CS181 Assignment 5

Ashok Cutkosky and Tony Feng

April 25, 2013

Problem 1.

(a) We fix the following notation for the probability distribution and the utility function.

- Let p_t denote the probability distribution upon pursuing target t, so $p_t(r)$ is the probability of scoring r points, conditioned on aiming for target t.

- Let $U(r, s)$ denote the utility from scoring r points given current score s.

Now, given current score s, the expected utility from pursuing target t is

$$\mathrm{E}(t, s) = \sum_r p_t(r) U(r, s).$$

Therefore, the optimal action is to aim for target t^* satisfying

$$t^* = \mathrm{argmax}_t \, \mathrm{E}(t, s) = \mathrm{argmax}_t \sum_r p_t(r) U(r, s).$$

(b) This utility function leads to very greedy play, in which the agent will always aim to score as many points as possible without overshooting the goal. We think it is a pretty good heuristic, but not optimal because the agent will tend to get stuck with a low score, at which point it must try for very particular values that it will hit with low probability.

 We expect an agent playing with this utility function to do well at the beginning of the game, and quickly cut down its score. However, as discussed above, this greedy utility function seems like it would tend to perform poorly at the end, when it must account for subtle considerations like the best small score value to land on. To use an analogy, this strategy is like a golfer who puts all his or her emphasis on driving, and none on putting.

Problem 3

(a) If PolicyIteration converges to π, then setting $\pi = \pi^{\mathrm{new}} = \pi^{\mathrm{old}}$ in the algorithm (p.6 of Lecture 19), we find

$$V(s) = R(s, \pi(s)) + \gamma \sum_{s'} P(s' \mid s, \pi(s)) V(s')$$

$$\pi(s) = \mathrm{argmax}_a \, R(s, a) + \gamma \sum_{s'} P(s' \mid s, a) V(s').$$

By the second equation, $\pi(s)$ is the a maximizing the expression

$$R(s, a) + \gamma \sum_{s'} P(s' \mid s, a) V(s')$$

so by definition

$$V(s) = \max_a \left[R(s, a) + \gamma \sum_{s'} P(s' \mid s, a) V(s') \right].$$

These are precisely the Bellman-Ford equations. So we know that π, V give a solution to the Bellman-Ford equations. But we also know that the solution is unique, so this must be it.

(b) This is a bit involved, but please read it through!

By definition of $\pi^{(2)}(s)$ as

$$\operatorname{argmax}_a Q(s, a) = \operatorname{argmax}_a R(s, a) + \gamma \sum_{s'} P(s' \mid s, a) V(s'),$$

and $V^{\pi^{(1)}}$ as satisfying the equation

$$V^{\pi^{(1)}}(s) = R(s, \pi^{(1)}) + \gamma \sum_{s'} P(s' \mid s, \pi^{(1)}) V^{\pi^{(1)}}(s')$$

we have the following inequality for each s:

$$V^{\pi^{(1)}}(s) \le R(s, \pi^{(2)}) + \gamma \sum_{s'} P(s' \mid s, \pi^{(2)}) V^{\pi^{(1)}}(s') \tag{1}$$

We want to show that

$$V^{\pi^{(1)}}(s) \le V^{\pi^{(2)}}(s) \text{ for each } s. \tag{2}$$

Define the affine transformation T on the vector space of values on the states by

$$T(V)(s) = R(s, \pi^{(2)}) + \gamma \sum_{s'} P(s' \mid s, \pi^{(2)}) V(s').$$

Then by definition, $V^{\pi^{(2)}}(s)$ is the (unique) fixed vector of T.

Lemma 1. *If $V(s) \le V'(s)$ for each s, then $T(V)(s) \le T(V')(s)$ for each s.*

Proof. Indeed,

$$(T(V') - T(V))(s) = \gamma \sum_{s'} P(s' \mid s, \pi^{(2)})(V'(s') - V(s')) \ge 0$$

because each $V'(s') - V(s') \ge 0$ and $P(s' \mid s, \pi^{(2)}) \ge 0$. $\qquad\qquad\square$

We want to show that

$$V^{\pi^{(1)}}(s) < V^{\pi^{(2)}}(s) = R(s, \pi^{(2)}) + \gamma \sum_{s'} P(s' \mid s, \pi^{(2)}) V^{\pi^{(2)}}(s'). \tag{3}$$

Notice that equation (1) says that for each s, $V^{\pi^{(1)}}(s) \leq T(V^{\pi^{(1)}})(s)$. Therefore, applying Lemma 1 repeatedly, we find

$$V^{\pi^{(1)}}(s) \leq T(V^{\pi^{(1)}})(s) \leq T^2(V^{\pi^{(1)}})(s) \leq T^{\circ 3}(V^{\pi^{(1)}})(s) \leq \ldots \leq T^{\circ n}(V^{\pi^{(1)}})(s) \leq \ldots$$

We claim that this sequence converges to $V^{\pi^{(2)}}(s)$, which will establish (3) as desired. As mentioned above, $V^{\pi^{(2)}}$ is defined to be a fixed point vector of the operator T, so by the contraction mapping theorem, it suffices to show that T is a contraction.

Lemma 2. T *is a contraction with factor $\gamma < 1$.*

Proof. Indeed, for all s we have

$$|(T(V') - T(V))(s)| = |\gamma \sum_{s'} P(s' \mid s, \pi^{(2)})(V'(s') - V(s'))|$$

$$\leq \gamma \sum_{s'} P(s' \mid s, \pi^{(2)})|V'(s') - V(s')|$$

$$\leq \gamma ||V' - V||$$

because $\sum_{s'} P(s' \mid s, \pi^{(2)}) = 1$. $\qquad\qquad\square$

By the above Lemma, and the Contraction Mapping Theorem,

$$\lim_{n \to \infty} T^{\circ n}(V^{\pi^{(1)}})$$

is the unique fixed point of T, and is equal to $V^{\pi^{(2)}}$. This completes the proof, by the reasoning discussed above.

(c) First note that if $V^{\pi^{(1)}}(s) = V^{\pi^{(2)}}(s)$ for each s, then the algorithm will converge. Indeed, if V does not change then π, which is defined in each step as an argmax with respect to a function depending only on the states, actions, and V, will not change either.

Otherwise, at each step $\sum_s V^\pi(s)$ strictly increases. However, there are only finitely many distinct possibilities for π, since π is a function from the finite set of states S to the finite set of actions A, and hence only finitely many distinct possibilities for V^π. Therefore, it can only strictly increase finitely many times, so it converges in finitely many steps. By (a), when it converges, it must return an optimal stationary policy.