Newcomblike Problems and Optimal Agents

Manuel Moertelmaier

Operalgo
Wels, Austria
manuel.moertelmaier@gmail.com

Abstract. This article discusses the family of Newcomblike problems in the context of reinforcement learning. This reframes the problem of rational decision making as one of obtaining maximal rewards in a wide range of environments. Newcomblike problems are characterized by correlations between agent and environment policies. Such correlations are likely if the environment contains other agents with similar architectures, which is a realistic assumption in practice. An optimal policy, taking into account these correlations, is given for known environments.

Keywords: Probability theory, Newcomb's problem, multi-agent, decision making, Nash equilibrium, planning, game theory, normative, universal semi-measure.

1 Newcomblike Problems

Newcomblike problems are a family of thought experiments designed to bring two intuitively appealing notions of rational decision making into mutual conflict:

> The *maximization of expected payoff* [1]: Condition the expected payoff on the set of possible actions, and choose the action that maximizes the expected payoff.

> The *principle of (strong) dominance* [2]: If action a will lead to a higher payoff than action b for all possible environments, then choose action a over action b.

The namesake, "Newcomb's Problem", invented by William Newcomb, and presented in 1969 by Robert Nozick [3], illustrates this conflict:

> Suppose a being in whose power to predict your choices you have enormous confidence. (One might tell a science-fiction story about a being from another planet, with an advanced technology and science, who you know to be friendly, etc.) [...] All this leads you to believe that almost certainly this being's prediction about your choice in the situation to be discussed will be correct.

> There are two boxes, (B1) and (B2). (B1) contains $ 1.000. (B2) contains either $ 1.000.000 ($ M), or nothing. What the content of (B2) depends upon will be described in a moment. ... You have a choice between two actions: (1) taking what is in both boxes (2) taking only what is in the second box. Furthermore, and you know this, the being knows that you know this, and so on: (I) If the being predicts you will take what is in both boxes, he does not put the $ M in the second box. (II) If the being predicts you will take only what is in the second box, he does put the $ M in the second box. ... The situation is as follows. First the being makes its prediction. Then it puts the $ M in the second box, or does not, depending upon what it has predicted. Then you make your choice. What do you do? ([3], reproduced from [4].)

The principle of dominance recommends taking both boxes, as this will provide an extra $ 1.000, no matter whether B1 is filled or not. However, the expected payoff of this strategy is just $ 1.000, as opposed to $ 1.000.000 for taking only B2, which is in conflict with the principle of payoff maximization.

A number of authors have defended both choices on grounds of various definitions of what constitutes rational behavior. Others have criticized the overall setting for its implausibility. Over the years, related thought experiments of different degrees of realism have been proposed, such as Kavka's toxin puzzle, the smoker's lesion, the psychopath button or Parfit's hitchhiker. (For a review see [4]).

2 Are Newcomblike Problems Relevant?

It is important to note that in the above scenario, taking only one box still results in higher returns even if the predictor is only slightly better than chance. More importantly, the (non-iterated) prisoner's dilemma can be identified as a Newcomblike problem, if strong correlation between the actions of both players is assumed [5]: While defecting dominates cooperation, it may increase the chance that the opponent, following a similar logic, also defects. Such correlation is not present in controlled laboratory experiments [6], but may be more pronounced in real-life settings, where global factors can nudge opponents towards a common preference. In general, whenever two or more agents with similar internal structure are presented with similar perceptions, one can expect correlation between their actions. For strategies that ignore such correlations, the optimal performance is reached in the Nash equilibrium [7]. However, as the prisoner's dilemma demonstrates, the Nash equilibrium is not necessarily Pareto optimal. Exploiting correlations in decision making can therefore lead to increased expected agent performance. Depending on the reward structure of the setting, the increase in performance can be substantial even for relatively small correlations, as can be seen in the original Newcomb's Problem. While Newcomblike problems are often discussed in rather outlandish thought experiments, the underlying principles are likely to manifest in realistic settings. For example, correlation between agent's behavior is expected to be pronounced in software multi-agent settings, where agents are based on a common software architecture. While rational decision making

is fundamental to the design of artificial intelligent agents, Newcomblike problems have so far received little attention from the AI community.

