Performance of Bounded-Rational Agents With the Ability to Self-Modify

Jakub Tětek, Marek Sklenka, Tomáš Gavenčiak

The question

How does self-modification and bounded rationality interact?

Known: Full rationality $\Rightarrow$ self-modification changes nothing (Everitt et al. 2016)

How do different types of bounded rationality compare? How much utility do bounded rational agents lose?

First, we have to define bounded rationality.

Our setting

Utility optimizer: Procedure chooses an action maximizing the expected future exponentially discounted perceived utility with respect to some belief about the world

Agent which:

- Is **not fully rational**
- Has the ability to **self-modify**

We care about how things develop in time.

What is bounded rationality?

Bounded optimizaton: Can take action with expected utility OPT - Ɛ

Misaligned: Utilities off by ≤ Ɛ

Ignorant: Probabilities of events off by ≤ Ɛ

Impatient: Discount factor γ instead of γ* (the correct one)

Our results

How does self-modification and bounded rationality interact?

Bounded-**optimization** agents **deteriorate exponentially** in time (in expectation)

Bounded-**knowledge** agents have **performance constant** in time

How do different types of bounded rationality compare? How much utility do bounded rational agents lose?

We **quantify** the **amount of utility lost**. Both in the **worst** and **average case**.
Our **results are tight**.

Conclusion

The culprit: Exponential discounting

⇒ Important question: Alternatives to exponential discounting

Feel free to contact me: j.tetek@gmail.com