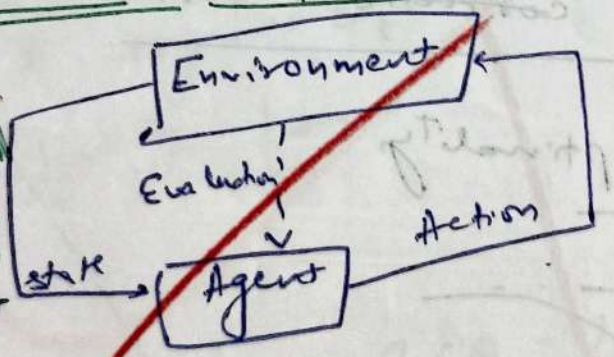
Chapter 1 :- Introduction to Reinforcement Learning

Learn through system by interacting with system

What is Reinforcement learning

- Learning about stimuli and action based on rewards and punishments alone.
- No detail supervision available
- Trial and error learning
- Delayed reward
- Sequence of actions required to obtain reward
- Associative learning required
  - Need to associate action to states
- Learn about policies not just actions
- Typically in a stochastic world.

Chapter 2 :- RL framework and application



Prediction at (t+1) is more accurate than t

Reward is scalar

- Learn from close interaction
- stochastic environment
- Noisy delayed scalar evaluation.
- Maximize a measure of long term performance

- Temporal

Chapter 3 Immediate Reinforcement

Learning problems long the way

out come is immediate

eg drug testing.

t R_t

a_t

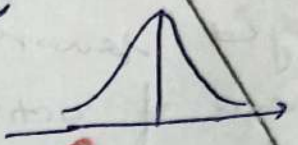
Exploration / Exploitation } dilemma

Exploitation / Exploration

[Reward / Payoff / cost] something

$A = [1, \ldots, n]$

for each of this action $q_*(a)$



$q_*(a)$ true expected payoff action a

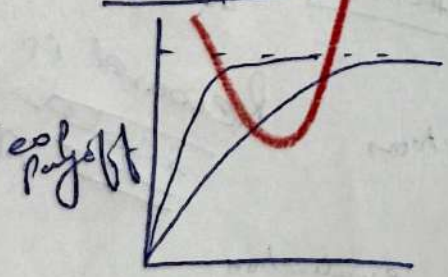
Chapter 4: Bandit Problems

multi arm Bandit Problems

Solution ?

1. Asymptotic ~~correction~~ correctness

2. Regret optimality



(3) PAC optimality

Probably Approximately correct

Probably

absolute guarantee

(ε, δ)-PAC

~~Probability~~

~~$P(q_*(a) \geq (1-\epsilon)\, q_*(a^{*1}) \geq (1-\delta)$~~

$P(q_*(a) > q_*(a^{*}) - \epsilon) \geq (1-\delta)$

Chapter 5 Value function Based Methods (Q_t)

- where we do not Know q_* and we try to predict q_*

$$Q_t(a) = \frac{\sum_{i=1}^{t} 1(a_i = a) \cdot R_i}{\sum 1(a_i = a)}$$ (Indicator function) This give average

Greedy $\arg\max_a Q_t(a)$ → take action that give maximum (exploitation)

Exploration Policy

ϵ-greedy

→ $P_r(1-\epsilon)$ $\arg\max_a Q_t(a)$

← uniformly from A

1-ε exploration

46-3

softmax

$$\frac{Q_t(a)}{\sum_{b=1}^{n} Q_t(b)} = P_t(Q_t = a)$$

[if difference is very very small than softmax used]

$$\frac{e^{Q_t(a)/\beta}}{\sum_{b=1}^{n} e^{Q_t(b)/\beta}}$$

[β is temperature paramiter]

$p(a_t = a^*) = 1 - \epsilon + \epsilon/A$

$Q_t(a) = \frac{Q_{t-1}(a)\, n_a + R_t \mathbb{1}(a_t = a)}{n_a + \mathbb{1}(a_t = a)}$

$\mathbb{1}(a_t = a)$

$= \frac{Q_t(n) n_a}{n_a + 1} + \frac{R_t}{n_a + 1}$

$= Q_{t-1}(a)\left(1 - \frac{1}{n_a + 1}\right) + \frac{R_t}{n_a + 1}$

$= Q_{t-1}(a) + \frac{1}{n_a + 1}\,(R_t - Q_{t-1}(a))$

α → learning rate; R_t → target; $Q_{t-1}(a)$ → actual output; $(R_t - Q_{t-1}(a))$ → error; $Q_t(a)$ → estimate

Non stationary problems

$Q_t(a) = Q_{t-1}(a) + \alpha [R_t - Q_{t-1}(a)]$

Chapter 6 UCB1 explanation

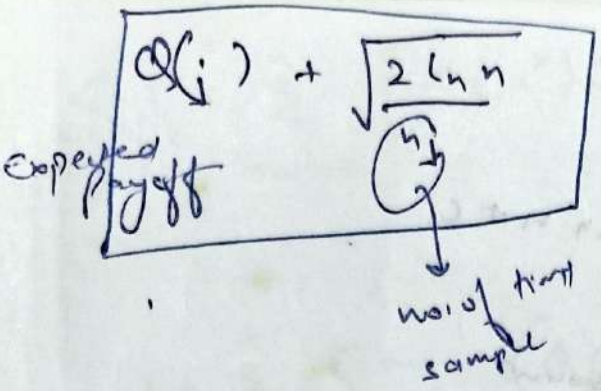
MA Bandit problem

Upper confidence Bound (UCB)

There are K arms

Init Play each arm once

Loop Play arm j that maximizes



Chapter 7: Concentration Bound

fairly trivial

arm index

Theorem: For all $K > 1$, if UCB1 is run on K arms having arbitrary reward distribution, with support in $[0,1]$ then its expected regret after any number of plays n is at most

magic constant

$$8 \sum_{i : q_*(i) < q_*(a^*)} \left(\frac{\ln n}{\Delta_i} \right) + \left(1 + \frac{\pi^2}{3}\right) \left(\sum_{j=1}^{K} \Delta_j \right)$$

$$\Delta_i = q_*(a^*) - q_*(i)$$

lost

no. of time arm i played in first n trials

$T_i(n)$ no. of time arm i played in first n trials

$$\text{Regret}_n = \sum_i \mathbb{E}[T_i(n)] \Delta_i$$

$$\mathbb{E}[X_{i,n}] = q_*(i)$$

we show that

$$E[T_j(n)] \leq \frac{8}{\Delta_j^2} \ln n + C$$

Chernoff - Hoeffding bound

Let $x_1, \ldots x_n$ be r.v with common range $[0,1]$ s.t. $E[x_t | x_{t-1}] = \mu$

Let $s_n = \frac{x_1 + \cdots + x_n}{n}$ then $\forall \epsilon > 0$

$$P(s_n \geq \mu + \epsilon) \leq e^{-2\epsilon^2 n}$$

$$P(s_n \leq \mu - \epsilon) \leq e^{-2\epsilon^2 n}$$

Chapter 8 : Theorm 1 proof (UCB1 theorm)

Proof of theorm Let $C_{n,s} = \sqrt{\frac{2 \ln n}{s}}$

$[I_n = i]$ Let l be an arbitrary (at most n plays and i)

n integer

$$T_i(n) = 1 + \sum_{m=K+1}^{n} [I_m = i]$$

$$\leq l + \sum_{m=K+1}^{n} \{ I_m = i , T_i(m-1) \geq l \}$$

$$\leq l + \sum_{m=k+1}^{n} \{ Q(a^*) C_{m,n}, T_{a^*}(m-1) \leq \phi(i)$$

$$+ C_{m-1}, T_i(m-1) \} \geq l]$$

$$\leq l + \sum_{m=k+1}^{n} \{ \min_{0<s<m} \left(Q_s(a^*) + C_{m-1}, T_{a^*}(s) \right)$$

$$\leq \max_{l \leq s_i \leq m} \left(Q_s(i) + C_{m-1}, T_i(s_i) \right)]$$

over count

$$\leq l + \sum_{m=1}^{\infty} \sum_{s=1}^{m-1} \sum_{s_i=l}^{m-1} \{ Q(a^*) + C_{m}, T_{a^*}(s) \leq$$

$$Q_{s_i}(i) + C_m, T_i(s_i)]$$

true value — confidence value

$$Q(a^*) \leq q_*(a^*) - C_m T_{a^*}(s)$$

$$Q_{s_i}(i) \geq q_*(i) + C_m, T_i(s_i)$$

$$q_*(a_*) < q_*(i) + 2 C_m, T_i(s_i)$$

$$Q(i) \leq l + \sum_{m=k+1}^{n} \{ \min_{0<s<m} \left(Q_s(a^*) + C_{m-1} \right)$$

$$\leq \max_{l \leq s_i \leq m} \left(Q_{s_i}(i) + C_{m-1}, T_i(s_i) \right)]$$

$q_*(i)$

[illegible] answer

① $P(Q_s(a^*) \le q_*(a^*) - c_{m,T_s(s)}) \le m^{-4}$ (m^-4)

② $P(Q_{s_i}(i) \ge q_*(i) + c_{m,T_s(s_i)}) \le m^{-4}$

③ For $l = \lceil 8(\ln(m)/\Delta_i^2 \rceil$ ③ is false!

