The Reinforcement Learning Problem

The supervised and unsupervised learning methods we've looked at are both very specialized compared to real-life learning by humans.

- We seldom learn based on a fixed "training set", but rather based on a continuing stream of information.
- We usually act on our knowledge so far during the course of learning, not just at the end.
- We usually don't take single actions, but rather sequences of actions.
- The effects of our actions depend on the state of the world.
- We obtain a "reward" that depends in a complex way on the state of the world and on our actions.

Formalizing the Reinforcement Learning Problem

We envision the world going through a seqence of *states*, s_0, s_1, $s_2, \ldots$, at integer times. I'll mostly assume that there are a finite number of possible state. (Of course, real time is continuous, and the real state of the world would be too complex to ever model.)

At every time, we take an *action* from some set (which I'll usually assume is finite). There might be a "do nothing" action. The sequence of actions taken is a_0, a_1, $a_2, \ldots$.

As a consequence of the state, s_t, and action, a_t, we receive some *reward*, r_{t+1}, at the next time step.

Our aim is to maximize the total reward we receive over time. Sometimes a future reward is *discounted* by γ^{k-1}, where k is the number of time-steps in the future when it is received. This is like interest payments — money arriving in the future is worth less than money arriving now.

Introducing Probabilities

The world may not operate deterministically; we may not as well. Even if the world is really deterministic, an imprecise model of it will need to be probabilistic.

We assume the *Markov property* — that the future depends on the past only through the present state (really a criterion for what the state needs to be.)

We can then describe how the world works by a transition/reward distribution, given by the following probabilities (assumed the same for all t):

$$P(s_{t+1} = s',\, r_{t+1} = r \mid s_t = s,\, a_t = a)$$

We can describe our own *policy* for taking actions by the probabilities (again, assumed the same for all t):

$$P(a_t = a \mid s_t = s)$$

Note that the policy is something we decide on, whereas the way the world works is beyond our control. Note also that in this formalism we're limited to deciding only on the basis of the current state, not previous states or rewards. (When we learn a policy, however, our actions will indirectly depend on past states.)

Exploration Versus Exploitation

If we knew exactly how the world worked, there would be no need to randomize our actions — we could just take the optimal action in each state. Randomizing would at best equal this.

But if we don't have full knowledge of the world, always taking what appears to be the best action might mean we never experience states and/or actions that could produce higher rewards. There's a tradeoff between immediate reward (exploitation) and gaining knowledge that might enable higher future reward (exploration).

In a full Bayesian approach to this problem, we would still find that at any point there's an optimal action, accounting for the value of gaining knowlege, but computing it might be infeasible. A practical approach is to randomize our actions, somtimes doing apparently sub-optimal things so that we learn more.

The Expected Reward from Following a Policy

If we fix some policy, π, which defines $P(a_t = a \mid s_t = s)$, we can define the value of a state under that policy, $V^\pi(s)$, as the expected discounted reward if we follow that policy starting from state s_0 at time 0:

$$V^\pi(s_0) \quad = \quad E\Big[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \;\Big|\; s_0\Big]$$

Since we've assumed probabilities are the same for all t, this tells us the value of a state at any other time too.

This value function will satisfy the following consistency equation:

$$V^\pi(s) \quad = \quad \sum_a P^\pi(a_t = a \mid s_t = s) \sum_{s'} \sum_r P(s_{t+1} = s',\, r_{t+1} = r \mid s_t = s,\, a_t = a)\, (r + \gamma V^\pi(s'))$$

We can also look at the expected value of a state if we perform a certain action, a, and then follow policy π, which is called $Q^\pi(s, a)$.