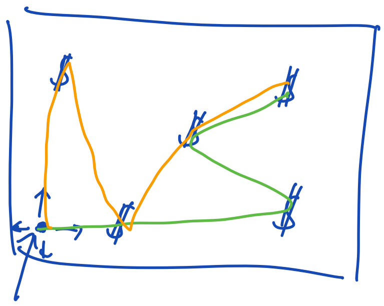
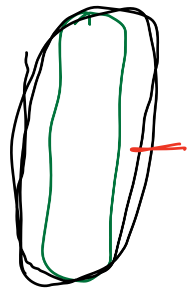
Markov Decision Processes (MDP)



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

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

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 $\pi_h(s) = \arg\max_a Q_h(s,a)$ 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 $h = H-1, H-2, \ldots, 1$

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

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

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

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

From now on $\{P_h\}$ are <u>unknown</u>.

Contextual-Bandits, treating each context individually as a bandit, is equal to MDP w/ $H=1$.

$s_1 \sim \nu$ (interpret s_1 as context)

playing action a_1 in response to s_1 achieves reward $r_{t_1} \sim \mathbb{E}[r_t(s_1, a_1) | a, s] = v(s,a)$

Goal $\max_\pi \mathbb{E}_{s_1 \sim \nu}[v(s_1, \pi(s_1))]$.

Given $\{\pi_h\}_{h=1}^H$ how can I estimate

$$V_0^\pi = \mathbb{E}_{s_1 \sim \nu}[V_1^\pi(s_1)] \; ?$$

To construct estimator $\hat{V}_0^{\pi}$ one idea is to collect data randomly:

Regardless of what state I'm in,

choose action uniformly at random.

Worked for CB ($H=1$). Work for MDP?

No.

Consider an MDP as follows:

A actions

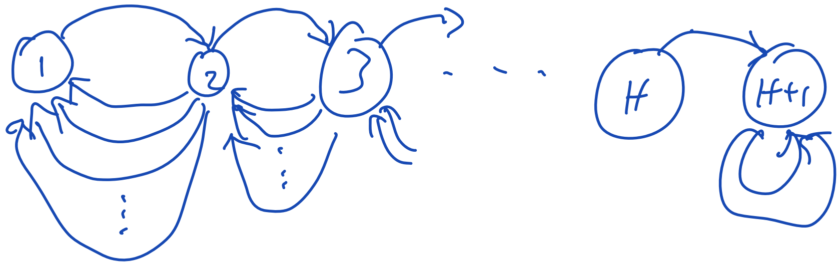
$H+1$ states

$S_1 = \{1\}$

$$P_h(s_{h+1} = \min\{H, i+1\} \mid s_h = i, a_h = 1) = 1$$

$$P_h(s_{h+1} = \max\{1, i-1\} \mid s_h = i, a_h = i) = 1 \quad \text{for } i > 1$$

$r_H(H+1, a) = 1$, $r_h(s,a) = 0$ otherwise.



$\pi_\star(s) = 1$ for all s, h

$V_0^{\pi_\star} = 1$, $V_0^{\pi} = 0 \quad \forall \pi \neq \pi_\star$

If only 1 action among A has to be taken H times in a row.

what is the prob. of random exploration discovering $\pi_\star$? $\quad A^{-H}$

Claim: There exists an algorithm that can identify $\pi^\star$ in just $O(SAH^4)$ episodes w/ constant prob.

UCB-VI (UCB value iteration)

UCB-VI for Reinforcement Learning

Input: deterministic reward functions $r_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ for all $h \in [H]$, $\delta \in (0,1)$

Initialize: For all $\ell \in \mathbb{N}$ let

$$n_h^\ell(s,a,s') = \textstyle\sum_{i=1}^{\ell-1} \mathbf{1}\{(s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\},$$
$$n_h^\ell(s,a) = \textstyle\sum_{i=1}^{\ell-1} \mathbf{1}\{(s_h^i, a_h^i) = (s,a)\},$$
$$\widehat{P}_h^\ell(s'|s,a) = n_h^\ell(s,a,s')/n_h^\ell(s,a)$$

for $k = 1, 2, \ldots, K$

$\quad \widehat{V}_{H+1}^k = \mathbf{0} \in \mathbb{R}^S$

$\quad$ **for** $h = H, H-1, \ldots, 1$

$$\widehat{Q}_h^k(s,a) = \min\left\{H,\ H\sqrt{\frac{\log(2KHSA/\delta)}{2n_h^k(s,a)}} + r_h(s,a) + \widehat{P}_h^{k(s,a)} \cdot V_{h+1}^k\right\}$$

$$\widehat{V}_h^k(s) = \max_a \widehat{Q}_h^k(s,a) \text{ and } \pi_h^k(s) \arg\max_a \widehat{Q}_h^k(s,a)$$

