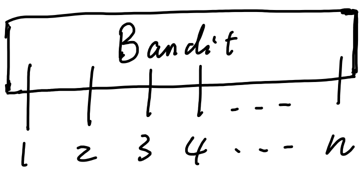
Supervised Learning $(x_1, y_1) \dots (x_n, y_n)$

$$\min_{w} L_n(w) = \frac{1}{n} \sum_{i=1}^{n} l_i(g(x_i, w), y_i)$$

Reinforcement Learning (S_t, a_t)

Multi-Armed Bandit Problem (MAB-Problem)



Arms $1, 2, \dots, n$

Rewards $D_1, D_2, \dots, D_n \in [0, 1]$

a-priori unknown

$\mu_i = \mathbb{E} D_i \quad i = 1, \dots, n$

Target Find $i^* = \arg\max_{i=1\cdots n} \{\mu_i\}$?

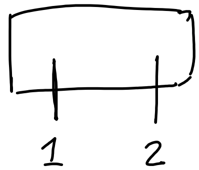
Arm sequence played : $i_1, i_2, \ldots, i_t, \ldots$

Reward sequence : $r_1, r_2, \ldots, r_t, \ldots$

$r_t =$ an i.i.d. copy of D_{i_t}

Total Reward $\sum_{t=1}^{T} r_t$

Suppose $n = 2$



D_1 $\quad$ D_2

$\uparrow$ $\quad$ $\uparrow$

$\mu_1 = \mathbb{E} D_1$ $\quad$ $\mu_2 = \mathbb{E} D_2$

Naive Algorithm

Step 1 play arm 1 $T/2$ times

play arm 2 $T/2$ times

$R_i^{(k)}$ are i.i.d. rewards each time (k) when the i-th arm is played

$R_i^{(k)}$ is a copy of D_i

$$\hat{\mu}_1 = \frac{R_1^{(1)} + R_1^{(2)} + R_1^{(3)} + \dots + R_1^{(T/2)}}{T/2} \xrightarrow[L.L.N]{T/2} \mathbb{E} R_1^{(1)} = \mu_1$$

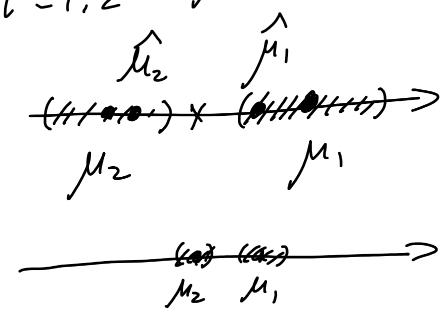
$$\boxed{|\hat{\mu}_i - \mu_i|}$$

similarly $\hat{\mu}_2 = \dfrac{R_2^{(T/2+1)} + \dots + R_2^{(T)}}{T/2} \xrightarrow{L.L.N.} \mathbb{E} R_2^{(1)} = \mu_2$

<u>Step 2</u> $\quad \hat{i}_T = \underset{i=1,2}{\arg\max} \{\hat{\mu}_i\}$

Assume $\mu_1 > \mu_2$

Arm 1 is best



intervals $\to 0$

as $T \to \infty$

"concentration"

$\Delta = |\mu_1 - \mu_2|$

Theorem Naive-Thinking Algorithm Outputs better arm $i^* = \arg\max_{i=1,2} \{\mu_i\}$

with probability $\geq 1 - 4\exp\left(-\frac{\Delta^2 T}{4}\right)$

(Hoeffding's Ineq.) If $X_1, X_2, \ldots X_n$ are i.i.d. r.v.'s

$X_i \in [a_i, b_i]$ for $\forall i \in \{1 \ldots n\}$

$$X = \sum_{i=1}^{n} X_i \qquad \mathbb{P}\left(X - \mathbb{E}X \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

$$\mathbb{P}\left[|\mu_i - \hat{\mu}_i| \le \frac{\Delta}{2} \right] = 1 - \mathbb{P}\left[|\mu_i - \hat{\mu}_i| > \frac{\Delta}{2} \right]$$

