OPTIMAL JUDGEMENT AGGREGATION

Working paper, december 2005

Jesús Zamora Bonilla

UNED, Madrid, jpzb@fsof.uned.es

In recent years, literature on 'judgement aggregation' has grown in an appreciable way, basically in connection with the analysis of the so called 'discursive dilemma', or 'doctrinal paradox'. This states that aggregation of individual judgements by majority vote can result in contradictory claims depending on whether the voting is made on the premises of an argument, or on its conclusion. Most discussions of this result have been ellaborated within a formal framework related to social choice theory, and a number of 'impossibility theorems' similar to those of Arrow, Sen or Gibbard and Satterwhite have been proved: in general, these theorems assert that no mechanism for judgement aggregation exists that simmultaneously satisfies certain minimal

requirements.[1] The approach that will be pursued in this paper will be different, however, for I will suggest to consider the problem from a *contractarian* point of view, and so, it is an approach more related to 'constitutional political economy' than to 'social choice'. The problem I suggest to begin with is the following: if you were a member of a group that has to adopt a set of collective judgements on the basis of its members' opinions, according to some aggregation rule, *what specific rule would you prefer*? Assuming that this decision about the aggregation mechanism has to be taken 'under a veil of ignorance' (i.e., the rule has to be chosen before knowing what issues it will be applied to), we can simplify our analysis by considering just the preferences of a single, 'representative' individual. I will also assume that only *one kind* of aggregation norms is considered by the agents: those selecting as their outcome *a consistent and deductively closed set of statements* (i.e., a 'theory', in the logico-mathematical sense). This guarantees that the chosen rule will not generate paradoxes once it becomes applied. A further simplifying assumption will be that every individual always has a *complete* and *consistent* set of beliefs. So, the rules we are looking for belong to the set of functions that map one profile of complete theories (the individual opinions) into one single, not necessarily complete theory (the 'collective' judgment). The next relevant question about these functions is whether there is some implementable voting mechanism that systematically has as its outcome the optimum theory.

I will analyse these questions by proposing two different types of voting mechanisms: *premise based majority voting* (PMV) and *theory based majority voting* (TMV). According to the former, the theories that are discussed by the members of a

[1] See, e.g., C. List and P. Pettit, "Aggregating Sets of Judgments: An Impossibility Result," *Economics and Philosophy* 18(1) (2002): 89-110, and "Aggregating Sets of Judgments: Two Impossibility Results Compared," *Synthese* 140(1-2) (2004): 207-235.

group or committee can be expressed as conjunctions of a finite set of atomic propositions or their negations (e.g., each row in a truth table), votes are cast on each atomic proposition independently, and an atomic proposition (or its negation) is accepted if and only it receives a majority of votes. If p^1, p^2, ..., p^k, are the atomic propositions, complete theories will have the form: $p = \pm p^1 \,\&\, \pm p^2 \,\&\, ... \,\&\, \pm p^k$, where each symbol '±' is to be replaced by a negation or by nothing. The theory accepted by individual i will be called $p_i = \pm p^1{}_i \,\&\, \pm p^2{}_i \,\&\, ... \,\&\, \pm p^k{}_i$. The next thing we need is an assumption about the preferences of i. I will suppose to begin with that agents have only 'epistemic' preferences, i.e., they only care about the 'distance' between the theory which is collectively adopted and the theory they personally favour. This distance can be measured in a very straightforward way:[2] $d(p,q) = (1/k)$*(nr. of mismatches between p and q)*, and hence I will assume that the utility i receives if theory q is collectively accepted is given by the formula $u_i(q) = 1 - d(q,p_i)$.

From these assumptions, two interesting theorems can be derived:

(1) The outcome of PMV maximises the sum of the individual utilities.

Proof: For each atomic proposition p^j, majority voting guarantees that the outcome has fewer mistmatches with the individual opinions about p^j than its negation.[3]

(2) PMV is non manipulable.

