal levels. This results in a total contribution midway between t_g and t_p. Assuming that they best respond to these observed levels, in the next period, we will see Gene's contribution decrease and Pat's increase by an equal amount, so that the total does not change. These adjustments will continue each period until Gene is donating nothing, and Pat is donating t_p. In fact, no matter how small the initial disagreement was, adjustments eventually lead Gene and Pat to donate 0 and t_p, respectively.

Clearly, the two players do not have to follow a Cournot best-response dynamics. But the concept of a Nash Equilibrium[4] is useful here. For selfish players to reach an equilibrium where they do not change their donations, they should both be choosing a best response to the other's decision. This happens only when Gene donates 0 and Pat donates t_p.

This type of game models many other situations, as well. For example, it can model a debate, where participants want to influence common opinion on an issue. It also relates to the polarization of parents disciplining a child,

where the resulting discipline is the sum of the levels imposed by the parents. Similarly, roommates may polarize over how clean the bathroom should be, with the "cleaner one" cleaning up for both. Business partners polarize over the conduct of their affairs, and politicians polarize over the conduct of the affairs of a nation.

The next example illustrates how costly and destabilizing polarization may be.

1.2 Example: Costly Household Polarization

Consider a weekly work-consumption-savings plan of a household consisting of two individuals. Viewed as a strategic game, each of the players, $i = 1, 2$, decides on his own work level w_i and his level of consumption of a frivolous good f_i, both measured in dollar units with $0 \leq f_i$ and $0 \leq w_i \leq 1600$. The income not consumed, $b = (w_1 + w_2) - (f_1 + f_2)$ is deposited as savings in the bank (b could be negative). In this example, an arbitrarily small disagreement in the necessary amount of savings will lead one player to work as hard as possible and spend no money on frivolous goods, while the other player will do no work and all of the spending, and both players will be equally miserable.

Case 1. Suppose both participants have the following identical utility function,

$$u_i = 2\min(b, 500) - (w_1 + w_2) + 0.1(f_1 + f_2).$$

In words, each participant gets two units of positive utility for each dollar in the bank, up to \$500, and no utility for extra money in the bank. On the other hand, a dollar's work (by either party) costs both players one unit of utility, while frivolous consumption (by either party) yields both players only 0.1 units of utility per dollar spent. Thus each participant would ideally like a total of \$500 of combined work and no frivolous consumption. Formally, when $w_1 = w_2 = 250$ and $f_1 = f_2 = 0$, the players are at a Nash Equilibrium with equal positive utilities of 500.

Case 2. Now suppose the players almost agree, with,

$$\begin{aligned} u_1 &= 2\min(b, 500) - (w_1 + w_2) + 0.1(f_1 + f_2) \\ u_2 &= 2\min(b, 500 + \epsilon) - (w_1 + w_2) + 0.1(f_1 + f_2) \end{aligned}$$

In this case, the only Nash Equilibrium has $w_1 = 0$, $w_2 = 1600$, $f_1 = 1100$, $f_2 = 0$, with both players equally miserable at $u_1 = u_2 = -490$.

Proof. At Equilibrium we cannot have $b > 500$, because player 1 could increase his utility by increasing f_1. Player 1 would gain utility from this frivolous consumption without losing any due to having less in the bank.

Similarly, we cannot have $b < 500$. If this is the case, either they have frivolous consumption or both players are not working to capacity. If they have frivolous consumption, then both players would benefit by 1.9 units of utility per dollar from reducing this consumption. Alternatively, if they do not both work to full capacity, then a unit increase in work by either player will increase utility by 1 unit.

Thus at Equilibrium $b = 500$. Now if $w_2 < 1600$ then player 2, who wants more than 500 in the bank, can improve by increasing w_2. So at Equilibrium we must have $b = 500$ and $w_2 = 1600$. Continuing with this type of reasoning, it is easy to see that $f_2 > 0$ is suboptimal for player 2, who wants more in the bank. With $w_2 = 1600$ and $f_2 = 0$, the only optimal response of player one is to have $w_1 = 0$ and $f_1 = 1100$. □

From a game-theoretic perspective, the above example illustrates drastic destabilization. In the full agreement game, with the identical utility functions, all the possible payoffs of the game are of the form (x, x) with $-2880 \leq x \leq 500$. The equilibrium discussed there yields the payoffs $(500, 500)$. So the equilibrium payoffs are greater, for both players simultaneously, than any other possible payoffs in the game. This is what Aumann and Sorin [5] call an equilibrium of common interest. The equilibrium is also strict in the sense that any deviation from the equilibrium strategies causes a strictly positive loss to the deviating player. Strict Nash equilibria of common interest survive most theoretical notions of equilibrium refinements and are considered highly stable. So it is surprising that, as we introduce an arbitrarily small perturbation to the payoff of either one of the players, this equilibrium completely collapses, and instead we have a unique equilibrium with payoffs $(-490, -490)$.