$$\geq 1 - 2\exp\left(\frac{-\Delta^2 T}{4}\right)$$

$$\left.\begin{array}{l}\mathbb{P}(A^c) \\ \mathbb{P}(B^c)\end{array}\right\} \le 2\exp\left(\frac{-\Delta^2 T}{4}\right)$$

$$A = \left\{ |\mu_1 - \hat{\mu}_1| \le \frac{\Delta}{2} \right\} \qquad B = \left\{ |\mu_2 - \hat{\mu}_2| \le \frac{\Delta}{2} \right\} \qquad \mathbb{P}(\underline{A \cap B})$$

$$\mathbb{P}(A \cap B) = 1 - \mathbb{P}\left((A \cap B)^c\right) = 1 - \mathbb{P}(A^c \cup B^c)$$

$$\geq 1 - \mathbb{P}(A^c) - \mathbb{P}(B^c)$$

$$\geq 1 - 4\exp\left(-\frac{\Delta^2 T}{4}\right)$$

$$\mathbb{P}(A^c \cup B^c) \le \mathbb{P}(A^c) + \mathbb{P}(B^c)$$

maximizing the expected overall rewards $\mathbb{E}\left(\sum_{t=1}^{T} r_t\right)$

Regret $\quad R_T = T\max\{\mu_1 \ldots \mu_n\} - \mathbb{E}\left(\sum_{t=1}^{T} r_t\right)$

$$= \mathbb{E}\left[\sum_{t=1}^{T}\left(\max\{\mu_1 \ldots \mu_n\} - r_t\right)\right]$$

Problem $\quad$ minimize regret ...

Example $\quad$ For Naive-Thinking Algorithm $\mu_1 > \mu_2$

$$R_T = \frac{T}{2}(\mu_1 - \mu_2) = \frac{T}{2}\Delta \sim O(T) \quad \begin{cases} O(T\ln T) \\ O(\ln T) \\ \ldots \end{cases}$$

Upper-Confidence-Bound (UCB) method

How to understand Regret?

Assume there are n-arms $\mu_1 > \mu_2 \geq \cdots \geq \mu_n$

$$\Delta_i = \mu_1 - \mu_i > 0$$

$$\sum_{t=1}^{T} r_t = \sum_{i=1}^{n} \sum_{k=1}^{T_i(T)} R_i^{(k)} = \sum_{i=1}^{n} T_i \left(\frac{\sum_{k=1}^{T_i(T)} R_i^{(k)}}{T_i} \right) = \sum_{i=1}^{n} T_i \hat{\mu}_{i,T}$$

$$\frac{\sum_{k=1}^{T_i(T)} R_i^{(k)}}{T_i} = \hat{\mu}_{i,T}$$

$T_i = T_i(t)$ = # of times that the i-th arm is played up to time t

Wednesday, August 5, 2020 9:45 AM

$$\text{Regret} = R_T = \mathbb{E}\left[\sum_{t=1}^{T} (\mu_1 - r_t)\right]$$

$$T = T_1 + \dots + T_n$$
$$= T_1(T) + \dots + T_n(T)$$

$$= T\mu_1 - \mathbb{E}\left(\sum_{t=1}^{T} r_t\right)$$

$$= T\mu_1 - \mathbb{E}\left(\sum_{i=1}^{n} T_i \hat{\mu}_{i,T}\right)$$

$$= (T_1 + \dots + T_n)\mu_1 - \mathbb{E}\left(\sum_{i=1}^{n} T_i \hat{\mu}_{i,T}\right)$$

$$= \mathbb{E}\left(\sum_{i=1}^{n} \underset{\Delta}{T_i} \underbrace{(\mu_1 - \hat{\mu}_{i,T})}_{\Delta}\right)$$

$\underline{\mu_1 - \hat{\mu}_{i,T}}$?

