

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**)

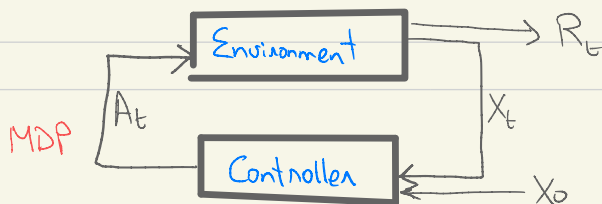
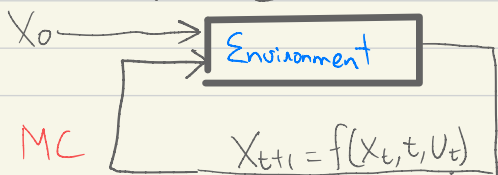
- Def: A **Markov Chain** is a stochastic process $(X_t)_{t=0}^{\infty}$ given by a **stochastic update** (i.e., randomized fn)

$$X_{t+1} = f(\overset{\text{state}}{X_t(t)}, \underset{\text{independent n.v. ('disturbance')}}{U_t})$$

where $U_1, U_2, \dots$ are iid $U[0,1]$ r.v. (recall: ANY r.v. 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 distrⁿ Π_0 over S)
 - **(time t) transition 'kernel'** $P_t(x|X_t) = P[X_{t+1}=x|X_t]$

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



Defn. A **Markov Decision Process** comprises of 3 interlocking sequences -

States $X_0, X_1, X_2 \dots \in S$ (State space)
Actions $A_0, X_1, X_2 \dots \in A$ (Action space)
Rewards $R_1, R_2, \dots$ (Reward)

These are related via two functions

$$\begin{aligned}
 X_{t+1} &= f(X_t, A_t, U_t) \quad (\text{Transition function}) \\
 R_{t+1} &= g(X_t, A_t, U_t) \quad (\text{Reward function})
 \end{aligned}$$

Notes

- The transitions can be represented via a **transition kernel**

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

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

$\underbrace{S, A}_{\text{state/action spaces}}, \underbrace{T, R}_{\text{transition/reward models}}, \Pi_0 \leftarrow \text{initial state (dist}^n \text{ of } X_0)$

- 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** $\pi = (\pi_1, \pi_2, \dots)$, $\pi_t: S \rightarrow A$ 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 :
$$V(x) = \mathbb{E}_x \left[\sum_{t=0}^{H-1} R_t(X_t, A_t = \pi_t(X_t)) \right]$$

$\uparrow$ Start at $X_0 = x$ $\leftarrow$ fixed horizon $\leftarrow$ pick actions from policy π

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

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

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

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

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

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

$$V(x) = \mathbb{E}_x \left[\sum_{t=0}^{H-1} R_t(X_t, A_t) \right]$$

$\leftarrow$ random horizon $\sim \text{Geometric}(\gamma)$

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

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

LP formulations of MDPs

- **Main Idea** - Can 'insulate' future from past decisions by using **state-action frequencies** as variables
- For finite horizon - Let $\eta_t(x, a) \triangleq \mathbb{E}[R_t(x_t=x, A_t=a)]$
Consider any policy π , and suppose we 'run' it over many episodes $j \in \{1, 2, \dots, J\}$.

i.e. $\frac{1}{J} \sum_{j=1}^J \mathbb{1}\{x_t^j = x, A_t^j = a\}$

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 $\pi \equiv V^\pi(x_0) = \sum_{t=0}^{H-1} \sum_{x \in \mathcal{S}} \sum_{a \in \mathcal{A}} \eta_t(x, a) q_t(x, a)$
- **Consistency (flow-balance)** - $q_0(x, a) = 0 \forall x \neq x_0, \sum_{a \in \mathcal{A}} q_0(x_0, a) = 1$

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

Putting it together we get the LP

Finite-horizon Primal

	$\max \sum_{t=0}^{H-1} \sum_{x \in \mathcal{S}} \sum_{a \in \mathcal{A}} q_t(x, a) \eta_t(x, a)$	
dual var	s.t.	$q_0(x, a) = 0 \quad \forall x \neq x_0$
$V_0(x_0)$		$\sum_{a \in \mathcal{A}} q_0(x_0, a) = 1$
$V_t(x)$		$\sum_{a \in \mathcal{A}} q_t(x, a) = \sum_{x' \in \mathcal{S}} \sum_{a \in \mathcal{A}} q_{t-1}(x', a) T_t(x x', a) \quad \forall x \in \mathcal{S} \quad \forall t \geq 1$
		$q_t(x, a) \geq 0 \quad \forall t, x, a$

We can also now look at the dual LP

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

• Note - If $X_0 \sim \pi_0$, we set $\sum_a q_0(x, a) = \pi_0(x) \quad \forall x$ in the primal, and $\min \sum_x \pi_0(x) V_0(x)$ as the objective in the dual

• We can simplify $\textcircled{*}$ in the dual to get

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

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 $\{V_t(x)\}_{x \in \mathcal{S}}$ are referred to as the **value function** at t . Any feasible value fn $V_t(x)$ induces a corresponding policy $\pi_t^V(x) = \arg \max_{a \in A} [r_t(x, a) + \sum_{y \in \mathcal{S}} \tilde{T}_t(y|x, a) V_{t+1}(y)] \quad \forall t < H, \forall x$

