Modelling Iterative Judgment Aggregation

Terzopoulou, Z.; Endriss, U.

Publication date
2018
Document Version
Final published version
Published in
Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, Thirtieth Innovative Applications of Artificial Intelligence Conference, Eighth Symposium on Educational Advances in Artificial Intelligence

The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Modelling Iterative Judgment Aggregation

Zoi Terzopoulou
Institute for Logic, Language and Computation
University of Amsterdam
The Netherlands
z.terzopoulou@uva.nl

Ulle Endriss
Institute for Logic, Language and Computation
University of Amsterdam
The Netherlands
ulle.endriss@uva.nl

Abstract

We introduce a formal model of iterative judgment aggregation, enabling the analysis of scenarios in which agents repeatedly update their individual positions on a set of issues, before a final decision is made by applying an aggregation rule to these individual positions. Focusing on two popular aggregation rules, the premise-based rule and the plurality rule, we study under what circumstances convergence to an equilibrium can be guaranteed. We also analyse the quality, in social terms, of the final decisions obtained. Our results not only shed light on the parameters that determine whether iteration converges and is socially beneficial, but they also clarify important differences between iterative judgment aggregation and the related framework of iterative voting.

1 Introduction

Judgment aggregation is a rich formal framework for modelling decision making by groups of autonomous agents in complex domains that may require agreement on several interdependent issues (List and Puppe 2009; Grossi and Pigozzi 2014; Endriss 2016). Due to its broad applicability as a modelling tool, judgment aggregation has received attention from legal scholars, philosophers, economists, and computer scientists alike. However, one important feature of group decision making has so far not been thoroughly explored in the literature on judgment aggregation, namely the fact that complex decisions typically are arrived at in an iterative fashion, with each agent having the opportunity to refine her own stance multiple times as she finds out about the positions taken by her peers. To fill this gap, in this paper we propose a simple model of iterative judgment aggregation and study some of its fundamental properties.

To illustrate the idea, consider a group of colleagues that have to decide on whether to start working on a new project (proposition p). All of them think that they should proceed with p if and only if (a) their old project has reached a good stage of progress and (b) the new project is promising. Hence, every agent expresses a binary (*yes* / *no*) judgment on these two criteria and her judgment on p follows logically. Then, fixing the judgments of the whole group, a collective decision is made via an aggregation rule. In practice, online applications such as Doodle enable the agents to submit their judgments and observe the behaviour of their colleagues in real time. Subsequently, every agent may modify her declared judgment, if such an act induces an outcome on p that is more desirable to herself.

In our model, in each round of the decision process, each agent (maybe partially) observes the judgments reported by her peers and decides whether or not to update her own reported judgment in view of this information. The first agent to indicate she wishes to update gets granted permission to do so. This process is repeated *ad infinitum* or until none of the agents want to make any further updates. If and when the process terminates, an aggregation rule is used to map the final profile of individual judgments into a collective decision. In our analysis, we address two broad types of questions:

- Under what circumstances will the process of iterative judgment aggregation converge to a stable outcome, and how many iterations are required in the worst case?
- What is the quality, in social terms, of the decisions obtained by means of iterative judgment aggregation?

Answers to these questions will depend on a number of parameters, most notably the aggregation rule used. We provide answers for two aggregation rules, the *premise-based rule* and the *plurality rule*. The former is probably the most widely studied judgment aggregation rule in the literature (Pettit 2001; Chapman 2002; Dietrich and List 2007; Dietrich and Mongin 2010; Hartmann and Sprenger 2012), while the latter is closely related to the most widely analysed rule in voting theory (Zwicker 2016). We have chosen these rules for two reasons: First, they are computationally simple—unlike most other popular aggregation rules, such as the *distance-based rule* (Endriss, Grandi, and Porello 2012; Lang and Slavkovik 2014; de Haan and Slavkovik 2017). And second, they are guaranteed to always return a logically consistent collective judgment, i.e., they are not subject to the famous "doctrinal paradox" (Kornhauser and Sager 1993)—which is not the case for most other computationally simple rules, notably the *majority rule* (List and Pettit 2002).

Our model of iterative judgment aggregation is inspired by a recent line of research on *iterative voting*, initiated by Meir et al. (2010). For instance, we adopt the notion of *truth bias* employed by Obraztsova, Markakis, and Thompson (2013) and we model *partial information* similarly to

Reijngoud and Endriss (2012). While the general scenario of iterative decision making is the same in iterative voting and in our work, judgment aggregation is the more expressive framework of the two. Indeed, as is well-known, voting and preference aggregation can be embedded into judgment aggregation (Endriss 2016). In our work, we use the premise-based rule to demonstrate effects that occur in judgment aggregation but cannot be studied in the less expressive framework of voting, and we use the plurality rule to illustrate similarities between the two frameworks. Further results pertaining to our model may be found in the Master's thesis of the first author (Terzopoulou 2017).

The remainder of this paper is organised as follows. In Section 2 we review the basic framework of judgment aggregation and we define our iterative model under full and partial information. In Section 3 we establish various requirements for the convergence of aggregation rules and we study its speed. Then, we present our findings concerning the benefits of iteration for a group of agents as a whole in Section 4. We conclude in Section 5.

2 The Model

In this section we first recall relevant concepts in judgment aggregation (List and Puppe 2009; Grossi and Pigozzi 2014; Endriss 2016) and then present our new model of iterative judgment aggregation.

2.1 Basic Notation and Terminology

Consider a finite set of *agents* $N = \{1, \ldots, n\}$, with $n \geqslant 2$, who make judgments on several issues. These issues are represented by formulas in propositional logic. The set of issues to be judged is called the *agenda*, a nonempty set of formulas of the form $\Phi = \Phi^+ \cup \{\neg\varphi \mid \varphi \in \Phi^+\}$, where the *pre-agenda* Φ^+ consists of non-negated formulas only.

