Stochastic Multi-Armed Bandit

$m \rightarrow$ No. of arms

$A = \{a_1, a_2, \ldots a_m\}$ Arms

Bernoulli (0,1) rewards

Mean ~~arm~~ reward of arm a_i $\odot = P_i \in [0,1]$.

T: Horizon/budget

Highest mean $= P_*$; optimal arm a_*.

Algorithm:-

a^0, r^0 $\quad r^0 \in$ Bernoulli (P_{a^0})

$a^1 \; r^1$ $\quad r^1 \in$ Bernoulli (P_{a^1})

$\vdots$

$a^{t-1} . r^{t-1}$ | Which arm to pull at t?

An algorithm is a mapping from the set of histories to the set of arms:

- Probability distribution over arms (A) (randomized)
- Deterministic.

① ε-first $\quad \varepsilon \in [0,1]$

Pull arms uniformly at random (or round robin) for εT steps.

- Identify a_{best} an arm with the highest empirical mean.
- For the remaining pulls, sample a_{best}

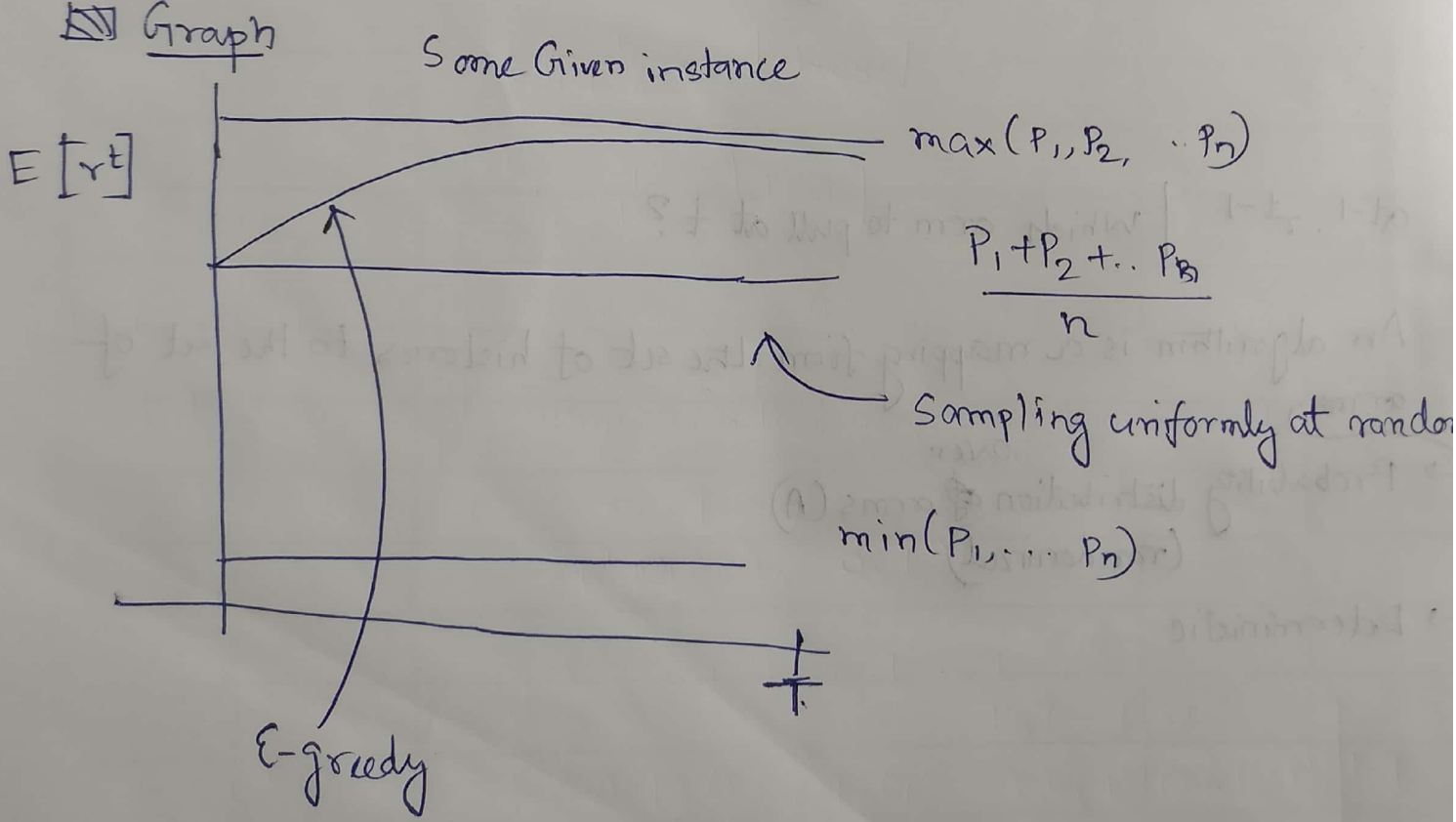
ε-greedy

For $t \in 0, 1, 3 \ldots$

With probability ε, pick an arm uniformly at random to pull.

With probability $1-\varepsilon$, pull the arm with the highest empirical mean.

Graph

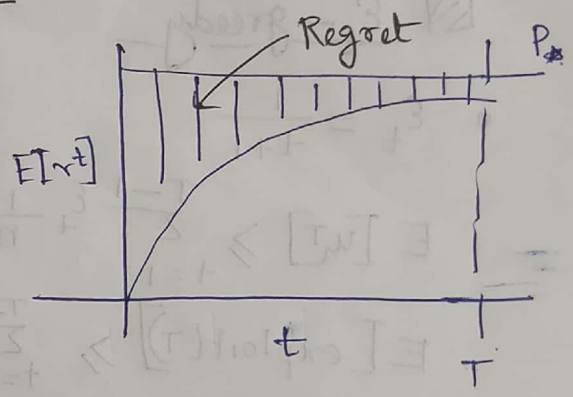


Will the ε-greedy line meet the max line at $t = \infty$

$$E(r_\infty) = \varepsilon\left(\frac{P_1 + P_2 + \ldots P_n}{n}\right) + (1-\varepsilon).\max(P_1, P_2, \ldots P_n)$$

Expected ~~cumulative~~ v. cumulative regret:

$$R^T = P_* T - \sum_{t=0}^{T-1} E[r^t]$$



Is it possible to achieve a sub-linear regret:

Can we design an algorithm for which $R^T = o(T)$ or equivalently,

$$\lim_{T \to \infty} \frac{R^T}{T} = 0$$

Let u_a^T denote the number of pulls of a in T pulls.

Let exploit(T) denote the number of times in T pulls that an arm with the highest empirical average get pulled.

For sub-linear regret we need:

$\forall a \in A$, $\lim_{T \to \infty} E[u_a^T] = 0$ [Doing Infinite Exploration]

$\lim_{T \to \infty} \frac{E[\text{exploit}(T)]}{T} = 1$. [Greedy in the Limit]

ε_t-greedy

- $\varepsilon_t = \frac{1}{t+1}$

$$E[u_a^T] \geq \sum_{t=1}^{T-1} \varepsilon_t \frac{1}{n} = \frac{1}{n} \sum_{t=0}^{T-1} \frac{1}{t+1} \geq c_1 \log T.$$

$$E[\text{exploit}(T)] \geq \sum_{t=0}^{T-1} (1-\varepsilon_t) + \frac{\varepsilon_t}{n}$$

$$\geq T - c_2 \log T$$

- $\varepsilon_t = e^{-t}$

$$E[u_a^T] \geq \sum_{t=0}^{T-1} \frac{e^{-t}}{n}$$

$$= \frac{1}{n} \left(1 + e^{-1} + e^{-2} + e^{-3} \ldots\right)$$

$$= \frac{1}{n} \cdot \frac{1}{1-e^{-1}} \quad [\text{Finite}]$$

4. Lower Bound (Lai and Robbins, 1985)

Let L be an algorithm such that for every bandit instance I and for every $\alpha > 0$, as $T \to \infty$