2 Polarization in aggregation games

In both of the above examples, an aggregate quantity determines the players' utilities. At all the Nash Equilibria, each player is either completely satisfied, meaning that no outcome could improve her utility, or is polarized, meaning that she is playing a strategy on the boundary of her feasible set of strategies. This is the polarization phenomenon.

We proceed to present simple conditions on the aggregation and utility functions that result in this phenomenon. Informally, the condition on the aggregation function is that any individual player, by changing his own strategies, can move the aggregate value in any direction (specifically in some open set), provided that player is not constrained by his own individual limitations, i.e. playing a strategy on the boundary of his feasible set. Also, each player's utility function must have no (interior) local maxima that aren't also global maxima. Under these two conditions, we argue that if a player is not completely satisfied, then she is not at a global or local maximum. Furthermore, if she is playing a best response, then she must actually be at a boundary strategy, otherwise she could move the aggregate to increase her utility. We formalize this argument as follows.

Player i in $\{1, 2, ..., n\}$ has a feasible set of strategies S_i. Let $S = \times_i S_i$ denote the set of strategy profiles. An aggregating function AGG: $S \longrightarrow A$ selects an outcome a from a set A for every strategy profile s. Each player i has a utility function u_i: $A \longrightarrow \mathcal{R}$ describing his preferences over the selected outcome. We assume only that the sets S_i and A are subsets of abstract topological spaces, but in all our examples they are simply subsets of Euclidean spaces.

An aggregation game consists of the simultaneous selection of individual strategies where players' payoffs are computed through the realized outcomes. With an abuse of notation, we denote this by $u_i(s) = u_i(\text{AGG}(s))$.

Next, we describe the notion that every player not limited by his own feasibility constrains, i.e. not playing a boundary strategy, can move the aggregate value within some neighborhood of its current value. Formally, the *range of influence* of player i at $s_{-i} = (s_1, s_2, ..., s_{i-1}, s_{i+1}, ..., s_n)$ is $\text{AGG}(s_1, ..., s_{i-1}, S_i, s_{i+1}, ..., s_n) \subseteq A$. An *individually responsive* aggregating function AGG has the property that, for every player i and strategy profile s, if $\text{AGG}(s)$ is on the boundary of i's range of influence at s_{-i}, then s_i is on the boundary of i's feasible set of strategies S_i.

A utility function has *no local maxima* if every (interior) local maximum is actually a global maximum. That is, if $u_i(a)$ is a maximum of u_i over a neighborhood of a, then it is a maximum over A.

Finally, player i is *completely satisfied* with a strategy profile s if $u_i(\text{AGG}(s))$ is a maximum of u_i over A. Player i is *polarized* at a strategy profile s if s_i is on the boundary of S_i. Based on these definitions, we have,

Polarization Lemma. *In any (pure strategy) Nash Equilibrium of an aggregation game with an individually responsive aggregating function and utility functions with no local maxima, every player who is not completely satisfied is polarized.*

Proof: Suppose not. Say we have a Nash Equilibrium with player i neither completely satisfied nor playing a boundary strategy. Let s be the equilibrium strategy profile and R_i be i's range of influence at s_{-i}. Since s_i is not on the boundary of i's feasible set of strategies and AGG is individually responsive, $\text{AGG}(s)$ is not on the boundary of R_i (to see this note that a restatement of individual responsiveness is that at any s_{-i} and an s_i interior to S_i the aggregation function must yield a value interior to i's range of influence). Thus, R_i is a neighborhood of $\text{AGG}(s)$. Furthermore, since i is not completely satisfied and u_i has no local maxima, $u(\text{AGG}(s))$ is neither a global nor local maximum. This means that there must be some a_i' in R_i with $u_i(a_i') > u_i(\text{AGG}(s))$. Since a_i' is in player i's range of influence, s_i is not a best response and we have a contradiction.

In fact, even a minimal amount of disagreement between the players' preferences leads to polarization for all but one player. To be precise,

Corollary. *If no single outcome is optimal for more than one player, then at any Nash Equilibrium of an aggregation game satisfying the conditions of the lemma, at least $n-1$ players are polarized.*

Remarks:

1. The scope of the lemma. The sufficient conditions used in the polarization lemma are quite general. Individually responsive aggregation functions, as described by the general topological property above, include many aggregation methods other than those obtained by adding

or averaging individual positions. Weighted averages, geometric averages, averages of monotonic functions of the players positions, are illustrations of individually responsive aggregation functions. If the players positions and the set of aggregate values are one dimensional, for example, any function which is strictly monotonic coordinatewise is individually responsive.

2. A similar observation is true for the no (interior) local maximum condition. Consider the cases that the set of possible aggregate values is convex. For any strictly convex utility function there are no interior local maxima, and the condition holds. Conversely, if an individual utility functions is concave, any interior maximum is a global maximum and the condition holds. But intermediary conditions are also possible. For example any function with the property that all its local maxima are global, e.g., multiple equal peaks, satisfies the condition.