Several restrictions can be imposed on the structure of an agenda, in order to better capture the essence of specific aggregation situations. For example, a *conjunctive agenda* consists of a set of *premises* and a single *conclusion*. The latter is understood to be satisfied if and only of all premises are. Formally, the pre-agenda of a conjunctive agenda is of the form $\Phi^+ = \{p_1, \ldots, p_k, c\}$, with the p_j being propositional variables and $c = p_1 \wedge \cdots \wedge p_k$ (Dietrich and List 2007). Conjunctive agendas naturally occur in situations in which a final collective judgment has to be made on a conclusion, but the reasons that lead to that decision, encoded by the premises, are also important. The project management example above uses a conjunctive agenda with two premises.

Every agent $i \in N$ has a *truthful judgment set* $J_i \subseteq \Phi$, the set of issues she accepts. We assume that every agent's judgment set is a *consistent* set of formulas in the standard sense of logic. We also assume that it is *complete*, i.e., that $\varphi \in J_i$ or $\neg\varphi \in J_i$ for every $\varphi \in \Phi^+$. The set of all consistent and complete subsets of φ is denoted by $\mathcal{J}(\Phi) \subseteq 2^\Phi$, where 2^Φ is the powerset of Φ. A *profile* $\boldsymbol{J} = (J_1, \ldots, J_n) \in \mathcal{J}(\Phi)^n$ is a vector of judgment sets, one for each agent. A (partial) profile $\boldsymbol{J}_{-i}$ is a vector of all the agent's judgment sets, except for agent i. We write $N_\varphi^{\boldsymbol{J}}$ for the set $\{i \mid \varphi \in J_i\}$ of agents who accept formula φ in profile $\boldsymbol{J}$.

2.2 Aggregation Rules

An *aggregation rule* is a function $F : \mathcal{J}(\Phi)^n \to 2^\Phi$ that maps every profile $\boldsymbol{J}$ of complete and consistent judgment sets to a single (not necessarily complete and consistent) judgment set $F(\boldsymbol{J})$. While many of the rules studied in the literature indeed can return outcomes that are not complete and consistent, the specific rules we analyse in this paper do not suffer from this deficiency. We stress that we focus on *resolute* rules, which always return a single judgment set. While many naturally defined aggregation rules are irresolute, permitting ties between several outcomes, in practice we usually require rules that return a definitive answer for every input. Whenever necessary, we enforce this by using a (lexicographic) tie-breaking rule.

An example for a widely used aggregation rule is the *premise-based rule* F^{pr}, which we define here on conjunctive agendas Φ only. Let $\Phi^+ = \{p_1, \ldots, p_k, c\}$. Given a profile $\boldsymbol{J} \in \mathcal{J}(\Phi)^n$, we compute $F^{pr}(\boldsymbol{J})$ as follows. First, a collective decision is made on the premises using the majority rule, i.e., for every $p \in \{p_1, \ldots, p_k\}$, we let $p \in F^{pr}(\boldsymbol{J})$ if $|N_p^{\boldsymbol{J}}| > \frac{n}{2}$, and $\neg p \in F^{pr}(\boldsymbol{J})$ otherwise. Then, $c \in F^{pr}(\boldsymbol{J})$ if $\{p_1, \ldots, p_k\} \subseteq F^{pr}(\boldsymbol{J})$, and $\neg c \in F^{pr}(\boldsymbol{J})$ otherwise.

Another aggregation rule, directly inspired by voting theory (Zwicker 2016), is the *plurality rule* $F^{p\ell}$. It returns one of the judgment sets most frequently reported by the individual agents, i.e., for $\boldsymbol{J} = (J_1, \ldots, J_n)$ we get:

$$F^{p\ell}(\boldsymbol{J}) \quad \in \quad \operatorname*{argmax}_{J \subseteq \varphi} |\{i \in N \mid J = J_i\}|$$

Ties are possible under this rule, so we always return the lexicographically first amongst the judgment sets maximising the plurality score. The plurality rule is a so-called *representative-voter rule* (Endriss and Grandi 2014), i.e., an aggregation rule for which the range of possible outcomes is limited to judgment sets occurring in the input profile. Note that this is not the case for the premise-based rule.

2.3 Preferences

Let us think of an aggregation problem in terms of a game: every agent chooses an action, which is the (truthful or untruthful) judgment set she reports, and the outcome is computed by the submitted profile of the group in accordance with a fixed aggregation rule. We assume that every agent i holds some (transitive and complete) preference order $\succsim_i$ over all the possible collective judgment sets in 2^Φ. Then, an agent's choice of action, as in every standard game, is directly connected to the agent's preferences over the outcomes. Similarly to Baumeister et al. (2015) and de Haan (2017), we assume that each agent i has a *desired set* $J_i^\heartsuit$, i.e., a set of issues that she most cares about, which is a consistent subset of her truthful judgment J_i. Then, the preferences are formed based on the formulas that occur in the agent's desired set.

One instance of complete preferences—which applies to conjunctive agendas and is justified when only conclusions carry consequences that an agent i cares about, i.e., when $J_i^\heartsuit = \{c\}$ or $J_i^\heartsuit = \{\neg c\}$—is the class of *conclusion-oriented* preferences (Dietrich and List 2007; List and Pettit 2011). More formally, $\succsim_i$ is called conclusion-oriented

if, for all $J, J' \in 2^{\Phi}$, it holds that $J \succsim_i J'$ if and only if (i) $c \in J_i \cap J'$ implies $c \in J_i \cap J$ and (ii) $\neg c \in J_i \cap J'$ implies $\neg c \in J_i \cap J$.

On the other hand, in settings where an agent i is expected to care equally about all the issues in the agenda, i.e., where $J_i^{\heartsuit} = J_i$, it is appropriate to assume *Hamming-distance* preferences. Although somewhat restrictive, this is the most widely used notion of preference in the literature on strategic behaviour in judgment aggregation to date (Dietrich and List 2007; Endriss, Grandi, and Porello 2012; Baumeister et al. 2015; Botan, Novaro, and Endriss 2016). The *Hamming distance* $H(J, J')$ between two judgment sets $J, J' \in 2^{\Phi}$ is defined as the number of formulas on which they disagree: $H(J, J') = |J \backslash J'| + |J' \backslash J|$. Then agent i with truthful judgment set J_i is said to have Hamming-distance preferences if it is the case that $J \succsim_i J'$ if and only if $H(J, J_i) \leqslant H(J', J_i)$, for all judgment sets $J, J' \in 2^{\Phi}$.

