LEARNING TO REASON ASSISTED BY AUTOMATED REASONING

Wolfgang Windsteiger

Research Institute for Symbolic Computation (RISC)
Johannes Kepler University Linz (JKU)

CICM 2022 — September 22, 2022 — Tbilisi, Georgia (virtual)

JOHANNES KEPLER
UNIVERSITY LINZ

INTRODUCTION

- "… Assisted by Automated Reasoning" $\rightsquigarrow$ We use the Theorema System …
- "Learning to Reason …" $\rightsquigarrow$
  - in teaching a logic course
  - at undergraduate university level
  - for computer science & AI students

A NEW MODERN LOGIC COURSE

- Modern topics in addition to traditional ones . . .
  - Module Propositional Logic + SAT
  - Module Predicate Logic
    + Pragmatics: How to specify problems? How to do real mathematical proofs?How to do real mathematical proofs?
  - Module Satisfiability Modulo Theories (SMT)
- Modern presentation by showing “logic in action” with logic software.
  - Limboole (SAT solver)
  - RISC-AL
  - TheoremaTheorema
  - Z3, Yices, CVC4, Boolector (SMT Solvers)
- Modern grading
  - Minitests, bonus exercises, lab exercises.
  - No final exam.

WHY AUTOMATED THEOREM PROVING IN THE COURSE?

- One of the teaching goals of the course (Module Predicate Logic):
  Students should be able to do (simple) mathematical proofs
  by hand correctly and completely.
- Main didactic hypothesis:
  For doing (correct and complete) proofs it is beneficial to first get acquainted with the rules of formal proving based on the formal language of predicate logic. Then learn how to translate (formal) proof trees into natural language proofs in mathematical style.
- Method:
  Use software (Theorema) as tutoring system for students
  on a voluntary basis in the frame of bonus exercises.

THEOREMA DEMO

THEOREM (DISTINCT MINIMAL HAS NO SMALLEST)

In[16]:= $\forall_{A}$

In[17]:= $\left(\exists_{a,b\in A} (a \neq b \wedge \text{minimal}[a, A] \wedge \text{minimal}[b, A])\right) \Rightarrow \neg \exists_{s\in A} \text{smallest}[s, A]$ (1)

The predicates used in the theorem are defined as follows:

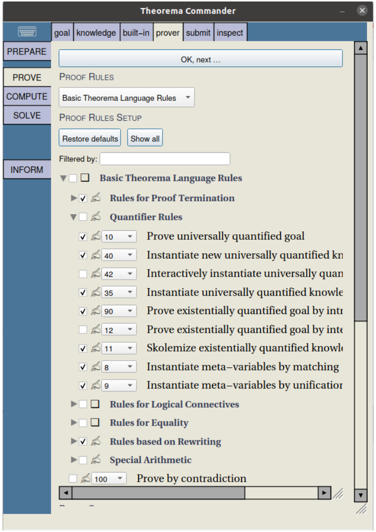
DEFINITION (MIN/SMALLEST)

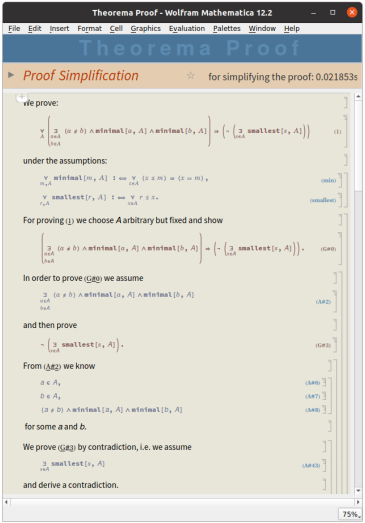
In[18]:= $\forall_{m,r,A}$

In[19]:= $\text{minimal}[m, A] :\Leftrightarrow \forall_{x\in A} (x \leq m \Rightarrow x == m)$ (min)

In[20]:= $\text{smallest}[r, A] :\Leftrightarrow \forall_{x\in A} r \leq x$ (smallest)

THEOREMA DEMO





NESTED STRUCTURE 1$^{\text{ST}}$-ORDER PREDICATE LOGIC B

	Week 0	Week 1	Week 2	Week 3	Week 4	Week 5
FOB1	FOB1L FOB1E* FOB1B*	FOB1E†	FOB1B† FOB1Q			L
FOB2		FOB2L FOB2E* FOB2B*	FOB2E†	FOB2B† FOB2Q		A
FOB3			FOB3L FOB3E* FOB3B*	FOB3E†	FOB3B† FOB3Q	B

FOB*n*B (bonus exercises, voluntary): students submit automated proofs for problems of exercise FOB*n*E, which they already did (or have to do) by hand.