$\quad$ Roll-out $\{\pi_h^k\}$ such that $s_1 \sim \nu$ and $a_h^k = \pi_h^k(s_h)$ and $s_{h+1} \sim P_h(\cdot|s_h^k, a_h^k)$ for all $h \in [H]$

$$Q_H^{\pi_*}(s,a) = r_H(s,a)$$

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

If satisfied then $V_{h+1}(s') = \max_{a'} Q_h(s',a')$

$$\widehat{P}_h(s,a) \in \Delta_S$$

$$[\widehat{P}_h(s,a)]_{s'} := \widehat{P}_h(s'|s,a)$$

Note: As $n_h(s,a) \to \infty$ $\quad \widehat{P}_h^k \to P_h \Rightarrow \widehat{V}_h^k \to V_h^*$

$$\text{Regret} = \mathbb{E}\left[K V_0^{\pi_*} - \sum_{k=1}^{K} \sum_{h=1}^{H} r_h(s_h^k, a_h^k) \right]$$

Want to show Regret $\leq O(\sqrt{K})$

$$\mathcal{E}_{\text{optimism}} = \bigcap_{s,a} \bigcap_{k=1}^{K} \bigcap_{h=1}^{H} \left\{ \left| \sum_{s'} \left(P_h(s'|s,a) - \hat{P}_h^k(s'|s,a) \right) V_{h+1}^{\pi_*}(s') \right| \leq H \sqrt{\frac{\log(2KHSA/\delta)}{2 n_h^k(s,a)}} \right\}$$

For any $V \in [0,H]^S$

$$n_h^k(s,a) \sum_{s'} \left(P_h(s'|s,a) - \hat{P}_h^k(s'|s,a) \right) V(s')$$

$$= \sum_{i=1}^{k-1} \underbrace{\mathbb{1}\{a_h^i = a, s_h^i = s\} \sum_{s'} \left(P_h(s'|s,a) - \mathbb{1}\{s' = s_{h+1}^i\} \right) V(s')}_{=: X_i}$$

$$\mathbb{E}[X_i \mid a_h^i, s_h^i] = 0, \quad |X_i| \leq H \mathbb{1}\{a_h^i = a, s_h^i = s\}$$

By martingale bounds w/ predictable seq the result follows. (Azuma-Hoeffding)

Lemma For any $V \in [0,H]^S$ we have for any (s,a,h,k)

$$\mathbb{P}\left(\left| \sum_{s'} (P_h(s'|s,a) - \hat{P}_h^k(s'|s,a)) V(s') \right| \le H \sqrt{\frac{\log(2/\delta)}{2 n_h^k(s,a)}} \right) \ge 1-\delta.$$

$\mathcal{E}_{\text{optimism}}$ follows by a union bound over all $S \times A \times [H] \times [K]$

Lemma On event $\mathcal{E}_{\text{optimism}}$ (which holds w.p. $\ge 1-\delta$) we have $\hat{V}_h^k(s) \ge V_h^{\pi_\star}(s)$ and $\hat{Q}_h^k(s,a) \ge Q_h^{\pi_\star}(s,a)$ for all h, s, a.

