Reinforcement Learning Leads to Risk Averse Behavior

Jerker C. Denrell (denrell@gsb.stanford.edu)
Graduate School of Business, Stanford University, 518 Memorial Way,
Stanford, CA 94305 USA

Keywords: Reinforcement Learning; Risk Taking; Adaptation; Exploration.

Animals and humans often have to choose between options with reward distributions that are initially unknown and can only be learned through experience. Recent experimental and theoretical work has demonstrated that such decision processes can be modeled using computational models of reinforcement learning (Daw et al, 2006; Erev & Barron, 2005; Sutton & Barto, 1998). In these models, agents use past rewards to form estimates of the rewards generated by the different options and the probability of choosing an option is an increasing function of its reward estimate. Here I show that such models lead to risk averse behavior.

Reinforcement learning leads to improved performance by increasing the probability of sampling alternatives with good past outcomes and avoiding alternatives with poor past outcomes. Such adaptive sampling is sensible but introduces an asymmetry in experiential learning. Because alternatives with poor past outcomes are avoided, errors that involve underestimation of rewards are unlikely to be corrected. Because alternatives with favorable past outcomes are sampled again, errors of overestimation are likely to be corrected (Denrell & March, 2001; Denrell, 2005; 2007; March, 1996). Due to this asymmetry, reinforcement learning leads to systematic biases in decision making (e.g. Denrell, 2005; Denrell & Le Mens, 2007).

In this paper I demonstrate formally that because of this asymmetry, reinforcement learning leads to risk averse behavior: among a set of uncertain alternatives with identical expected value, the learner will, in the long run, be most likely to choose the least variable alternative.

In particular, suppose that

1) In each period, the learner must choose one of N alternatives, each with a normally distributed reward, $r_{i,t}$.

2) The learner uses a weighted average of past experiences to form a reward estimate, $y_{i,t}$, for each alternative. Specifically, the reward estimate of alternative i is: $y_{i,t+1} = (1-b)y_{i,t} + br_{i,t+1}$.

3) The learner chooses among alternatives according to a logit choice rule: the probability that alternative i is chosen in period t is $Exp(Sy_{i,t})/[\sum_{j=1}^{N} Exp(Sy_{j,t})]$.

This model leads to risk averse behavior: asymptotically the probability that alternative i is chosen is

$$\lim_{t\to\infty} P_{i,t} = \frac{Exp[S\mu_i - \frac{S^2 b}{2(2-b)}\sigma_i^2]}{\sum_{j=1}^{N} Exp[S\mu_j - \frac{S^2 b}{2(2-b)}\sigma_j^2]},$$

where μ_i & σ_i^2 are the expected reward and the variance of alternative i.

This probability is an increasing function of the expected reward, but a decreasing function of the variance. Moreover, these choice probabilities are identical to that of a decision maker who knows the probability distributions, prefers alternatives with high mean and low variance, and chooses between options according to a logit choice rule. Thus, the learning model generates choice probabilities identical to a random utility model assuming mean-variance preferences.

I prove that the result that reinforcement learning leads to risk averse behavior generalizes to a large class of probability distributions and several other belief-updating rules and choice rules.

If the reward distributions are not symmetric, I show the learning model can generate behavior consistent with a preference for alternatives with a reward distribution with low variance and positive skew. I also show that a modified logit choice rule can generate behavior consistent with an s-shaped utility function.

References

Daw, N. D., O'Doherty, J. P., Dayan , P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, 441, 876-879.

Denrell, J. (2007). Adaptive learning and risk taking. *Psychological Review*, 114, 177-187.

Denrell, Jerker. (2005). Why most people disapprove of me: Experience sampling in impression formation. *Psychological Review*, 112, 951-978.

Denrell, J. and G. Le Mens (2007). Interdependent sampling and social influence. *Psychological Review*, 114, 398-422.

Denrell, J. & March, J. G. (2001). Adaptation as information restriction: The hot stove effect. *Organization Science*, 12, 523-538.

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. Psychological Review, 112, 912-931.

March, J. G. (1996). Learning to be risk averse. *Psychological Review*, 103, 309-319.

Sutton, R. & Barto, A. G. (1998). *Reinforcement Learning*. Cambridge, MA: The MIT Press.