Relation to ratifiability

May 7, 2018

Setup

ToDo: I expect all fo this to be re-done once we integrate all of our work into a more unified paper.

For now, I'll only consider games without anthropics and without different situations. The agent submits a probability distribution once, an action is sampled from it and then the environment behaves in some way depending on that action and probability distribution.

Our decision problems have a set of actions A and a set of observations O and a set of (hidden) states that give rise to observations. The decision problem has to be such that no anthropic uncertainty can arise. That is, no matter the policies and actions of the agent, it is impossible to get an observation twice during a single run of the decision problem. (Otherwise, the Q-values become non-trivial to define and the law of large numbers doesn't give us the convergence of the Q values to expected values anymore.)

Let P_i denote the sequence of strategies $O \rightsquigarrow A$. Note that the P_i are random variables. Let $Q_i \in \mathbb{R}^{O\times A}$ be the sequence of empirical EVs a.k.a. Q-values, again random variables. Let $U(a, o, p)$ be the actual EV given an action a upon observation o and strategy p, where U is continuous in p. (Defining U such that it also assigns expected utilities to actions that are assigned zero probability is important for discussing the behavior of U in the limit.) (Unfortunately, this function cannot be as easily defined for problems involving anthropics and so forth.)

We say a sequence of random variables $(X_i)_{i\in\mathbb{N}}$ converges almost surely to x, or $X_i \underset{\text{a.s.}}{\to} x$, if

$$P(\lim_{n\to\infty} X_n = X) = 1 \tag{1}$$

(see `https://en.wikipedia.org/wiki/Convergence_of_random_variables#Almost_sure_convergence`). Conversely, we say that the sequence converges to x with positive probability, or $X_i \underset{\text{w.p.p.}}{\to} x$, if $P(\lim_{i\to\infty} X_n = x) > 0$.

A policy π is a (deterministic) function that maps natural number representing the time and a set of Q-values for each pair of an action and an observation onto strategies. Instead of $\pi(i, q)$, we write $\pi_i(q)$.

Definitions of ratifiability

Let π be a "one-shot policy".

- A strategy p is *weakly ratifiable modulo π-exploration* if for all $o \in O$ there are $b_{o,1}, ..., b_{o,n}$ such that $b_{o,j} = U(a_j, o, p)$ if $a_j \in supp(p(o)) = \{x \in A \mid p(o, x) > 0\}$ and $b_{o,j} \le \min_{a_k\in supp(p(o))} U(a_k, o, p)$ otherwise such that for all o and $a_j \in supp(p(o))$, it is

$$p(a_j, o) = \pi\left((b_{r,i})_{r\in O, i=1,...,n}\right)(a_j). \tag{2}$$

- A strategy p is *weakly ratifiable* if for all $o \in O$, $U(a, o, p)$ is constant over $a \in supp(p(o))$.
- A probability distribution over actions p is *strongly ratifiable modulo π-exploration* if for all $o \in O$ and all $a_j \in A$, it is

$$p(a_j, o) = \pi(U(a_i, r, p)_{a_i \in A, r \in O})(a_j). \tag{3}$$

- A probability distribution over actions p is *strongly ratifiable* if for all $o \in O$, $U(a, o, p)$ is constant over $a \in supp(p(o))$ and lower than that constant for $a \notin supp(p(o))$.

ToDo: explain relationship to existing ratifiability proposals and give references on ratifiability (including ones related to the tickle defense)

Results

Theorem 1. *Let π be a policy s.t. π_i is continuous for all $i \in \mathbb{N}$. For each strategy q and each $j \in \{1, ..., n\}$ let*

$$\pi_i(q) \to \pi_\infty(q). \tag{4}$$

Assume that π_∞ "other things equal always gives higher probabilities to the actions with higher utility". Furthermore, let $P_i \underset{w.p.p.}{\to} \mathbf{p}$ and let $U(a, o, p)$ be continuous in p around $\mathbf{p}$ for each $a \in A$ and $o \in O$. Then $\mathbf{p}$ is weakly ratifiable modulo π-exploration.