$q_*(a^*) - q_*(i) - 2c_{m,T_i(s_i)}$

$= q_*(a^*) - q_*(i) - \Delta_i = 0$

$E[T_i(n)] \le \left\lceil \frac{8 \ln n}{\Delta_i^2} \right\rceil + \sum_{m=1}^{t-1} \sum_{s=1}^{\lceil 8(\ln m/\Delta_i^2)\rceil} 2m^{-4}$

$\le \frac{8 \ln n}{\Delta_i^2} + 1 + \frac{\pi^2}{3}$

Chapter 9 PAC Exploration and Naive Algo Proof (PAC Bounds)

Chernoff-Hoeffding Bound: Let $x_1, \ldots x_n$ be r.v with common range $[0,1]$, s.t

$E[x_t | x_1, \ldots x_{t-1}] = \mu$. Let $s_n = \frac{x_1 + \ldots + x_n}{n}$

then for $\epsilon > 0$,

$$P[s_n \geq \mu + \epsilon] \leq e^{-2\epsilon^2 n}$$

$$P[s_n \leq \mu - \epsilon] \leq e^{-2\epsilon^2 n}$$

Asymptotic Regret

Markov Inequality

$$P_r(X \geq a) \leq \frac{E(x)}{a}$$

Union Bound

$$P(\bigcup_i A_i) \leq \sum_i P(A_i)$$

media action elimination

PAC Bounds for MAB

Naive

Input $\epsilon > 0$, $\delta > 0$

output "An arm"

for each $a \in A$ do

Sample 'a' l times,

$$l = \frac{2}{\epsilon^2} \cdot \ln\left(\frac{2k}{\delta}\right)$$

Let $Q(a)$ be the average reward

for a arm a

end

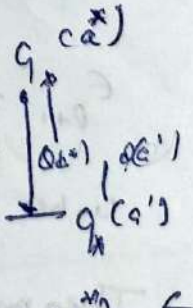
output $a = \underset{a \in A}{\text{argmax}}\, Q(a)$

Google

Cycle action elimination of Reinforcement learning

Theorem 1:- Naive (ϵ, δ) is an (ϵ, δ) - PAC algo, with arm sample complexity

$$O\left(\left(k/\epsilon^2\right) \cdot \log\left(k/\delta\right)\right)$$



Proof Let a' be an arm s.t $q_*(a') < q_*(a^*) - \epsilon$

$$P(Q(a') > Q(a^*)) \leq P(Q(a') > q_*(a') + \epsilon/2$$

$$\text{OR } Q(a^*) < q_*(a^*) - \epsilon/2)$$

$\leq P(\hat{p}(a') > p_{l+1}(a') + \epsilon_l/2) + P(\hat{p}(a^*) < p_{l+1}(a^*) - \epsilon_l/2)$

$\leq 2e^{-2\frac{\epsilon^2}{4}l}$

Subst. $P(\hat{p}(a') > \hat{p}(a^*)) \leq \frac{\delta}{k}$

Summing over all a',

Prob. of failure $\leq (k-1)\cdot\frac{\delta}{k} \leq \delta$

Lecture 10 — Median elimination Algo

Input $\epsilon > 0$, $\delta > 0$; process in rounds.

Median elimination

Output An arm

Set $S_1 = A$, $\epsilon_1 = \epsilon/4$, $\delta_1 = \delta/2$, $l = 1$

repeat

Sample every arm $a \in S_l$ for $\frac{1}{(\epsilon_l/2)^2}\cdot\log\left(\frac{3}{\delta_l}\right)$ ~~let~~ let $\hat{p}_l(a)$ denote ~~in value~~ $\frac{1}{\left(\frac{\epsilon_l}{2}\right)^2}\cdot\log\left(\frac{3}{\delta_l}\right)$ ← $\frac{1}{\frac{\epsilon_l^2}{2}}\cdot\log\left(\frac{3}{\delta_l}\right)$

find ~~meaning~~ the median of $\hat{p}(a)$, denoted $\log m_l$

$S_{l+1} \leftarrow S_l \setminus \{a : \hat{p}_l(a) < m_l\}$

$\epsilon_{l+1} = \frac{3}{4}\epsilon_l$; $\delta_{l+1} = \frac{\delta_l}{2}$, $l = l+1$

Until $|S_l| = 1$

Theorem The MEA(ϵ,δ) is a (ϵ,δ) PAC Algo w. sample complexity $O\left(\frac{K}{\epsilon^2}\cdot\log\left(\frac{1}{\delta}\right)\right)$

Lemma For every phase l in MET, we have

$$\Pr\left[\max_{j \in S_l} q_*^*(j) \leq \max_{i \in S_{l+1}} q_*(i) + \epsilon_l\right] \geq 1 - \delta_l$$

Proof:- w.l.g., consider $l = 1$ $\quad E_1 = \left[Q(a_1^*) < q_*(a_1^*) - \epsilon_{1/2}\right]$

$$E_1 = \left[Q(a_1^*) < q_*(a_1^*) - \epsilon_{1/2}\right]$$

True best arm in 1^{st} level

$$P[E_1] \leq \delta_{1/3}$$

$$P\left[Q_1(j) \geq Q_1(a_1^*) \,\middle|\, Q_1(a_1^*) < q_*(a_1^*) - \epsilon_{1/2}\right] \leq \delta_1/3$$

j^{th} arm i, not ϵ_1 optimal

Let # bad, be the no. of bad arms from which (1) holds

$$E\left[\#\text{bad} \,\middle|\, Q_1(a_1^*) \geq q_*(a_1^*) - \epsilon_{1/2}\right] < |S_1|$$

$$P\left[\#\text{bad} \geq \frac{|S_1|}{2} \,\middle|\, Q_0(a_1^*) \geq q_*(a_1^*) - \epsilon_{1/2}\right]$$

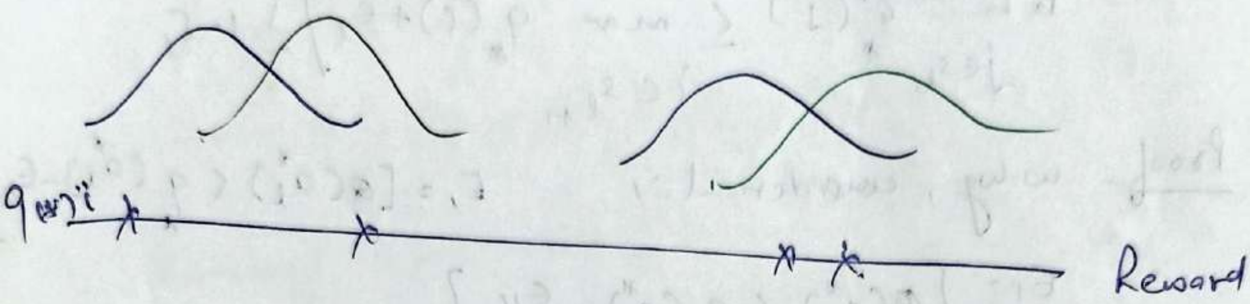
$$\leq \frac{|S_1|(\delta_1/3)}{|S_1|/2}$$