Proof By induction. Trivially we have $\hat{Q}_H^k(s,a) \ge Q_H^{\pi_\star}(s,a)$.

$$\begin{aligned} \hat{V}_H^k(s) = \max_a \hat{Q}_H^k(s,a) &\ge \hat{Q}_H^k(s, \pi_\star(s)) \\ &\ge Q_H^{\pi_\star}(s, \pi_\star(s)) \\ &= V_H^{\pi_\star}(s) \end{aligned}$$

$$\Longrightarrow \hat{V}_H^k(s) \ge V_H^{\pi_\star}(s).$$

So assume $\hat{V}_{h+1}^k(s) \ge V_{h+1}^{\pi_\star}(s)$ and show this implies $\hat{V}_h^k(s) \ge V_h^{\pi_\star}(s)$.

$$Q_h^{\pi_*}(s,a) = r_h(s,a) + \sum_{s'} P_h(s'|s,a) V_{h+1}^{\pi_*}(s')$$

$$= r_h(s,a) + \hat{P}_h^k(s,a)^T V_{h+1}^{\pi_*} + \left(P_h(s,a) - \hat{P}_h^k(s,a)\right)^T V_{h+1}^{\pi_*}$$

($\mathcal{E}_{conf}$)

$$\leq r_h(s,a) + \hat{P}_h^k(s,a)^T V_{h+1}^{\pi_*} + H\sqrt{\frac{\log(2HKSA/\delta)}{2n_h^k(s,a)}}$$

(induction)

$$\leq r_h(s,a) + \hat{P}_h^k(s,a)^T \hat{V}_{h+1}^k + H\sqrt{\frac{\log(2HKSA/\delta)}{2n_h^k(s,a)}}$$

$$= \hat{Q}_h^k(s,a)$$

$$\Rightarrow \quad \hat{Q}_h^k(s,a) \geq Q_h^{\pi_*}(s,a).$$

By same seq. of steps for Q_H, it follows that $\hat{V}_h^k(s) \geq V_h^{\pi_*}(s)$. ∎

Use optimism to bound regret @ episode k:

$$V_0^{\pi_*} - V_0^{\pi_k} = \mathbb{E}_{s_1 \sim \gamma}\left[V_1^{\pi_*}(s_1) - V_1^{\pi_k}(s_1) \right]$$

$$\leq \mathbb{E}_{s_1}\left[\hat{V}_1^k(s_1) - V_1^{\pi_k}(s_1) \right]$$

$$\hat{V}_h^k(s) = \max_a \hat{Q}_h^k(s,a) = \hat{Q}_h^k(s, \pi^k(s))$$

$$= \mathbb{E}_{s_1}\left[\hat{Q}_1^k(s_1, \pi_k(s_1)) - r_1(s_1, \pi_k(s_1)) - P_1(s_1, \pi_k(s_1))^T V_2^{\pi_k} \right]$$

$s_2 \sim P_1(s_1, \pi_k(s_1))$ w/ $\mathbb{P}(s_2 = s') = P_1(s' | s_1, \pi_k(s_1))$

and expected reward of $V_2^{\pi_k}(s_2)$

so $\mathbb{E}\left[V_2^{\pi_k}(s_2) \right] = P_1(s_1, \pi_k(s_1))^T V_2^{\pi_k}$

$$= \mathbb{E}_{s_1}\left[H\sqrt{\frac{\log(2HKSA/\delta)}{2n_1^k(s_1, \pi_k(s_1))}} + \hat{P}_1^k(s_1, \pi_k(s_1))^T \hat{V}_2^k - P_1(s_1, \pi_k(s_1))^T V_2^{\pi_k} \right]$$

$$= \mathbb{E}_{s_1}\left[H\sqrt{\frac{\log(2HKSA/\delta)}{2n_1^k(s_1, \pi_k(s_1))}} + \left(\hat{P}_1^k(s_1, \pi_k(s_1)) - P_1(s_1, \pi_k(s_1)) \right)^T \hat{V}_2^k + P_1(s_1, \pi_k(s_1))^T \left(\hat{V}_2^k - V_2^{\pi_k} \right) \right]$$

$$\mathbb{E}_{s_1}\left[P_1(s_1, \pi_k(s_1))^T \left(\hat{V}_2^k - V_2^{\pi_k} \right) \right] = \mathbb{E}_{s_1, s_2 \sim P_1(s_1, \pi_k(s_1))}\left[\hat{V}_2^k(s_2) - V_2^{\pi_k}(s_2) \right]$$

$$= \underline{\sum_{h=1}^{H} \mathbb{E}_{\pi_k}\left[H\sqrt{\frac{\log(2HKSA/\delta)}{2n_h^k(s_h, \pi_k(s_h))}} + \left(\hat{P}_h^k(s_h, \pi_k(s_h)) - P_h(s_h, \pi_k(s_h)) \right)^T \hat{V}_h^k \right]}$$