Notice that both the conclusion-oriented and the Hamming-distance preferences are *closeness-respecting* preferences, as defined by Dietrich and List (2007). We write $J \succ_i J'$ whenever $J \succsim_i J'$ but not $J' \succsim_i J$.

2.4 Iteration under Full Information

Let F be an aggregation rule. We consider its iteration as follows. In each round t the rule F prescribes a (temporary) collective outcome based on the submitted profile $\boldsymbol{J}_t = (J_{1,t}, \ldots, J_{n,t})$. All the agents observe the judgments of their peers as well as the collective decision of round t and the first who communicates a wish to alter her judgment proceeds with doing so. We assume that an agent has an incentive to change her stance when a new judgment can induce a more desirable outcome for herself. This behaviour is called *strategic behaviour* or *manipulation* (Dietrich and List 2007). We shall also assume that in every round the agents choose a reaction with regard to the information that is available to them in that specific round only, that is, they are *memoryless*. In addition, they are *myopic*, in the sense that they only aim at a better outcome in the next round—said differently, they treat every round as if it were the last one. These assumptions are common in the literature on iterative voting (Meir et al. 2010; Obraztsova et al. 2015).

We say that the judgment set J_i' *dominates* the judgment set J_i for agent i in round t if $F(J_i', \boldsymbol{J}_{-i,t}) \succ_i F(J_i, \boldsymbol{J}_{-i,t})$. Then, $J_{i,t+1}$ is a *better response* of agent i in round t if it dominates agent i's current judgment $J_{i,t}$. A better response $J_{i,t+1}$ is a *best response* of agent i in round t if additionally it is *undominated*. The best responses constitute available *improvement steps* of agent i. Moreover, if agent i has no best response in round t but it is the case that her truthful judgment J_i induces an equally desirable collective outcome as the untruthful judgment $J_{i,t}$ that she currently submits, i.e., if $F(J_i, \boldsymbol{J}_{-i,t}) \sim_i F(J_{i,t}, \boldsymbol{J}_{-i,t})$, then agent i has two options: either to stick to her insincere opinion, or to switch to her sincere one. The latter move will be considered an improvement step if and only if the agent is *truth-biased* (Meir et al. 2010; Obraztsova, Markakis, and Thompson 2013). Finally, we shall assume that an agent wants to change her judgment in round t if and only if she has some improvement step available in that round. When an agent is the one selected to submit a new judgment and there is more than one opportunity for improvement steps, if her truthful judgment is one of them, then she will choose to be honest. Otherwise, she will perform one improvement step at random. Furthermore, an agent being *outcome-focused* captures a reasonable assumption: whenever some improvement step of hers consists in directly submitting a preferable judgment set and turning it into the collective outcome, her priority is to do so, instead of manipulating the result indirectly. If an agent is not outcome-focused, we call her *unrestricted*.

Depending on the profile that the agents submit in the first round of the iterative process and the order in which they modify their judgments, different *improvement paths* are created. We say that the iteration *converges to a stable state* (or an *equilibrium*) if every improvement path terminates after a finite number of rounds. In other words, an equilibrium is a profile where no agent can profit from a unilateral deviation.

We stress that our model of iterative judgment aggregation, in which agents move strategically, differs from the model introduced by Slavkovik and Jamroga (2016), in which agents cooperate by moving towards the positions taken by their peers so as to reach consensus.

2.5 Iteration under Partial Information

So far we have been making a rather strong assumption following the literature to date in judgment aggregation: that the agents constantly know everything about the judgments of their peers. However, this may often not be the case, for example when the aggregation involves confidential issues or a large number of agents. Indeed, the study of iterative rules under *partial information* in voting has been pursued recently, e.g., by Reijngoud and Endriss (2012) and Endriss et al. (2016). We now develop a novel model of partial information in the context of judgment aggregation.

Consider an iterative process where the aggregation rule F is applied. In every round t, the *information set* $\mathcal{W}_{i,t}$ of agent i contains all the (partial) profiles $\boldsymbol{J}_{-i} \in \mathcal{J}(\Phi)^{n-1}$ that agent i views as possible to be submitted by the group in round t. For instance, if agent i is fully informed about the judgments of her peers in round t, then $\mathcal{W}_{i,t} = \{\boldsymbol{J}_{-i,t}\}$; if agent i is fully ignorant, then $\mathcal{W}_{i,t} = \mathcal{J}(\Phi)^{n-1}$. Under partial information, we say that the judgment set J_i' *dominates* the judgment set J_i for agent i in round t if (i) there is a possible scenario $\boldsymbol{J}'_{-i,t} \in \mathcal{W}_{i,t}$ where agent i is able to achieve a strictly preferable outcome for herself by submitting J_i' instead of J_i (i.e., if $F(J_i', \boldsymbol{J}'_{-i,t}) \succ_i F(J_i, \boldsymbol{J}'_{-i,t})$) and (ii) there is no scenario $\boldsymbol{J}''_{-i,t} \in \mathcal{W}_{i,t}$ under which J_i' will lead the agent to a strictly less desirable outcome (i.e., $F(J_i, \boldsymbol{J}''_{-i,t}) \succ_i F(J_i', \boldsymbol{J}''_{-i,t})$). This means that the agents are *risk-averse* (Reijngoud and Endriss 2012). Again, $J_{i,t+1}$ is a *better response* of agent i in round t if it dominates her current judgment $J_{i,t}$, and it is a *best response* if it is also *undominated*. The agents choose among their best responses—which also constitute their improvement steps—by prioritising their truthful judgments, similarly to the case of full information. In addition, if agent i has no best re-

sponse available in round t but her truthful judgment J_i induces an equally desirable collective outcome as her current untruthful judgment $J_{i,t}$ in all scenarios that the agent considers possible, i.e., if $F(J_i, \boldsymbol{J}_{-i,t}) \sim_i F(J_{i,t}, \boldsymbol{J}_{-i,t})$ for all $\boldsymbol{J}_{-i,t} \in \mathcal{W}_{i,t}$, then a *truth-biased* agent i will want to switch to her honest judgment in the next round $t+1$. Note that all the notions of this section refine the case of full information.