$$R^T(L, I) = o(T^\alpha)$$

Then, for every bandit instance I as $T \to \infty$

$$R^T(L, I) \geq \left[\sum_{a \in A : P_a(I) \neq P_*(I)} \frac{P_* T - P_a(I)}{KL(P_a(I), P_*(I))} \right] \ln T$$

where
$KL(x,y) = x \ln \frac{x}{y} + (1-x) \ln \frac{1-x}{1-y}$

UCB Algorithm

- Sample every arm at least once.
- For $t = n, n+1, \ldots$

Let $\hat{p}_a^t$ denote the empirical mean of arm a
Let u_a^t be a's number of pulls
Define $ucb_a^t = p_a^t + \sqrt{\frac{2}{u_a^t} \ln t}$

Pull $\underset{a \in A}{\operatorname{argmax}}\ ucb_a^t$

$$R^T = O\left(\left(\sum_{\substack{a \\ P_a \neq P^*}} \frac{1}{P^* - P_a}\right) \log T\right) \quad \text{for UCB}$$

Hoeffding's Inequality (1967)

Let X be a random variable that takes values in $[a,b]$, with $E[X] = p$. Fix $u \geq 1$ and $\varepsilon > 0$. Let $r^1, r^2, \ldots r^u$ be samples drawn i.i.d from X and let $\hat{p} = \frac{\sum_{i=1}^{u} r^i}{u}$

Then:

$$P\{\hat{p} \geq p + \epsilon\} \leq e^{-\frac{2u\varepsilon^2}{(b-a)^2}} \quad \text{and} \quad P\{\hat{p} \leq p - \epsilon\} \leq e^{-\frac{2u\varepsilon^2}{(b-a)^2}}$$

③ KL-UCB (2011)

Run UCB as above, but use:

$ucb_a^t \stackrel{def}{=}$ the largest element $q \in [\hat{p}_a^t, 1]$ that satisfies

$$KL(\hat{p}_a^t, q) \leq \frac{\ln t + 3 \ln \ln t}{u_a^t}$$

$$R^T = O\left(\sum_{a: p_a \neq p^*} \frac{p^* - p_a}{KL(p_a, p^*)} \ln T\right)$$

$q \leq \hat{p} + \sqrt{\frac{2}{u} \ln t}$ this also holds.

$KL(p_a, p^*) \geq 2(p^* - p_a)^2$

Pinsker's Inequality

Thompson Sampling (1933, 2011-12)

For $t = 0, 1, 2, \ldots$

For $a \in A$

Let s_A^t denote the number of "successes". (1-reward)

Let f_a^t denote the number of "failures". (0-reward)

Draw $x_a^t \sim Beta(s_a^t + 1, f_a^t + 1)$

Pull $\underset{a \in A}{\operatorname{argmax}}\ x_a^t$

Mean $\frac{\alpha}{\alpha+\beta}$, Variance $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

$\Delta_a \overset{def}{=} P_* - P_a$

Z_a^t be the event that arm a is pulled at time t.

Let z_a^t be a random variable that takes value 1 if arm a is pulled a time t and 0 otherwise.