$$= \frac{2\delta_1}{3}$$

Chapter 11 Thompson Sampling

OR

Posterior sampling



Chapter 12 Policy search

state to action

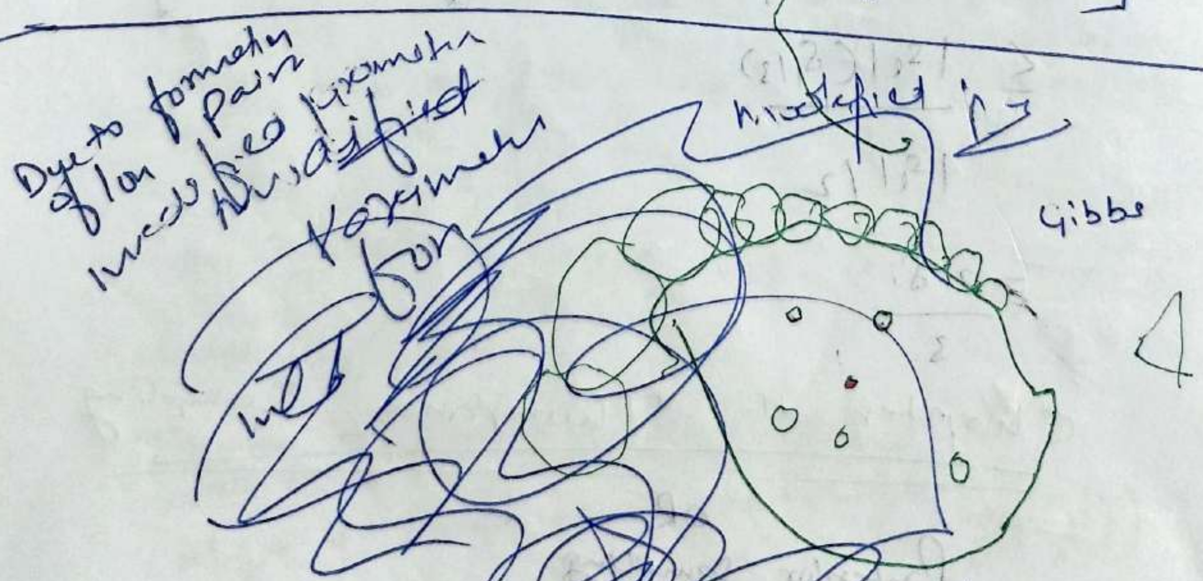
$\pi_t(a) \triangleq Pr(a_t = a)$

Binary bandit (only two outcomes 0 or one)

R_t 0, 1

If $R_t = 1$

$$\pi_{t+1}(a_t) = \pi_t(a_t) + \alpha[1 - \pi_t(a_t)]$$



$$\pi_{t+1}(a') = \pi_t(a')(1-\alpha)$$

$a' \neq a_t$

If $R_t = 0$?

$$\pi_{t+1}(a_t) = \pi_t(a_t)(1-\beta)$$

$$\pi_{t+1}(a') = \pi_t(a') + \beta[1 - \pi_t(a')]$$

if $\alpha = \beta$, L_{R-P}

$\alpha >> \beta$, $L_{R-\epsilon P}$

$\beta = 0$, L_{R-I} ⟶ inaction

Policy Gradient (Parametrization Policy gradient)

- Depends on some set of parameters, θ
- Preferences
- weights of ANN

Modified parameters system (or)

Chapter 13 Reinforce $\pi(\cdot;\theta)$

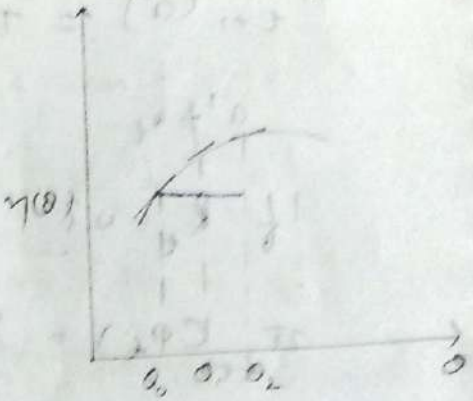
$\eta(\theta) = E[R_t]$

$\sum_a q_*(a), \pi(a,\theta)$

$\theta \leftarrow \theta + \alpha \nabla \eta(\theta)$

$$\nabla_\theta \eta(\theta) = \sum_a q_\pi(a) \nabla_\theta \pi(a|\theta)$$

$$= \sum_a \left(q_\pi(a) \frac{\nabla_\theta \pi(a,\theta)}{\pi(a,\theta)} \right) \pi(a,\theta)$$



$$E_{\pi(\cdot;\theta)} \left[\frac{q_\pi(a) \nabla_\theta \pi(a,\theta)}{\pi(a,\theta)} \right]$$

$$\approx \frac{1}{N} \sum_{t=1}^{N} R_t \frac{\nabla_\theta \pi(a_t,\theta)}{\pi(a_t,\theta)}$$ [gradient]

Incremental version: $\theta_{t+1} = \theta_t + \Delta\theta_t$

$$\Delta\theta_t = \alpha_t R_t \frac{\nabla \pi(a_t,\theta_t)}{\pi(a_t,\theta_t)}$$

Reward direction of good gradient otherwise backward in direction opposite to gradient

$$\Delta\theta_t = \alpha_t R_t \frac{\partial \ln \pi(a_t,\theta_t)}{\partial\theta}$$

$$\Delta\theta_t = \alpha_t (R_t - b_t) \frac{\partial \ln \pi(a_t,\theta_t)}{\partial\theta}$$

Characteristic eligibility:- which have the highest gradient

Reinforce

Two actions arbitrary rewards

$$\pi(a,\theta) = \begin{cases} \theta & a=1 \\ 1-\theta & \text{if } a=0 \end{cases}$$

$$\frac{\partial \ln \pi}{\partial\theta} = \frac{a-\theta}{\theta(1-\theta)}; \quad \alpha = \rho(\theta(1-\theta)), \; b=0$$

$$\Delta\theta_t = \rho(a_t-\theta) \cdot R_t$$

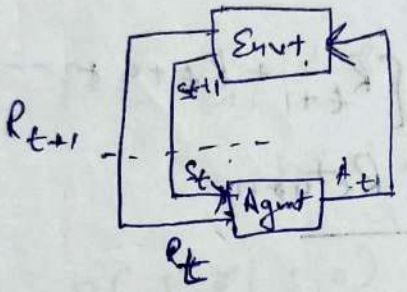
$$\pi(a, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-(a-\mu)^2/2\sigma^2}$$

Chapter 14 Contextual Bandits

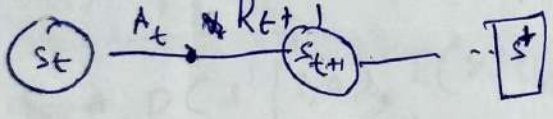
- action can have parameter
- Parameterized Bandit

$$Q(a) = \alpha_1 \theta_1 + \alpha_2 \theta_2 + \alpha_3 \theta_3 \cdots$$

Chapter 15: Full RL introduction



R_t Reward signal from environment



$t=1$ decision 1

$t=2$ decision 2 . . .

Chapter 3 in the book

Reward

- Scalar quantites :- Just a number
- ouside ~~the~~ direct control of agent
- ~~Bounded~~
- frequent

Policy

$$\pi_t(a|s) \doteq P_r(A_t = a \mid S_t = s)$$

$$\pi_1 = \pi_2 = \cdots \pi$$

$$P_r(A_1 = a_1 \mid S_1 = 1)$$

$$= P_r(A$$

Chapter 16: Returns, Value functions and MDPs

Return value function and MDP

Return:

G_t — total return

$$G_t = R_{t+1} + R_{t+2} + \dots + R_T$$

$$G_t \equiv R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$$

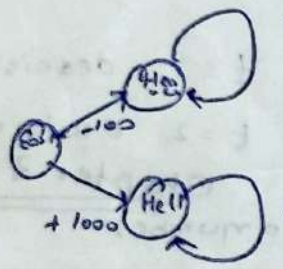
Discounted Return $0 \le \gamma < 1$

near sighted agent ← → far sighted.

$$G_t \equiv \lim_{N \to \infty} \frac{1}{N} \left[R_{t+1} + R_{t+2} + \dots R_{t+N} \right]$$

Level of accuracy to γ^k correlation

Average reward Return:

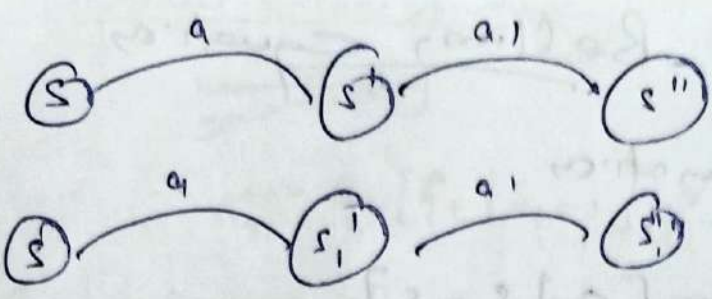


Value function:-

v^π is value function associated with policy

$$v^\pi(s) = E_\pi \{ G_t \mid S_t = s \}$$

$$q^\pi(s,a) = E_\pi \{ G_t \mid S_t = s, A_t = a \}$$



How could you characterized the problem.

