Modeling bounded rationality with markets

Abhimanyu Pallavi Sudhir

25 January 2023

Abstract

We extent Garrabrant induction into a model of bounded rationality.

1 Introduction

Computer scientists and economists have long recognized, at least since calls by Good [1] and Simon [2], that perfect rationality (Bayes-rational expected utility maximization) is computationally impossible. While this is not a problem for classical paradigms of game theory and economics that only seek to prove the existence of equilibrium, more dynamics-centric frameworks, as well as theoretical discussions of artificial general intelligence, create the demand for a more realistic yet useful (for describing "goals") model of agents. Developing such a model is the problem of *bounded rationality* [3].

A brief overview of classical literature on bounded rationality follows:

- Early approaches: explicit modeling of specific heuristics (reviewed in [4]), as-if theories which studied perfect rationality with some information constraints (reviewed and criticized for their lack of predictive usefulness in [5, 6]) and modeling of agents as finite automata (reviewed in [7]).
- Definitions of boundedly rational programs, although without means to construct them, e.g. bounded optimality, which defines optimality given computational constraints [8–10], and machine games, which are games with programs for choices, which have equilibria under weak assumptions, but no framework to effectively compute them [11–13].
- Thermodynamic rationality [14–16], which appeals to the underlying physics of computation to describe bounded rationality.

In this paper, we study an alternative approach to modeling boundedly rational behaviour: *markets*. The basic approach was originally introduced in [17], an algorithm we will hereby call Garrabrant induction, which defines the probabilities of logical sentences as their prices in a prediction market.

There are several reasons it is attractive to think of markets as an appropriate framework for modeling bounded rationality. Like agents, markets hold

beliefs about their model of the world and make decisions – unlike perfectly rational agents, markets may hold incomplete beliefs or even inconsistent beliefs (in the form of undiscovered arbitrage), and are only optimal "conditional on" computational constraints (indeed, the reason that as-if theories [18–20] fail is they model agents as limited only by the scarcity of statistical, rather than algorithmic, information). Markets are fundamentally ensembles of algorithms, and naturally capture the notion of "integrating all available algorithmic information".

In this paper, we will [propose some extensions of the Garrabrant framework /// generalize the Garrabrant framework to a full theory of BR /// ??]

Notation. The sets $\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}$ mean what they always do, with $0 \notin \mathbb{N}$. $f : A \to_\circ B$ is a function with finite support, we use $f : (a : A) \to B$ as an alternative to lambda notation (and we will not bother to distinguish functions and dependent types), and the set of functions $A \to B$ may also be denoted as B^A. Types may be left implicit as __ when it is understood from context what sort of object is being referred to.

2 Logical induction

First an informal sketch of the framework from [17]. The basic construction is a prediction market of logical sentences – a market of "stocks" in sentences that pay out \$1 when proven by some unquestionable authority (a theorem enumerator). For any given trader, the price it neither buys nor sells at can be interpreted as its subjective probability assignment; analogously, the price at which the aggregate of all traders places no orders (i.e. the equilibrium price) can be interpreted as the subjective probability assigned by the consensus of all traders, i.e. taking into account all of their algorithmic information.

The key in integrating information from different algorithms is in determining which algorithms' to take into account. One may argue that the traders are "incentivized" to trade well, but these are in general just programs, with no understanding of incentive. At the end of the day, one needs some notion of "learning", and for markets this is achieved with budgeting: traders can only trade their own money, and those who make more money can trade more, while those who go bankrupt cannot trade.

The remaining part of this section details the precise construction.

Constants. $\mathcal{S}$ is a set of "sentences" in some language; $\mathcal{T} : (t : \mathbb{N}) \to (\mathcal{S} \to_\circ \{0, 1\})$ is an enumerator of "theorems"; $\mathcal{P} = \mathcal{S} \to ([0, 1] \cap \mathbb{Q})$ is the "type of prices" (in particular $\mathcal{T}(t) \in \mathcal{P}$); $\mathcal{Q} = \mathcal{S} \to_\circ \mathbb{Q}$ is the "type of quantities"; $\mathcal{Q}' = \mathcal{Q} \times \mathbb{Q}$ is the type of quantities plus a cash term. "Addition" $(+)$ between quantities and cash amounts, and the dot product $(\cdot) : \mathcal{P} \times \mathcal{Q} \to \mathbb{Q}$ are defined in the obvious way, while $(\cdot) : \mathcal{P} \times \mathcal{Q}' \to \mathbb{Q}$ is defined as $P \cdot (Q + c) = P \cdot Q + c$.