$$E[z_a^t] = P\{Z_a^t\}(1) + (1 - P\{z_a^t\}(0) = P\{Z_a^t\}$$

$$u_a^t \overset{def}{=} \sum_{i=0}^{t-1} z_a^t$$

$$\overline{u}_a^t \overset{def}{=} \left[\frac{8}{(\Delta_a)^2} \ln(T)\right]$$

$$R^T = T P_* - \sum_{t=0}^{T-1} E[r^t]$$

$$= T P_* - \sum_{t=0}^{T-1} \sum_{a \in A} P(z_a^t) \cdot E[r^t \mid z_a^t]$$

$$= T P_* - \sum_{t=0}^{T-1} \sum_{a \in A} E[z_a^t] P_a$$

$$= \left(\sum_{a \in A} E[u_a^T]\right) P_* - \sum_{a \in A} E[u_a^T] P_a$$

$$= \sum_{\substack{a \in A \\ P_a \neq P_*}} E[u_a^T] (P_* - P_a)$$

$$= \sum_{\substack{a \in A \\ P_a \neq P_*}} E[u_a^T] \Delta_a$$

$$E[u_a^T] = \sum_{t=0}^{T-1} E[z_a^t]$$

$$= \sum_{t=0}^{T-1} \mathbb{P}\{z_a^t \text{ and } u_a^t < \bar{u}_a^t)\} + \sum_{t=0}^{T-1} \mathbb{P}\{z_a^t \text{ and } (u_a^t \geq \bar{u}_a^T)\}$$

$$= A + B$$

$$A = \sum_{t=0}^{T-1} \mathbb{P}\{z_a^t \text{ and } (u_a^t < \bar{u}_a^T)\}$$

$$= \sum_{t=0}^{T-1} \sum_{m=0}^{\bar{u}_a^T - 1} \mathbb{P}\{z_a^t \text{ and } (u_a^t = m)\}$$

$$= \sum_{m=0}^{\bar{u}_a^T - 1} \sum_{t=0}^{T-1} \mathbb{P}\{z_a^t \text{ and } (u_a^t = m)\}$$

$$= \sum_{m=0}^{\bar{u}_a^T - 1} \mathbb{P}\{(z_a^0 \text{ and } (u_a^0 = m)) \text{ or } (z_a^1 \text{ and } (u_a^1 = m)) \ldots\}$$

$$\leq \sum_{m=0}^{\bar{u}_a^T - 1} 1$$

$$\leq \bar{u}_a^T$$

$$B = \sum_{t=0}^{T-1} \mathbb{P}\{z_a^t \text{ and } (u_a^t \geq \bar{u}_a^T)\}$$

$$\leq \sum_{t=0}^{T-1} \mathbb{P}\left\{\left(\hat{p}_a^t + \sqrt{\frac{2}{u_a^t} \ln(t)} \geq \hat{p}_a^t + \sqrt{\frac{2}{u_a^t} \ln(t)}\right) \text{ and } (u_a^t \geq u_a^T)\right\}$$

$$\leq \sum_{t=0}^{T-1} \sum_{x=\bar{u}_a^T}^{t} \sum_{y=1}^{t} \mathbb{P}\left\{\hat{p}_a(x) + \sqrt{\frac{2}{x} \ln t} \geq \hat{p}_a(y) + \sqrt{\frac{2}{y} \ln(t)}\right\}$$

$$\hat{P}_a(x) + \sqrt{\frac{2}{x}\ln t} \geq \hat{P}_*(y) + \sqrt{\frac{2}{y}\ln(t)}$$

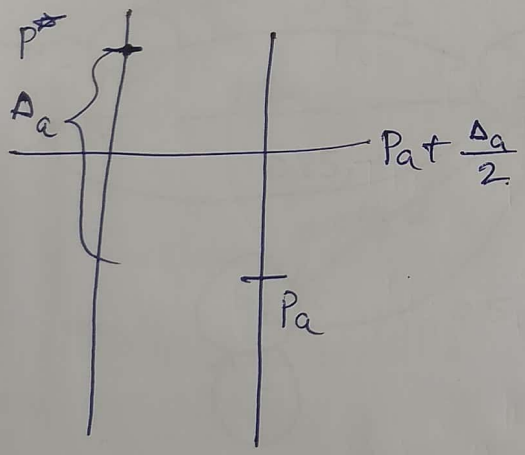
$$\Rightarrow \left(\hat{P}_a(x) + \sqrt{\frac{2}{x}\ln(t)} \geq P_*\right) \text{ or } \left(\hat{P}_*(y) + \sqrt{\frac{2}{y}\ln(t)} < P_*\right)$$