$Pr(S_{t+1}, R_{t+1} | S_t, A_t, \ldots S_0)$

$= Pr(S_{t+1}, R_{t+1} | S_t, A_t)$

~~$Pr(S_{t+1})$~~ $P(S_{t+1} = s', R_{t+1} = r | S_t = s, A_t = a)$

$= p(s', r | s, a)$

$S, A, p(s', r | s, a), r$

$S, A, p(s' | s, a), E(r | s, a, s')$

Markov Decision Process

Chapter 17 :- MDP Modeling

state, actions, transition probability

$S, A, P(s' | s, a), E[r | s, a, s']$

$r_{can} = +1$

$r_{out\ of\ battery} = -3$

$r_{recharge} = 0$

State
- Charge : low, high
- No. of cans

Action
- Actively seek cans
- Wait for cans
- Recharge.

Chapter 18 Bellman equations

Value functions

$$v^{\pi}(s) = E_{\pi}[G_t \mid s_t = s]$$

$$q^{\pi}(s,a) = E_{\pi}[G_t \mid s_t = s, A_t = a]$$

$$v^{\pi}(s) = E_{\pi}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \mid s_t = s]$$

$$= E_{\pi}[R_{t+1} + \gamma v^{\pi}(s') \mid s_t = s]$$

$$v^{\pi}(s) = \sum_a \pi(a|s) \sum_{s'} p(s'|s,a) \{E[r|s,a,s'] + \gamma v^{\pi}(s')\}$$

Bellman eqn for v^{π}, q^{π}

$$q^{\pi}(s,a) = \sum_{s'} p(s'|s,a) \{E[r|s,a,s'] + \gamma \sum_{a'} \pi(a'|s') q^{\pi}(s',a')\}$$

$$v^{\pi}(s) = \sum_a \pi(a|s) \cdot q^{\pi}(s,a)$$

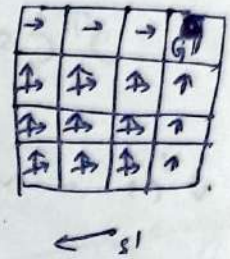
Chapter 19 Bellman optimality eqn

Optimal values function:-

want π best of all π

~~$\max_{\pi} E[R_t]$~~

$$\max_{\pi} E_{\pi}\left[R_t \mid S_t = s\right] \quad \forall s$$



$\leftarrow s'$

$$\pi^* = \arg\max_{\pi} v^{\pi}(s) \quad \forall s$$

optimal policy

$$v^{\pi^*} \triangleq v^* = \max_{\pi} v^{\pi}(s) \quad \forall s$$

↑ optimal value function for

$$q^*(s,a) = \max_{\pi} \sum_{s'} p(s'|s,a) \cdot \left[E[r|s,a,s'] + \gamma v^{\pi}(s')\right]$$

$$= \sum_{s'} p(s'|s,a) \cdot \left[E[r|s,a,s'] + \gamma \max_{\pi} v^{\pi}(s')\right]$$

$$v^*(s) = \max_{a} q^*(s,a)$$

$$v^*(s) = \max_{a} \sum_{s'} p(s'|s,a)\left[E[r|s,a,s'] + \gamma v^*(s')\right]$$

Bellman optimality eq for v^*

$$\pi^*(s) = \arg\max_{a} \sum_{s'} p(s'|s,a)\left[E[r|s,a,s'] + \gamma v^*(s')\right]$$

$$\pi^*(s) = \arg\max_{a} q^*(s,a)$$

Chapter 20 Cauchy sequence and Green's equation

MDPs $S, A, p(s'|s,a), E[r|s,a,s'] \gamma$

Bellman eqn:

$$v^\pi(s) = \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)\left[E[r|s,a,s'] + \gamma v^\pi(s')\right]$$

$$v^*(s) = \max_a \sum_{s'} p(s'|s,a)\left[E[r|s,a,s'] + \gamma v^*(s')\right]$$

Finite MDPs

$\pi(a|s)$, $\pi(s) \rightarrow$ Det. policy

v^π is a vector with $|S|$ components

$\|v\| = \sup_{s \in S} |v(s)|$

$\|x\| = 0$ iff $x = 0$

$\|\alpha x\| = \alpha \|x\|$

$\|x+y\| \leq \|x\| + \|y\|$

Complete Normed Vector space

Cauchy sequence $x_1, x_2, x_3, \ldots$

for every ϵ real $\epsilon > 0$,

$\exists N \in Z^+$ s.t. $\forall m, n > N$

$$\|x_m - x_n\| < \epsilon$$

If every cauchy sequence is convergent then the vector space is complete

$$q_\pi(s) = \sum_a \pi(a|s) \cdot \sum_{s'} p(s'|s,a) E[r|s,a,s']$$

$$\left\{ = \sum_{s'} p(s'|s,\pi(s)) \cdot E[r|s,\pi(s),s'] \right\}$$

$$p_\pi(j|s) = \sum_a \pi(a|s) \cdot p(j|s,a)$$

$$\{ = p(j|s,\pi(s))$$

γ_π is a $|S|$ dimensional vector p_π is a $|S| \times |S|$ dim stochastic matrix

$0 \le \gamma < 1$

$$\gamma_\pi + \gamma p_\pi \cdot v = v^\pi$$

$$\Rightarrow v^\pi = (I - \gamma p_\pi)^{-1} \gamma_\pi$$

Relationship of eigen value to invertibility of matrix (Google it)

$\pi_1 = a_1$ | $v^\pi(1) = \dots$

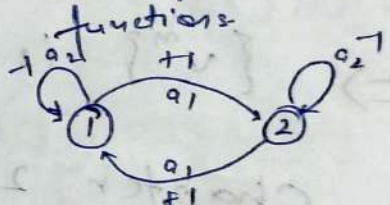
$\pi_2 = a_2$ | $v^\pi(1) = 1 + \gamma v^\pi(1)$

Chapter 2

simplest non vector space is Banach space

Banach fixed pt. theorem

V = space of all value functions



$L_\pi : V \longrightarrow V$

$L_\pi v \triangleq r_\pi + \gamma p_\pi v$

$L_\pi v^\pi = v^\pi \Rightarrow v^\pi$ is a "fixed point" of L_π

Banach fixed pt. theorem

Suppose U is a Banach space and $T : U \longrightarrow U$ is a contraction mapping ($\|Tu - Tv\| \le \gamma \|u - v\|$) then $\exists$ an unique v^* in U s.t. $Tv^* = v^*$ and for arbitrary v^0 in U, the sequence $\{v^n\}$ defined by $v^{n+1} = Tv^n = T^{n+1} v^0$, converges to v^*

T is a contraction, if $\|Tu - Tv\| \le \lambda \|u - v\|$, $0 \le \lambda < 1$, $\forall u, v \in U$ ∴ apply in fixed point space

Proof

$$\|v^{n+m} - v^n\| \le \|v^{m+n} - v^{n+m/2}\| + \|v^{n+m/2} - v^n\|$$

$$\|v^{n+m} - v^n\| \le \sum_{k=0}^{m-1} \|v^{n+k+1} - v^{n+k}\|$$

$$= \sum_{k=0}^{m-1} \|T^{n+k} v^1 - T^{n+k} v^0\|$$

$$\le \sum_{k=0}^{m-1} \lambda^{n+k} \|v^1 - v^0\|$$

$$= \frac{\lambda^n (1 - \lambda^m)}{1 - \lambda} \|v^1 - v^0\|$$

$\to 0$

$\Rightarrow \{v^n\}$ is cauchy

Chapter 22 Convergence Proof

$$0 \le \|Tv^* - v^*\| \le \|Tv^* - v^n\| + \|v^n - v^*\|$$ use triangle inequality

$$= \|Tv^* - Tv^{n-1}\| + \|v^n - v^*\|$$

$$\le \lambda \|v^* - v^{n-1}\| + \|v^n - v^*\|$$

$\|v^* - v^n\| \to 0$ as $n \to \infty$

$0 \le \|Tv^* - v^*\| \le 0$

$\Rightarrow Tv^* = v^*$

Let u^*, v^* be 2 fixed points

$$\|Tu^* - Tv^*\| \leq \lambda \|u^* - v^*\|$$

$$\Rightarrow u^* = v^*$$

Let $u \& v$ be in V

$$L_\pi u(s) = r_\pi(s) + \sum_{j \in s} \gamma^k p_\pi(j|s) \cdot u(j)$$

$$L_\pi v(s) = r_\pi(s) + \sum_{j \in s} \gamma^k p_\pi(j|s) v(j)$$

Let $L_\pi v(s) = L_\pi(v u(s))$

$$0 \leq L_\pi v(s) - L_\pi u(s) \leq r_\pi(s) + \gamma^k \sum p_\pi(j|s) v(j) - r_\pi(s) - \gamma^k \sum p_\pi(j|s) \cdot u(j)$$

$$= \gamma^k \sum_j p(j|s) [v(j) - u(j)]$$

$$\leq \gamma^k \|v - u\| \sum_j p(j|s)$$

$$= \gamma^k \|v - u\|$$

other way (o.w)

$$|L_\pi v(s) - L_\pi u(s)| \leq \gamma^k \|v - u\| \quad \forall s$$

$\Rightarrow L_\pi$ is a contraction

Chapter 23! Lpi Convergence

L_π is a contraction.