Definition 2.1. A trader is a computable function $\alpha : (t : \mathbb{N}) \to \mathcal{P}^t \to \mathcal{Q}$; we denote its type as $\mathbb{A}$.

(Interpretation: on each day $t = 1, 2, \dots$, it looks at the historical prices $P(1), P(2) \dots P(t-1)$ and the prevailing price $P(t)$ of all traded sentences, and outputs the quantities of some sentences to buy and sell at the prevailing price.)

Definition 2.2. Given a trader α, a price history $(P(1), \dots)$ and an external income stream $(B(1), \dots)$, the budgeted version $[\alpha]_B$ is given by:

$$[\alpha]_B(t) = \begin{cases} 0 \cdot \alpha(t) & \text{if } \mathcal{T}(t) \cdot [\alpha]_{B,H}(t-1) \leq 0 \\ \frac{\mathcal{T}(t) \cdot [\alpha]_{B,H}(t-1)}{-\mathcal{T}(t) \cdot [\alpha]_E(t)} \cdot \alpha(t) & \text{if } \mathcal{T}(t) \cdot [\alpha]_{B,H}(t-1) > 0 \text{ but} \\ & \quad \mathcal{T}(t) \cdot ([\alpha]_{B,H}(t-1) + [\alpha]_E(t)) \leq 0 \\ \alpha(t) & \text{else} \end{cases}$$

Where

- $[\alpha]_E(t) = B(t) + \alpha(t) - P(t) \cdot \alpha(t)$ is the change in assets of the unbudgeted trader α on day t; $[\alpha]_{B,E}(t) = [\alpha]_B(t) - P(t) \cdot [\alpha]_B(t)$ for the budgeted trader.
- $[\alpha]_{B,H}(t) = \sum_{s=0}^{t} [\alpha]_{B,E}(t)$ are the holdings of the budgeted trader on day t.
- Dotting with $\mathcal{T}(t)$ gives the worst-case possible cash value of the trader's assets – i.e. the amount that the trader is "good for".

(Interpretation: Each trader can hold negative assets – i.e. debt – but only as much as it is "good for", i.e. the amount that it can repay even in the worst case where all its unproven beliefs are proven false. If a trade would push it into debt it is not good for, it scales down the trade; if it is already in debt it is not good for, it is considered bankrupt and cannot trade.)

Definition 2.3. An agent-producer is a computable function $\mathbf{m} : (t : \mathbb{N}) \to (\mathbb{A} \to_\circ \mathbb{Q})$ ($\mathbf{m}(t, \alpha)$ is the external income of agent α at time t). Given a price history $(P(1), \dots)$, we may define the associated "aggregate trader":

$$\mu(t) = \sum_{\alpha \in \mathbb{A}} [\alpha]_B(t)$$

(Interpretation: we want to allow possibly infinite sets of traders, but only if they are finite at any given time, and if they can be generated algorithmically.)

Now we are interested in an algorithm to compute these prices $P(t)$ from the agents' orders for the day: a market-maker. Ideally each price will be the equilibrium price – this is troublesome for various reasons (an equilibrium may not exist; if it does exist, approximating it may be computationally expensive – indeed, the entire point of markets in the real world is to have this computation done in a distributed fashion), but for now let us just abstract away the function of the market maker.

Definition 2.4. A market-maker is a computable function $\mathfrak{m} : (t : \mathbb{N}) \to (\alpha : \_\_) \to \mathcal{P}$.

We will assume there exists a $\mathfrak{m}$ such that $\alpha(t, \mathfrak{m}(t, \alpha))$ is arbitrarily small on all sentences – [17] shows this is possible for any continuous trader α, although this $\mathfrak{m}$ is very slow as an algorithm (it simply brute force searches through all possible rational prices until it finds one good enough).

Definition 2.5.

Algorithm 1 Garrabrant inductor

Require:

Definition 2.6.

There are two questions to consider:

- What should be the price of a sentence like "The price of this asset on Day 5 is less than 0.5"? If such sentences exist in our language, then no equilibrium will exist.
- The equilibrium price, if it exists, may not be computable (or even in $\mathbb{Q}$).

The actual approach taken in [17] is to restrict trading to continuous traders, so nothing like "buy 100 if price under 0.5, else sell 100". This seems quite odd: it is as if we are refusing to aggregate information from non-continuous traders. Indeed, the resulting prices could be exploited by a non-continuous trader, if it were allowed to trade on it.

But this restriction can be seen in the more general light of price discrimination and market incompleteness: we may consider the

References

[1] Irving John Good. *Probability and the Weighing of Evidence*. Charles Griffin, 1950. URL: https://books.google.co.uk/books?id=k9g0AAAAMAAJ.

