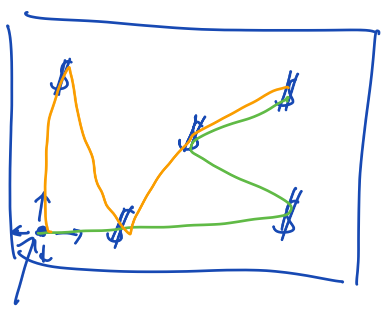
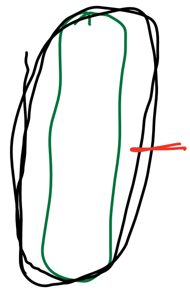
Markov Decision Processes (MDP)



Start at time $h=1$



Driving a car around a "wet" track.

Actions can move 1 unit left, right, down, up

For H timesteps you want to collect as much money as possible

---

MDP is defined by tuple $(\mathcal{S}, \mathcal{A}, \{P_h\}, \{r_h\}, H, \nu)$

- $\mathcal{S}$ state space, $S = |\mathcal{S}|$ is finite.
- $\mathcal{A}$ action space, $A = |\mathcal{A}|$ is finite
- Transition function $P_h : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$. At time $h \in [H]$ if I play action a_h is state s_h then

  $P_h(s' \mid s_h, a_h)$ is the probability that $s_{h+1} = s'$

- Reward function $r_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ At time $h \in [H]$ if I play action a_h is state s_h then I receive reward $r_h(s_h, a_h)$. Assumed known.
- Horizon length $H \in \mathbb{N}$

- Initial state dist. $\nu \in \Delta_S$ s.t. $s_1 \sim \nu$

A policy determines action given state and time h.

- Deterministic policy $\pi = \{\pi_h\}_{h=1}^H$, $\pi_h : \mathcal{S} \to \mathcal{A}$, $a_h = \pi_h(s_h)$
- Randomized policy $\pi = \{\pi_h\}_{h=1}^H$, $\pi_h : \mathcal{S} \to \Delta_{\mathcal{A}}$, $a_h \sim \pi_h(s_h)$

To evaluate a policy we can "roll it out"

- Draw $s_1 \sim \nu$, $a_h \sim \pi_h(s_h)$ for all $h \in [H]$

$$s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$$

Value of a policy, for any s, h

$$V_h^\pi(s) = \mathbb{E}\left[\sum_{t=h}^{H} r_t(s_t, a_t) \,\Big|\, \pi, s_h = s\right]$$

where expectation is taken wrt random transitions and potentially randomized policy.

$V_0^\pi = \mathbb{E}_{s_1 \sim \nu}\left[V_1^\pi(s_1)\right]$, Goal $\max_\pi V_0^\pi$.

note: $V_h^\pi(s) \in [0, H]$ for all h since $r_h(s,a) \in [0,1]$.

Define state-action value function of π @ h

$$Q_h^\pi(s,a) = \mathbb{E}\left[\sum_{t=h}^{H} r_t(s_t,a_t) \,\middle|\, \pi, s_h = s, a_h = a\right]$$

start a time h in state s, and play action a, but $t > h$ play $\pi_t(s_t) = a_t$.

Also note: $Q_h^\pi(s,a) \in [0,H]$.

<u>Theorem</u> (Bellman Optimality Equations) Define

$$Q_h^*(s,a) = \sup_\pi Q_h^\pi(s,a)$$

where sup over all randomized policies. For some function $Q_h : S \times A \to \mathbb{R}$, we have that $Q_h = Q_h^*$ for all $h \in [H]$ if and only if for all $h \in [H]$

$$Q_h(s,a) = r_h(s,a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)}\left[\max_{a'} Q_{h+1}(s',a')\right]$$

where $Q_{H+1}(s,a) = 0$. Furthermore $\pi_h(s) = \operatorname{argmax}_a Q_h(s,a)$ is an optimal policy.

Great, how do we find such a Q_h?

Value iteration:

- Set $Q_H(s,a) = r_H(s,a)$
- For $h = H-1, H-2, \dots, 1$

$$Q_h(s,a) = r_h(s,a) + \mathbb{E}_{s' \sim P_h(\cdot | s,a)}\left[\max_{a'} Q_{h+1}(s',a')\right]$$

Infinite horizon MDP, w/ discounts.

Fix $\gamma \in (0,1)$ the discounted value

$$V^\pi(s) = \mathbb{E}\left[\sum_{h=1}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, \pi, s_1 = s\right]$$

Optimality equation

$$Q(s,a) \overset{(*)}{=} r(s,a) + \gamma\, \mathbb{E}_{s' \sim P(\cdot | s,a)}\left[\max_{a'} Q(s',a')\right]$$

Value iteration:

- Init $Q^0(s,a)$ arbitrarily
- $Q^{k+1}(s,a) = r(s,a) + \gamma\, \mathbb{E}_{s'}\left[\max_{a'} Q^k(s',a')\right] =: T(Q^k)$

By defn Q^* satisfies $(*)$

so $T(Q^*) = Q^*$.

Can show $|T(Q^k) - Q^*| = |T(Q^k) - T(Q^*)|$

$$\leq \gamma\, |Q^k - Q^*|$$

$$\leq \gamma^k |Q^0 - Q^*|$$