CONTENT OF UNITS IN MODULE FOB

	FOBnE/FOBnQ	FOBnB
FOB1	pattern-based proof search procedure with hypothetical inference rules, first-order proofs without quantifiers	first-order proofs without quantifiers from FOB1E
FOB2	first-order proofs with quantifiers	first-order proofs with quantifiers from FOB2E
FOB3	first-order proofs with quantifiers and informal natural language presentation referring to concrete mathematical concepts introduced by definitions; induction proofs	concrete mathematical proofs from FOB3E

EVALUATION OF USING SOFTWARE IN THE COURSE

Two-fold evaluation:

1. Personal impression of students
   - ☐ Filling out a questionnaire is required for bonus-submission
   - ☐ Category A: Theorema-proof successful $\rightsquigarrow$ groups A.1–A.9
   - ☐ Category B: Theorema-proof failed $\rightsquigarrow$ groups B.10–B.16
2. Performance in the quizzes
   - ☐ Influence of doing the bonus or not doing it
   - ☐ Correlations to groups A.1–B.16

QUESTIONNAIRE: SUCCESSFUL PROOF (CATEGORY A)

A.1 I did not try or was not able to do the examples by hand, but now I think would be able to do them.

A.2 I did not try or was not able to do the examples by hand. I think I would still not be able to do such proofs.

A.3 I had no problems doing the proofs by hand. However, they are different from the Theorema proofs and I'm confused now whether my proofs are wrong.

A.4 I had no problems doing the proofs by hand. However, they are slightly different from the Theorema proofs because Theorema uses certain rules that I did not know. Still, I think my proofs are fine.

A.5 I had no problems doing the proofs by hand. However, they are slightly different from the Theorema proofs and in the future I would do my proofs differently.

A.6 I had no problems doing the proofs by hand. After doing the proofs with Theorema I realized that at least one of my original proofs was wrong.

A.7 I had a hard time doing the proofs by hand. However, I think when doing the next proof by hand, it will be equally difficult, doing the proof with Theorema did not help me for improving my own skills.

A.8 I had a hard time doing the proofs by hand. After doing the proof with Theorema I understand much better how all of this works. I feel that my own skills improved by using Theorema.

A.9 I don't see any connection between the examples from the exercises and the Bonus Exercise with Theorema

QUESTIONNAIRE: PROOF FAILURE (CATEGORY B)

B.10 I did not try or was not able to do these examples by hand. I wanted to see how Theorema does the proofs, but I failed to produce a compete proof.

B.11 I did not try or was not able to do these examples by hand. Theorema is much too complicated for me to use it for such exercises.

B.12 I had no problems doing the proofs by hand. Unfortunately, I failed to produce a complete proof with Theorema. It would have been interesting to compare.

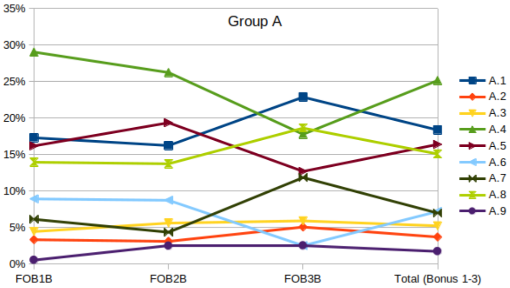
B.13 I had no problems doing the proofs by hand. I'm not interested how an automated proof looks, I have done them by hand anyway.

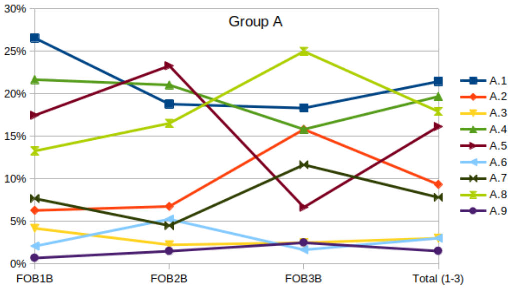
B.14 I had a hard time doing the proofs by hand. Unfortunately, I failed to produce a complete proof with Theorema. It would have been interesting to compare.