[2] Herbert Simon. "Models of man: social and rational". In: New York: John Wiley and Sons, 1957.

[3] Gerd Gigerenzer. "Towards a Rational Theory of Heuristics". In: *Minds, Models and Milieux*. Palgrave Macmillan UK, 2016, pp. 34–59. DOI: 10.1057/9781137442505_3. URL: https://doi.org/10.1057/9781137442505_3.

[4] Gerd Gigerenzer and Reinhard Selten, eds. *Bounded rationality: an adaptive toolbox*. Dahlem Workshop Reports. London, England: MIT Press, July 2002.

[5] Kenneth J. Arrow. "Is bounded rationality unboundedly rational? Some ruminations." In: Models of a man: Essays in memory of Herbert A. Simon. Cambridge, MA, US: MIT Press, 2004, pp. 47–55. ISBN: 0-262-01208-1.

[6] Daniel Friedman et al. *Risky curves*. London, England: Routledge, May 2017.

[7] Robert J. Aumann. "Rationality and Bounded Rationality". In: *Games and Economic Behavior* 21.1-2 (Oct. 1997), pp. 2–14. DOI: 10.1006/game.1997.0585. URL: https://doi.org/10.1006/game.1997.0585.

[8] Richard L. Lewis, Andrew Howes, and Satinder Singh. "Computational Rationality: Linking Mechanism and Behavior Through Bounded Utility Maximization". In: *Topics in Cognitive Science* 6.2 (2014), pp. 279–311. DOI: 10.1111/tops.12086.

[9] Stuart J. Russell and Devika Subramanian. "Provably Bounded-Optimal Agents". In: *CoRR* cs.AI/9505103 (1995). URL: https://arxiv.org/abs/cs/9505103.

[10] Shlomo Zilberstein. "Metareasoning and Bounded Rationality". In: *Metareasoning*. The MIT Press, Mar. 2011, pp. 27–40. DOI: 10.7551/mitpress/9780262014809.003.0003. URL: https://doi.org/10.7551/mitpress/9780262014809.003.0003.

[11] Joseph Y. Halpern and Rafael Pass. "Algorithmic Rationality: Game Theory with Costly Computation". In: *CoRR* abs/1412.2993 (2014). arXiv: 1412.2993. URL: http://arxiv.org/abs/1412.2993.

[12] Joseph Y. Halpern and Rafael Pass. "I Don't Want to Think About it Now:Decision Theory With Costly Computation". In: *CoRR* abs/1106.2657 (2011). arXiv: 1106.2657. URL: http://arxiv.org/abs/1106.2657.

[13] Moshe Tennenholtz. "Program equilibrium". In: *Games and Economic Behavior* 49.2 (Nov. 2004), pp. 363–373. DOI: 10.1016/j.geb.2004.02.002. URL: https://doi.org/10.1016/j.geb.2004.02.002.

[14] Pedro A. Ortega et al. *Information-Theoretic Bounded Rationality*. 2015. DOI: 10.48550/ARXIV.1512.06789. URL: https://arxiv.org/abs/1512.06789.

[15] Pedro A Ortega and Daniel A Braun. "Thermodynamics as a theory of decision-making with information-processing costs". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 469.2153 (2013), p. 20120683. URL: https://arxiv.org/abs/1204.6481.

[16] Pedro A. Ortega and Daniel A. Braun. "Information, Utility & Bounded Rationality". In: *CoRR* abs/1107.5766 (2011). arXiv: 1107.5766. URL: http://arxiv.org/abs/1107.5766.

[17] Scott Garrabrant et al. "Logical Induction". In: *CoRR* abs/1609.03543 (2016). arXiv: 1609.03543. URL: http://arxiv.org/abs/1609.03543.

[18] Xavier Gabaix and David Laibson. *Myopia and Discounting*. Tech. rep. Mar. 2017. DOI: 10.3386/w23254. URL: https://doi.org/10.3386/w23254.

[19] Xavier Gabaix. *Behavioral Inattention*. Working Paper 24096. National Bureau of Economic Research, Dec. 2017. DOI: 10.3386/w24096. URL: http://www.nber.org/papers/w24096.

[20] Christopher A. Sims. "Implications of rational inattention". In: *Journal of Monetary Economics* 50.3 (2003), pp. 665–690. ISSN: 0304-3932. DOI: https://doi.org/10.1016/S0304-3932(03)00029-1. URL: https://www.sciencedirect.com/science/article/pii/S0304393203000291.