$x \geq \overline{u}_a T$, we have $\sqrt{\frac{2}{x}\ln t} \leq \sqrt{\frac{2}{\overline{u}_a T}\ln(t)} \leq \frac{\Delta_a}{2}$ and so,

$$\leq \sqrt{\frac{2\Delta_a^2}{8_4 \ln T}\ln t}$$

$$\leq \frac{\Delta_a}{2}$$

$$\Rightarrow \hat{P}_a(x) + \sqrt{\frac{2}{x}\ln(t)} \geq P_* \Rightarrow \hat{P}_a(x) \geq P_a + \frac{\Delta_a}{2}$$



In summary:-

$$\hat{P}_a(x) + \sqrt{\frac{2}{x}\ln t} \geq \hat{P}_*(y) + \sqrt{\frac{2}{y}\ln(t)} \geq \left(\hat{P}_a(x) \geq P_a + \frac{\Delta_a}{2}\right)$$
$$\text{or } \left(\hat{P}_*(y) < P_* - \sqrt{\frac{2}{y}\ln(t)}\right)$$

$$B \leq \sum_{t=0}^{T-1} \sum_{x=u_a T}^{t} \sum_{y=1}^{t} \mathbb{P}\left\{\hat{P}_a(x) \geq P_a + \frac{\Delta a}{2}\right\} + \mathbb{P}\left\{\hat{P}_{a^*}(y) \leq P_{a^*} - \sqrt{\frac{2}{y}\ln t}\right\}$$

$$\leq \sum_{t=0}^{T-1} \sum_{x=u_a T}^{t} \sum_{y=1}^{t} \left(e^{-2x\left(\frac{\Delta a}{2}\right)^2} + e^{-2y\left(\sqrt{\frac{2}{y}\ln t}\right)^2}\right)$$

$$\leq \sum_{t=0}^{T-1} \sum_{x=u_a T}^{t} \sum_{y=1}^{t} \left(e^{-4\ln t} + e^{-4\ln t}\right)$$

$$\leq \sum_{t=0}^{T-1} t^2 \left(\frac{1}{t^4}\right)$$

$$\leq \sum_{t=0}^{T-1} \frac{1}{t^2} = \frac{\pi^2}{3}$$

$$1 + \frac{1}{2^2} + \frac{1}{3^2} + \ldots = \frac{\pi^2}{3}$$

Markov Decision Problems

S — set of states

A — set of Actions — $A = \{a_1, a_2\}$

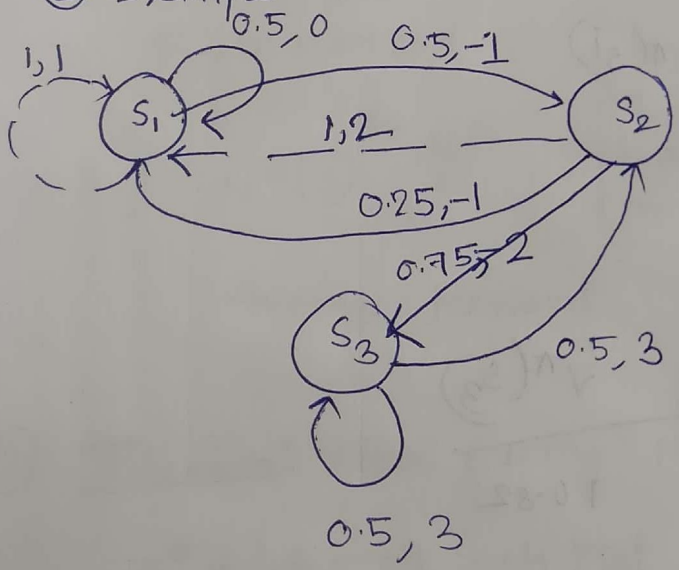
T — transition function

T: $S \times A \longrightarrow$ set of probability distribution over S.