3 Optimality and Rationality

After 40 years, decision theorists remain divided over possible solutions to Newcomb's problem. The field can be roughly carved up into adherents of one or the other of the above principles of rational decision making. This paper proposes to reframe the problem of rationality in the terms of reinforcement learning [8]. Here, an agent takes actions on the environment, and receives perceptions and rewards as feedback. Rational behavior, in this context, is defined as maximizing the overall expected reward, in similar spirit as the maximization of expected payoff. However, reinforcement learning puts greater emphasis on an agent's performance across a wide range of situations, whereas the discussion of Newcomblike problems in the existing literature is generally focused on individual (counter)examples, supposedly proving or disproving the superiority of particular approaches.

4 Formal Description of Agent-Environment Interaction

Agent **a** and environment **e** interact through their respective outputs y (actions) and x (perceptions). Without much loss of generality, it is assumed that the interaction is taking place in discrete time-steps, within a finite time-horizon m, with x_t and y_t, t=1,2,3...m denoting environment and agent output at time t. The non-negative rewards r_t at time t are either directly provided by the environment, as part of the agents perceptions, or are computed internally by the agent. In any case, r_t, and its sum over time, are assumed to be deterministic functions of the interaction history $y_1x_1,y_2x_2,\ldots,y_tx_t =: yx_{1..t}$ such that $r_1+r_2+\ldots+r_t =: \mathbf{r}(yx_{1..t})$. Actions and perceptions at time t are assumed to be the result of agent and environment policies **p** and **q** dependent on the interaction history $y_t=\mathbf{p}(yx_{1..t-1})$, $x_t=\mathbf{q}(yx_{1...t-1},y_t)$.

What distinguishes Newcomblike problems from other decision-theoretic problems is that **p** and **q** are not distributed independently. This means that the choice of the agent policy **p** may result in a different environment **q**. This in turn results in different expected responses x_t to actions y_t even for identical interaction histories. As mentioned above, some factors contributing to such a correlation are, e.g. correlation between agent architectures, correlations in agent's interaction histories with their respective environment, or situations where the environment contains a model of the agent.

Resulting from this correlation is a discrepancy between the expected *overall* reward $r_1+r_2+\ldots+r_m$ of an agent policy, and its expected *future* reward at time t $r_t+r_{t+1}+\ldots+r_m$. A policy that optimizes the first can be called *planning-optimal*, one that optimizes the second, *action-optimal* [9]. An action-optimal policy is not necessarily planning-optimal, as it may be correlated with unfavorable environments. Colloquially, an action-optimal policy makes the best of each situation, but may tend to end up in unfavorable situations. A planning-optimal policy is one that both acts efficiently, and is correlated with favorable environments. If there is no correlation between **p** and **q**, action-optimality and planning-optimality coincide. This paper uses the total accumu-

lated reward of an agent as a measure of rationality, and therefore focuses on planning-optimal agent policies.
For a representation of agent and environment that allows for potential correlations, a joint probability distribution for interaction histories $\boldsymbol{\mu}(yx_{1\ldots m})$ is used, as proposed in [10]. An agent policy $\mathbf{p}$ corresponds to a true subset $\mathbf{Z^p}$ of all possible interaction histories $\{yx_{1\ldots m}\}$ such that for all t, $y_t=\mathbf{p}(yx_{1..t-1})$. The expected reward V from following p in the context of μ is

$$V_\mu^p := \frac{\sum_{yx_{1..m}\in Z^p} \boldsymbol{r}(yx_{1..m})\mu(yx_{1..m})}{\sum_{yx_{1..m}\in Z^p} \mu(yx_{1..m})} (1)$$

For a known environment μ, the planning optimal policy p* is therefore

$$p_\mu^* := \arg\max_p \frac{\sum_{yx_{1..m}\in Z^p} \boldsymbol{r}(yx_{1..m})\mu(yx_{1..m})}{\sum_{yx_{1..m}\in Z^p} \mu(yx_{1..m})} (2)$$

