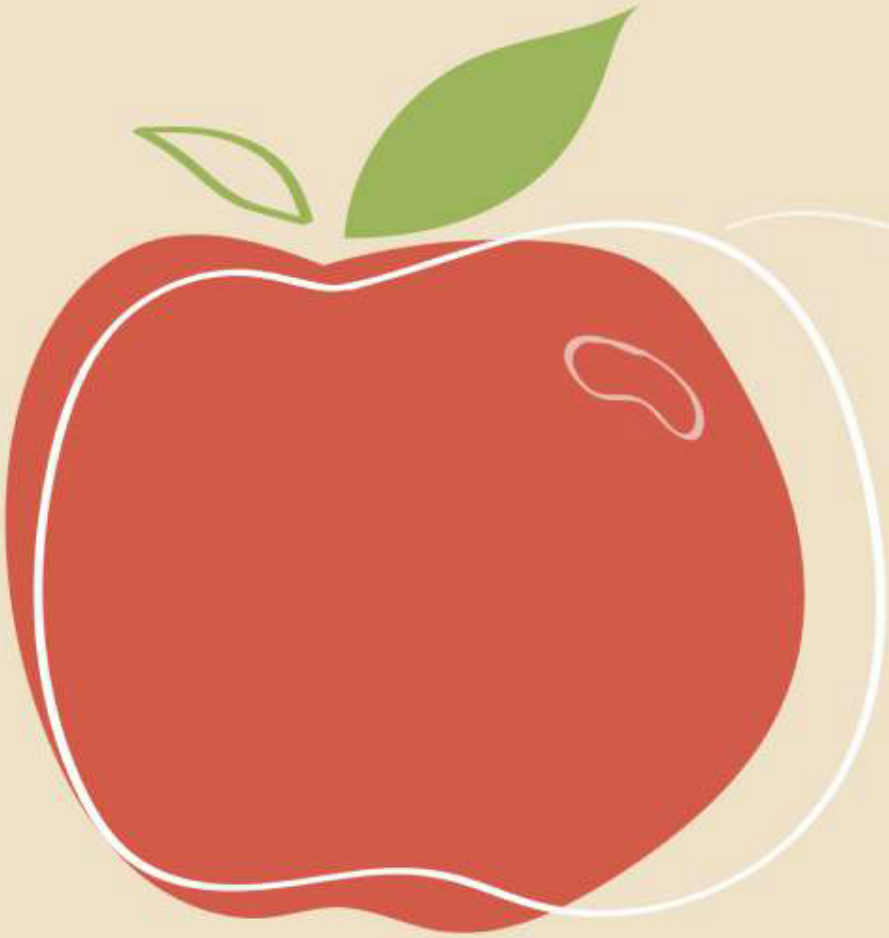
IE 620
Reinforcement learning algorithms



Ref: Reinforcement Learning by Richard S. Sutton and Andrew G. Barto

7TH JAN:

① single-agent algorithms ← algorithm itself is the agent
- Q-learning
- Actor-critic methods (with fundamental approximation)

② multi-agent algorithms (MARL) ← designing one algorithm for every agent
- non-cooperative
- cooperative (more complex but same as a single pseudo agent)

we look for equilibrium, can these agents learn nash equilibrium ↑ can be ported to a single agent algo

- de-centralised algorithms ← Nature of this overlaps with both non-coop and coop

decentralised models:

common objective
ant independency
consensus (there should be some agreement)

Eg: different company of logistic providers, algorithm common goal is to minimise total cost

many logistic service providers supply a raw material to a large firm ← $\sum_{i=1}^{n} f_i$ ← common objective

Eg: $\min(f_1(x) + f_2(x)) = \min_x f_1(x) + \min_y f_2(y)$ x and y are arguments

this is called separability / decomposibility

Note: In separable case, we will get many single agent and then we use single agent ideas

Eg: $\min\{f_1(x,y) + f_2(x,y)\}$ ← there is a consensus, it's needed but complex and so we cannot trasfer this to single agent

Sequential decision making models:

decision made at decision epochs/periods

Eg: routine decisions in networks (logistic / communication networks / ordering in decisions)

Eg: supply / value chain
like operating large dams, drinking, industrial use, agriculture

rainfall is very involved here as this is future (uncertainity) needs, evaporation (also uncertainites)

seepage
walls of reservoir
Exposed area

Note: dynamic system is an uncertain environment, it is about state evaluation in prev dam part

'state' $\in$ {state space}
↖ even time state space will change

uncertainity affects states

X_{n+1} = state at $n+1$

$X_{n+1} = X_n + (\quad) + w_n$
aggregate effects term
uncertainity
X_n: what we started with
Random variables