R — Reward function — $R: S \times A \times S \longrightarrow [-R_{max}, R_{max}]$

γ — Discount factor — $\gamma \in [0, 1)$

② Example



③ Policy $\pi: S \longrightarrow A$ → Not a part of the MDP.

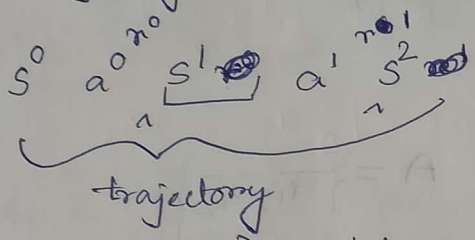
Deterministic

Markovian :- Action depends on only State

Stationary :- Not changing with time

~~Determini~~

④ Trajectory over time



π

$a^t = \pi(s^t)$ $\quad s^{t+1} \sim T(s^t, a^t)$ $\quad r^t = R(s^t, a^t, s^{t+1})$

⑤ Value function of π

$$V^\pi : S \longrightarrow \mathbb{R}$$

$$V^\pi(s) \stackrel{def}{=} \mathbb{E}\left[r^0 + \gamma r^1 + \gamma^2 r^2 + \cdots \right]$$

$s^0 = s, \forall i = 0, 1, 2, \ldots \quad : a^i = \pi(s^i)$

⑥ Evaluation

π	$V^\pi(s_1)$	$V^\pi(s_2)$	$V^\pi(s_3)$
111	4.45	6.55	10.82
112	−5.61	−5.75	−4.05
121	2.76	4.48	9.12
122	2.76	4.48	3.48
211	10	9.34	13.1
212	10	7.25	10
221	10	11	14.45
222	10	11	10

Continuing tasks and episodic tasks

i) $V^{\pi}(s) \overset{\text{def}}{=} E_{\pi}[r^0 + \gamma r^1 + \gamma^2 r^2 + \dots \infty \mid s^0 = s]$

Infinite discounted reward.

ii) $V^{\pi}(s) \overset{\text{def}}{=} E_{\pi}[r^0 + r^1 + \dots \mid s^0 = s]$

Total reward.

Episodic tasks → Have a final state

iii) $V^{\pi}(s) \overset{\text{def}}{=} E_{\pi}[r^0 + r^1 + \dots r^{H} \mid s^0 = s]$

Horizon : specified along with the MDP

Finite horizon

iv) $V^{\pi}(s) \overset{\text{def}}{=} E\left[\lim_{m \to \infty} \frac{r^0 + r^1 + \dots r^{m-1}}{m}\right]$

Average Reward

(SATRγ)

② Theorem: For every MDP, there exists a policy π^*

$\pi^* : S \to A$ such that $\nexists s \in S, \forall \pi \in \Pi$.

$$V^{\pi^*}(s) \geq V^{\pi}(s)$$

π^* is an optimal policy. There can be many optimal policies, but they all have the same value function.

Finite horizon reward :-

$\pi^* : S \times \{1, 2, \dots H\} \to A$

Policy Evaluation

$$V^\pi(s) = E_\pi\left[r^0 + \gamma r^1 + \dots + \gamma^n \dots \mid s^0 = s\right]$$

$$= E_\pi\left[r^0 + s^0 \mid = s\right] + \gamma E\left[r^1 + \gamma r^2 + \dots \mid s^0 = s\right]$$

$$= \sum_{\bar{s} \in S} T(s, \pi(s), \bar{s}) \cdot R(s, \pi(s), \bar{s}) + \gamma \sum_{s \in S} T(s, \pi(s), \bar{s}) E_\pi\left[r^1 + \gamma r^2 + \gamma^2 r^3 + \dots \mid s^0 = s, s_1 = \bar{s}\right]$$

$$= \sum_{\bar{s} \in S} T(s, \pi(s), \bar{s}) \cdot R(s, \pi(s), \bar{s}) + \gamma \sum_{s \in S} T(s, \pi(s), \bar{s}) E_\pi\left[r^1 + \gamma r^2 + \gamma^2 r^3 \mid s^1 = \bar{s}\right]$$

$$= \sum_{\bar{s} \in S} T(s, \pi(s), \bar{s}) R(s, \pi(s), \bar{s}) + \gamma \sum_{s \in S} T(s, \pi(s), \bar{s}) E_\pi\left[r^0 + \gamma r^1 + \gamma^2 r^2 + \dots \mid s^0 = s\right]$$

