Markov Decision Processes

- An MDP is a general way to model an online decision-making problem where any uncertain parameter is modelled in a Bayesian manner (i.e., as being drawn via some known stochastic process)

- MDPs can be defined over continuous spaces, and with continuous-time updates. We will focus (for now...) on discrete time updates, and discrete (finite/countable) states (This is sometimes called a tabular MDP)

- Defn: A Markov Chain is a stochastic process $(X_t)_{t=0}^{\infty}$ given by a stochastic update (i.e., randomized fn)

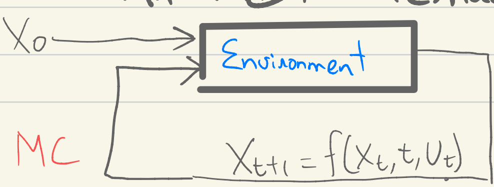
$$X_{t+1} = f(\overbrace{X_t, t}^{\text{state}}, \underbrace{U_t}_{\text{independent r.v. ('disturbance')}})$$

where $U_1, U_2, \ldots$ are iid $U[0,1]$ rv (recall: ANY rv Y with cdf F can be constructed as $Y = F^{-1}(U)$)

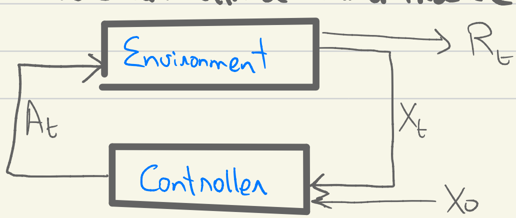
  - Any Markov chain comprises of the following 'inputs'
    - State space S
    - Initial state X_0 (or initial distn π_0 over S)
    - (time t) transition 'kernel' $P_t(x|X_t) = \mathbb{P}[X_{t+1} = x | X_t]$

- An MDP interlaces a Markov chain with a 'control' module

MC



MDP



Defn. A Markov Decision Process comprises of 3 interlacing sequences -

States $X_0, X_1, X_2 \ldots \in \mathcal{S}$ (State space)

Actions $A_0, X_1, X_2 \ldots \in \mathcal{A}$ (Action space)

Rewards $R_1, R_2, \ldots$ (Reward

These are related via two functions

$$X_{t+1} = f(X_t, A_t, U_t) \quad \text{(Transition function)}$$

$$R_{t+1} = g(X_t, A_t, U_t) \quad \text{(Reward function)}$$

· Notes

- The transitions can be represented via a transition kernel

$$T_t(x \mid X_t, A_t) = \mathbb{P}[X_{t+1} = x \mid X_t, A_t]$$

- Rewards are sometimes written as $R(X_t, A_t, X_{t+1})$, or just $R(X_t, A_t)$
- Like with MCs, the inputs for an MDP are

$$\underbrace{\mathcal{S}, \mathcal{A}}_{\substack{\text{state/action} \\ \text{spaces}}}, \underbrace{T, R}_{\substack{\text{transition/reward} \\ \text{models}}}, \pi_0$$

initial state (distⁿ of X_0)

- To model time varying processes, have t included in state-space

  To model state-dependent action spaces, define T and R appropriately

  To model 'terminal rewards' in some 'final' state, include dummy actions...

- (Basically, can model everything this way!)
- Policy $\pi = (\pi_1, \pi_2, \ldots)$, $\pi_t : \mathcal{S} \to \mathcal{A}$ is a collection of mappings (one for each) time from states to actions

- **Optimality Criteria** (ie, 'flavors' of MDPs)

MDPs come in different flavors depending on their objective

- Finite-horizon (Episodic) - Given known 'horizon' $H \geq 1$, for any starting state $X_0 = x$, objective is to maximize over all policies π:

$$V(x) = \mathbb{E}_x\Big[\sum_{t=0}^{H} R_t\big(X_t, \underbrace{A_t = \pi_t(X_t)}_{\text{pick actions from policy } \pi}, U_t\big)\Big]$$

← fixed horizon

Start at $X_0 = x$

- Shortest Path problem - Given (terminal) subset $U \subset \mathcal{S}$, let $\tau_U = \inf\{t \geq 1 \mid X_t \in U\}$. The objective is to minimize

$$C(x) = \mathbb{E}_x\Big[\sum_{t=0}^{\tau_U - 1} R_t(X_t, A_t)\Big]$$

'cost' - sometimes $R_t(X_t, X_{t+1})$

- Discounted Reward - Given discount factor $\gamma \in (0,1)$, objective is

$$V(x) = \mathbb{E}_x\Big[\sum_{t=0}^{\infty} \gamma^t R_t(X_t, A_t)\Big]$$

Equivalently, given an independent, random horizon $H \sim \mathrm{Geom}(\gamma)$

← random horizon ~ Geometric(γ)

$$V(x) = \mathbb{E}_x\Big[\sum_{t=0}^{H} R_t(X_t, A_t)\Big]$$

- (Infinite horizon) Average Reward - Objective is to maximize over all (A_t)

$$V(x) = \limsup_{H \to \infty} \frac{1}{H} \mathbb{E}_x\Big[\sum_{t=0}^{H} R_t(X_t, A_t)\Big]$$

LP formulations of MDPs

- Main Idea - Can 'insulate' future from past decisions by using **state-action frequencies** as variables

