Formal notes on Exploration and Time Preference

Alexander S. Rich

Assume an agent is making repeated choices among several actions in an unchanging world in which the agent has uncertainty about the reward distributions of the actions. This situation can be framed as a reinforcement learning problem in which the agent's state at time t is expressed by its history of outcomes H_t, which describes all of the information it has gained about the world. (H_t could also be replaced by a set of sufficient statistics of its past experience.) Upon choosing an action, the agent receives an outcome h_{t+1} which is incorporated into the information state $H_{t+1} = H_t h_{t+1}$. At time t an optimal action by selecting the action with the highest value of

$$Q(H_t, a) = \sum_{h_t} P^a_{H_t h_{t+1}} [R^a_{H_t h_{t+1}} + \delta V(H_t h_{t+1})]$$

where

$$V(H_t) = \max_a Q(H_t, a)$$

P and R describe the probabilities and rewards, respectively, associated with each action a and information-state transition from H_t to $H_t h_{t+1}$, and δ is the exponential discount rate of rewards from future outcomes.

Within this optimal choice framework, each action a in a given information state H_t can be defined as exploitative or exploratory based on the components of its Q value. An action is exploitative to the extent that it maximizes $\sum_{h_t} P^a_{H_t h_{t+1}} R^a_{H_t h_{t+1}}$, that is, the expected immediate reward from the current choice. An action is exploratory to the extent that it maximizes $\sum_{h_t} P^a_{H_t h_{t+1}} \delta V(H_t h_{t+1})$, the expected value of the next state, which is a measure of the expected future rewards over all future choices.

This definition formalizes the intuition that exploitation is driven totally by immediate reward (R), and is not directly impacted by the information received (h_{t+1}), while exploration is driven totally by how the information received affects future value ($V(H_{t+1})$), independent of immediate reward. It also makes clear that the balance between exploration and exploitation depends on how much the agent cares about the future, measured by δ in the case of an optimal agent.

The recursive expression of the Bellman equation for the optimal value function, and its use to determine the optimal action at any time point, depends on temporal discounting being exponential. However, there is strong evidence that people do not discount time exponentially. Instead, they have some form of present bias, which warps their time discounting to discount steeply over the near future, and subsequently less steeply for the more distant future.

There are several different theories of the form of the present-biased discount function, and of people's level of awareness of their own present bias. Two commonly used discount functions are the beta-delta function and the hyperbolic function. In the beta-delta function, adopted largely for mathematical convenience, the present time point has a discount factor of 1, while a time point t steps into the future has a discount factor of $\beta\delta^t$, where $\beta < 1$ and δ is the standard expenential discount rate. In this model, the steep discounting over the near future is expressed totally by the bias β for immediate rewards over rewards with any delay, and all subsequent discounting is exponential in nature.

In the hyperbolic model, the discount factor at time t is $\frac{1}{1+tk}$, where k is a constant describing the speed of discounting. In this model, rewards are discounted steeply but not discontinuously as they move from immediacy into the future, and are gradually discounted at a slower rate as they move farther into the future.

Along with modeling people as either beta-delta or hyperbolic discounters, one must model people as either naive or sophisticated with regard to their present bias. A naive person is unaware that they are present biased, and will select an action in the present under the belief that the future sequence of actions will be optimal and time-consistent. A sophisticated person is aware of their own time-inconsistency and that it will continue in the future, and acts accordingly in the present.

The mathematically simplest model of time-inconsistent behavior is a naive beta-delta model. In this model, The value function can be calculated using the standard Q values as in the optimal model above. But, when actually making a choice, present-biased utility is maximized by choosing the highest Q_β, where

$$Q_\beta(H_t, a) = \sum_{h_t} P^a_{H_t h_{t+1}} [R^a_{H_t h_{t+1}} + \beta\delta V(H_t h_{t+1})]$$

In this equation, the value function V has been calculated under the assumption of exponential discounting at rate δ, but at actual decision time the value of future rewards is discounted an additional amount β.