Eg: Remote mountains, a pond is left alone, only seepage / evaporation occurs, this is not an MDP / RL problem

when there is allocation we get reward and so we need to design them by MDP / RL

X_n or S_n is state, a_n is action

$r_n(s_n, a_n)$ = reward function

we have sequential random rewards

$r_1(\cdot,\cdot), \ldots r_n(s_n, a_n)$ ← $\pi(s_n)$

if we fix one policy we will get particular reward stream

we want to compare different series of rewards

Basic case:

action depends on policy π, s_n also depends on policy

$\forall$, n^{th} period $\begin{cases} r_n(s_n, \pi(s_n), w_n) \\ r_n(s_n, \tilde{\pi}(s_n), w_n) \end{cases}$

comparing random variables, so errors all in comparing r^{π}, r^{*}

There are multiple dimensions to compare rewards

Eg: $\begin{pmatrix}1\\1\end{pmatrix}$; $\begin{pmatrix}1\\-1\end{pmatrix}$; $\begin{pmatrix}-1\\1\end{pmatrix}$; $\begin{pmatrix}-1\\-1\end{pmatrix}$ → No sort of ordering, on real line, there is order

9th Jan:

S = set of states

A = set of actions

$A_s, s \in S$ → action set per state

Eg: AOI $A_S \in \{A, R\}$ $S \neq N$

$A_N = \{R\}$ $S = N$ ← full capacity of buffer

Markov decision models:

under a given policy, the queueing system is a Markov chain

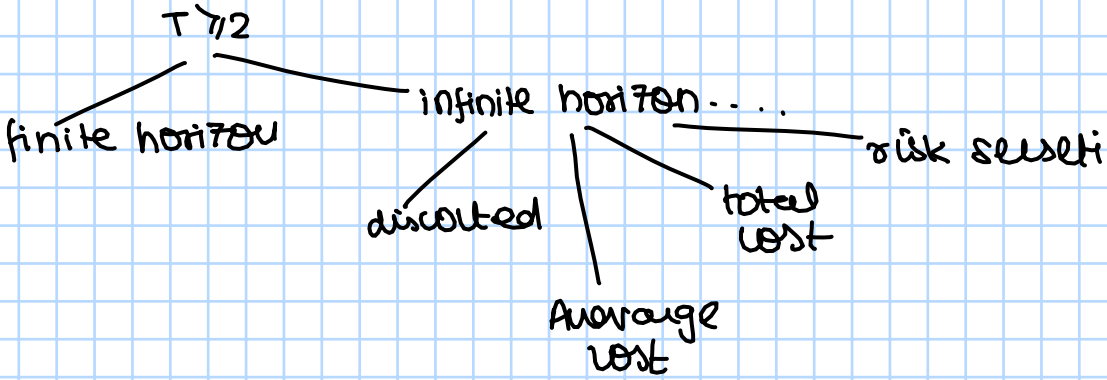
under action a in state s → state moves to s' in next period with prob $P(s'|s,a)$

→ system/algo has a reward $r(s,a)$ → $r(s,a,s')$

$$r(s,a) = \sum_{s' \in S} P(s'|s,a)\, r(s,a,s')$$

$\{P(j|s,a)\}$ → dynamics

planning horizon:



discounted models:

$\lambda \in (0,1)$ discount factor

$$\sum_{n=1}^{\infty} \lambda^{n-1} \mathbb{E}^{\pi}[r(X_n, Y_n)] = v_\lambda^{\pi}(X_1) \in \mathbb{R}$$

r has to be money

$\pi' v_\lambda^{\pi}(X_1) \in \mathbb{R}$

now we can compare them

when objective functions are money streams

setup is that, discounted model has money streams

λ is related to the RBI's prevailing interest rate, r ← transient behavior

Note: path following robots use $\lambda = 0.9$ as their context is to approximate finite horizon to infinite stuff as tail does not matter (same for LAN)

Average cost model:

$\{r(X_n, Y_n)\}$

basic case: finite state-action model

$$\lim_{T\to\infty} \frac{1}{T} \sum_{n=1}^{T} \mathbb{E}^{\pi}[r(X_n, Y_n)]$$

(policy not stationary dist)

$$= \sum_{S\in\mathcal{S}} \pi(S) r(S, \pi(a))$$

↑ this is stationary distribution of markov chain

Time average (late Ronald Wolff, IEOR UCB)

$$\lim_{T} \frac{1}{T} \sum_{n=1}^{T} r(X_n, Y_n) \to \text{a r.v}$$

Sample: $\{X_1, X_2, \ldots\}$

$$S_n := \sum_{t=1}^{n} X_t(\omega)$$

$$\mathbb{E}[X_1] = \frac{S_n}{n} \quad a.s \quad (LLN)$$

Note: in this case the tail will determine the value

so, average cost criteria captures tail behaviour, relevant for asymptotic growth, non monotonic growth like e.g pareto dropped, etc

