First-Order Logic Formalisation of Arrow's Theorem

Umberto Grandi and Ulle Endriss u.grandi@uva.nl ulle.endriss@uva.nl

Institute for Logic, Language and Computation, University of Amsterdam

Arrow's Theorem is a central result in social choice theory. It states that, under certain natural conditions, it is impossible to aggregate the preferences of a finite set of individuals. We formalise this result in the language of first-order logic, reducing it to a statement saying that a set of formulas does not possess a finite model.

In the long run, we hope that this formalisation can serve as the basis for a fully automated proof of Arrow's Theorem and similar results in social choice theory. We prove that this is possible in principle, at least for a fixed number of individuals, and we report on initial experiments with automated reasoning tools.

Arrow's Theorem

Let I be a set of individuals expressing preferences over a set A of alternatives. For every $i \in I$ represent these preferences as a linear order P_i and call $\mathcal{L}(A)$ the set of all linear orders on A.

A social welfare function (SWF) for A and I is a function $w : \mathcal{L}(A)^I \longrightarrow \mathcal{L}(A)$

A SWF w associate to every preference profile $\underline{P} = (P_1, \dots, P_n)$ a "social order" $w(\underline{P})$.

Theorem 1 (Arrow, 1950)**.** *If A and I are finite and non-empty, and $|A| \geqslant 3$, then there is no social welfare function for I and A that satisfies* **UN**, **IIA** *and* **NDIC**.

Lin and Tang (2008) present an inductive proof of the theorem, proving the base case automatically. We generalise one of their lemmas to cover the case of an infinite number of alternatives:

Lemma 1. *If there exists a SWF for $|A| \geqslant 3$ and I that satisfies* **UN**, **IIA** *and* **NDIC** *then there exists a SWF for $|A'| = 3$ and I that satisfies the same properties.*

F. Lin and P. Tang. Computer-Aided Proofs of Arrow's and other Impossibility Theorems. AAAI 2008.

First-Order Logic

First-order logic is a natural language to talk about linear orders and first-order automated theorem provers are more developed than for other systems. At first sight Arrow's conditions contain several universal quantifications over preference profiles: a second-order quantification over linear orders. Following Lin and Tang (2008), we solve this problem introducing a set of situations, to be used as "names" for preference profiles, denoting with $\underline{P}^s$ the preference profile associated to a situation s.

The first-order signature we introduce is thus composed by:

- three unary relations to mark individuals $I(z)$, alternatives $A(x)$ and situations $S(u)$;
- constant symbols a_1, a_2, a_3 for 3 alternatives, i_1 and s_1 for an individual and a situation;
- a relation $p(z, x, y, u)$ to represent the linear order P^u_z of z in situation u;
- a relation $w(x, y, u)$ to represent the social outcome $w(\underline{P}^u)$ for every situation u.

$$\mathcal{L} = \{a_1,\ a_2,\ a_3,\ i_1,\ s_1,\ I^{(1)},\ A^{(1)},\ S^{(1)},\ w^{(3)},\ p^{(4)}\}$$

Arrow's Conditions in First-Order Logic

- Unanimity (**UN**): if aP_ib for every $i \in I$ then $aw(\underline{P})b$;
- Independence of Irrelevant Alternatives (**IIA**): given two preference profiles $\underline{P}$ and $\underline{Q}$, if aP_ib if and only if aQ_ib for every $i \in I$, then $aw(\underline{P})b$ if and only if $aw(\underline{Q})b$;
- Non-dictatorship (**NDIC**): there is no individual i such that for every profile $\underline{P}$ the social order $w(\underline{P}) = P_i$.

$\Rightarrow$

- **UN**: $S(u) \wedge A(x) \wedge A(y) \to [(\forall z\ \ I(z) \to p(z, x, y, u)) \to w(x, y, u)]$
- **IIA**: $S(u_1) \wedge S(u_2) \wedge A(x) \wedge A(y) \to$ $[(\forall z\ \ I(z) \to (p(z, x, y, u_1) \leftrightarrow p(z, x, y, u_2))) \to (w(x, y, u_1) \leftrightarrow w(x, y, u_2))]$
- **NDIC**: $I(z) \to [\exists x,\ y,\ u\ \ A(x) \wedge A(y) \wedge (x \neq y) \wedge S(u) \wedge p(z, x, y, u) \wedge w(y, x, u)]$

