Reinforcement Learning in Repeated Interaction Games: An Extension to Score-Based Reinforcement Processes

Jonathan Bendor, Dilip Mookherjee and Debraj Ray

In Bendor, Mookherjee and Ray [2001] (henceforth BMR), we study reinforcement learning in repeated interaction games. In these notes, we show how the model in BMR can be easily be adapted to the learning version of the probabilistic choice model of Luce (1959), with a suitable redefinition of the state variable. Denoting the actions for a player $i = 1, 2, \ldots, m$, Luce's model specifies the existence of a scaling function $v_i, i = 1, \ldots, m$ representing the "utility" or "score" of different actions at any given date, such that the probability of choosing action i at that date is

$$p_i(v) = \frac{v_i}{\Sigma_j v_j} \tag{1}$$

where v denotes the vector of scores $(v_1, \ldots, v_n)$. The scores form the state variable in this model, whose evolution from one play to the next is given by a linear rule of the form

$$v_i(t+1) = \beta_i v_i(t) + \gamma_i$$

where β_i, γ_i are parameters that depend on the action-outcome pair (a, f, F).[1]

Here the learning rule is defined in terms of the evolution of the *scores* determining choice probabilities, rather than these probabilities themselves. Another example of such a formulation is the Erev and Roth (1995, 1998) version of the Luce model:

$$v_{j,t+1} = \max\{\nu, (1-\phi)v_{j,t} + \gamma_j(f_t - F)\} \tag{2}$$

where $v_{j,t}$ denotes the score assigned to action j at round t, ν is a small nonnegative number, ϕ is a 'forgetting' parameter between 0 and 1, and γ_j is the score reinforcement function depending continuously on the gap between payoff f_t at round t and aspiration F.

To adapt the model in BMR to this setting, reformulate the state variable to consist of the scores themselves. We will soon check that the scores are nonnegative for every action, and that their sum is positive. The state then determines the choice probabilities via Luce's rule (1). The Erev-Roth score reinforcement functions can be calibrated as follows. Introduce a matrix of functions γ_{ji}, describing how action j is reinforced when

[1] This is the so-called Gamma model in Luce (1959, Chapter 4), obtained under axioms of positivity and boundedness of scores, independence of units, and independence from irrelevant alternatives. A special case of this is the Beta model, where γ_i is set equal to zero.

action i happened to be chosen at round t. The reinforcement γ_{ii} of the chosen action itself is set to be a continuous increasing function with $\gamma_{ii}(0) = 0$. Likewise, for $j \neq i$, take γ_{ji} to be a continuous decreasing function with $\gamma_{ji}(0) = 0$. So then (2) reduces to the following: at date t, if action i is chosen with payoff f_t, then

$$v_{j,t+1} = \max\{\nu, (1-\phi)v_{j,t} + \gamma_{ji}(f_t - F)\} \tag{3}$$

Indeed, it is possible to allow the reinforcements γ_{ji} to also depend on the current score $v_{j,t}$ (e.g., the reinforcements could be proportional to the current score), though this does not form part of the Erev-Roth model.

Start the process with scores that are nonnegative and aggregate to a positive number. Then by construction $v_{j,t} \geq 0$ for all j and $\sum_j v_{j,t} > 0$ for all $t \geq 1$ with probability one.[2] So the process is well-defined.

Next, we proceed as follows. Since we want to allow for the possibility that some actions are chosen with zero probability, we set $\nu = 0$. The following properties (taken as assumptions in BMR) can now be checked:

(a) *Compact state space:* This follows from the fact that reinforcements are bounded above by $\max_{i,j} \max\{\gamma_{ii}(\bar{f} - F), -\gamma_{ji}(\underline{f} - F)\}$, where $\bar{f}$ and $\underline{f}$ respectively denote the highest and lowest payoff in the game. So scores lie in some compact interval $[0, V]$ with $V < \infty$.

(b) *Norman's DD property (SDD version):* follows from the construction.

(c) *Positive Reinforcement:* If i is chosen and $f_t \geq F$ then $\gamma_{ii} \geq 0$ and $\gamma_{ji} \leq 0$ (for all $j \neq i$), implying $v_{i,t+1} \geq (1-\phi)v_{i,t}$ and $v_{j,t+1} \leq (1-\phi)v_{j,t}$ for $j \neq i$. Consequently,

$$\frac{v_{i,t+1}}{\sum_j v_{j,t+1}} \geq \frac{v_{i,t+1}}{\sum_{j \neq i} v_{j,t}(1-\phi) + v_{i,t+1}} \geq \frac{v_{i,t}}{\sum_j v_{j,t}}$$

which verifies (PR).

(d) *Negative Reinforcement:* If i is chosen and $f_t < F$ then $\gamma_{ji} > 0$ for all $j \neq i$, implying that all such actions will receive a positive score at $t+1$.

As in the model defined on the choice probabilities, the reinforcement rule has to be modified subsequently by adding inertia and trembles. Using a parallel formulation, inertia is modeled as modifying the updating rule: write as an "interim score" the payoff in (3):

$$w_{i,t+1} = \max\{0, (1-\phi)v_{i,t} + \gamma_{ii}(f_t - F)\} \tag{4}$$

[2] The first property is obvious. The second is also obvious if $\nu > 0$. So consider the case where $\nu = 0$. Then note that if $f_t > F$ then $v_{i,t+1} > 0$. If $f_t = F$ then $v_{j,t+1}$ has the same sign as $v_{j,t}$ for all j. And if $f_t < F$ then $v_{j,t+1} > 0$ for any $j \neq i$.

and then add inertia:

$$v_{j,t+1} = (1-\epsilon)w_{j,t+1} + \epsilon\delta_{i(t)} \tag{5}$$

where δ_i denotes the unit vector with one in the ith component and zero elsewhere, and $\epsilon \in (0,1)$.

It is easily verified that all the properties required by our theory continue to be satisfied: in particular, properties (a)-(d) continue to hold for the untrembled process. Finally, the following properties can also be checked:

(e1) *Inertia:* the probability weight on i is bounded away from zero, since the scores lie in a compact interval, and

(e2) *Inertia-cum-PR:* causes the probability that the most recently chosen action will *not* be chosen will go down at least at some geometric rate bounded away from zero. This follows from the fact that the combination of inertia and PR implies that

$$\frac{v_{i,t+1}}{\sum_j v_{j,t+1}} \geq \frac{v_{i,t}+\kappa}{\sum_j v_{j,t}+\kappa}$$

where $\kappa \equiv \frac{\epsilon}{(1-\epsilon)(1-\phi)}$. Defining $\zeta \equiv \frac{\kappa}{\epsilon+\kappa}$, this implies that $(1-p_{i,t+1}) \leq \zeta(1-p_{i,t})$, where $p_{i,t}$ denotes the probability that i will be chosen at t.

Properties (a)-(d), (e1) and (e2) of the induced stochastic process over choice probabilities are all that are required by our theory.[3] With trembles added to the scores in a manner analogous to the way they were added to the process defined directly over choice probabilities themselves, all our results extend.

References

J. Bendor, D. Mookherjee and D. Ray, "Reinforcement Learning in Repeated Interaction Games," *Advances in Economic Theory* **1**.

[3] Property (e2) is used in the proof of Lemma 1 in BMR.