T_i

$$\text{Regret} = R_T = \mathbb{E}\left[\sum_{i=1}^{n} T_i \left(\mu_1 - \hat{\mu}_{i,T}\right)\right] = \mathbb{E}\left[\sum_{i=1}^{n} T_i(\mu_1 - \mu_i)\right] + \mathbb{E}\left[\sum_{i=1}^{n} T_i(\mu_i - \hat{\mu}_{i,T})\right]$$

$$\mu_1 > \mu_2 \geq \cdots \geq \mu_n \qquad \mathbb{E}\,\hat{\mu}_{i,T} = \mu_i \in [0,1]$$

$$\hat{\mu}_{i,T} = \frac{1}{T_i} \sum_{k=1}^{T_i} R_i^{(k)}$$

$$\mathbb{E}(\mu_1 - \hat{\mu}_{i,T}) = \mu_1 - \mu_i \qquad \mu_1 > \mu_i$$

To make R_T small want to play arm 1 as many as you can

But this make T_1 large!

Hoeffding's Inequality

$$\mathbb{P}\left[\left| \mu_i - \hat{\mu}_{i,T} \right| \le \sqrt{\frac{\ln(2Tn)}{T_i}} \right]$$

$$= 1 - \mathbb{P}\left[\left| \mu_i - \hat{\mu}_{i,T} \right| > \sqrt{\frac{\ln(2Tn)}{T_i}} \right]$$

$$= 1 - \mathbb{P}\left[\left| \mu_i - \frac{\sum_{k=1}^{T_i} R_i^{(k)}}{T_i(t)} \right| > \sqrt{\frac{\ln(2Tn)}{T_i}} \right]$$

$$\ge 1 - 2\,\mathbb{P}\left[\frac{\sum_{k=1}^{T_i} R_i^{(k)}}{T_i} - \mu_i > \sqrt{\frac{\ln(2Tn)}{T_i}} \right]$$

$$X_k = \frac{R_i^{(k)}}{T_i} \in \left[0, \frac{1}{T_i}\right]$$

$$1 - 2\exp\left(- \frac{2 \cdot \frac{\ln(2Tn)}{\cancel{T_i}}}{\sum_{k=1}^{T_i} \frac{1}{T_i^2} = \frac{1}{\cancel{T_i}}} \right)$$

$$\mathbb{P}(|a-b| > c)$$

$$= \mathbb{P}(a-b>c \text{ or } b-a>c)$$

$$\le \mathbb{P}(a-b>c) + \mathbb{P}(b-a>c)$$

Monday, August 10, 2020 9:02 AM

$$\mathbb{P}\left[\left|\mu_i - \hat{\mu}_{i,T}\right| \le \sqrt{\frac{\ln(2Tn)}{T_i}}\right] \ge 1 - 2\cdot\left(\frac{1}{2Tn}\right)^2 = 1 - \frac{1}{2T^2n^2}$$

"concentration"

i.e. $\hat{\mu}_{i,T} \in \left[\mu_i - \sqrt{\frac{\ln(2Tn)}{T_i}},\ \mu_i + \sqrt{\frac{\ln(2Tn)}{T_i}}\right]$ confidence interval

with probability $\ge 1 - \frac{1}{2T^2n^2}$

UCB-Arm $\quad i_t = \arg\max_i \left\{ \underline{\hat{\mu}_{i,t-1}} + \underbrace{\sqrt{\frac{\ln(2Tn)}{T_i(t-1)}}} \right\}$

Set $\quad E_i = \left\{ |\mu_i - \hat{\mu}_{i,T}| \le \sqrt{\frac{\ln(2Tn)}{T_i}} \right\}$