T space of value functions

$v \in V$

v/s bounded function

$$v^* = \max_\pi \{ r_\pi + \gamma P^\pi v^* \} \triangleq L v^*$$

$$v^*(s) = \max_a \{ E[r|s,a] + \gamma \sum_j p(j|s,a)\, v^*(j) \}$$

$$Lv \triangleq \max_\pi \{ r_\pi + \gamma P^\pi v \}$$

Claim L is a contraction

Let $a_s^* \in \arg\max_a \{ E[r|s,a] + \gamma \sum p(j|s,a)\, v(j) \}$

$$0 \le Lv(s) - Lu(s) \le E[r|s,a_s^*] + \gamma \sum p(j|s,a_s^*)\, v(j) - \{ E[r|s,a_s^*] + \gamma \sum p(j|s,a_s^*)\, u(j) \}$$

$$= \gamma \sum_j p(j|s,a_s^*) [v(j) - u(j)]$$

$$\le \gamma \|v-u\| \sum_j p(j|s,a_s^*) \le \gamma \|v-u\| \cdot 1$$

$$0 \le Lv(s) - Lu(s) \le \gamma \|v-u\|$$

$$|Lv(s) - Lu(s)| \le \gamma \|v-u\| \quad \forall s$$

$\Rightarrow L$ is contraction

Value iteration

Value iteration

1. Select $\vartheta^0 \in V$, pick an $\epsilon > 0$ Set $n=0$
2. for each $s \in S$ compute $\vartheta^{n+1}(s)$ by

$$\vartheta^{n+1}(s) = \max_a \left\{ E(r|s,a) + \gamma \sum_{s'} p(s'|s,a) \vartheta^n(s') \right\}$$

$$\left\{ \vartheta^{n+1} = L\vartheta^n \right\}$$

3. If $\| \vartheta^{n+1} - \vartheta^n \| < \epsilon (1-\gamma)/2\gamma$ go to step 4 wo increment n by 1 go to step 2
4. for each $s \in S$

$$\pi(s) = \arg\max_a \left\{ E[r|s,a] + \gamma \sum_{s'} p(s'|s,a) \cdot \vartheta^{n+1}(s') \right\}$$

stop

Chapter 24 Value iteration proof

Theorem: Let $\vartheta^0 \in V$ $\epsilon > 0$, $\{\vartheta^n\}$ be derived from

$$\vartheta^{n+1} = L\vartheta^n$$

Then

a. ϑ^n converges in norm to ϑ^*

b. $\exists$ a finite N at which (3) is met for $\forall n > N$

c. π defined by (4) is ϵ optimal $\forall$

d. $\| \vartheta^{n+1} - \vartheta^* \| \leq \epsilon/2$ when (3) holds

(3) suppose (2) is met for some η, α, π satisfies (C,1) then

$$\|v^{\pi} - v^{*}\| \leq \|v^{\pi} - v^{\eta+1}\| + \|v^{\eta+1} - v^{*}\|$$

$L_\pi v^{\eta+1} - E[r|s,\pi(s)] + \gamma \sum_{s'} p(s'|s,\pi(s)) v^{\eta+1}(s')$

$L v^{\eta+1} = L_\pi v^{\eta+1}$

$$\rightarrow \|v^{\pi} - v^{\eta+1}\| = \|L_\pi v^{\pi} - v^{\eta+1}\|$$

$$\leq \|L_\pi v^{\pi} - L v^{\eta+1}\| + \|L v^{\eta+1} - v^{\eta+1}\|$$

$$= \|L_\pi v^{\pi} - L_\pi v^{\eta+1}\| + \|L v^{\eta+1} - L v^{\eta}\|$$

$$\leq \gamma \|v^{\pi} - v^{\eta+1}\| + \gamma \|v^{\eta+1} - v^{\eta}\|$$

$$\|v^{\pi} - v^{\eta+1}\| \leq \frac{\gamma}{1-\gamma} \|v^{\eta+1} - v^{\eta}\|$$

$$\|v^{\eta+1} - v^{*}\| \leq \sum_{k=0}^{\infty} \|v^{\eta+k+2} - v^{\eta+k+1}\|$$

$$= \sum_{k=0}^{\infty} \|L^{k+1} v^{\eta+1} - L^{k+1} v^{\eta}\|$$

$$\leq \sum_{k=0}^{\infty} \gamma^{k+1} \|v^{\eta+1} - v^{\eta}\|$$

$$= \frac{\gamma}{1-\gamma} \|v^{\eta+1} - v^{\eta}\|$$

google. Linear convergence with rate of gamma

Rates of convergence for different algorithms

Chapter 25: Policy iteration

select π_0, set $n=0$

Policy evaluation. $v^{\pi_n} = (I - \gamma P_{\pi_n})^{-1} r_{\pi_n}$

3. Policy Improvement choose π_{n+1}

$$\pi_{n+1} \in \arg\max_{\pi} \{ r_\pi - \gamma P_\pi v^{\pi_n} \}$$

4. stop if $\pi_{n+1} = \pi_n$, $\pi_n = \pi^*$

Let π_{n+1} satisfy (3)

$$r_{\pi_{n+1}} + \gamma P_{\pi_{n+1}} v^{\pi_n} \geq r_{\pi_n} + \gamma P_{\pi_n} v^{\pi_n} \cdot v^{\pi_n}$$

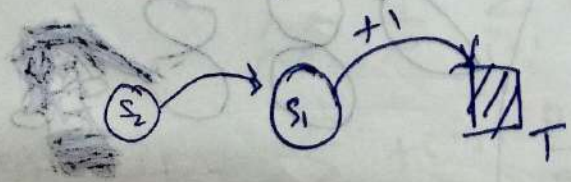
$$r_{\pi_{n+1}} \geq (I - \gamma P_{\pi_{n+1}}) v^{\pi_n}$$

$$(I - \gamma P_{\pi_{n+1}})^{-1}$$

$$v^{\pi_{n+1}} \geq v^{\pi_n}$$

Chapter 26 Dynamic Programming

$$v^{n+1}(s) = \max_a \{ E[r \mid s,a] + \gamma \sum_{s'} p(s' \mid s,a) \cdot v^n(s') \}$$



Asynchronous DP

RTDP (Real time dynamic programming)

GPI — way the error

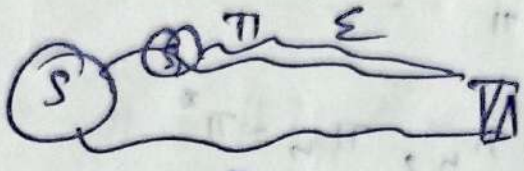
PE : Policy evaluation

PI : Policy improvement

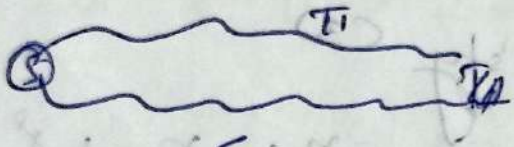
Chapter 27 Monte carlo

when we have no access to MDP

$$v^{\pi}(s) = E_{\pi}[G_t \mid s_t = s]$$



Ev



Σ

Monte carlo methods for policy evaluation

Sample models

Exploring Model

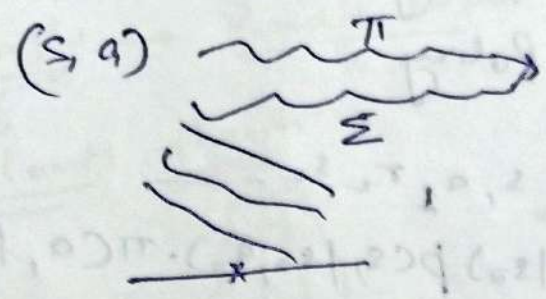
First visit MC

Every visit MC

Chapter 28 Control of Monte carlo

GPI (Generalized policy iteration)

