Markov Decision problems: (processes)!

discount

$$(S, A, P, R, \gamma, D)$$

Distribution of initial state

finite set of states

Reward function
$R: S \to \mathbb{R}$, bounded.

actions:
$A: S \to$ some discrete set

Probability ~~of that next state is s~~

model $P(S_{i+1} = s \mid S_i = \hat{s}, a(s_i) = \hat{a})$

Policy

$$\pi : S \to A$$

(i.e. in state s, do this)

General, very useful model for action
+ control

+ Can ~~to~~ extend.
+ Note moral similarity to LQR.
(Where we ~~cost~~ had states
actions ≐ controls
rewards
But transitions were deterministic

Model

- Start in $S_0 \sim D$, $\begin{bmatrix}\text{total}\\ \text{reward}\end{bmatrix} = 0$
- Choose a_i
- $S_{i+1} \sim P(S_{\cancel{k}} \mid S_i, a_i)$
- get reward $R(S_{i+1})$
- ~~$\begin{bmatrix}\text{total}\\ \text{reward}\end{bmatrix} \to \gamma \begin{bmatrix}\text{total}\\ \text{reward}\end{bmatrix} + R(S_{i+1})$~~

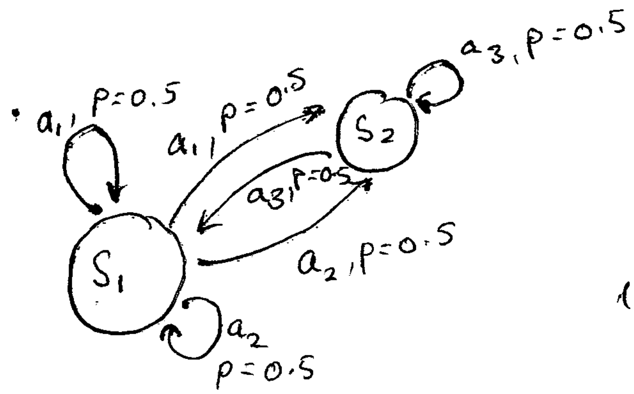
Notice effect of discount — long ago rewards have less weight.

want to max $\sum_{t=0}^{\infty} \gamma^t R(S_t)$ ← discounted total reward

Natural Question:

- What is the best policy π^* for a given problem?
  (This might not be unique)

eg



$\pi(s_1) = a_1$ is as good as $\pi(s_1) = a_2$

- ## Subordinate Question:

assume policy is π; what is Value of states?

i.e. $V_\pi(s) = E\left[\begin{array}{l}\text{Reward to be obtained by} \\ \text{following } \pi \text{ from } s\end{array}\right]$

Where expectation is over P.

Value is important, because its linked to policy. Imagine we know $V_{\pi^*}(s)$.
then we can recover π^*

π^* : choose action with best expected value of $V_{\pi^*}(s)$.

Notice a_1 has expected value $\sum_s P(s|\hat{s}, a_1) V_{\pi^*}(s)$

Notice: $V_{\pi^*}(s)$ is usually more than $R(s)$ because you get rewards from future actions

~~But $\gamma = 0 \Rightarrow V_{\pi^*}(s) = R(s)$~~

$\therefore$ get $V_{\pi^*}(s) = R(s) + \gamma \max_a \sum P(s'|s, a) V_{\pi^*}(s')$

AND $\pi^*(s) = \underset{a}{\operatorname{argmax}} \sum_{s'} P(s'|s, a) V_{\pi^*}(s')$

Bellman equation

Estimating the optimal value fn.

Value iteration:

init $V^{(0)}(s) = 0$

update $V^{(n+1)}(s) = R(s) + \gamma \max_{a} \sum_{s'} P(s'|s,a) V^{(n)}(s)$

Thm $V^{(k)} \longrightarrow V^{\pi^*}$ as $k \to \infty$

Proof:

write $B[V] = R(s) + \gamma \max_{a} \sum_{s'} P(s'|s,a) V(s)$

$\uparrow$

Bellman Backup Operator.

lemma below shows

$$\max_s \left| B[V_1](s) - B[V_2](s) \right| \leq \gamma \max_s \left| V_1(s) - V_2(s) \right|$$

Note that

$$B[V_{\pi^*}] = V_{\pi^*}$$

assume lemma

then $\left| B[V^{(n)}] - B[V_{\pi^*}] \right|_\infty$

$$= \left| B[V^{(n)}] - V_{\pi^*} \right|_\infty$$

$$\leq \gamma \left| V^{(n)} - V_{\pi^*} \right|_\infty$$

and $0 \leq \gamma < 1$

So $V^{(n+1)} = B[V^{(n)}]$ is closer to V_{π^*} than $V^{(n)}$ $\square$

Lemma:

$$\left| B[V_1] - B[V_2] \right|_\infty$$

$$= \gamma \left| \left[\max_a \sum_{s'} P(s'|s,a) V_1(s')\right] - \left[\max_a \sum_{s'} P(s'|s,a) V_2(s)\right] \right|_\infty$$

$$\le \gamma \max_a \left| \sum_{s'} P(s'|s,a) V_1(s') - \sum_{s'} P(s'|s,a) V_2(s') \right|_\infty = A$$

