April 27

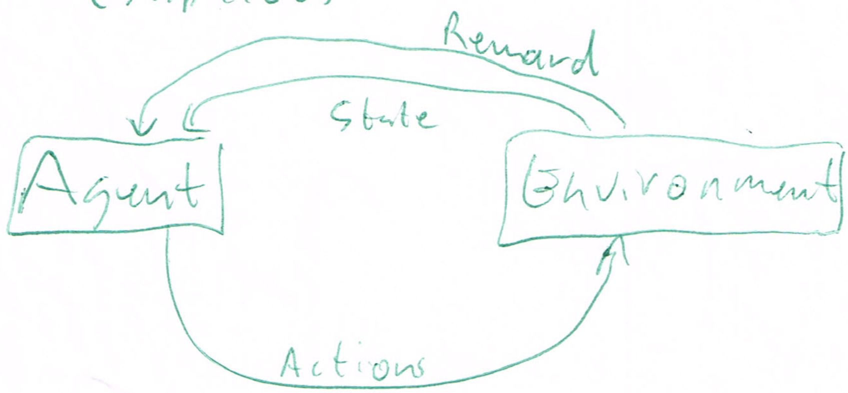
Hw 4 out today

Reinforcement Learning

Intuition: learning to play backgammon

- Play game
- Win / Lose
- "Reinforce" decisions that lead to a win
  "Punish" decisions that lead to a loss
  - Credit assignment problem
    - which moves are responsible

(Tesauro, Neurogammon)
(AlphaGo)



S: States
A: Actions $A: S \to S$
R: Reward $R: S \to \mathbb{R}$
$R: S \to \{0, 1, -1\}$

Markov Decision Process

Actions are probabilistic

$P(s' | s, a)$ $s \in S$ $a \in A$

π: Policy $\pi: S \to A$

Cumulative Reward (Utility)

$$U_h([s_0, \ldots, s_t, \ldots]) = \sum_{t=0}^{\infty} R(s_t)$$

Cumulative Discounted Reward (Utility)

$0 \le \gamma \le 1$

$$U_h([s_0, \ldots, s_t, \ldots]) = \sum_{t=0}^{\infty} \gamma^t R(s_t)$$

Expected Utility of a policy

$$U^\pi(s_0) = E \sum_{t=0}^{\infty} \gamma^t R(s_t)$$

where for $t>0$ $\quad s_t = a_t(s_{t-1})$

$a_t = \pi(s_{t-1})$

Optimal Policy π^*

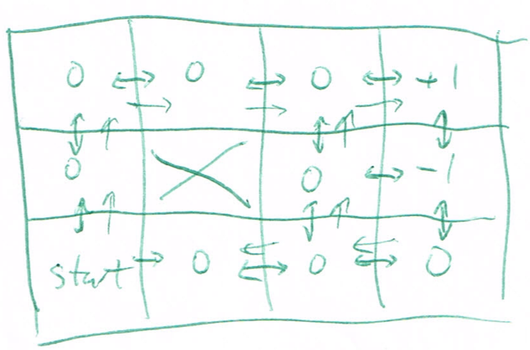
$$\pi^*(s) = \underset{\pi}{\operatorname{argmax}}\, U^\pi(s) = U(s)$$

$$= \underset{a \in A}{\operatorname{argmax}} \sum_{s' \in S} P(s'|s,a)\, U(s')$$

$$= R(s) + \gamma \max_{a \in A} \sum_{s' \in S} P(s'|s,a)\, U(s')$$

Bellman Equation

Example



$P(\text{going up} \mid \text{Up action}) = .8$

$P(\text{going L} \mid \text{Up}) = .1$

$P(\text{going R} \mid \text{Up}) = .1$

and so on

.812	.868	.918	1
.762	X	.660	−1
.705	.655	.611	.388

$\gamma = 1$

$R(s) = -.04$ at non terminal states

Bellman Equation

$$u(1,1) = -.4 + \gamma \max(.8\,u(1,2) + .1\,u(2,1) + .1\,u(1,1),\; .9\,u(1,1) + .1\,u(2,1),\; .8\,u(2,1) + .1\,u(1,2) + .1\,u(1,1))$$

Value iteration

$$U_0(s) = 0 \quad \forall s$$

Repeat

$$U_{i+1}(s) = R(s) + \gamma \max_{a \in A} \sum_{s' \in S} P(s'|s,a)\, U_i(s')$$

(for all s)

Until <stopping condition>

(Assumes you maintain all states)

At each state s

$$\pi^*(s) = \underset{a \in A}{\operatorname{argmax}} \sum_{s' \in S} P(s'|s,a)\, U(s')$$

(Assumes you know $P(s'|s,a)$)