3 Convergence to Equilibria

In this section we utilise the framework of iterative judgment aggregation, by analysing in depth two common aggregation rules: the premise-based rule and the plurality rule.

3.1 The Premise-based Rule Iterated

Dietrich and List (2007) showed that when the premise-based rule is applied, incentives for manipulation arise for agents who have conclusion-oriented preferences. We also restrict attention to this case. We begin with investigating fully-informed agents who are truthful in the first round. We prove that an equilibrium is then guaranteed to be reached after at most one round of iteration:

Theorem 1. *For a conjunctive agenda Φ and fully-informed agents with conclusion-oriented preferences, the iterative premise-based rule converges from the truthful profile in at most one round, independently of truth-bias assumptions.*

Proof. In the first round an agent has an opportunity for an improvement step if and only if (i) she rejects the conclusion c, (ii) the group collectively accepts c, and (iii) the agent can make the group reject c in the next round by untruthfully rejecting some premise. Suppose that there is such an agent who performs an improvement step. In the second round all the agents that accept c still have no way to affect the outcome and all the other agents already obtain their desired result. This is an equilibrium. □

What if fully-informed agents initially submit some non-truthful profile of judgments? We prove that the iterative F^{pr} always reaches an equilibrium, but it may take a linear number of rounds with respect to the number of the agents n (Theorems 3 and 4). Lemma 2 is quite intuitive:

Lemma 2. *For a conjunctive agenda Φ and the iterative premise-based rule, a conclusion-oriented agent who truthfully accepts the conclusion performs an improvement step only when she is untruthful and switches to truthfulness.*

Theorem 3. *For a conjunctive agenda Φ and fully-informed, non-truth-biased agents with conclusion-oriented preferences, the iterative premise-based rule converges from any initial profile in at most n rounds.*

Proof. Call A and R the sets of agents who truthfully accept and reject the conclusion, respectively. By Lemma 2, an agent in A can perform an improvement step at most once. On the other hand, an agent in R can perform an improvement step if she submits a judgment that makes a previously accepted conclusion be rejected by the group. Since the agents in A and R only perform improvement steps alternatively, an equilibrium is reached after at most n rounds. □

Theorem 4. *For a conjunctive agenda Φ and fully-informed, truth-biased agents with conclusion-oriented preferences, the iterative premise-based rule converges from any initial profile in at most $\frac{3n}{2}$ rounds.*

The bound of Theorem 4 is higher than that of Theorem 3, because truth-biased agents who truthfully reject the conclusion may also perform improvement steps when the group already rejects the conclusion. The details of the proof are omitted in the interest of space.[1]

How does partial information affect the strategic acts of the agents and the convergence of an aggregation rule? This is the question we shall focus on next. We examine the extreme case of full ignorance for the premise-based rule and we show that an equilibrium is reached in at most as many rounds as the number of the agents n.

Theorem 5. *For a conjunctive agenda Φ and fully-ignorant agents with conclusion-oriented preferences, the iterative premise-based rule converges from any initial profile (including the truthful one) in at most n rounds, independently of truth-bias assumptions.*

Proof. First, the agents who truthfully accept the conclusion have an improvement step available only if they are insincere and they move to their truthful judgment (similarly to Lemma 2). Moreover, every agent who truthfully rejects the conclusion can make a unique best improvement step: since she does not know whether the conclusion is collectively accepted and which may be the critical premises that can alter the result, the option that dominates all her other options is to reject all the premises. Thus, all the agents may perform an improvement step at most once. □

Overall, we conclude that withholding information from the agents can be both damaging to and beneficial for the convergence speed of the premise-based rule, depending on whether sincerity in the first round is to be expected or not.

3.2 The Plurality Rule Iterated

Contrary to the premise-based rule, the plurality rule may be applied on any kind of agenda. Hence, we now work with agents who hold the more general Hamming-distance preferences. Since the plurality rule is well established in voting, we wish to also illuminate its features in judgment aggregation. Interestingly, as far as convergence is concerned, we observe that the similarities between the two frameworks prevail. First, we study fully-informed agents and state two main results that hold immediately from the relevant literature on voting, as their proofs have a direct translation to our model: By Meir et al. (2010), the plurality rule converges from any initial profile for non-truth-biased, outcome-focused agents. Furthermore, for iterations that start from the truthful profile, $F^{p\ell}$ always converges even when the agents are unrestricted (Reijngoud 2011).

Recall that the premise-based rule was shown in the previous section to always converge, independently of whether the agents are truth-biased. Nevertheless, the truth-bias assumption is proven to be critical for the plurality rule:

[1] For a proof of a slightly weaker result see Terzopoulou (2017).

Proposition 6. *The iterative plurality rule does not always converge for truth-biased agents with Hamming-distance preferences.*

Proof. Consider the agenda Φ with $\Phi^+ = \{p, q, r, s\}$ and the judgment sets $J_1 = \{p, q, \neg r, \neg s\}, J_2 = \{p, \neg q, \neg r, \neg s\}, J_3 = \{p, q, r, s\}, J_4 = \{\neg p, \neg q, \neg r, \neg s\}$. Then, take the group of agents $N = \{1, \ldots, 6\}$, where agents 3 and 4 truthfully hold the judgments J_3 and J_4 and they have the Hamming-distance preferences $J_3 \succ_3 J_1 \succ_3 J_2 \succ_3 J_4$ and $J_4 \succ_4 J_2 \succ_4 J_1 \succ_4 J_3$ respectively.