- For finite horizon - Let $r_t(x,a) \triangleq \mathbb{E}[R_t(X_t=x, A_t=a)]$

Consider any policy π, and suppose we 'run' it over 'many' episodes

Now define $q_t(x,a)=$ fraction of runs which end up in state x at time t and policy π plays action a

- Expected reward of $\pi \equiv V^{\pi}(x_0) = \sum_{t=0}^{H} \sum_{x\in S} \sum_{a\in A} r_t(x,a)\, q_t(x,a)$

- Consistency (flow-balance) - $q_0(x,a)=0 \;\forall x\neq x_0$, $\sum_{a\in A} q_0(x_0,a)=1$

and $\forall t\geq 1, \forall x\in S$: $\underbrace{\sum_{a\in A} q_t(x,a)}_{\text{'flow out' of } x,t} = \underbrace{\sum_{x'\in S}\sum_{a'\in A} q_{t-1}(x',a')\,T_{t-1}(x|x',a')}_{\text{'flow in' to } x,t}$

Putting it together we get the LP

Finite-horizon Primal

$$\boxed{\begin{array}{l|lr}
\text{dual var} & \max \sum_{t=0}^{H} \sum_{x\in S} \sum_{a\in A} q_t(x,a)\, r_t(x,a) & \\
V_0(x) & \text{s.t. } q_0(x,a) = 0 & \forall x\neq x_0 \\
V_0(x_0) & \sum_{a\in A} q_0(x_0,a) = 1 & \\
V_t(x) & \sum_{a\in A} q_t(x,a) = \sum_{x'\in S}\sum_{a'\in A} q_{t-1}(x',a')\,T_{t-1}(x|x',a') & \forall \begin{array}{c} t\geq 1 \\ x\in X \end{array} \\
 & q_t(x,a) \geq 0 & \forall t,x,a
\end{array}}$$

We can also now look at the dual LP

$$\begin{aligned} \min \quad & V_0(x_0) \\ \text{s.t.} \quad & V_t(x) - \sum_{x' \in S} \tilde{T}_t(x'|x,a) V_{t+1}(x') \geq r_t(x,a) \quad \forall t<H, \forall x,a \\ & V_t(x) \geq 0 \quad \forall t<H, \forall x,a \end{aligned}$$

- Note - If $X_0 \sim \Pi_0$, we set $\sum_a q_0(x,a) = \Pi_0(x) \ \forall x$ in the primal, and $\min \sum_x \Pi_0(x) V_0(x)$ as the objective in the dual

- We can simplify ⊛ in the dual to get

$$\boxed{V_t(x) \geq \max_{a \in A} \left[r_t(x,a) + \sum_{y \in S} \tilde{T}_t(y|x,a) V_{t+1}(y) \right] \quad \forall t<H, \ \forall x \in S}$$

Finite-horizon HJB eqns

- This is called the Bellman optimality condition (and is a special condition of the more general Hamilton-Jacobi-Bellman or HJB equation).

- The variables $\{V_t(x)\}_{x \in S}$ are referred to as the value function at t. Any feasible value fn $V_t(x)$ induces a corresponding policy

$$\Pi_t^V(x) = \arg\max_{a \in A} \left[r_t(x,a) + \sum_{y \in S} T_t(y|x,a) V_{t+1}(y) \right] \quad \forall t<H, \forall x$$

Similarly any policy $\Pi = (\Pi_t(x))$ induces a corresponding value fn

$$V_t^{\Pi}(x) = r_t(x, \Pi_t(x)) + \sum_{y \in S} T_t(y|x, \Pi_t(x)) V_{t+1}^{\Pi}(y) \quad \forall t<H, \forall x$$

(we need as input 'terminal' rewards $V_H^{\Pi}(x)$ in both cases)

LP formulations for other criterion

The advantage of the state-action frequency LP is that it naturally extends to the other flavors of MDPs.

- Discounted rewards

- Consider a time-invariant MDP, i.e, with $R_t = R$ and $T_t = T$
  - Claim - the opt policy can also be taken to be time invariant
- Suppose we run a policy π over many trials $j \in \{1,2,\dots\}$, where each trial terminates after $H^j \sim \text{Geom}(\gamma)$ rounds
- As before, $r(x,a) = \mathbb{E}[R(x,a)]$

  Define $q(x,a) = $ avg # of times action a played in state x

  ← avg over each trial

  Also assume $X_0 \sim \pi_0$

  Then the MDP ≡ following LP

Discounted MDP primal

$$\max \sum_{x\in S}\sum_{a\in A} r(x,a)\, q(x,a)$$

s.t.

$$\pi_0(x) + \sum_{y\in S}\sum_{a\in A} q(y,a)(1-\gamma)\, T(x|y,a) = \sum_{a\in A} q(x,a) \quad \forall x\in X$$

$$q(x,a) \geq 0 \quad \forall\, x,a$$

↑ dual var $V(x)$

and its dual

$$\min \sum_{x\in X} \pi_0(x)\, V(x)$$

s.t.

$$V(x) \geq r(x,a) + (1-\gamma)\sum_{y\in S} T(y|x,a)\, V(y) \quad \forall x, \forall a$$

$$V(x) \geq 0 \quad \forall x$$

equivalently

$$\boxed{\forall x\in S \quad V(x) \geq \min_{a\in A}\Big[r(x,a) + (1-\gamma)\sum_{y\in S} T(y|x,a)\, V(y)\Big]}$$

discounted reward HJB eqn