If the true environment is unknown, a possible strategy is to assume that it is contained in a class $\mathcal{M}$ of potential environments. In order to maximize generality, as in [11], $\mathcal{M}$ can be chosen to be the set of all enumerable semimeasures ν. Individual ν are assigned weights w_ν based on their Kolmogorov complexity K(ν)

$$w_\nu \sim 2^{-K(\nu)} (3)$$

These weights can be interpreted as a priori beliefs that ν is the true environment. The expected reward of a policy with respect to the resulting mixture ξ [12] over environments

$$V_\xi^p = \sum_{\nu\in M} \frac{\sum_{yx_{1..m}\in Z^p} \boldsymbol{r}(yx_{1..m})\nu(yx_{1..m})}{\sum_{yx_{1..m}\in Z^p} \nu(yx_{1..m})} w_\nu (4)$$

is, however, not enumerable. While it is tempting to attempt to define the planning-optimal policy for ξ as:

$$p_\xi^* := \arg\max_p V_\xi^p (5)$$

this policy is itself not enumerable, which contradicts the original assumption that interaction histories are drawn from enumerable semimeasures [10].

Existing publications on Newcomblike problems often attempt to give precise descriptions of the overall setting, so p_μ^* can serve as a normative model of rational behavior. For unknown environments, Monte-Carlo based optimization ap-

proaches similar in spirit to [13] may yield useful approximations to optimal agents. Finally, for known environments, the performance of the planning-optimal policy p_μ^* can be compared to the action-optimal policy AIμ described in [11]. While p_μ^* takes correlations between agent and environment policies into account, AIμ doesn't. The relative difference, in terms of reward achieved, between the two models can serve as a quantitative measure of the "Newcomblikeness" of a given problem.

5 Summary

The reinforcement learning framework allows to define Newcomblike problems as settings where agent and environment policies are correlated. This correlation results in action-optimal policies (maximizing future reward) not automatically being planning-optimal (maximizing total reward). A planning-optimal policy can be given, if the probability of interaction histories is known. However, the universally action-optimal agent for unknown environments AIXI defined in [10] cannot be straightforwardly extended to Newcomblike problems.

Discussing Newcomblike problems in the reinforcement learning framework can benefit the fields of philosophy and economics, as this allows to move away from appeals to (counter)examples and intuitive notions of rational behavior towards summarization over environments and the quantitative criterion of optimality.

This discussion can also benefit the field of AI, as Newcomblike problems are real and relevant wherever agent behavior is correlated in a multi-agent setting. Such correlations can arise from, e.g, common architectures, common input/output histories, or mutual modeling. Planning-optimal agents that take such correlations into account can systematically outperform action-optimal agents.

References

1. Jeffrey, R. C. The logic of decision, McGraw-Hill Book Company (1965)
2. Neumann, J., Morgenstern, O. Theory of Games and Economic Behavior, Princeton University Press (1944)
3. Nozick, R. Newcomb's Problem and Two principles of Choice. In: Essays in Honor of Carl G. Hempel, eds. N. Rescher (1969)
4. Ledwig, M. Newcomb's Problem. University of Konstanz, Konstanz (2000)
5. Lewis, D. Prisoner's Dilemma Is a Newcomb Problem. Philosophy and Public Affairs vol. 8, pp. 235-240 (1979)
6. Tversky, A., Shafir, E. Thinking through uncertainty: Nonconsequential reasoning and choice. Cognitive Psychology vol. 24, pp. 449-474 (1992)
7. Nash, J. Non-cooperative games. The annals of mathematics vol. 54, pp. 286-295 (1951)
8. Sutton, R. and Barto, A. Reinforcement Learning: An Introduction, MIT Press (1998)
9. Aumann, R., Hart, S., Perry, M. The Absent-Minded Driver. Games and Economic Behavior, vol. 20, pp. 102-116 (1997)
10. Hutter, M. A gentle introduction to the universal algorithmic agent AIXI. Manno-Lugano, Switzerland: IDSIA (2003)

11. Hutter, M. Towards a Universal Theory of Artificial Intelligence based on Algorithmic Probability and Sequential Decisions. Paper presented at the Proceedings of the 12th European Conference on Machine Learning (ECML-2001) (2000)
12. Zvonkin, A., and Levin, L. The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. Russian Mathematical Surveys vol. 25, pp. 83-124 (1970)
13. Veness, J., Ng, K., Hutter, M., Uther, W. and Silver, D. A Monte-Carlo AIXI Approximation, JAIR vol. 40, pp. 95-142 (2011)