[2] The notion of 'logical distance' between propositions or states of affairs has been particularly expoited within the literature on 'truth approximation' or 'verisimilitude' (cf. I. Niiniluoto, *Truthlikeness*, Dordrecht, D. Reidel, 1987).

Proof: For each atomic proposition, no individual can attain a higher utility level by voting the negation of the proposition she accepts.

As it is clear from discussions on the 'discursive paradox', PMV can lead the group to accept many conclusions (different than atomic propositions) that a majority of its members would reject. What our results show is that this has not to be taken as someting that is necessarily 'bad' from the point of view of the members of the group: in order to guarantee logical consistency, they abstain from voting the conclusions (hence, the mechanism violates the requirement of 'systematicity'), but, within this restriction, the mechanism selects an optimal theory, in the sense that the utility a member can expect to get by the systematic application of the rule is higher than with any other rule which selects consistent and complete theories. The adoption of PMV could be seen, then, as a reasonable 'epistemic constitutional agreement' (some problems will be discussed in a moment, though). The theory that is selected in each case is, so to say, the 'center of gravity' of the individual opinions, and, as long as the group members only care about the logical distance between the collective and their private opinion, they cannot do better if they want to preserve the consistency and completeness of the collective judgements. This is parallel to what happens in any kind of negotiation: the final outcome will very likely not be preferred by any of the negotiators (everyone would vote 'no' if she had right to make the choice by herself), but all can be 'satisfied' with it.

[3] Here, and in the rest of the paper, the simplifying assumption is made that the group has an odd number of members.

One problem with PMV has to do with the fact that theeories can have different axiomatisations, or, stated in a different way, the same states of affairs can be described by means of a different set of atomic propositions. It can be proved that logical distances are essentially dependent on the set of elementary sentences that is chosen.[4] So, our mechanism will only keep its appealing properties if the members of the group have previously agreed on what propositions are to be taken as atomic; this will certainly happen many times, but the choice of a 'language' can create opportunities to the strategic manipulation of the voting agenda. Nevertheless, from a contractarian perspective it is natural to assume that the prevalence of a linguistic framework can be the outcome of a previous agreement, and this suggests that optimality conditions for the choice of a language could also be analysed. This will be left for further work, nevertheless.

A more serious problem concerns our assumption that the individual utility funtions depended only on the logical distances between propositions. As the examples employed in the discussions on the doctrinal paradox show, this is not usually the case: agents can have a much stronger interest in the *conclusions* of the social deliberation process than in the premises collectively adopted to justify them, surely not because of the epistemic properties of those statements, but because of the *actions* that the group will take on the basis of the adopted conclusion. In order to tackle this problem, I will modify my former hypothesis about individual utilities, as well as the assumption that the group or committee is forced to choose a *complete* theory; instead, I will assume that it can choose as the 'collective opinion' any consistent theory, be it complete or not (the assumption that the outcome must be deductively closed is preserved). This entails

[4] This is the problem known as 'language variance' in the literature on truthlikeness, and

that, now, the options are not necessarily mutually inconsistent in a logical sense; technically, the options are all the possible subsets of rows in a truth table, save the empty set; the subsequent discussion will not depend on the hypothesis that the theories can be expressed in a propositional language. The new assumptions about individual utilities will be the following (again, $u_i(A)$ is the utility get by i if theory A is collectively adopted, and p_i is the complete theory accepted by i):

(3) (a) If $p_i \vdash A \vdash B$, then $u_i(B) \leq u_i(A)$.

(b) If $A \vdash B \vdash \neg p_i$, then $u_i(A) \leq u_i(B)$.

(c) $\forall A \forall i \forall j$, if $p_i \vdash A$, and $p_j \vdash \neg A$, then $u_i(A) = - u_j(A)$.

(d) $\forall A \ \forall i \forall j$, if $i, j \vdash A$, then $u_i(A) = u_j(A)$.

