# 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)

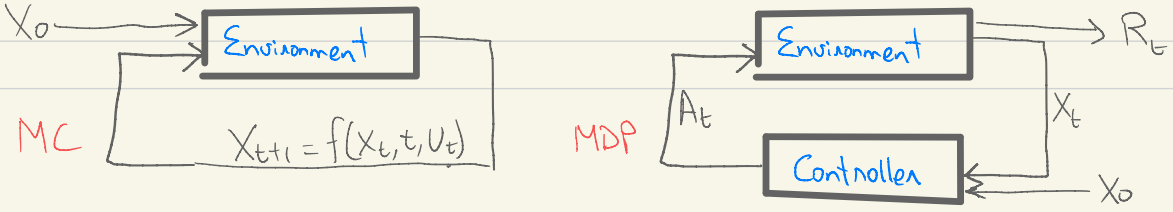
$$X_{t+1} = f(X_t(t), U_t)$$

($X_t(t)$: state; $U_t$: independent r.v. ('disturbance'))

where $U_1, U_2, \ldots$ are iid $U[0,1]$ rvs (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 $\pi_0$ 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



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

States $X_0, X_1, X_2 \ldots \in S$ (state space)
Actions $A_0, X_1, X_2 \ldots \in A$ (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
    $S, A$ (state/action spaces), $T, R$ (transition/reward models), $\pi_0$ (initial state (dist$^n$ 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** $\pi = (\pi_1, \pi_2, \ldots)$, $\pi_t : S \to A$ 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 $\pi$:

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

($H-1$: fixed horizon; $X_0 = x$: start at $X_0 = x$; $A_t = \pi_t(X_t)$: pick actions from policy $\pi$)

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

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

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

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

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

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

($H$: random horizon $\sim$ Geometric($\gamma$))

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

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

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

# LP formulations of MDPs

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

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

Consider any policy $\pi$, 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 $\pi$ plays action $a$

- Expected reward of $\pi \equiv V^\pi(x_0) = \sum_{t=0}^{H+1} \sum_{x\in\mathcal{S}} \sum_{a\in\mathcal{A}} r_t(x,a)\, q_t(x,a)$

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

Putting it together we get the LP

Finite-horizon Primal

$$\max \sum_{t=0}^{H} \sum_{x\in\mathcal{S}} \sum_{a\in\mathcal{A}} q_t(x,a)\, r_t(x,a)$$

| dual var | | |
|---|---|---|
| $V_0(x)$ | s.t. $q_0(x,a) = 0$ | $\forall x \neq x_0$ |
| $V_0(x_0)$ | $\sum_{a\in\mathcal{A}} q_0(x_0,a) = 1$ | |
| $V_t(x)$ | $\sum_{a\in\mathcal{A}} q_t(x,a) = \sum_{x'\in\mathcal{S}} \sum_{a'\in\mathcal{A}} q_{t-1}(x',a') T_{t-1}(x|x',a')$ | $\forall t \geq 1, x \in \mathcal{X}$ |
| | $q_t(x,a) \geq 0$ | $\forall t,x,a$ |

We can also now look at the dual LP

$$\min \quad V_0(x_0)$$
$$\text{s.t.} \quad \circledast \; V_t(x) - \sum_{x' \in \mathcal{S}} T_t(x'|x,a)\, V_{t+1}(x') \geq r_t(x,a) \quad \forall t<H, \forall x,a$$
$$V_t(x) \geq 0 \quad \forall t<H, \forall x,a$$

- 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 $\circledast$ in the dual to get

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

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. Any feasible value fn $V_t(x)$ induces a corresponding policy

$$\pi_t^V(x) = \operatorname*{argmax}_{a \in \mathcal{A}} \left[ r_t(x,a) + \textstyle\sum_{y \in \mathcal{S}} T_t(y|x,a)\, V_{t+1}(y) \right] \quad \forall t<H, \forall x$$

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

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

(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, ie, 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 $\pi$ over many trials $j \in \{1, 2, \dots, J\}$, where each trial terminates after $H^j \sim \text{Geom}(1-\gamma)$ rounds
- As before, $r(x,a) = \mathbb{E}[R(x,a)]$

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

  ↖ ie, $\frac{1}{J} \sum_{j=1}^{J} \sum_{t=0}^{H^j - 1} \mathbb{1}\{x_t^j = x, A_t^j = a\}$

  Also assume $X_0 \sim \pi_0$

  Then the MDP $\equiv$ following LP

Discounted MDP primal

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

s.t.

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

dual on $V(x)$

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

and its dual

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

s.t.

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

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

equivalently

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

discounted reward HJB eqn

- **Avg Reward** – Again we consider time-invariant MDPs. Given any stationary policy $\pi$, suppose we run it for a long time (i.e., $H \to \infty$)
  - Define $q(x,a) = \frac{1}{H}\sum_{t=1}^{H} \mathbb{1}\{x_t = x, A_t = a\}$
    $=$ Avg freq of playing action $a$ in state $x$
  - As before $r(x,a) = \mathbb{E}[R(x,a)]$. Also $X_0 \sim \pi$ does not matter by MC ergodic thm

Now the LP becomes

avg reward primal

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

s.t.

$V(x)$ | $\sum_{a\in\mathcal{A}} q(x,a) - \sum_{y\in\mathcal{S}} \sum_{a\in\mathcal{A}} q(y,a)\, T(x|y,a) = 0 \quad \forall x \in \mathcal{S}$

$\lambda$ | $\sum_{x\in\mathcal{S}} \sum_{a\in\mathcal{A}} q(x,a) = 1$

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

The corresponding dual LP is

$$\min \lambda$$

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

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

$\lambda$ 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

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

avg reward HJB equation

# Interpreting the HJB equations

From above we get the 3 HJB equations

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

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

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

avg reward HJB equation

Also, since in each case we minimize $V_0(x_0)$ (or $\lambda$), its 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 $\pi$ to be a mapping from $S$ to $A$ (in time invariant, else $\pi = \{\pi_t\}, \pi_t : S \to A$). Then

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

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

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

return time