Proof. If indeed $P_i \to \mathbf{p}$, then $Q_i(a, o)$ converges almost surely for all $a \in A$ and $o \in O$. For all $a_j \in A$ and $o \in O$, let $Q_i(a_j, o) \to b_{j,o}$. In particular, if $a \in supp(\mathbf{p}(o))$ (or, in fact, more generally if a is taken infinitely many times in response to o almost surely), then

$$Q_i(a, o) \underset{\text{a.s.}}{\to} U(a, o, \mathbf{p}) \tag{5}$$

according to the strong law of large numbers, the fact that $U(a, p)$ is continuous in p around $\mathbf{p}$ and the standard theorem about the limit of composite functions (TODO maybe `http://elib.mi.sanu.ac.rs/files/journals/tm/22/tm1211.pdf` and `http://www.math.uconn.edu/~stein/math115/Slides/math115-130notes.pdf`).

Because π_∞ prefers better actions, it has to be for all o

$$a_j \notin supp(\mathbf{p}(o)) \implies b_j \leq \min_{a_k \in supp(\mathbf{p}(o))} U(a_k, \mathbf{p}(o)). \tag{6}$$

If $Q_i(a_j, o) \to b_{j,o}$, then

$$P_i = \pi_i(Q_i) \to \pi_\infty(b), \tag{7}$$

because π_i and π_∞ are continuous.

From the premises of the theorem, we have now inferred that w.p.p. it is not only $P_i \to \mathbf{p}$ but at the same time also $P_i \to \pi_\infty(b)$. Hence, it must be $\mathbf{p} = \pi_\infty(b)$, where the b_i satisfy the necessary conditions for weak ratifiability modulo π_∞-exploration.

□

ToDo: example / graph

Corollary 2 (Weak Ratifiability)**.** *Same conditions as for theorem 1, but also assume that π_∞ doesn't explore, i.e. that*

$$\pi_\infty(v)(a_j, o) > 0 \iff j \in \arg\max_k v_{k,o}. \tag{8}$$

Then $\mathbf{p}$ is weakly ratifiable.

ToDo: example / graph

Theorem 3. *Same conditions as theorem 1. Assume also that π almost surely explores all actions infinitely many times. Then $\mathbf{p}$ is strongly ratifiable modulo π_∞-exploration.*

Proof. Analogous to the proof of theorem 1. □

ToDo: example / graph

Corollary 4 (Strong Ratifiability)**.** *Same conditions as theorem 3. Assume also that π_∞ doesn't explore. Then $\mathbf{p}$ is strongly ratifiable.*

Proof. Follows directly from theorem 3 and the definitions of strong ratifiability modulo π_∞ and strong ratifiability period. □

ToDo: example / graph

We won't give the details here, but if infinite exploration is not given, then in some problems there is a positive probability that an action will, based on bad luck, be severely underestimated, such that the agent then stops to take that action. Thus, the algorihtm might converge on some $\mathbf{p}$ under which action $a \notin \mathit{supp}(\mathbf{p})$ should be taken.

Notes for generalization:

- ~~Probably, I could easily generalize this to expected values conditional on some observation.~~
- ~~If you explore all options infinitely often almost surely, you almost surely converge to a strongly ratifiable solution. If you don't explore all options infinitely often almost surely, there is a positive probability that it doesn't converge to a strongly ratifiable option.~~
- ~~If it doesn't converge, then there is still ratifiability of some frequency construct, perhaps?~~
- ~~Include anthropic cases. What utilities do the Q-values converge to in the anthropic cases? Define two different kinds of Q-values. Then add a lemma showing that they converge to the respective expected utilities.~~
- ~~Instead of Q-values, one could use "forgetful Q-values" as long these converge toward U at constant probabilities.~~
- ~~Do I need continuity (for U, π_i and π_∞) or just sth like continuity almost everyhwere?~~
- What if π takes other information into account (like Gittins-indices, and UCB) but converges to ignoring that information in the limit?

Ratifiability of frequencies

Even if the probabilities do not converge at all, the frequencies of actions over many turns usually do. In fact, they often converge to ratifiable ones. E.g., in Death in Damascus, even if the probabilities do not converge, the frequencies converge to the ratifiable 50-50. TODO example with graph. TODO refer to Linda's section.