Similarly any policy $\pi = (\pi_t(x))$ induces a corresponding value fn

$$V_t^\pi(x) = r_t(x, \pi_t(x)) + \sum_{y \in \mathcal{S}} \tilde{T}_t(y|x, \pi_t(x)) V_{t+1}^\pi(y) \quad \forall t < H, \forall x$$

(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 $j \in \{1, 2, \dots, J\}$, where each trial terminates after $H^j \sim \text{Geom}(\gamma)$ rounds

- As before, $\eta(x, a) = \mathbb{E}[R(x, a)]$

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

Also assume $X_0 \sim \pi_0$

Then the MDP $\equiv$ following LP

$\text{i.e., } \frac{1}{J} \sum_{j=1}^J \sum_{t=0}^{H^j-1} \mathbb{1}\{x_t^j = x, A_t^j = a\}$

$$\begin{aligned} & \max \sum_{x \in \mathcal{S}} \sum_{a \in \mathcal{A}} \eta(x, a) q(x, a) && \text{Discounted MDP primal} \\ & \text{s.t.} \\ & \pi_0(x) + \sum_{y \in \mathcal{S}} \sum_{a \in \mathcal{A}} q(y, a) \gamma T(x|y, a) = \sum_{a \in \mathcal{A}} q(x, a) \quad \forall x \in \mathcal{X} \\ & q(x, a) \geq 0 \quad \forall x, a \end{aligned}$$

$\uparrow$
dual var $V(x)$

and its dual

$$\begin{aligned} & \min \sum_{x \in \mathcal{X}} \pi_0(x) V(x) \\ & \text{s.t.} \quad V(x) \geq \eta(x, a) + \gamma \sum_{y \in \mathcal{S}} T(y|x, a) V(y) \quad \forall x, \forall a \\ & \quad \quad V(x) \geq 0 \end{aligned}$$

equivalently

$$\forall x \in \mathcal{S} \quad V(x) \geq \min_{a \in \mathcal{A}} \left[\eta(x, a) + \gamma \sum_{y \in \mathcal{S}} T(y|x, a) V(y) \right]$$

discounted reward HJB eqn

- Avg Reward - Again we consider time-invariant MDPs. Given any stationary policy π , suppose we run it for a long time (i.e., $H \rightarrow \infty$)

- Define $q(x,a) = \frac{1}{H} \sum_{k=1}^H \mathbb{1}\{x_k=x, A_k=a\}$

= Avg freq. of playing action a in state x

- As before $\eta(x,a) = \mathbb{E}[R(x,a)]$. Also $X_0 \sim \pi$

does not matter by MC ergodicity

Now the LP becomes

$\max \sum_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}} q(x,a) \eta(x,a)$		avg reward primal
s.t.		
$V(x)$	$\sum_{a \in \mathcal{A}} q(x,a) - \sum_{y \in \mathcal{S}} \sum_{a \in \mathcal{A}} q(y,a) T(x y,a) = 0 \quad \forall x \in \mathcal{S}$	
$\rightarrow$	$\sum_{x \in \mathcal{S}} \sum_{a \in \mathcal{A}} q(x,a) = 1$	
<u>duals</u>	$q(x,a) \geq 0 \quad \forall x,a$	

The corresponding dual LP is

min γ

s.t. $V(x) \geq \eta(x,a) + \sum_{y \in \mathcal{S}} T(y|x,a) V(y) - \gamma \quad \forall x \in \mathcal{S}$

$V(x), \gamma \in \mathbb{R}$

γ is called the long-run average reward of the MDP, and $V(x)$ is now the relative value-fn of state x . Simplifying we get

$$V(x) \geq \max_{a \in \mathcal{A}} \left[\sum_{y \in \mathcal{S}} T(y|x,a) V(y) + \eta(x,a) - \gamma \right] \quad \forall x$$

avg reward HJB equation

Interpreting the HJB equations

From above we get the 3 HJB equations

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

Finite-horizon HJB eqns

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

discounted reward HJB eqn

$$V(x) \geq \max_{a \in A} \left[\sum_{y \in S} T(y|x, a) V(y) + r(x, a) - \gamma \right] \quad \forall x \in S$$

avg reward HJB equation

Also, since in each case we minimize $V_0(x_0)$ (or γ), it's easy to argue that the above inequalities can be replaced by equality.

More importantly, $V(x)$ in each case has an associated probabilistic interpretation. Recall we defined a policy π to be a mapping from S to A (in time invariant, else $\pi = \{\pi_t\}, \pi_t: S \rightarrow A$). Then

$$V_t^{FH}(x) = \min_{\{\pi_t\}} \mathbb{E} \left[\sum_{t'=t}^{H-1} r_{t'}(X_{t'}, \pi_{t'}(X_{t'})) \mid X_t = x \right]$$

$$V^{Disc}(x) = \min_{\pi} \mathbb{E} \left[\sum_{t=0}^{H-1} \gamma^t r(X_t, \pi_t(X_t)) \mid X_0 = x \right], \text{ with } H \sim \text{Geo}(r)$$

$$V^{Avg}(x) = \min_{\pi} \mathbb{E} \left[\sum_{t=0}^{\tau_x-1} r(X_t, \pi_t(X_t)) - \gamma \mid X_0 = x \right], \text{ with } \tau_x = \min\{t > 0 \mid X_t = x\}$$

return time