9. $\pi(s_i)$
(s_i'', a_i)



used q function for controlled

Epsilon soft policy

off policy

$$E_{x\sim p}[f(x)] = \sum_x p(x) f(x) \approx \frac{1}{N} \sum_{i=1}^{N} f(x_i)$$

$$= \sum_x p(x) \cdot \frac{q(x)}{q(x)} \cdot f(x)$$

$$= \sum_x \frac{p(x)}{q(x)} \cdot f(x) \cdot q(x)$$

$$= E_{x\sim q}\left[\frac{p(x)}{q(x)} \cdot f(x)\right] \approx \frac{1}{N} \sum_i f(x_i) \frac{p(x_i)}{q(x_i)}$$

Importance factor / weight

Chapter 29 off policy MC

$$E_{x\sim p}[f(x)] = E_{x\sim q}\left[f(x) \frac{p(x)}{q(x)}\right]$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} f(x_i) \cdot \frac{p(x_i)}{q(x_i)}$$

$$\text{weighted IS} = \frac{\sum_{i=1}^{N} f(x_i) \frac{p(x_i)}{q(x_i)}}{\sum_{i=1}^{N} \frac{p(x_i)}{q(x_i)}}$$

π :- Estimation policy

μ :- Behavior Policy

x_i : s_0 a_0 r_1 s_1 a_1 r_2 s_2 ---

$$p(x_i) = \overset{p(s_0)}{1}\pi(a_0|s_0)\,p(s_1|s_0,a_0)\cdot\pi(a_1|s_1)\,p(s_2|s_1,a_1)\cdots$$

$$q(x_i) = p(s_0)\cdot\mu(a_0|s_0)\,p(s_1|s_0,a_0)\,\mu(a_1|s_1)\,p(s_2|s_1,a_1)\cdots$$

$$\frac{p(x_i)}{q(x_i)} = \frac{\pi(a_0|s_0)\,\pi(a_1|s_1)\cdots}{\mu(a_0|s_0)\,\mu(a_1|s_1)\cdots}$$

$$= \frac{\prod_{j=0}^{T}\pi(a_j|s_j)}{\prod_{j=0}^{T}\mu(a_j|s_j)} = \prod_{j=0}^{T}\frac{\pi(a_j|s_j)}{\mu(a_j|s_j)}$$

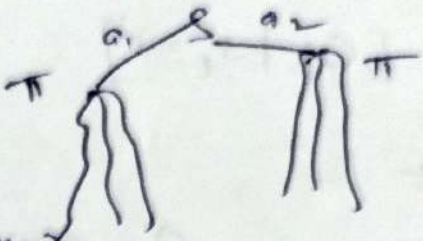
$$v^{\pi}(s) = \frac{\sum_{i=1}^{n} q_t^{(i)} \prod_{j=0}^{T}\frac{\pi(a_j^{(i)}|s_j)}{\mu(a_j^{(i)}|s_j^{(i)})}}{\sum_{i=1}^{n}\prod_{j=0}^{T}\frac{\pi(a_j^{(i)}|s_j^{(i)})}{\mu(a_j^{(i)}|s_j^{(i)})}}$$

Chapter 30 UCT

Google → Bandit based Monte carlo planing

↓

Rollouts theorm



search (state depth)

If terminal (state) then return 0 // Terminal reward

If leaf (state, depth) then return Eval (state)

action = select Action (state depth)

(next state ~~reward occurred~~) = Simulate (state action)

q = reward + γ search (next state depth+1)

Update Values (state, action, q, depth)

return q

Chaos lectures

$c_p \sqrt{\frac{2 \ln t}{n}}$

$c_p \sqrt{\frac{\ln t}{n}}$

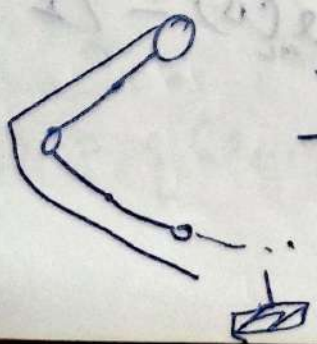
Chapter 31 TD(0)

(Temporal Difference method)

full backup

Bootstrap

MC



- Sample backups
- No Bootstrap

TD — Sample Backup

— Bootstrap.

MC $E[G_t | S_t = s]$

G_t

$$V(s)_{new} = \underbrace{V_{old}(s)}_{\text{current}} + \alpha \underbrace{[G_t - V_{old}(s)]}_{\text{error}}$$

target

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots$$

$$= R_{t+1} + \gamma G_{t+1}$$

$$\approx R_{t+1} + \gamma V(S_{t+1})$$

$$V_{new}(S_t) = V_{old}(S_t) + \alpha [R_{t+1} + \gamma V_{old}(S_{t+1}) - V(S_t)]$$

TD(0) algorithm

$$V_{old}(S_t) + \alpha [R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+2} + \gamma^3 V_{old}(S_{t+3})]$$

A, 1, B, 0, ~~B, 0, B, 0~~

A, 0, 1, B, 0

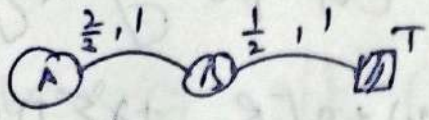
B, 1

~~B, 1, B, 1~~

A, 1, B, 1

$V_{mc}(A) = 1/2$

$V_{mc}(B) = 1/2$



$V_{TD}(B) = \frac{1}{2}$

$V_{TD}(A) = \frac{2}{3} + \gamma V(B) \quad \gamma = 1$

$= \frac{2}{3} + \frac{1}{2} = \frac{7}{6}$

Chapter 32 TD(0) control

- Evaluation
- Sample backup & bootstrap

Control

$$q_{new}(s,a) = q_{old}(s,a) + \alpha \underbrace{\left[R_{t+1} + \gamma \underbrace{q_{old}(s_{t+1}, a)}_{v_{old}(s_{t+1})} - q_{old}(s,a) \right]}_{\text{Error}}$$

Target — current

$TD_{error} \equiv \delta$

Start with arb., $q(\cdot,\cdot)$

Pick a start s_0, $t \Rightarrow$

Pick action a_t in s_t, acc. to $q(s_t, \cdot)$, ϵ-greedy

→ Apply a_t and sample s_{t+1}, R_{t+1}

- Pick a_{t+1}, acc. to $q(s_{t+1})$ (greedy)

$q(s_t, a_t) = q(s_t, a_t) + \alpha \left[R_{t+1} + \gamma q(s_{t+1}, a_{t+1}) - q(s_t, a_t) \right]$

$t \leftarrow t+1$

SARSA – On policy Algo.

* $q(s_t, a_t) = q(s_t, a_t) + \alpha [R_{t+1} + \gamma \sum_{a'} \pi(s_{t+1}, a') - q(s_t, a_t)]$

Chapter 33 Q-Learning :-

$q^*(s_t, a_t) =$

$E\left[R_{t+1} + \gamma v^*(s_{t+1})\right)$

$q^*(s_t, a_t) E\left[R_{t+1} + \gamma v^*(s_{t+1})\right]$

$= E[R_{t+1} \gamma \max_{a'} q^*(s_{t+1}, a')]$

$q(s_t, a_t) = q(s_t, a_t) + \alpha [R_{t+1} \gamma \max q(s_{t+1}, a') - q(s_t, a)]$

- start with arb $q(\cdot, \cdot)$

Pick a state s_0, $t=0$

Pick action a_t in s_t {acc to $q(s_t, \ )$; ϵ greedy}

Apply at a sample s_{t+1}, R_{t+1}

$q(s_t, a_t) = q(s_t, a_t) + \alpha [R_{t+1}, \gamma \max_{a'} q(s_{t+1}, a') - q(s_t, a_t)]$