(e) $\forall i \ u_i(Taut) = 0$.

(3.a-b) assert that, amongst two theories that i accepts, she will prefer as the collective opinion the one with more content, and, amongst two theories she does not accept, she will prefer the less contentful; i.e., in principle, people prefers that the collective opinion reflects as more as possible their individual beliefs. This is a very reasonable assumption. On the other hand, (3.c-e) are only assumed by analytical convenience (*'Taut'* stands by the tautology), though the fourth one seems reasonable within the discussion about a choice made 'under a veil of ignorance'.[5] From the point fo view of the participants in such a choice, the most important fact is that, in each collective choice situation, *there will be some theory for which the sum of individual*

---

first identified by David Miller.

utilities attains a maximum, and agents would like to have an aggregation procedure which systematically leads the group to accept that theory, or at least, that on average produces a satisfactory output. I will argue that such a mechanism does exist.

By *theory based majority voting* (TMV) I will refer to a process in which the members of a group can form coallitions that propose a theory, A, which is then voted. If a majority of the members of the full group vote in favour of A, it becomes the collective opinion. If no theory attains a majority, then the group suspends judgment, what results in everybody having a utility equal to 0. A further difference is made between *simple* majority voting and *qualified* majority voting; in the latter case, some predetermined percentage w (> 0.5) of the group must vote in favour of the proposed theory if it is to be socially accepted (for simplicity, in the case of simple majority voting we take $w = 0.5$). I will say that a theory A is *w-defeatable* if and only if there is another theory B such that the set of members for which $u_i(A) < u_i(B)$ constitutes a w-majority. The following proposition state some very basic properties of this voting procedure (again, p_i represents the complete theory accepted by i; S represents the outcome of applying the w-TMV aggregation procedure):

(4) (a) $\exists i, j, ..., l,\ S = p_i \vee p_j \vee ... \vee p_m$

(b) S is non-w-defeatable.

Proof: (4.a) follows from (3.a), for, the individuals voting in favour of S will always prefer it to any other theory entailed by it. (4.b) follows from the fact that, if S were w-defeatable, other coallition would propose some theory which defeats S (this is

[5] (3.b) can be derived from (3.a) and (3.c), but for clarity I have preferred to state

equivalent to saying that S is the theory which maximises u_i for those individuals belonging to the winning majority).

Let S^* be the theory which maximises the sum of individual utilities. Then the most important result is the following theorem:

(5) There is some qualified majority level, w^*, such that S^* is the outcome of w^*-TMV.

Proof: (3.c-d) entail that S^*, having a possitive sum of individual utilities, will have a majority of members in favour of it (i.e., for which $u_i(S^*) > u_i(\neg S^*)$). Let w^* be the proportion of members in favour of S^*. If this theory is w^*-defeatable, there will exist another theory, *S'*, such that at least the same number of members prefer *S'* to S^*, but this, together with (3.c-d), entails that *S'* has a bigger aggregated utility than S^*, contrarily to the definition of S^*.

The next relevant question is whether S^* can be reached by simple majority voting (i.e., whether $w^* = 0.5$). It is easy to see that, in general, this will not be the case. (3.c-d) entail that the social utility associated to simple majority is equal to the utility of just *one* of the individuals voting for the winner theory (since there are $2n + 1$ individuals, $n+1$ of them will get $u_i(S)$, and the remaning n will get $-u_i(S)$). Let S^1 be the winning theory if w is set equal to $(n+2)/(2n+1)$; in this case the total utility attained by the group is $3u_i(S^1)$ (from a similar argument), and so, simple majority

---

explicitely the two first assumptions.