	J_1	J_2	J_3	J_4	
round 1:	$\underline{2}$	2	1	1	4
round 2:	2	$\underline{3}$	1	0	3
round 3:	$\underline{3}$	3	0	0	4
round 4:	$\underline{3}$	2	0	1	3
round 5:	$\underline{2}$	2	1	1	4
	$\vdots$				

In every row of the table above we depict the number of agents that submit each judgment in that round. The underlined numbers denote that the respective judgment set is the (temporary) collective decision, and at the right side of each row we see the agent who makes an improvement step in that round. The profile of the fifth round is the same as the profile of the first round, so a cycle is created. □

An extended analysis of the iterative plurality rule must also take into account settings of partial information. We now pursue this direction. Remarkably, it is the case that under complete lack of information the agents never have an incentive to strategise when the plurality rule is applied (Terzopoulou 2017), which means that every profile is an equilibrium (see the work of Conitzer, Walsh, and Xia (2011) for an analogous fact in voting). We have thus established that the iterative plurality rule always converges after a finite number of rounds when non-truth-biased agents are in one of the two extremes of the information-spectrum. Next, we ponder: Is convergence guaranteed for any type of intermediate information that the agents may hold? Notably, we show that under a very natural type of information, where the agents are only informed about the current collective decision in each round, an equilibrium may never be reached.

Formally, when agent i is only informed about the collective outcome $F^{p\ell}(J_{i,t}, \boldsymbol{J}_{-i,t})$ in round t, she considers all (partial) profiles $\boldsymbol{J'}_{-i}$ that would induce this outcome possible to have been submitted by the group. That is, $\mathcal{W}_{i,t} = \{\boldsymbol{J'}_{-i} \in \mathcal{J}(\Phi)^{n-1} : F^{p\ell}(J_{i,t}, \boldsymbol{J}_{-i,t}) = F^{p\ell}(J_{i,t}, \boldsymbol{J'}_{-i,t})\}$.

Proposition 7. *The iterative plurality rule does not always converge for agents with Hamming-distance preferences that only know the current collective judgment, independently of their truth bias and initial truthfulness.*

Proof. Consider the agenda Φ with $\Phi^+ = \{p, q, r, s\}$ and the judgment sets $J_1 = \{p, q, \neg r, \neg s\}, J_2 = \{p, q, \neg r, s\}, J_3 = \{p, q, r, s\}, J_4 = \{\neg p, \neg q, r, s\}$. Then, take a group of agents $N = \{1, \ldots, 5\}$, where agents 3 and 4 truthfully hold the judgments J_3 and J_4 and they have the Hamming-distance preferences $J_3 \succ_3 J_2 \succ_3 J_4 \sim_3 J_1$ and $J_4 \succ_4 J_3 \succ_4 J_2 \succ_4 J_1$ respectively.

	J_1	J_2	J_3	J_4	
round 1:	$\underline{2}$	1	1	1	3
round 2:	$\underline{2}$	2	0	1	4
round 3:	$\underline{2}$	2	1	0	3
round 4:	$\underline{2}$	1	2	0	4
round 5:	$\underline{2}$	1	1	1	3
	$\vdots$				

The profile depicted in the table above creates a cycle. To read the table, consult the proof of Proposition 6. □

Hence, it is worth stressing that less information can both bring about and prevent the convergence of the plurality rule.

4 The Social Benefits of Iteration

Does individual strategic behaviour profit a group of agents as a whole? This question has been formulated in its full generality by algorithmic game theorists (Koutsoupias and Papadimitriou 2009) and was recently investigated in an iterative voting framework by Brânzei et al. (2013). In this section we initiate the discussion of *social welfare* in judgment aggregation. To put our analysis into context, we first wonder how profitable sticking to the truth actually is for a group of agents *en masse*. To that end, we define and use a notion analogous to the well-known *Price of Anarchy* (PoA) (Papadimitriou 2001), namely the *Price of Truth* (PoT). Then, in order to study to what extent strategising is able to improve the outcome in social terms, we employ the *Dynamic Price of Anarchy* (DPoA).

Fix an agenda Φ and a group of agents $N = \{1, \ldots, n\}$. Considering an aggregation rule F and a truthful profile $\boldsymbol{J}$, the optimal social outcome is achieved when the collective decision maximises the *proportional* agreement with the agents' desired sets, viz., the *social welfare*. We define the *Price of Truth of F for profile $\boldsymbol{J}$* as the ratio between the social welfare of the optimal outcome and the social welfare of the outcome obtained under the sincere profile:

$$\mathrm{PoT}(F, \boldsymbol{J}) = \frac{\max\limits_{\boldsymbol{J'} \in \mathcal{J}(\Phi)^n} \sum\limits_{i \in N} \frac{|F(\boldsymbol{J'}) \cap J_i^{\heartsuit}|}{|J_i^{\heartsuit}|}}{\sum\limits_{i \in N} \frac{|F(\boldsymbol{J}) \cap J_i^{\heartsuit}|}{|J_i^{\heartsuit}|}}$$

Then, the *Price of Truth* of an aggregation rule F is its maximum Price of Truth for all possible initial profiles:

$$\mathrm{PoT}(F) = \max_{\boldsymbol{J} \in \mathcal{J}(\Phi)^n} \mathrm{PoT}(F, \boldsymbol{J})$$

The higher the PoT of an aggregation rule F is, the more socially harmful telling the truth is when F is used. But the result can change if we allow for repeated aggregation. Naturally, we wish to evaluate the social benefits of iteration in cases where the agents are initially sincere. To that end, we study converging aggregation rules and we take into account their worst-case equilibria, regarding non-trivial iterations

where at least one improvement step takes place when starting from the truthful profile. We define the *Dynamic Price of Anarchy of F for profile $\boldsymbol{J}$*:

$$\mathrm{DPoA}(F, \boldsymbol{J}) = \frac{\max\limits_{\boldsymbol{J}' \in \mathcal{J}(\Phi)^n} \sum\limits_{i \in N} \frac{|F(\boldsymbol{J}') \cap J_i^{\heartsuit}|}{|J_i^{\heartsuit}|}}{\min\limits_{\boldsymbol{J}'' \in EQ} \sum\limits_{i \in N} \frac{|F(\boldsymbol{J}'') \cap J_i^{\heartsuit}|}{|J_i^{\heartsuit}|}}$$

Here EQ is the set of all equilibrium profiles $\boldsymbol{J}''$, such that there is an improvement path (of length at least one) starting from profile $\boldsymbol{J}$ and reaching $\boldsymbol{J}''$.

