Q-learning

General scenario: We are an *agent* in some *state*. Have sensations, perform actions, get rewards.

Q-learning: Goal to maximize discounted reward.

Today: explore basic mathematical issues involved.

Simplest case

- sensation = state.
- next state = f(current state, action).
- reward = f'(current state, action).

Q-learning

Goal is to maximize discounted reward:

$$r_0 + \gamma r_1 + \gamma^2 r_2 + \gamma^3 r_3 + \ldots, \quad (\gamma < 1)$$

Define: $Q(s,a) = r_s(a) + \sum_{i=1}^{\infty} \gamma^i \, r_i,$

where $r_s(a)$ is reward you got for doing action a in state s, and $r_1, r_2, \ldots$ are rewards for optimal action sequence after a.

Thinking about Q-values

- $Q(s,a) = r_s(a) + \sum_{i=1}^{\infty} \gamma^i r_i$.

  (r_i are rewards for doing optimal from then on)

- Why $\gamma < 1$?

- Why γ^i, and not, e.g., $1/i^2$?

  One answer: makes it *time independent*. In other words the best action to take in state s doesn't depend on when you get there.

Checkerboard example

Red to black gives $+1$ reward. Black to red gives -1 reward.

Equivalent defn of Q

- $Q(s,a) = r_s(a) + \gamma \max_{a'} Q(s_a, a')$.
- $V(S) = \max_a Q(s,a)$.
  so, $V(s) = \max_a [r_s(a) + \gamma V(s_a))]$.

Why is this OK?

A: Only one solution.

Can see this either by proving by contradiction, or noticing that if you're off by some amount Δ on the right-hand-side, then you'll be off by only $\gamma\Delta$ on the left-hand-side.

Probabilistic setting

- $R_s(a) =$ expected reward for doing a in s.
- Goal: maximize *expected* discounted reward.
- transition function is probabilistic.

Q learning algs

1. In state s, do a, go to s', get reward r:

$$\hat{Q}(s,a) \longleftarrow r + \gamma \max_{a'} \hat{Q}(s',a')$$

2. Probabilistic setting: on the tth update of $\hat{Q}(s,a)$:

$$\hat{Q}(s,a) \longleftarrow (1-\alpha_t)\hat{Q}(s,a) + \alpha_t[r + \gamma \max_{a'} \hat{Q}(s',a')]$$

Idea: dampen the randomness.

$\alpha_t = 1/t$ or similar. With $\alpha_t = 1/t$, you get a fair average of all the rewards received for doing a in state s. If you make more slowly decresing, you favor more recent r's.

Proof of convergence

Proof of convergence when world is deterministic and you see each $< s, a >$ infinitely often.

Proof: Start with some $\hat{Q}$ values. For the initial $\hat{Q}$'s, let:

$$D_0 = \max_{s,a} |\hat{Q}(s,a) - Q(s,a)|.$$

Define a "Q-interval" to be an interval in which every $< s, a >$ is tried at least once. What we will show is that after the ith Q-interval,

$$\max_{s,a} |\hat{Q}(s,a) - Q(s,a)| \leq D_0 \gamma^i.$$

In addition, during the ith Q-interval, this maximum difference will be at most $D_0 \gamma^{i-1}$.

Prove by induction. Base case (the very beginning) is OK. After an update in interval i of $\hat{Q}(s,a)$, (let's say that action a takes you from s to s') we have:

$$\begin{aligned}
|\hat{Q}(s,a) - Q(s,a)| &= |(r + \gamma \max_{a'} \hat{Q}(s',a')) - (r + \gamma \max_{a'} Q(s',a')| \\
&= \gamma |\max_{a'} \hat{Q}(s',a') - \max_{a'} Q(s',a')| \\
&\leq \gamma \max_{a'} |\hat{Q}(s',a') - Q(s,a)| \\
&\leq \gamma \cdot \gamma^{i-1} D_0.
\end{aligned}$$

This proves the theorem.