voting will be collectively better than *w*-majority voting only if $u_i(S)/u_i(S^1) > 3$. Hence, if individual utility decreases 'slowly' from the level it attains with the outcome of simple majority voting to the level attained under unanimity (i.e., when *w* equals 1), S^* will necessarily correspond to a majority level higher than 0.5. An elementary calculation shows that, if individual utility decreases linearly, then w^* will be at least 0.75 (depending on whether the utility attained under unanimity is 0 or higher; in spite of (3.e), this is not necessarily 0, but, taking into account (4.a), that utility level corresponding to the disjunction of all the complete theories that are *actually* accepted by somebody). But it will be still higher if individual utility decreases more intensely as *w* tends to *1*, i.e., *if the utility of getting a social outcome closer and closer to your own beliefs has decreasing marginal returns*, which seems a reasonable assumption (cf. figure 1, that represents the maximum utility an individual can get if a theory which is entailed by her own beliefs is collectively accepted, as a function of the minimum size of the group that would vote for that theory; if the utility function is linear, the area of the smaller shaded box will be proportional to the utility associated to the optimum theory, whereas if the utility function is concave, the corresponding area will be that of the bigger shaded box).

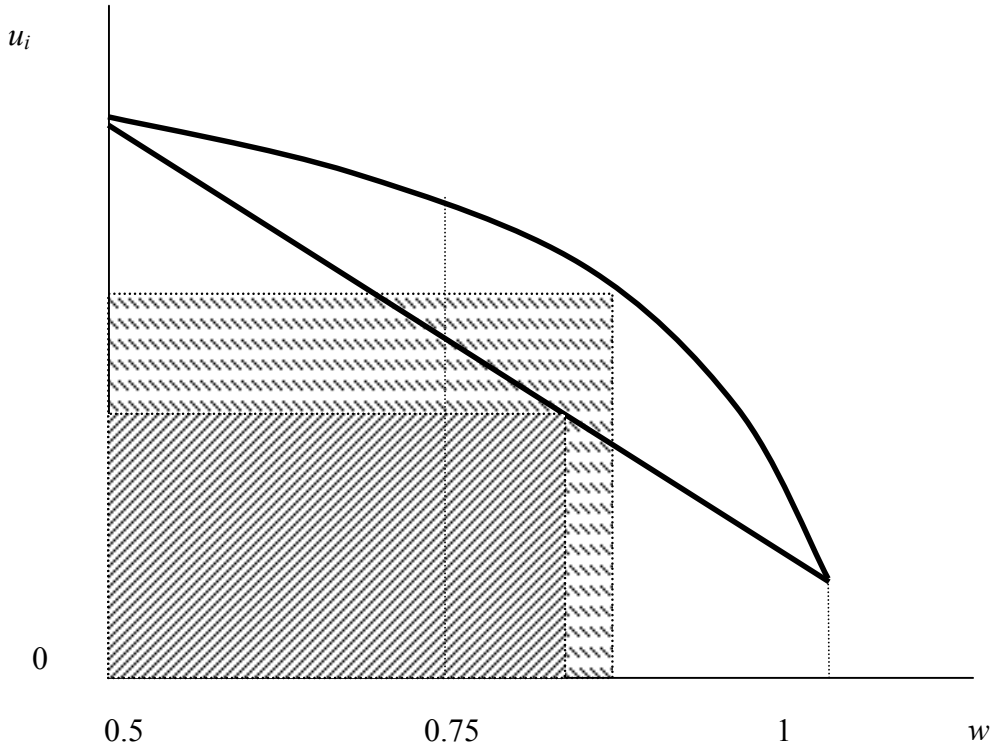

Figure 1

In conclusion, setting a *w*-TMV with *w* of 0.75 or slighty higher seems to be a nice strategy for the members of a group or committee to reach satisfactory cognitive agreements. Obviously, the optimum majority levels can be expected to change from a situation to another, but the group can set a *fixed* majority level that works well on the average. The theory that will be socially accepted in each case will be equal to the *disjunction* of all the theories accepted by the winning coallition (4.a). The combination of the unanimous choice of a qualified majority rule, and the capacity to negotiate in order to constitute a winning coallition, are the factors that allow individuals to perform satisfactorily in the task of making collective epistemic choices.