Q-learning

ϵ greedy $\epsilon = 0.1$

Q-Learning off-policy

$$R_{t+1} + \gamma R_{t+2} + \gamma^2 \max_{a} \cdots$$

Chapter 34 Afterstate

Chapter 35 Eligibility traces

To reduce sample complexity

One way trying to speed up convergence

How you can debug the code

$TD(0) \longrightarrow TD(\lambda)$

$$G_t^{mc} = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \gamma^{T-t-1} R_T = G_t$$

$$G_t^{TD} = R_{t+1} + \gamma V(S_{t+1}) \equiv G_t^{(1)}$$ Truncated return

$$= R_{t+1} + \gamma R_{t+2} + \gamma^2 V(S_{t+2}) \equiv G_t^{(2)}$$

$G_t^{(n)}$

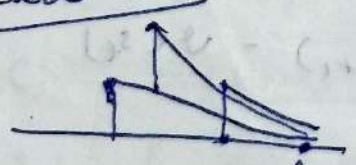
$$\Delta V_t(s) = \alpha \left[G_{t:n}^{(n)} - V_t(s) \right]$$

On-line $\rightarrow V_{t+1}(s) = V_t(s) + \Delta V_t(s)$

off-line $\rightarrow V(s) = V(s) + \sum_{t=0}^{T} \Delta V_t(s)$

no changes until an episode come to end

Backward view



Forward view

R

$$G_t^{\lambda} = 0.5 G_t^{(1)} + 0.5 G_t^{(2)}$$

Eligibility mechanism

Immediate updates

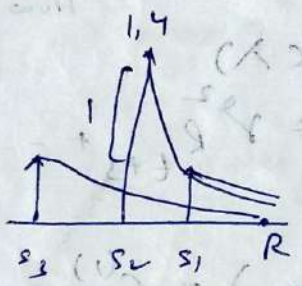
trace mechanism

$$G_t^{\lambda} = \cancel{0.33 G_t^{(1)} + \dots}$$

$$G_t^{\lambda} = (1-\lambda)\left[G_t^{(1)} + \lambda G_t^{(2)} + \lambda^2 G_t^{(3)} + \dots\right]$$

$$G_t^{\lambda} = (1-\lambda) \sum_{n=1}^{T-t} \lambda^{n-1} G_t^{(n)} + \lambda^{T-t-1} G_t$$

Chapter 3 6 Backward view of eligibility traces



$$e_t(s) = \begin{cases} \gamma\lambda\, e_{t-1}(s) & \text{if } s_t \neq s \\ \gamma\lambda\, e_{t-1}(s) + 1 & \text{if } s_t = s \end{cases}$$

Accumulating trace

Read section 7.4 first edition for derivation

$\mathbb{1}$ if $s_t = s$

Replacing traces.

$$\delta_t = R_{t+1} + \gamma V(s_{t+1}) - V(s_t)$$

$$\Delta V_t(s) = \alpha \cdot \delta_t \cdot e_t(s)$$

$$TD(0) = TD(1)$$

Dutch trace

Read sect 7.4

$$E_t(s) = \begin{cases} (\gamma\lambda)\, E_{t-1}(s) & \text{if } s_t \neq s \\ \beta(\gamma\lambda)\, E_{t-1}(s) + 1 & \text{if } s_t = s \end{cases}$$

Chapter 37 :- Eligibility trace control

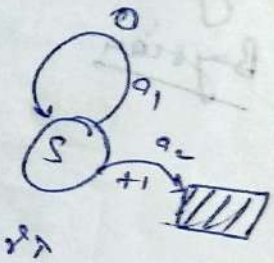
TD(λ) – Estimation

control

SARSA(λ)

$$E_t(s,a) = \begin{cases} \gamma\lambda E_{t-1}(s,a) + 1 & \text{if } s_t = s \text{ \& } A_t = a \\ \gamma\lambda E_{t-1}(s,a) & \text{o.w} \\ 0 & \text{if } s_t = s,\ A_t = a \\ 0 & \text{if } s_t = s\ A_t \neq a \\ \gamma\lambda E_{t-1}(s,a) & s_t \neq s \end{cases}$$

$\sum_{t=1}^{k} \gamma\lambda^t$



$(1+\gamma\lambda)\,\gamma\lambda$

$$\Delta Q_t(s,a) = \alpha \cdot \delta_t \cdot E_t(s,a)$$

$$\delta_t = R_{t+1} + \gamma Q_t(s_{t+1}, A_{t+1}) - Q_t(s_t, A_t)$$

Q(λ) Watkins & lambda

$$\delta_t = R_{t+1} + \gamma \max_{a'} Q_t(s_{t+1}, a') - Q_t(s_t, A_t)$$

By exiguting greed policy executing

$s \xrightarrow{a} s' \xrightarrow{\max a'} s'' \xrightarrow{\max a''}$

$E_t(s,a) = \gamma\lambda E_{t-1}(s,a) + 1$ if $s_t = s$ $A_t = a$

$\& A_t = \arg\max_{a'} Q_t(s_t, a')$

$= \gamma\lambda e_{t-1}(s,a)$ if $A \neq a$ $\& A_t = \arg\max_{a'} Q_t(s_t, a')$

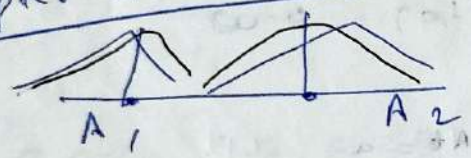
$= 0$ if $A_t \neq \arg\max Q_t(s_t, a')$

Off Policy TD(λ) | $\rho_t = \frac{\pi(s_t, a_t)}{\mu(s_t, a_t)}$

$E_t(s) = \rho_t(\gamma\lambda e_{t-1}(s) + 1)$

$= \rho_t(\gamma\lambda) e_{t-1}(s)$ if $s_t = s$

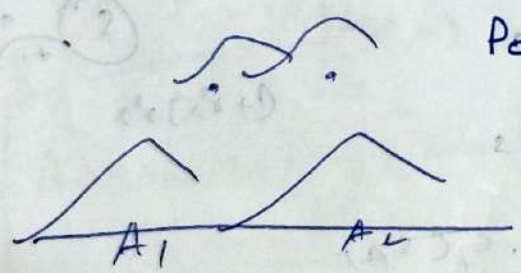
Chapter 8 Thompson Sampling



Bayesian

— Posterior sampling

Posterior sampling



Chapter 9 function approximation

Lookup table

— Memory

— ~~continuous state~~

— continuous state

— Generalize

Soln parameterized representation

— function Approximation

How out put depend on t

$v^\pi \quad \pi \quad \hat{v}(\cdot, \theta)$

Least square

$\sum_s d^\pi(s) \left[\hat{v}(s, \theta) - v^\pi(s) \right]^2$

Mean square

$\frac{1}{n} \sum_{t=1}^{n} \left[\hat{v}(s_t, \theta) - v^\pi(s_t) \right]^2$

$= \frac{1}{n} \sum_{t=1}^{n} \left[v^\pi(s_t) - \hat{v}(s_t, \theta) \right]^2$

G_t

$R_{t+1} + \gamma \hat{v}(s_{t+1}, \theta)$

Regr:- $\langle x, f(x) \rangle \longrightarrow \hat{f}(x)$

RL

$s_t \; A_t, R_{t+1} \quad s_{t+1}, A_{t+1}, R_{t+2}, \ldots$

$\hat{v}(s; \theta)$

$\langle s_t, R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots \rangle$

$\langle s_{t+1}, R_{t+2} + \gamma R_{t+3} \ldots \rangle$

$\langle s_t, R_{t+1} + \gamma \hat{v}(s_{t+1}, \theta_{old}) \rangle$

$\langle s_{t+1} R_{t+1} + \gamma \hat{v}(s_{t+2}, \theta) \rangle$

$\theta_{t+1} \leftarrow \theta_t + \alpha \nabla \left(R_{t+1} + \gamma \hat{v}(s_{t+1}, \theta_t) - \hat{v}(s_t, \theta_t) \right)^2$

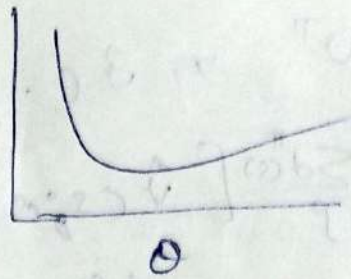
Chapter 4: Linear parameterization

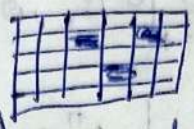
$s \rightarrow \langle \phi_1(s), \phi_2(s), \phi_3(s), \ldots, \phi_k(s) \rangle$