3. Mixed strategies. The polarization lemma holds for pure strategy Nash Equilibria but not necessarily for mixed strategies. Consider the charity example where each player has a target value of 10%. A mixed strategy equilibrium exists where each player chooses either 4% or 6% with equal probability. In this case, neither player is polarized, and neither is completely satisfied. However, if each player's utility is a strictly concave function of his own strategy (keeping the opponents fixed), it is easy to see that there are only pure strategy Nash equilibria, and thus the polarization lemma applies in general.

4. General best response. The polarization lemma can be applied to an individual player in a game. For example, fix any strategies for players $2, 3, ..., n$ and consider a best response strategy of player 1. If player 1's utility function has no local maxima and the aggregation function is individually responsive to her, then any of her best response strategies will result in her complete satisfaction or polarization. This means that even in a world where some of the players are irrational, all the players that optimize relative to beliefs about opponents strategies must be either completely satisfied or polarized. This observation may also be useful in studying other best response based notions, such as rationalizability [6, 7] and Cournot best response dynamics.

5. The lemma as a computational device. The polarization lemma may help in computing solutions that are based on the notion of best response. For example, in aggregation games that satisfy the hypotheses of the corollary, the corollary offers a severe reduction of the possible equilibria of the game. As an extreme case, notice that if the strategy sets are unbounded (or simply have no boundary), then any disagreement on the best choice implies the nonexistence of an equilibrium. Similar simplification are possible in the analysis of Cournot best response dynamics.

3 An example without polarization: political debates

Despite the generality of the polarization lemma, most real life aggregation games do not end in polarization. To discuss the possible reasons, it is useful to introduce another example.

Consider a social decision maker who has to decide on the allocation of money to m budget items. Say the game is played by several interested advocates who are trying to influence his decision.[2] Let B, a closed convex subset of $\mathcal{R}^m$, denote the set of possible budget allocations. Simultaneously, every advocate chooses a point from B as his proposal for the final budget, and after hearing these proposals the social decision maker chooses a point in B to be the selected budget.

Assuming that the aggregation rule of the decision maker and the preferences of the advocates satisfy the hypotheses of the corollary in the previous section, we should expect at least $n-1$ advocates to offer budgets on the boundary of the set of feasible budget. Why is this often not the case?

First, the usual assumption of game theory that players maximize utility does not hold in many situations. Second, the players may not have complete information about opponents preferences assumed in the lemma. It would be interesting to investigate polarization in games of incomplete information.

Third, the assumption of individual responsiveness to advocates' positions

[2] There are several examples from politics that fall into this category. For example the advocates may be lobyists trying to influence a budget committee or political parties trying to influence the aggregate opinion of the voters.

does not hold in many situations. For example if the set of possible budgets is large, the opinion of an advocate who goes to "too far an extreme" may be totally ignored, and the social decision mechanism may stop responding to it.

Finally, the lemma applies to an isolated one-shot game. In practice, the game or similar games may be repeated. For example, an advocate may find himself to the left of one set of advocates and thus strategically want to choose an extreme left position, while among other advocates, find himself to the right. If the system has memory, it may be costly for him to flip flop his recommendation from one extreme to another. Thus a cautious non-myopic advocate may prefer to moderate his position or even act honestly in any given game.

4 Conclusions

Polarization can be a costly but natural phenomenon. In some examples, an arbitrarily small disagreement in preferences destabilizes a generally positive equilibrium to one where all players are miserable. In addition, we have generalized this to a class of aggregation games.

Polarization, of course, does not always occur and mechanisms may be chosen to actively avoid it. In politics, for example, adopting the choice of the median voter[8], rather then the average voter, results in an aggregation function that is not individually responsive and thus people do not necessarily polarize. Similarly, in the study of arbitration, it is recommended that the arbitrator use final-offer-arbitration, where she chooses an outcome recommended by one of the disputing parties, and not the average positions of the disputing parties [9]. Final-offer-arbitration does not satisfy the individual responsiveness condition. And, in households or other organizations, people often delegate decisions. This also breaks the individual responsiveness condition and thus avoids frustrating and costly polarization.

References

[1] Axelrod, Robert. (1984) The Evolution of Cooperation. New York: Basic Books.

[2] Cournot, August. (1963) The Mathematical Principles of the Theory of Wealth. Irwin Paperback Classics in Economics, selected and edited by Lloyd Reynolds and William Fellner.

[3] Griffin, E. M. (1997) A first look at communication theory. New York: McGraw-Hill.

[4] Nash, John F. (1951) Non Cooperative Games. Annals of Mathematics 54, 289-295.

[5] Aumann, R. J. and S. Sorin (1989), "Cooperation and Bounded Recall," Games and Economic Behavior, 1, 5-39.

[6] Bernheim, B.D. (1984), "Rationalizable Strategic Behavior," Econometrica 52, 1007-28.

[7] Pearce, D.G. (1984), "Rationalizable Strategic Behavior and the Problem of Perfection," Econometrica 52, 1029-50.

[8] Black, D. (1958), The Theory of Committees and Elections, London, Cambridge University Press.

[9] Crawford, V.P. (1985), "The Role of Arbitration and the Theory of Incentives," in Game Theory Models of Bargaining, A. Roth editor, Cambridge University Press, 363-39.