The maximum value of the DPoA of a rule over all initial truthful profiles constitutes its *Dynamic Price of Anarchy*:

$$\mathrm{DPoA}(F) = \max_{\boldsymbol{J} \in \mathcal{J}(\Phi)^n} \mathrm{DPoA}(F, \boldsymbol{J})$$

The smaller and closer to 1 the DPoA of an aggregation rule F is, the more socially profitable F is in iteration. Moreover, if $\mathrm{PoT}(F) > \mathrm{DPoA}(F)$ holds, it means that strategic behaviour benefits the group more than sincerity for F.

4.1 Benefits under the Premise-based Rule

We examine the interesting case for the premise-based rule F^{pr}, which concerns conclusion-oriented agents. Our first observation is that telling the truth can unfortunately be infinitely detrimental for the group as a whole:

Proposition 8. *There exist an agenda Φ and a group of agents N for which PoT(F^{pr}) is infinite.*

Proof. Consider the conjunctive agenda Φ with $\Phi^+ = \{p_1, p_2, p_3, c\}$. There is a truthful profile $\boldsymbol{J}$ where all the agents reject the conclusion, but the rule still accepts it:

	p_1	p_2	p_3	c
Agent 1:	Yes	Yes	No	**No**
Agent 2:	Yes	No	Yes	**No**
Agent 3:	No	Yes	Yes	**No**
F^{pr}	Yes	Yes	Yes	**Yes**

Then, $\sum\limits_{i \in N} |F^{pr}(\boldsymbol{J}) \cap J_i^{\heartsuit}| = 0$. □

Luckily, iteration provides a solution to this problem. Specifically, it can guarantee the reach of the optimal social result, both under full information and under full ignorance:

Theorem 9. *For a conjunctive agenda Φ and fully-informed conclusion-oriented agents, DPoA$(F^{pr}) = 1$.*

Proof. Recall that we consider iterations that start from the truthful profile of the agents. Then, an improvement step can be made only by an agent who rejects the conclusion and sees that at least half of her peers also reject it. Thus, the result on the conclusion in an equilibrium has to agree with the majority's desired sets at least, and will be optimal. □

Theorem 10. *For a conjunctive agenda Φ and fully-ignorant conclusion-oriented agents, DPoA$(F^{pr}) = 1$.*

Proof. The dominant option of the agents who accept the conclusion under full ignorance on a conjunctive agenda is to be truthful, and that of the agents who reject the conclusion is to reject all the premises. Hence, once an equilibrium is reached, either the majority of the agents accepts all the premises (making the conclusion accepted) or the majority rejects them, making the conclusion rejected. In either case, the optimal outcome is achieved. □

Overall, when the premise-based rule is applied and the agents behave strategically, the group is always better off as a whole than in cases where everyone is simply truthful:

Corollary 11. *Take a conjunctive agenda φ, fully-informed or fully-ignorant conclusion-oriented agents, and any truthful profile $\boldsymbol{J}$ that induces at least one round of iteration. Then, PoT$(F^{pr}, \boldsymbol{J}) >$ DPoA$(F^{pr}, \boldsymbol{J})$.*

4.2 Benefits under the Plurality Rule

Back to the analysis of the plurality rule and speaking intuitively, a plurality outcome does not capture the logical complexity of the individual judgments. Thus, it is reasonable to guess that the PoT as well as the DPoA of the plurality rule for agents with Hamming-distance preferences can be rather high. We formally verify this intuition (note that all the results of this section refer to fully-informed agents):

Proposition 12. *There exist an agenda Φ and a group of agents N with Hamming-distance preferences for which PoT$(F^{p\ell})$ is $\Omega(|\mathcal{J}(\Phi)|)$.*

Proof. The following technical comment is important. We denote by $100\ldots0$ the judgment set that contains a designated formula φ_1 in the pre-agenda Φ^+ and the negations of all the other formulas, etc. By Dokow and Holzman (2009), for every nonempty subset X of $\{0,1\}^m$ there exists an agenda Φ with $\Phi^+ = \{\varphi_1, ..., \varphi_m\}$ such that $\mathcal{J}(\Phi) = X$.

Consider an agenda Φ with $|\Phi^+| = m$, $|\mathcal{J}(\Phi)| = k$ and $k \leqslant m$ and a sufficiently large group of agents N with Hamming-distance preferences. We first construct an instance of a truthful profile $\boldsymbol{J} = (J_1, \ldots, J_n)$ that contains all the judgment sets in $\mathcal{J}(\Phi)$. To ease the demonstration we assume that n is divisible by k, but this is not critical.

$$\begin{array}{lll}
J_1 = J_{k+1} & = \ldots = J_{(x-1)k+1} & = 11\ldots1\ldots1 \\
J_2 = J_{k+2} & = \ldots = J_{(x-1)k+2} & = 00\ldots0\ldots0 \\
J_3 = J_{k+3} & = \ldots = J_{(x-1)k+3} & = 10\ldots0\ldots0 \\
J_4 = J_{k+4} & = \ldots = J_{(x-1)k+4} & = 01\ldots0\ldots0 \\
 & \ldots & \\
J_k = J_{2k} & = \ldots = J_{xk} & = 00\ldots1\ldots0
\end{array}$$

The optimal social welfare for the profile $\boldsymbol{J}$ is obtained when the collective decision is J_2. So, $\max\limits_{\boldsymbol{J}' \in \mathcal{J}(\Phi)^n} \sum\limits_{i \in N} \frac{|F^{p\ell}(\boldsymbol{J}') \cap J_i|}{m} = \frac{x}{m}(0 + m + (k-2)(m-1))$.