$E = \bigcap_{i=1}^{n} E_i$

$$\mathbb{P}(E) \geq 1 - \sum_{i=1}^{n} \mathbb{P}(E_i^c)$$

$$\geq 1 - \sum_{i=1}^{n} \frac{1}{2T^2n^2} = 1 - \frac{1}{2T^2n}$$

$$\mathbb{P}(E^c) \le \frac{1}{2T^2n}$$

Let t_i be the last time that the arm i is played

on E

$$\mu_i + 2\sqrt{\frac{\ln(2Tn)}{T_i(t_i-1)}} \overset{\text{on } E}{\geq} \hat{\mu}_{i,t_i-1} + \sqrt{\frac{\ln(2Tn)}{T_i(t_i-1)}}$$

$$\overset{\text{UCB}}{\geq} \hat{\mu}_{1,t_i-1} + \sqrt{\frac{\ln(2Tn)}{T_1(t_i-1)}}$$

$$\overset{\text{on } E}{\geq} \mu_1$$

$$\mu_i + 2\sqrt{\frac{\ln(2Tn)}{T_i(t_i-1)}} \geq \mu_1 > \mu_i \quad \text{on } E$$

$$\Leftrightarrow \quad \frac{\ln(2Tn)}{T_i(t_i-1)} \geq \left(\frac{\mu_1-\mu_i}{2}\right)^2$$

$\mu_1 - \mu_i = \Delta_i$

$$\Leftrightarrow \quad T_i \leq \frac{4\ln(2Tn)}{(\mu_1-\mu_i)^2} \sim \frac{\ln(Tn)}{\Delta_i^2}$$

Monday, August 10, 2020 9:19 AM

$\mu_i \in [0,1]$

$R_i^{(k)} \in [0,1]$

$$R_T^{UCB} = \mathbb{E}\left[\sum_{i=1}^{n} T_i\,(\mu_1 - \hat{\mu}_{i,T})\right]$$

$$\leq \mathbb{E}\left[\sum_{i=1}^{n} T_i\,(\mu_1 - \hat{\mu}_{i,T}) \,\middle|\, E\right] + \mathbb{P}(E^C)\cdot T$$

$$\boxed{R_T^{UCB} \lesssim O(\sqrt{nT\ln T})}$$

$\mathbb{E}\,\hat{\mu}_{i,T} = \mu_i$

$\Delta_i = \mu_1 - \mu_i$

$$\leq \mathbb{E}\left[\sum_{i=2}^{n} T_i \Delta_i \,\middle|\, E\right] + \mathbb{P}(E^C)\cdot T$$

$$\lesssim \mathbb{E}\left[\sum_{i=2}^{n} \sqrt{T_i}\cdot\sqrt{\frac{\ln(nT)}{\cancel{\Delta_i^2}}}\,\cancel{\Delta_i} \,\middle|\, E\right] + \underbrace{\mathbb{P}(E^C)}\cdot \cancel{T}$$

$$\frac{1}{2T^{2}n}\cdot \cancel{T} \sim O\left(\frac{1}{T}\right)$$

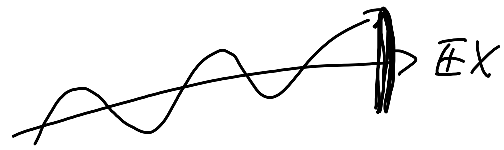
$$= \mathbb{E}\left[\sum_{i=2}^{n} \sqrt{T_i \ln(nT)} \,\middle|\, E\right]$$

T

$$\leq \mathbb{E}\left[(n-1)\sqrt{\ln(nT)\frac{\sum_{i=2}^{n} T_i}{n-1}} \,\middle|\, E\right] \sim \sqrt{(n-1)\ln(nT)\,T} \sim \sqrt{nT\ln T}$$

"Exploration - Exploitation"

"online learning"

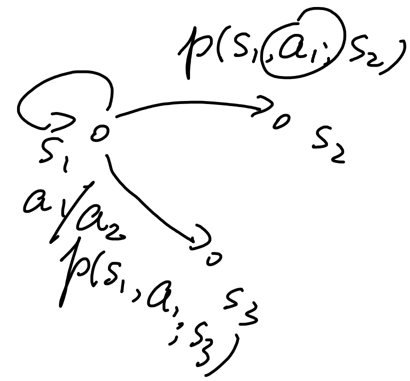


Badit

$a \in \{1, 2, 3 \dots, n\}$ — control

s $\qquad R(a) = D_a$

Markov Decision Processes
(MDP)



