

Relation to ratifiability

May 4, 2018

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.

Let P_i denote the sequence of probability distributions over A . Note that the P_i are random variables. Let $Q_i \in \mathbb{R}^A$ be the sequence of empirical EVs, again random variables. Let $U : A \times [0, 1]^A \rightarrow \mathbb{R}$ be the actual EV given an action a and probability distribution 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 $X_i \xrightarrow{\text{a.s.}} x$, if

$$P(\lim_{n \rightarrow \infty} X_n = x) = 1 \quad (1)$$

(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 $X_i \xrightarrow{\text{w.p.p.}} x$, if

$$P(\lim_{i \rightarrow \infty} X_n = x) > 0.$$

Definitions:

- A probability distribution over actions p is *weakly ratifiable modulo π -exploration* if there are b_1, \dots, b_n such that $b_j = U(a_j, p)$ if $a_j \in \text{supp}(p) = \{x \in A \mid p(x) > 0\}$ and $b_j \leq \min_{a_k \in \text{supp}(p)} U(a_k, p)$ otherwise such that for all $a_j \in \text{supp}(p)$, it is

$$p(a_j) = \pi(b_1, \dots, b_n)(a_j). \quad (2)$$

- A probability distribution over actions p is *weakly ratifiable* if $U(a, p)$ is constant over $a \in \text{supp}(p)$.
- A probability distribution over actions p is *strongly ratifiable modulo π -exploration* if

$$p(a_j) = \pi(U(a_1, p), \dots, U(a_n, p))(a_j). \quad (3)$$

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

Theorem 1. Let π be a policy s.t. $\pi(i, \cdot)$ is continuous for all $i \in \mathbb{N}$. For each $q \in \mathbb{R}^A$ and each $j \in \{1, \dots, n\}$ let

$$\pi(i, q) \rightarrow \pi^\infty(q). \quad (4)$$

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

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

$$Q_i(a) \xrightarrow{\text{a.s.}} U(a, \mathbf{p}) \quad (5)$$

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

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

If $Q_i(a_j) \rightarrow b_j$, then

$$P_i = \pi(i, Q_i) \rightarrow \pi^\infty(b_1, \dots, b_n), \quad (7)$$

because $\pi(i, \cdot)$ and π^∞ are continuous.

From the premises of the theorem, we have now inferred that w.p.p. it is at the same time $P_i \rightarrow \mathbf{p}$ and $P_i \rightarrow \pi^\infty(b_1, \dots, b_n)$. Hence, $\mathbf{p} = \pi^\infty(b_1, \dots, b_n)$, where the b_i satisfy the necessary conditions for weak ratifiability modulo π^∞ -exploration. \square

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

$$\pi^\infty(v_1, \dots, v_n)(a_j) > 0 \iff j \in \arg \max_k v_k. \quad (8)$$

Then \mathbf{p} is weakly ratifiable.

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. Proof is mostly analogous to the proof of theorem 1, except that one does not have to differentiate between $a \in \text{supp}(\mathbf{p})$ and $a \notin \text{supp}(\mathbf{p})$. \square

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. \square

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 algorithm might converge on some \mathbf{p} under which action $a \notin \text{supp}(\mathbf{p})$ 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 everywhere?
- 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.

Some references on ratifiability

Some versions of the tickle defense are based on ratifiability, including

Old proof of theorem 1

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

$$Q_i(a_j) \xrightarrow{\text{a.s.}} U(a_j, \mathbf{p}) \quad (9)$$

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

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

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

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

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

$$P_i \rightarrow \pi^\infty(b_1, \dots, b_n). \quad (11)$$

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

$$P_i(a_j) \rightarrow \pi^\infty(b_1, \dots, b_n)(a_j). \quad (12)$$

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

$$\mathbf{p}(a_j) = \pi^\infty(b_1, \dots, b_n)(a_j), \quad (13)$$

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

$$\pi : \mathbb{N} \times \mathbb{R}^A \rightsquigarrow A = \{a_1, \dots, a_n\} \quad (14)$$

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:

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

where u_e is the utility gained in episode e .

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

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

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

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