B.15 I had a hard time doing the proofs by hand. I'm not interested how an automated proof looks, I have done them by hand anyway.

B.16 I don't see any connection between the examples from the exercises and the Bonus Exercise with Theorema.

SELF-ASSESSMENT: GROUP SIZES W20 VS W21





- Top 4 vs. rest always 3:1
- A.1: not able to do the proofs by hand but feel capable after using Theorema
- A.8: hard time doing the proofs by hand but improvement through using Theorema
- A.5: no problems by hand but will do proofs differently after having used Theorema

PERFORMANCE IN QUIZZES

In each quiz, we record . . .

- Average scores and standard deviations
- p-values of a two-sided Student T-Test testing for equal mean values, i.e., $p \leq 0.05$ says that mean values differ statistically significantly

and compare . . .

All: all students in FOBnQ.

FOBnB: those students in FOBnQ who did bonus exercise FOBnB successfully.

FOB*B: those students in FOBnQ who did FOB1B–FOBnB successfully.

FOB0B: those students who did no bonus exercise successfully.

PERFORMANCE IN QUIZ 1

	$\mu \pm \sigma$	All	FOB0B
All (294)	4.50 ± 0.81	—	—
FOB0B (187)	4.36 ± 0.93	0.0943	—
FOB1B (107)	4.74 ± 0.49	0.0003	5.65×10^{-6}

- Population of groups (in parentheses) high $\rightsquigarrow$ no random numbers!
- Group FOB1B is better than all others.

PERFORMANCE IN QUIZ 2

	$\mu \pm \sigma$	All	FOB0B	FOB2B
All (290)	3.30 ± 1.29	—	—	—
FOB0B (166)	2.99 ± 1.24	0.0102	—	—
FOB2B (109)	3.79 ± 1.21	0.0006	2.41×10^{-7}	—
FOB*B (91)	3.87 ± 1.20	0.0002	8.43×10^{-8}	0.6353

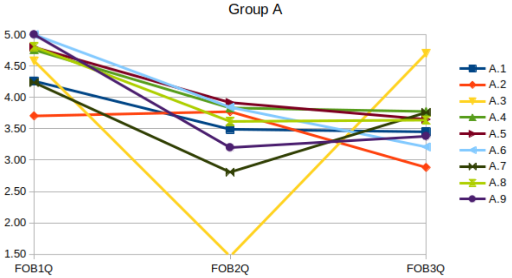
- Those who do bonus are significantly better than others, also than average.
- Those who do no bonus are significantly under average.

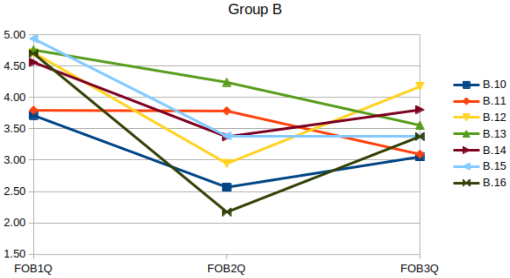
PERFORMANCE IN QUIZ 3

	$\mu \pm \sigma$	All	FOB0B	FOB3B
All (282)	3.46 ± 1.05	—	—	—
FOB0B (147)	3.30 ± 1.04	0.1329	—	—
FOB3B (97)	3.58 ± 1.07	0.3560	0.0474	—
FOB*B (64)	3.68 ± 1.10	0.1529	0.0215	0.5620

- Those who do bonus are significantly better than those who do not.
- Both “better than average” and “worse than average” are not significant.

SELF-ASSESSMENT VS. PERFORMANCE





- Neglect A.3 because it is too small.
- FOB2Q more difficult than FOB1Q: explains decline.
- Strange A.7: Software did not help $\rightsquigarrow$ still significant improvement.
- Strange A.8: Feel improvement $\rightsquigarrow$ performance stays constant.
- B.14 and B.15 from FOB2Q to FOBQ3: Equal in FOB2Q. Those interested in software improve, the others remain.

CONCLUSION

- Classroom experiment using the automated theorem proving software Theorema in the teaching of logic.
- Software is applied to aid the learning process of students.
- Tutoring-by-software correlates with students' performance.
- Students' experiences being tutored by software not always corresponds to performance.
- Correlations are not causalities!
- Theorema can be applied in a reasonable way in education with a big group of first-semester students.

JOHANNES KEPLER
UNIVERSITY LINZ