Fix s, a, h, k

$$\sup_{V \in [0,H]^S} \left| \left(\hat{P}_h^k(s,a) - P_h(s,a) \right)^T V \right|$$

$$= \max_{V \in \{0,H\}^S} \left| \left(\hat{P}_h^k(s,a) - P_h(s,a) \right)^T V \right|$$

$$\leq H \sqrt{\frac{\log(2^S \cdot 2/\delta)}{2 n_h^k(s,a)}}$$

where 2^S comes from union bound over all $V \in \{0,H\}^S$.

$$\leq H \sqrt{\frac{S \log(2/\delta)}{2 n_h^k(s,a)}}$$

$$\mathcal{E}_{\text{complex}} = \bigcap_h \bigcap_{k=1}^K \bigwedge_{s,a} \left\{ \sup_{V \in [0,H]^S} \left| \left(\hat{P}_h^k(s,a) - P_h(s,a) \right)^T V \right| \leq H \sqrt{\frac{S \log(2HKSA/\delta)}{2 n_h^k(s,a)}} \right\}$$

$$\mathbb{P}(\mathcal{E}_{\text{complexity}}) \geq 1 - \delta.$$

Conclude that

$$V_0^{\pi_*} - V_0^{\pi_k} \overset{(\mathcal{E}_{\text{opt}})}{\leq} \sum_{h=1}^H \mathbb{E}_{\pi_k} \left[H \sqrt{\frac{\log(2HKSA/\delta)}{2 n_h^k(s_h, \pi_k(s_h))}} + \underline{\left(\hat{P}_h^k(s_h, \pi_k(s_h)) - P_h(s_h, \pi_k(s_h)) \right)^T} \hat{V}_h^k \right]$$

$$\overset{(\mathcal{E}_{\text{comp}})}{\leq} \sum_{h=1}^H \mathbb{E}_{\pi_k} \left[H \sqrt{\frac{2 S \log(2HKSA/\delta)}{n_h^k(s,a)}} \right].$$

$$\text{Regret} \le \sum_{k=K} V_0^{\pi_*} - V_0^{\pi_k} \qquad \sum_{i=1}^{n} \frac{1}{\sqrt{i}} \le 2\sqrt{n}$$

$$\le \sum_{k=K}^{H} \sum_{h=1}^{H} \mathbb{E}_{\pi_k}\left[H \sqrt{\frac{2S\log(2HKSA/\delta)}{n_h^k(s_h^k, a_h^k)}} \right]$$

$$= H\sqrt{2S\log(2HKSA/\delta)} \sum_{h=1}^{H} \mathbb{E}_{\pi_k}\left[\sum_{k=1}^{K} \frac{1}{\sqrt{n_h^k(s_h^k, a_h^k)}} \right]$$

$$= H\sqrt{2S\log(2HKSA/\delta)} \sum_{h=1}^{H} \mathbb{E}_{\pi_k}\left[\sum_{s,a} \sum_{k=1}^{K} \mathbb{1}\{(s,a) = (s_h^k, a_h^k)\} \frac{1}{\sqrt{n_h^k(s,a)}} \right]$$

$$= H\sqrt{2S\log(2HKSA/\delta)} \sum_{h=1}^{H} \mathbb{E}_{\pi_k}\left[\sum_{s,a} \sum_{i=1}^{n_h^K(s,a)} \frac{1}{\sqrt{i}} \right]$$

$$\le 2H\sqrt{2S\log(2HKSA/\delta)} \sum_{h=1}^{H} \mathbb{E}_{\pi_k}\left[\sum_{s,a} \sqrt{n_h^K(s,a)} \right]$$

$$\le 2H\sqrt{2S\log(2HKSA/\delta)} \sum_{h=1}^{H} \mathbb{E}_{\pi_k}\left[\sqrt{SA \cdot \underbrace{\sum_{s,a} n_h^K(s,a)}_{=K}} \right]$$

$$\le \underline{H^2 S \sqrt{8A}} K \log(2HKSA/\delta)$$

$$\sum_{i=1}^{n} a_i b_i = \langle a, b \rangle \leq \sqrt{\left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right)}$$

Turns out, same algorithm but tighter analysis yields $H^2\sqrt{SAK} + S^2 A \,\mathrm{poly}(H)$.