$\forall s \in S$:

$$V^\pi(s) = \sum_{\bar{s} \in S} T(s, \pi(s), \bar{s}) \left[R(s, \pi(s), \bar{s}) + \gamma V^\pi(\bar{s})\right]$$

$|s|$ equations in $|s|$ unknown

Bellman's Equations

$$AX = B$$

Bellmann's equations are guaranteed to have solutions.

Action value function

$$q^{\pi}: S \times A \longrightarrow \mathbb{R}$$

$$q^{\pi}(s,a) = \sum_{s' \in S} T(s,a,s') \{ R(s,a,s') + \gamma V^{\pi}(s') \}$$

① Bellman's Equations :-

$$\forall s \in S, \forall \pi \in \Pi$$

$$V^{\pi}(s) = \sum_{s' \in S} T(s,\pi(s),s') \{ R(s,\pi(s),s') + \gamma V^{\pi}(s') \}$$

$$\forall s, \forall a \in A, \forall \pi \in \Pi.$$

$$q^{\pi}(s,a) = \sum_{s' \in S} T(s,a,s') \{ R(s,a,s') + \gamma V^{\pi}(s') \}$$

Policy Evaluation

② Bellman's Optimality Equations

$$\forall s \in S$$

$$V^{*}(s) = \max_{a \in A} \sum_{s' \in S} T(s,a,s') \{ R(s,a,s') + \gamma V^{*}(s') \}$$

$$q^{*}(s,a)$$

$|S| = n$ equations in n unknowns

$$V^{*} \longrightarrow q^{*} \longrightarrow \pi^{*}$$

$$\pi^{*} \longrightarrow V^{*} \longrightarrow q^{*}$$

Bellman Operator

For $\pi \in \Pi$, $B^{\pi} : (S \to \mathbb{R}) \to (S \to \mathbb{R})$

is defined as follows. Let $X : S \to R$

Then

$$(B^{\pi}(X))(s) \stackrel{\text{def}}{=} \sum_{s' \in S} T(s, \pi(s), s') \{ R(s, \pi(s), s') + \gamma X(s') \}$$

V^{π} is the fixed point of the Bellmann Operator.

Bellman Optimality Operator

$B^{*} : (S \to \mathbb{R}) \to (S \to \mathbb{R})$ is defined as follows.

Let $X : S \to \mathbb{R}$

$$(B^{*}(X))(s) \stackrel{\text{def}}{=} \max_{a \in A} \sum_{s' \in S} T(s, a, s') \{ R(s, a, s') + \gamma X(s') \}$$

Value Iteration

$$\lim_{t \to \infty} (B^{*})^{t}(X) = V^{*}$$

Let $(V, \|\cdot\|)$ be a normal vector space. A mapping $T: V \to V$ is called a contraction mapping if there exists $L < 1$ such that

$\forall u, v \in V$

$\|Tu - Tv\| \leq L\|u - v\|$.

- **Banach's Fixed Point Theorem:-**

Let $(V, \|\cdot\|)$ be a "complete" normed vector space and $T: V \to V$ a contraction mapping with constant L. Then T has a unique fixed point v. For $v_0 \in V$ and

$v_{i+1} = Tv_i$, $i = 0, 1, \ldots$

$$\|v_i - v\| \leq L^i \|v^i - v_0\|$$

Hence, $\lim_{i \to \infty} v_i = v$

B^π and B^* are contraction mappings in $\|\cdot\|_\infty$ with constant γ.