LIN$_p$: p is a linear order for every individual in every situation – $I(z) \wedge S(u) \wedge A(x) \wedge A(y) \to (p(z, x, y, u) \vee p(z, y, x, u) \vee x = y)$ – $I(z) \wedge S(u) \wedge A(x) \to \neg p(z, x, x, u)$ – $I(z) \wedge S(u) \wedge A(x_1) \wedge A(x_2) \wedge A(x_3) \wedge p(z, x_1, x_2, u) \wedge$ $\wedge p(z, x_2, x_3, u) \to p(z, x_1, x_3, u)$	**LIN**$_w$: w is a linear order in every situation: – $S(u) \wedge A(x) \wedge A(y) \to (w(x, y, u) \vee w(y, x, u)) \vee x = y$ – $S(u) \wedge A(x) \to \neg w(x, x, u)$ – $S(u) \wedge A(x_1) \wedge A(x_2) \wedge A(x_3) \wedge w(x_1, x_2, u) \wedge w(x_2, x_3, u) \to w(x_1, x_3, u)$
DEF: the arguments of p and w are of the correct type: – $p(z, x, y, u) \to (I(z) \wedge A(x) \wedge A(y) \wedge S(u))$ – $S(u) \wedge A(x_1) \wedge A(x_2) \wedge A(x_3) \wedge w(x_1, x_2, u) \wedge w(x_2, x_3, u) \to w(x_1, x_3, u)$	**MIN**: A and I are non-empty and there are at least 3 alternatives – $A(a_1) \wedge A(a_2) \wedge A(a_3) \wedge I(i_1) \wedge S(s_1)$ – $\neg(a_1 = a_2) \wedge \neg(a_1 = a_3) \wedge \neg(a_2 = a_3)$ **INJ**: two different situations encode different orders – $S(u) \wedge S(v) \wedge (u \neq v) \to \exists z,\ x,\ y\ \ [I(z) \wedge A(x) \wedge A(y) \wedge$ $\wedge p(z, x, y, u) \wedge p(z, y, x, v)]$
PERM: a hidden hypothesis is the condition of universal domain: – $p(z, x, y, u) \to \exists v\, \{S(v) \wedge p(z, y, x, v) \wedge$ $\forall x_1 [p(z, x, x_1, u) \wedge p(z, x_1, y, u) \to p(z, x_1, x, v) \wedge p(z, y, x_1, v)] \wedge$ $\forall x_1 [(p(z, x_1, x, u) \to p(z, x_1, x, v)) \wedge (p(z, y, x_1, u) \to p(z, y, x_1, v))] \wedge$ $\forall x_1 \forall y_1 [(x_1 \neq x) \wedge (y_1 \neq y) \wedge p(z, x_1, y_1, u) \to p(z, x_1, y_1, v)] \wedge$ $\forall z_1,\ x,\ y\ \ [(z_1 \neq z) \wedge I(z_1) \wedge A(x) \wedge A(y) \to (p(z_1, x, y, u) \leftrightarrow p(z_1, x, y, v))]\}$	**PART**: I, A and S form a partition – $A(x) \to (\neg I(x) \wedge \neg S(x))$ – $I(x) \to (\neg A(x) \wedge \neg S(x))$ – $S(x) \to (\neg I(x) \wedge \neg A(x))$ – $A(x) \vee I(x) \vee S(x)$

Table 1: Axioms of T_{SWF}

Axiomatizability Results

To every SWF w for $|A| \geqslant 3$ and I we can associate a model $\mathcal{M}_w$ of T_{SWF} (see Table 1); if the set A is finite this model is unique.

Proposition 1 (Completeness)**.** *$\mathcal{M}$ is a model of T_{SWF} if and only if there exist two non empty sets A and I, with $|A| \geqslant 3$, and a SWF w for A and I such that $\mathcal{M} = \mathcal{M}_w$.*

Define the theory T_{ARROW} by adding Arrow's conditions **UN**, **IIA** and **NDIC** to T_{SWF}. Arrow's Theorem can now be restated as:

Theorem 2. *T_{ARROW} has no finite models.*

Dealing with the Infinite

In order to use automated reasoning techniques we look for a sentence that can be derived formally from our theory and represent Arrow's Theorem. The main difficulty is that:

If I is infinite then there exists a SWF for I and $|A| \geqslant 3$ that satisfies **UN**, **IIA** and **NDIC** (Fishburn, 1970)

$\Downarrow$

T_{ARROW} is consistent.