Then, assuming that the lexicographic tie-breaking rule ranks J_1 at the top, in which case J_1 will be the truthful outcome of $F^{p\ell}$, the social welfare is $\frac{x}{m}(m+0+k-2)$. Thus, $\mathrm{PoT}(F^{p\ell}, \boldsymbol{J}) = \frac{(k-1)m+2-k}{m+k-2}$, which is linear in k. □

Proposition 13. *There exist an agenda Φ and a group of agents N with Hamming-distance preferences for which DPoA$(F^{p\ell})$ is $\Omega(|\mathcal{J}(\Phi)|)$.*

Proof. Analogous to the proof of Proposition 12. Consider the following initial profile:

$$\begin{array}{lll}
J_1 = J_{k+1} & = \ldots = J_{(x-1)k+1} & = 11111\ldots1\ldots1111 \\
J_2 = J_{k+2} & = \ldots = J_{(x-1)k+2} & = 01111\ldots1\ldots1111 \\
J_3 = J_{k+3} & = \ldots = J_{(x-1)k+3} & = 11111\ldots1\ldots1100 \\
J_4 = J_{k+4} & = \ldots = J_{(x-1)k+4} & = 11111\ldots1\ldots1000 \\
J_5 = J_{k+5} & = \ldots = J_{(x-1)k+5} & = 00000\ldots0\ldots0000 \\
J_6 = J_{k+6} & = \ldots = J_{(x-1)k+6} & = 10000\ldots0\ldots0000 \\
J_7 = J_{k+7} & = \ldots = J_{(x-1)k+7} & = 01000\ldots0\ldots0000 \\
J_8 = J_{k+8} & = \ldots = J_{(x-1)k+8} & = 00100\ldots0\ldots0000 \\
 & \ldots & \\
J_k = J_{2k} & = \ldots = J_{xk} & = 00000\ldots1\ldots0000
\end{array}$$

It suffices to note that there is an iteration terminating with J_3 winning. □

Bear in mind that when the plurality rule is applied in voting, strategic agents always achieve outcomes that are very close to the optimal one (Brânzei et al. 2013). The divergence between this result and ours mirrors some of the significant consequences that collective decision making in the richer framework of judgment aggregation carries.

Since neither sincerity nor iterated strategic behaviour can guarantee a good outcome for the agents *en masse* when the plurality rule is applied, an immediate next question arises: Does iteration at least bring the group *closer* to the optimal outcome? The answer is positive for agendas with three complete and consistent subsets, but negative in general:

Theorem 14. *Consider an agenda Φ with $|\mathcal{J}(\Phi)| = 3$ and a sufficiently large group of agents with Hamming-distance preferences. Then, PoT$(F^{p\ell}, \boldsymbol{J}) \geqslant$ DPoA$(F^{p\ell}, \boldsymbol{J})$ for every $\boldsymbol{J} \in \mathcal{J}(\Phi)$.*

Proof. Let $\mathcal{J}(\Phi) = \{J_1, J_2, J_3\}$ and $|J_1| = |J_2| = |J_3| = m$. Denote $|J_1 \cap J_2| = x_{12}$, $|J_1 \cap J_3| = x_{13}$ and $|J_2 \cap J_3| = x_{23}$. Since we only look at truthful profiles $\boldsymbol{J}$ where at least one agent has an opportunity to perform an improvement step, the following two cases partition the set of all truthful profiles that are of interest to us.

Case 1: There are (at least) two judgment sets, say J_1, J_2, that are submitted by the same number of agents σn, where $\sigma < \frac{1}{2}$ is a positive rational number. Moreover, the tie-breaking rule selects J_1, while an agent i who truthfully submits J_3 prefers J_2 to J_1. Then, after agent i makes her improvement step, J_2 becomes the collective outcome and the process terminates. We have that $\frac{\text{PoT}(F^{p\ell},\boldsymbol{J})}{\text{DPoA}(F^{p\ell},\boldsymbol{J})} = \frac{\sigma nm+\sigma nx_{12}+(n-2\sigma n)x_{23}}{\sigma nx_{12}+\sigma nm+(n-2\sigma n)x_{13}}$. Since agent i truthfully has the judgment J_3 and prefers J_2 to J_1, it holds that $x_{13} < x_{23}$, hence $\frac{\text{PoT}(F^{p\ell},\boldsymbol{J})}{\text{DPoA}(F^{p\ell},\boldsymbol{J})} > 1$.

Case 2: The winning judgment set is J_1 without loss of generality, having been submitted by σn agents, where $\sigma \leqslant \frac{1}{2}$ is a positive rational number. Moreover, profile J_2 is truthfully held by $\sigma n - 1$ agents and the tie-breaking rule ranks J_2 above J_1, while an agent i who truthfully submits J_3 prefers J_2 to J_1. Subsequently, after agent i performs her improvement step, J_2 becomes the collective outcome and the process terminates. For n sufficiently large, we have that $\frac{\text{PoT}(F^{p\ell},\boldsymbol{J})}{\text{DPoA}(F^{p\ell},\boldsymbol{J})} = \frac{\sigma nx_{12}+(\sigma n-1)m+(n-2\sigma n+1)x_{23}}{\sigma nm+(\sigma n-1)x_{12}+(n-2\sigma n+1)x_{13}} > 1$ if n is odd and $\frac{\text{PoT}(F^{p\ell},\boldsymbol{J})}{\text{DPoA}(F^{p\ell},\boldsymbol{J})} = 1$ if n is even and $\sigma = \frac{1}{2}$. □

Proposition 15. *There exist an agenda Φ with $|\mathcal{J}(\Phi)| > 3$ and an arbitrarily large group of agents N with Hamming-distance preferences such that DPoA$(F^{p\ell}, \boldsymbol{J}) >$ PoT$(F^{p\ell}, \boldsymbol{J})$, for some $\boldsymbol{J} \in \mathcal{J}(\Phi)$.*

Proof. Consider an agenda Φ with $|\mathcal{J}(\Phi)| = 4$ and the profile $\boldsymbol{J}$, where all the judgment sets $J_1 = 0000, J_2 = 0001, J_3 = 1110$ and $J_4 = 1100$ in $\mathcal{J}(\Phi)$ are submitted by an equal number of agents. Suppose that the tie-breaking rule selects J_1, which is also the socially optimal outcome. This means that PoT$(F^{p\ell}, \boldsymbol{J}) = 1$. However, assume that an iteration takes place, where at first an agent switches from J_4 to J_3 (which she prefers with regard to the current collective decision J_1). Then, an agent who truthfully holds J_1 moves to J_2, which is more desirable for her than J_3. Afterwards, another agent switches from J_4 to J_3, and so on, until the process terminates with J_2 winning. It is straightforward to measure that DPoA$(F^{p\ell}, \boldsymbol{J}) \geqslant \frac{5}{4} >$ PoT$(F^{p\ell}, \boldsymbol{J})$. □

5 Conclusion

We have developed a general yet simple model of iterative judgment aggregation and we have studied the effects of repeated strategic behaviour on two widely used aggregation rules: the premise-based rule and the plurality rule. We have seen that the former rule converges under any kind of reasonable assumptions and its iteration always achieves the optimal outcome for the group as a whole. On the other hand, our findings regarding the plurality rule are more complex: its convergence depends on various parameters such as the information of the agents and the initially submitted judgments of the group. Moreover, iteration may damage the social outcome when the available judgments of the agents are more than three. Our analysis has brought to light several similarities and differences between the framework of voting and the richer one of judgment aggregation.