MDP $= \mathrm{MDP}(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$

$\mathcal{S}$ — set of states $|\mathcal{S}| = S$

$\mathcal{A}$ — set of actions $|\mathcal{A}| = A$

H — number of steps in each episode "finite horizon"

$\mathbb{P}$ — transition matrix "under control by $\mathcal{A}$"

$\mathbb{P}_h(\cdot \mid x, a)$ is the transition probabilities at state x when action a is taken at step $h \in [H]$

$r_h : \mathcal{S} \times \mathcal{A} \to [0,1]$
is a deterministic reward at step $h \in [H]$

Dynamics of our MDP

In each episode of MDP initial state x_1 is picked arbitrarily

at step $h \in [H]$ of episode

agent observes $x_h \in \mathcal{S}$

picks action $a_h \in \mathcal{A}$

receive reward $r_h(x_h, a_h)$

then transit to next state x_{h+1}

drawn from $\mathbb{P}_h(\cdot | x_h, a_h)$

episode ends when x_{H+1} is reached

Policy $\pi = \{\pi_h : \mathcal{S} \to \mathcal{A}\}_{h \in [H]}$

"at step h take policy π_h take $a_h = \pi_h(x_h) \in \mathcal{A}$.

Value function $V_h^\pi : \mathcal{S} \to \mathbb{R}$ is the value function at step h under policy π

$$V_h^\pi(x) = \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}(x_{h'}, \pi_{h'}(x_{h'})) \;\middle|\; x_h = x \right]$$

Optimal Policy π^* $V_h^{\pi^*}(x) \equiv V_h^*(x) = \sup_\pi V_h^\pi(x)$ for all $x \in \mathcal{S}$ and $h \in [H]$

Regret Agent is to play this MDP of K episodes

$k = 1 \dots K$

suppose we pick a starting state x_1^k for each episode k

$$\text{Regret}(K) = \sum_{k=1}^{K} \left[V_1^*(x_1^k) - V_1^{\pi_k}(x_1^k) \right]$$

$$\pi_k = \{ \pi_{k,h} : \mathcal{S} \to \mathcal{A} \}$$

Key Q-value function at step h

$$Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$$

$$Q_h^\pi(x, a) = r_h(x, a) + \mathbb{E}\left[\sum_{h'=h+1}^{H} r_{h'}(x_{h'}, \pi_{h'}(x_{h'})) \,\middle|\, \begin{array}{l} x_h = x \\ a_h = a \end{array}\right]$$

$$\pi_h^*(x) = \arg\max_a Q_h^\pi(x, a)$$

π_h^* .

Bellmann's equations

$$\begin{cases} V_h^\pi(x) = Q_h^\pi(x, \pi_h(x)) \\ Q_h^\pi(x,a) = r_h(x,a) + \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot|x,a)} V_{h+1}^\pi(x') \\ V_{H+1}^\pi(x) = 0 \quad \forall x \in \mathcal{S} \end{cases}$$

Dynamical Programming

$$Q_h^\pi(x,a) : \underbrace{[H]}_{= \{1 \ldots H\}} \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}$$

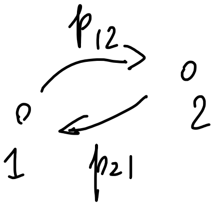
π^* — optimal policy

Bellmann's Optimality Equations

$$\begin{cases} V_h^*(x) = \max\limits_{a \in \mathcal{A}} Q_h^*(x,a) \\ Q_h^*(x,a) = r_h(x,a) + \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot \mid x,a)} V_{h+1}^*(x') \\ V_{H+1}^*(x) = 0 \qquad \forall x \in \mathcal{S} \end{cases}$$