P. Fishburn, Arrow's Theorem: Concise Proof and Infinite Voters. Journal of Economic Theory, 1970.

Fix the number of individuals

Define the theory T^n_{SWF} adding new constants $i_1, \dots, i_n$ and axioms:

- $i_k \neq i_j$ for every $k \neq j$ and $I(i_1) \wedge \dots \wedge I(i_n)$;
- $I(z) \to (z = i_1) \vee \dots \vee (z = i_n)$.

A completeness result analogous to Proposition 1 holds. Using Lemma 1 we can prove the following:

Proposition 2. *If w is a SWF for A and I with $|A| \geqslant 3$ and $|I| = n$ then $\mathcal{M}_w \models \neg$*(**UN**$\wedge$**IIA**$\wedge$**NDIC**)*. Therefore for every n:*

$$T^n_{\text{SWF}} \vdash \neg(\textbf{UN} \wedge \textbf{IIA} \wedge \textbf{NDIC})$$

Drawback: possibly different proofs for different n.

Kirman-Sondermann

Kirman and Sondermann (1972) proved the following generalisation of Arrow's Theorem:

If a SWF satisfies **UN** and **IIA**, then the collection of "winning coalitions", those subsets $J \subseteq I$ such that if xP_jy for every $j \in J$ then $x\, w(\underline{P})\, y$, is an ultrafilter.

We can translate this statement into a set of first-order formulas proved by T_{ARROW}, and conclude by noting that the condition of non-dictatorship corresponds to requiring the ultrafilter to be free; if the set of individuals is finite this is an unsatisfiable requirement. This finally formalises the argument of Fishburn (1970): if a SWF satisfies **UN**, **IIA** and **NDic**, then the number of individuals must be infinite.

A. Kirman and D. Sondermann. Arrow's Theorem, many agents, and invisible dictators. Journal of Economic Theory, 1970.

Automated Reasoning

We proved that in principle Arrow's Theorem can be proved automatically, despite not in its full generality:

- fixing the number of individuals and proving $T^n_{\text{SWF}} \vdash \neg$(**UN** $\wedge$ **IIA** $\wedge$ **NDIC**), or
- proving the axioms of the Kirman-Sondermann Theorem from T_{ARROW}.

We easily implemented our axiomatisation in Prover 9 (successor of Otter) syntax:

```
%LINp
(I(z) & S(u) & A(x) & A(y)) -> (p(z,x,y,u)|p(z,y,x,u)|x=y).
(I(z) & S(u) & A(x)) -> -p(z,x,x,u).
(I(z) & S(u) & A(x) & A(y) & A(v) & p(z,x,y,u) & p(z,y,v,u) ) -> p(z,x,v,u).
...
%UN
(S(u) & A(x) & A(y)) -> (( all z (I(z) -> p(z,x,y,u))) -> w(x,y,u)).
```

We ran several experiments using both Prover9 and E theorem prover:

- Negative results: even the easiest case of 3 alternatives and 2 individuals exceeds the search space limits;
- Positive results: we obtained a basic proof of a property called *non-imposition* from the unanimity condition on an instantiated domain (without using the axiom of permutation).

A Remark and Future Work

First-order formalisations of Arrow's Theorem already exist: Nipkow and Wiedijk (2008, 2007) used higher-order automated theorem checker to formalise different proofs in the finite case. However, these systems require all axioms of set theory and on the practical side have a very limited level of automation.

T. Nipkow, Social Choice Theory in HOL. JAR, 2008. *F. Wiedijk, Arrow's Impossibility Theorem. Formalized Mathematics, 2007.*

This work can be extended in a number of ways:

- Experiment with other automated provers (E, Vampire..)
- Formalise other (im)possibility results (Sen's Liberal Paradox, Gibbard-Satterthwaite's and Black's Theorem...). Test unknown (im)possibility results automatically on a weaker version of our axioms.

Decision, Games & Logic, 15-17 June 2009, Lausanne.