While the beginning has been made, further research is required to determine the patterns that govern strategic behaviour in iterative judgment aggregation, by examining a greater number of aggregation rules. At the same time, it would be worthwhile to consider additional types of agents, such as agents with memory, who keep track of the behaviour of their peers, and to investigate the reachability of equilibria under that kind of sophistication.

References

Baumeister, D.; Erdélyi, G.; Erdélyi, O. J.; and Rothe, J. 2015. Complexity of manipulation and bribery in judgment

aggregation for uniform premise-based quota rules. *Mathematical Social Sciences* 76:19–30.

Botan, S.; Novaro, A.; and Endriss, U. 2016. Group manipulation in judgment aggregation. In *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 411–419.

Brânzei, S.; Caragiannis, I.; Morgenstern, J.; and Procaccia, A. D. 2013. How Bad is Selfish Voting? In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI)*, 138–144.

Chapman, B. 2002. Rational aggregation. *Politics, Philosophy and Economics* 1(3):337–354.

Conitzer, V.; Walsh, T.; and Xia, L. 2011. Dominating manipulations in voting with partial information. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI)*, 638–643.

Dietrich, F., and List, C. 2007. Strategy-proof judgment aggregation. *Economics and Philosophy* 23(03):269–300.

Dietrich, F., and Mongin, P. 2010. The premiss-based approach to judgment aggregation. *Journal of Economic Theory* 145(2):562–582.

Dokow, E., and Holzman, R. 2009. Aggregation of Binary Evaluations for Truth-functional Agendas. *Social Choice and Welfare* 32(2):221–241.

Endriss, U., and Grandi, U. 2014. Binary aggregation by selection of the most representative voter. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI)*.

Endriss, U.; Obraztsova, S.; Polukarov, M.; and Rosenschein, J. S. 2016. Strategic Voting with Incomplete Information. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, 236–242.

Endriss, U.; Grandi, U.; and Porello, D. 2012. Complexity of judgment aggregation. *Journal of Artificial Intelligence Research* 45:481–514.

Endriss, U. 2016. Judgment aggregation. In Brandt, F.; Conitzer, V.; Endriss, U.; Lang, J.; and Procaccia, A. D., eds., *Handbook of Computational Social Choice*. Cambridge University Press.

Grossi, D., and Pigozzi, G. 2014. *Judgment Aggregation: A Primer*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers.

de Haan, R., and Slavkovik, M. 2017. Complexity results for aggregating judgments using scoring or distance-based procedures. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

de Haan, R. 2017. Complexity results for manipulation, bribery and control of the Kemeny procedure in judgment aggregation. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 1151–1159.

Hartmann, S., and Sprenger, J. 2012. Judgment aggregation and the problem of tracking the truth. *Synthese* 187(1):209–221.

Kornhauser, L. A., and Sager, L. G. 1993. The one and the many: Adjudication in collegial courts. *California Law Review* 81(1):1–59.

Koutsoupias, E., and Papadimitriou, C. 2009. Worst-case equilibria. *Computer Science Review* 3(2):65–69.

Lang, J., and Slavkovik, M. 2014. How hard is it to compute majority-preserving judgment aggregation rules? In *Proceedings of the 21st European Conference on Artificial Intelligence (ECAI)*, 501–506.

List, C., and Pettit, P. 2002. Aggregating sets of judgments: An impossibility result. *Economics and Philosophy* 18(1):89–110.

List, C., and Pettit, P. 2011. *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press.

List, C., and Puppe, C. 2009. Judgment aggregation: A survey. In Anand, P.; Pattanaik, P.; and Puppe, C., eds., *Handbook of Rational and Social Choice*. Oxford University Press.

Meir, R.; Polukarov, M.; Rosenschein, J. S.; and Jennings, N. R. 2010. Convergence to equilibria in plurality voting. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI)*, 823–828.

Obraztsova, S.; Markakis, E.; Polukarov, M.; Rabinovich, Z.; and Jennings, N. R. 2015. On the convergence of iterative voting: How restrictive should restricted dynamics be? In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI)*, 993–999.

Obraztsova, S.; Markakis, E.; and Thompson, D. R. 2013. Plurality voting with truth-biased agents. In *Proceedings of the 6th International Symposium on Algorithmic Game Theory (SAGT)*, 26–37.

Papadimitriou, C. 2001. Algorithms, games, and the Internet. In *Proceedings of the 33rd Annual ACM Symposium on Theory of Computing (STOC)*, 749–753.

Pettit, P. 2001. Deliberative democracy and the discursive dilemma. *Philosophical Issues* 11(1):268–299.

Reijngoud, A., and Endriss, U. 2012. Voter Response to Iterated Poll Information. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 635–644.

Reijngoud, A. 2011. Voter response to iterated poll information. Master's thesis, ILLC, University of Amsterdam.

Slavkovik, M., and Jamroga, W. 2016. Iterative judgment aggregation. In *Proceedings of the 22nd European Conference on Artificial Intelligence (ECAI)*, 1528–1536.

Terzopoulou, Z. 2017. Manipulating the manipulators: Richer models of strategic behavior in judgment aggregation. Master's thesis, ILLC, University of Amsterdam.

Zwicker, W. S. 2016. Introduction to the theory of voting. In Brandt, F.; Conitzer, V.; Endriss, U.; Lang, J.; and Procaccia, A. D., eds., *Handbook of Computational Social Choice*. Cambridge University Press.