{ because $\left|\max_x f(x) - \max_x g(x)\right| \le \max_x \left|f(x)-g(x)\right|$ }

$$A = \gamma \max_a \sum_{s'} P(s'|s,a) \left| V_1(s') - V_2(s') \right|$$

$$\le \gamma \max_s \left| V_1(s) - V_2(s) \right|$$

(P is non neg, and sums to 1)

Convergence :

assume rewards bounded by R_{max}

$$V^*(s) \leq \sum_{t=0}^{\infty} \gamma^t R_{max} = \frac{R_{max}}{1-\gamma}$$

$$\text{so} \quad \max_{s \in S} \left| V^k(s) - V^*(s) \right| \leq \frac{\gamma^k R_{max}}{1-\gamma}$$

(i.e. ~~fast~~ linear ; faster for small γ)

Asynchronous Version :

- notice what is above assumes <u>all</u> states are updated together
- i.e compute $\hat{V}(s) = B[V^{(n)}]$

  then $V^{(n+1)} = \tilde{V}(s)$
- <u>but</u> you could visit states in turn, updating each
  - Converges
  - usually better

8a

Policy iteration:

- initialize w/ policy, perhaps random $\pi^{(0)}$
- Compute V_π

  - Notice this involves solving a linear system

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

    $$V_\pi = R + \gamma T V_\pi$$

    $$(I - \gamma T) V_\pi = R.$$

    $$V_\pi = [I + \gamma T + \gamma^2 T^2 + \gamma^3 T^3 \ldots] R$$

    Neumann series

- update π to be greedy wrt V_π

  i.e. $\pi(s) = \arg\max_a \sum_{s'} P(s' | s, a) V_\pi(s')$

- Iterate

Policy iteration converges
- to exact optimal policy
- eventually.

In practice, policy iteration is usually preferred, particularly if transition probs are sparse

MDP's as LP's :

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

implies ~~equiv~~ to

$$V(s) \geq R(s) + \gamma \sum_{s' \in S} P(s'|s,a) V(s')$$

one constraint for each A

So we can seek

$$\min \sum_s V(s)$$

$$\text{s.t.} \quad V(s) \geq R(s) + \gamma \sum_{s'} p(s'|s,a)\, V(s') \quad \Big| \leftarrow \text{one for each } s, a$$

and this is the optimal value fn.

Write

$$1^T V$$

$$\text{s.t.} \quad (I - T_a) V - R \geq 0 \qquad a = 1 \ldots \#A$$

V is #S vector
one T_a per a.

Now consider the <u>dual</u>

λ_a
$\uparrow$ LM, #S dimensional, 1 per a

$$\max \sum_a \lambda_a^T R$$

$$\text{s.t.} \quad \sum_a (I - T_a)^T \lambda_a + 1 = 0$$

$$\lambda_a \geq 0$$

What are dual vars?

1) Notice one useful interpretation of γ.

$$(1-\gamma) = P[\text{MDP stops}]$$

$$
\begin{aligned}
E[\text{reward}] &= \cancel{\text{reward after stopp}}\ \text{reward at start} \\
&\quad (\text{reward at 1st state})\, P(\text{1 step}) \\
&\quad + (\text{reward at 2nd state})\, P(2) \\
&\quad + \quad \vdots \\
&= R(s_0) + R(s_1)\gamma + R(s_2)\gamma^2 + \cdots
\end{aligned}
$$

2) write Primal objective

$$\frac{1^T V}{|S|} \leftarrow \#\text{states}.$$

Dual becomes

$$
\begin{aligned}
\max \quad & \sum_a \lambda_a^T R \\
\text{s.t.} \quad & \sum_a (I - \mathbb{P}_a)^T \lambda_a + \frac{1}{|S|} = 0 \\
& \lambda_a \geq 0
\end{aligned}
$$

Now consider the expected number of times we are in s, take a

$= N(s,a)$.

$\cancel{N(s,a) = \frac{1}{|S|}}$

expected # of times we ~~are in~~ arrive in s

is

$$\frac{1}{|S|} + \gamma \sum_{s'} \sum_{a} P(s \mid a, s') \, N(s', a)$$

Because we have a uniform prob first state is s

~~joint~~ Prob $P(s \mid s', a)$

Notice:

$\sum_a N(s,a)$ = expected number of times we leave from s

But

$$\sum_a N(s,a) = \frac{1}{|S|} + \gamma \sum_{s'} \sum_{a} P(s' \mid s, a) \, N(s,a)$$

these are the constraints on Dual Vars!

So

$$\lambda_a = \lambda_a(s) = N(s,a).$$

These encode the optimal policy:

$\pi_*(s)$ = take the most common action at s
$= \underset{a}{\operatorname{argmax}}\, N(s,a).$

Natural follow up questions

- what happens if you don't know
  P, R ?
  (Reinforcement learning:
  - roughly: act, see what happens
  keep notes)

- what happens if you dont know
  P, R BUT see a
  skilled actor ?
  (our problem)