$\pi^* \longrightarrow Q_h^{\pi^*}(x,a)$

Ex

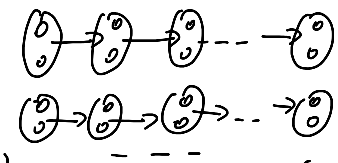


$S_1, S_2, \ldots, S_n, \ldots$

$$\pi_1 = \frac{p_{21}}{p_{12}+p_{21}} \qquad \pi_2 = \frac{p_{12}}{p_{12}+p_{21}}$$

Monte-Carlo

$$\pi_i \approx \frac{\#\text{ of visits to state } i}{n} \qquad i = 1, 2$$

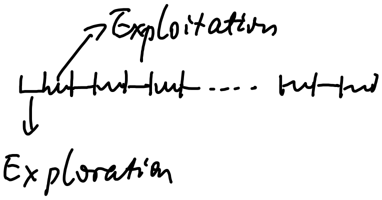


Monte-Carlo Dynamical Programming

$$Q_h^*(x,a) \approx r_h(x,a) + \underbrace{\frac{V_{h+1}^*(x_{h+1}^{(1)}) + V_{h+1}^*(x_{h+1}^{(2)}) \cdots + V_{h+1}^*(x_{h+1}^{(N)})}{N}}_{SS}$$

Exploration Exploitation

$$\mathbb{E}_{x' \sim \mathbb{P}_h(\cdot|x,a)} V_{h+1}^*(x').$$



Q- learning.

$$\frac{r_1+r_2+\dots+r_n}{n} = Q_n \approx \underline{\underline{\mathbb{E}\,V}}$$

$$\frac{r_1+r_2+\dots+r_{n+1}}{n+1} = Q_{n+1}$$

$$Q_{n+1} = \frac{r_{n+1}}{n+1} + \frac{r_1+\dots+r_n}{n+1} = \frac{r_{n+1}}{n+1} + \frac{n}{n+1}\left(\frac{r_1+\dots+r_n}{n}\right)$$

$$Q_{n+1} = \frac{n}{n+1}Q_n + \frac{r_{n+1}}{n+1} = Q_n + \underbrace{\frac{1}{n+1}}_{\alpha_n}\left(\underbrace{r_{n+1}-Q_n}_{\text{"incremental"}}\right) \to TD$$

Q-learning iteration

for episode $k = 1 \ldots K$ do

receive x_1

for step $h = 1 \ldots H$ do

$a_h \leftarrow \arg\max_{a'} Q_h(x_h, a')$ observe x_{h+1}

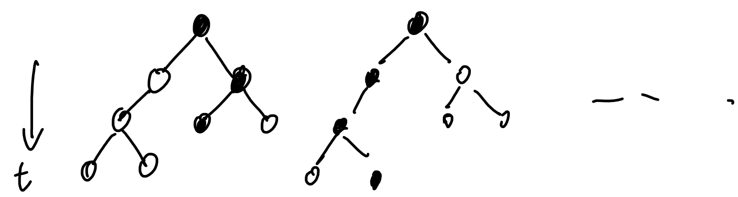
$t = N_h(x_h, a_h) \leftarrow N_h(x_h, a_h) + 1$

$Q_h(x_h, a_h) \leftarrow Q_h(x_h, a_h) + \alpha_t \left[r_t(x_h, a_h) + V_{h+1}(x_{h+1}) - Q_h(x_h, a_h) \right]$

$V_h(x_h) \leftarrow \min\left\{ H, \max_{a' \in A} Q_h(x_h, a') \right\}$

UCB → action ⇝ UCT (Upper Confidence Tree)

1 2 . . . n



Jin C. Zhu. Z.A. Bubeck. S & Jordan. M. Is Q-learning provably efficient? NeuIPS 2018

$$Q_h(x_h, a_h) \leftarrow Q_h(x_h, a_h) + \alpha_t \left[r_h(x_h, a_h) + V_{h+1}(x_{h+1}) - Q_h(x_h, a_h) + b_t \right]$$

$\sqrt{\frac{1}{t}}$ s