Take $X: S \to \mathbb{R}$ and $Y: S \to \mathbb{R}$

$$\|B^\pi(X) - B^\pi(Y)\|_\infty \leq \gamma \|X - Y\|_\infty$$

$$\|B^\pi(X) - B^\pi(Y)\|_\infty = \max_{s \in S} |B^\pi(X)(s) - B^\pi(Y)(s)|$$

$$= \max_{s \in S} \left| \sum_{s' \in S} T(s, \pi(s), s') \gamma (X(s') - Y(s')) \right|$$

$$\leq \gamma \max_{s \in S} \sum_{s' \in S} T(s, \pi(s), s') |X(s') - Y(s')|$$

$$\leq \gamma \max_{s \in S} \sum_{s' \in S} T(s, \pi(s), s') \|X - Y\|_\infty$$

$= \gamma \|x - y\|$

☒ $\|V_t - V_{t-1}\| \leq \varepsilon$

If V_t is close to V_{t-1}, is it close to V^*?

Ans → $\|V_t - V_{t-1}\| \leq \varepsilon$

$\|V_{t+1} - V_t\| \leq \gamma \varepsilon$

$\|V_{t+2} - V_{t+1}\| \leq \gamma^2 \varepsilon$

$\vdots$

$\|V_T - V_{T-1}\| \leq \gamma^{T-t} \varepsilon$

$\|V_T - V_{T-1}\| + \|V_{T-1} - V_{T-2}\| + \ldots \|V_{t+1} - V_t\|$

$\leq \varepsilon (\gamma + \gamma^2 + \ldots \gamma^{T-t})$

$\Rightarrow \|V_T - V_t\| \leq \varepsilon \gamma (1 + \ldots \gamma^{T-t-1})$

$= \varepsilon \gamma \cdot \frac{1 - \gamma^{T-t}}{1 - \gamma}$

$\Rightarrow \lim_{t \to \infty} \|V_T - V_t\| \Rightarrow \|V^* - V_t\| = \frac{\gamma \varepsilon}{1 - \gamma} (-1)$

$= \frac{\gamma \varepsilon}{\gamma - 1}$

Linear Programming

Minimize $\sum_{s \in S} V(s)$

subject to $V(s) \geq \sum_{s' \in S} T(s,a,s')\left(R(s,a,s') + \gamma V'(s)\right) \quad \forall s, \forall a \in$

We can also minimize ~~$5V_1 + 2V_2 + 3V_3$~~ $5V(1) + 2V(2) + 3V(3)$

Any positive weight will do the work.

For $X : S \to \mathbb{R}$ and $Y : S \to \mathbb{R}$, we define $X \geq Y$ if $\forall s \in S$: $X(s) \geq Y(s)$ and we define $X \gneq Y$ if $X \geq Y$ and $\exists s \in S : X(s) > Y(s)$

If $X \geq Y$, then $B^{\pi}(X) \geq B^{\pi}(Y)$

$B^{\pi}(X)(s) - B^{\pi}(Y)(s)$

$$= \sum_{s' \in S} T(s, \pi(s), s') \left\{ R(s, \pi(s), s') + \gamma X(s') \right\}$$

$$- \sum_{s' \in S} T(s, \pi(s), s') \left\{ R(s, \pi(s), s') + \gamma Y(s') \right\}$$

$$= \gamma \sum_{s' \in S} T(s, \pi(s), s') \left\{ X(s') - Y(s') \right\} \geq 0$$

Proof of Policy Improvement theorem :-

Observe that for $\pi, \pi' \in \Pi$, $\forall s \in S$, $B^{\pi'}(V^{\pi})(s) = Q^{\pi}(s, \pi'(s))$

Howard's Policy Iteration

- Switch all with arbitrary action switch

Best time complexity → $O\left(\frac{|K|^n}{n}\right)$

Other variants →

Deterministic $O(K^{0.72n})$

Randomised $(O(\log K))^n$