$\phi(s)$

ϕ_1

$\hat{v}(s, \theta) = \theta^T \phi$





values of features are linearly independent

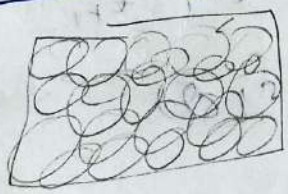
Linearly independent

for Linear

$\theta_{t+1} \leftarrow \theta_t$

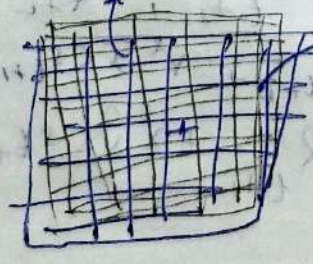
Chapter 41

State Aggregation method

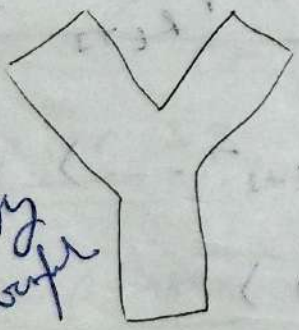


Coarse coding

Tile coding



multiple step grid



significantly powerful

* CMAC :- Cerebellar Model Actuator Control

Chapter 42 Approximation and eligibility

Traces

$$\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$$

$$\vec{e}_t = \gamma\lambda \vec{e}_{t-1} + \nabla_{\theta_t} V(S_t)$$

$$\vec{\theta}_{t+1} = \vec{\theta}_t + \alpha\delta_t e_t$$

control w. f.a

$$q_t(s,a)$$

SARSA

$\phi(s,a)$ (Small no. of action)

$$(\phi_a^T(s)\,\theta_a) = Q(s,a)$$

Large no. of actions:

- $\phi(s,a)$ - joint features
- After states

SARSA(λ)

$$\vec{e}_t = \gamma\lambda \vec{e}_t + \nabla_{\theta_t} \theta_1(s_t, a_t)$$

$$S - \phi(S) = \langle 0\,0\,1 \mid 1\,0\,1 \rangle$$

$$e_t^{(1)} = \gamma\lambda e_{t-1}(1)$$

$$e_t(4) = \gamma\lambda e_{t-1}(4)$$

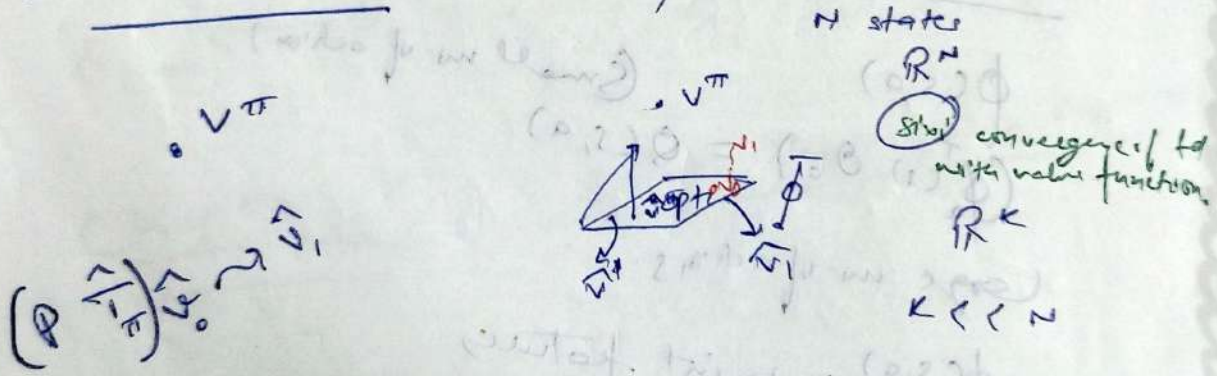
$$e_t(3) = 1 + \gamma\lambda e_{t-1}(3)$$

$$e_t(7) = 1 + \gamma\lambda e_{t-1}(7)$$

$$e_t(5) = \gamma\lambda e_{t-1}(5)$$

$$e_t(6) = 1 + \gamma\lambda e_{t-1}(6)$$

Chapter 43 LSTD & LSTDQ



$$\langle A_1 \dots, x_k), Y \rangle_{s=1}$$

$$\langle c\phi_1(s), \phi_k(s)\, u(s) \rangle$$

$$v(s) = \sum_{t=0}^{T} \gamma^t r(s)$$

$$(I - \gamma P_\pi)\hat{v}^\pi = \hat{v}^\pi$$

$$\hat{v}^\pi = (R + \gamma P^\pi \hat{v}^\pi)$$

$$\hat{v}^\pi = \bar{\phi}(\bar{\phi}^T\bar{\phi})^{-1}\bar{\phi}^T(R^\pi + \gamma P^\pi \hat{v}^\pi)$$

$$\hat{q}^\pi = \bar{\phi}(\bar{\phi}^T\bar{\phi})^{-1}\bar{\phi}^T(R^\pi + \gamma P^\pi \hat{q}^\pi)$$

$$\bar{\phi}\theta^T = \bar{\phi}(\bar{\phi}^T\bar{\phi})^{-1}\bar{\phi}(R^\pi + \gamma P^\pi \bar{\phi}\theta^\pi)$$

~~θ^T~~ = weighted version

$$\underbrace{\bar{\phi}^T\omega(\bar{\phi} - \gamma P^\pi \bar{\phi})\theta^\pi}_{A} = \underbrace{\bar{\phi}^T\omega R^T}_{b}$$

$$A\theta^\pi = b$$

$$A = \bar{\phi}^T\omega(\bar{\phi} - \gamma \vec{P}\bar{\phi})$$

$$= \sum_s \phi(s)\cdot\omega(s)\cdot\left(\phi(s) - \gamma\sum_{s'} P(s,\pi(s),s')\phi(s')\right)^T$$

$$= \sum_s \omega(s) \sum_{s'} P(s,\pi(s),s')\left[\phi(s)(\phi(s) - \gamma(\phi(s) - \gamma\phi(s'))\right]$$

$$D = \{(s_i, a_i, r_i, s_i') \mid_{i=1,2,\cdots L}\}$$

$$\tilde{A} = \frac{1}{L}\sum_{i=1}^{L}\left[\phi(s_i)(\phi(s_i) - \gamma\phi(s_i'))^T\right]$$

$$b = \bar{\phi}^T\omega R^\pi$$

$$= \sum_s \phi(s)\cdot\omega(s)\cdot\sum_{s'} P(s,\pi(s),s')R(s,\pi(s),s')$$

$$= \sum_s \omega(s)\sum_{s'} P(s,\pi(s),s')\left[\phi(s)\cdot R(s,\pi(s),s')\right]$$

$$\tilde{b} = \frac{1}{L} \sum_{i=1}^{L} \phi(s_i) \, r_i$$

$$\tilde{A} = \frac{1}{L} \sum_{i=1}^{L} \left[\phi(s_i, a_i) \left(\phi(s_i, a_i) - \gamma \phi(s_i', \pi s_i') \right)^T \right]$$

$$\tilde{b} = \frac{1}{L} \sum_{i=1}^{L} \phi(s_i, a_i) \, r_i$$

Chapter 44 LSPI and fitted-Q

LSTD / LSTDQ

$$\pi_{L+1}(s) = \arg\max_a \hat{\phi}^T \pi_T (s, a)$$

$$\pi_{t+1} \arg\max \phi \, \theta^{\pi_t}$$

$$a_j - \phi_1(s_1) \cdots \phi_k(s_1) \begin{pmatrix} s_1, \ldots, s_j \\ \vdots \\ a_1, \ldots, a_j - \pi(s_1) \end{pmatrix}_{t+1}$$

$$a_j \; \phi_1(s_0') \cdots \phi_k(s_0)$$

$$\hat{P}_0 \quad \phi(s_i, a_i) \, r_i + \gamma \max_a \phi_0(s_i', a)$$

Chapter 45 DQN and fitted Q Iteration

fitted Q-iteration

$\langle s_i, a_i, s_i, r_i \rangle$

$\langle \bar{Q}(s_i, a_i), target_i \rangle$

$target_i = r_i + \gamma \max_a \hat{Q}(\tilde{s}_i', a)$

Neural fitted Q. ~~FF~~ NFQ

DQN

Input — $\bar{Q}$ — $\hat{Q}$

18

"Online", off-policy algo

Replay Memory $\langle s_i, a_i, \tilde{s}_i, r_i \rangle$ efficient use

Transition replay