But this does not seem to be true in general, even if the other prerequisites of the theorem are met. Roughly, the reason is the following: the frequencies arising from applying the learning algorithm are based on the success of actions for the success probabilities, rather than that frequency itself. So, the frequencies can converge to 50-50 based on how the actions behave if the probability is far removed from 50-50, even if at a (hypothetical) probability of 50-50, one of the actions is better than the other. Again, TODO example with graph.

Obsolete stuff

Old proof of theorem 1

1. For all $a_j \in \mathit{supp}(\mathbf{p})$, if indeed $P_i \to \mathbf{p}$, then

$$Q_i(a_j) \underset{\text{a.s.}}{\to} U(a_j, \mathbf{p}) \tag{9}$$

because $U(a,p)$ is continuous in p. Hence we define for $a_j \in \mathit{supp}(\mathbf{p})$: $b_j = U(a_j, \mathbf{p})$

2. If $P_i \to \mathbf{p}$, then because π prefers better options in the limit, there must be an N such that for all $i > N$ and all $a \notin \mathit{supp}(\mathbf{p})$ it must be

$$Q_i(a) \leq \min_{a_k \in \mathit{supp}(\mathbf{p})} U(a_k, \mathbf{p}). \tag{10}$$

3. As $P_i \to \mathbf{p}$, $Q_i(a_j)$ almost surely converges to some value even for $a_j \notin \mathit{supp}(\mathbf{p})$. Because of step 2, these values are smaller than $\min_{a_k \in \mathit{supp}(\mathbf{p})} U(a_k, \mathbf{p})$. Hence, we will use these limits as b_j.

4. If $Q_i(a_j) \to b_j$ for all $a_j \in A$, then

$$P_i \to \pi_\infty(b_1, ..., b_n). \tag{11}$$

5. From the conditions of the hypothesis and steps 1–4, it follows that with positive probability, it is both $P_i \to \mathbf{p}$ and for all $a_j \in \mathit{supp}(\mathbf{p})$

$$P_i(a_j) \to \pi_\infty(b_1, ..., b_n)(a_j). \tag{12}$$

Hence it must be for all $a_j \in \mathit{supp}(\mathbf{p})$

$$\mathbf{p}(a_j) = \pi_\infty(b_1, ..., b_n)(a_j), \tag{13}$$

where the b_j satisfy the claims made in the hypothesis.

Old introduction with anthropics

A (potentially Newcomb-like) decision problem consists of

- a set of actions A,
- a set of (deterministic) decision trees T over A with end-points in $\mathbb{R}$,
- a set of observations O,
- a function $f : \mathcal{P}^O \rightsquigarrow T$.

A policy

$$\pi : \mathbb{N} \times \mathbb{R}^A \rightsquigarrow A = \{a_1, ..., a_n\} \tag{14}$$

is a function mapping a time step and a mapping of empirical expected values / Q-values onto probability distributions over actions.

The outcome of an individual run of the decision problem is obtained as follows. For each observation $o \in O$, π submits a probability distribution based on the current Q-values. Based on that probability distribution, the environment (non-deterministically) chooses a decision tree (using f). Then the decision tree is traversed, sampling from the probability distributions given by π for the respective observations.

Note that because the Q-values are only updated once the decision problem is fully run, the agent will submit the same probability distributions in all nodes of the decision tree in which it faces the same observation o.

Definition of the Q-values:

$$Q_{SSA}(a,o) = \sum_{\text{episode } e \text{ with } o \to a} u_e, \tag{15}$$

where u_e is the utility gained in episode e.

$$Q_{SIA}(a,o) = \sum_{\text{episode } e \text{ with } o \to a} \#(a \text{ in } e) \cdot u_e, \tag{16}$$

where $\#(a$ in $e)$ denotes the number of times action a is taken in episode e.

$$EU_{SIA}(o \to a, p) = \sum_{tree \in T} P(tree \mid p) \sum_{node \in tree \text{ with} o} P_{SIA}(node \mid p) \sum_{\text{terminal } t \in tree} P(t \mid a \text{ in } node, p) u(t) \tag{17}$$

$$EU_{SSA}(o \to a, p) = \sum_{tree \in T} P(tree \mid p) \sum_{node \in tree \text{ with} o} P_{SIA}(node \mid p) \sum_{\text{terminal } t \in tree} P(t \mid a \text{ in } node, p) u(t) \tag{18}$$