Total cost:

board game, shotest paths, ...

$$\mathbb{E}^{\pi}\left[\sum_{}^{\tau_\pi} r(X_n, Y_n)\right]$$

(τ_π ← random variable)

RL for this with multiagents

Note: the total cost criteria is general case of the other two cases

Discounted models $\Leftrightarrow$ LPs
Average models $\Leftrightarrow$ LPs
} via optimality equations

Total cost $\Rightarrow$ LP

Finite Horizon:

have there is a DP backward induction

$N \to$ $1 \le k \le N-1$ ← Backward induction for all

↑ initialisation

14th Jan:

we did MDP review, 3 cost criteria
Discounts, average, total cost and their differe

↑ planning horizon is finite but random
or shortest-path problem
(no LP)
- stopping time

$$\mathbb{E}^{\pi}\left[\sum_{1}^{\tau} R_n^{\pi}\right]$$

Finite horizon model:

the DP recursion

$\$, A, r_t(\cdot,\cdot)$, $N \rightarrow$ planning horizon (total periods)

$t=N$ $\quad v_N^{\pi}(h_N) = r_N(s_N) \quad \forall h_N = (h_{N-1}, s_{N-1}, a_{N-1})$

$yt=1$ stop and return policy

compute $t-1 \leftarrow t$ $\quad u_t^{\pi}(h_t) = r_t(\quad) + \sum p_t(\cdot|\cdot)\, u_{t+1}^{\pi}(\cdot)$

end output optimal policy $(\pi_1, \pi_2, \ldots \pi_{N-1})$

↖ take action till N-1 periods

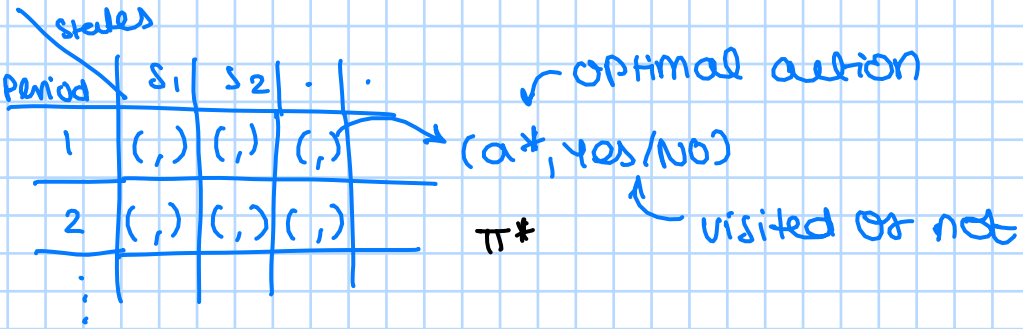
$1 \ldots -N-1, N$

↖ no actions in N period

Theorm: (correctness DP recursion) $v^{\pi^*}(s_1) = u(s_1) \quad \forall s_1 \in \$$

$$v^{\pi^*}(s_1) = \mathbb{E}^{\pi^*}\left[\sum_{t=1}^{N} r_t(\cdot,\cdot)\right]$$

Note: $\pi_t^* = \underset{a \in A_t}{\operatorname{argmin}}\, U_t^{\pi}$ ← if we have cost to go



this is lookup table (implementation viewpoint)

K = # states
L = # actions in each state

at period t: $K^{t+1} L^t$

so history model: $K^2 \sum_{0}^{N-1} (KL)^t$ # policies

markov model: $(N-1)K^2$ # policies

Optimality equations:

$$u_t(h_t) = \underset{a_t}{\text{or min}}\ \max \left(r_t(s_t, a_t) + \sum_{s \in S} p(s \mid s_t, a_t) u_{t+1}(h_t, a_t, s) \right)$$

Note: there are unique $V^{\pi^*}(s_1)$ but not policies

$$\underbrace{\pi_t^*(s_t)}_{a^*(s_t)} = \underset{a \in A}{\operatorname{argmin}}\ u_t^{\pi}$$

we are solving 2 period problem

$$\underbrace{r_t(s_t, a_t)}_{1 \text{ period}} + \underbrace{\sum_{s \in S} p(s \mid s_t, a_t) u_{t+1}(h_t, a_t, s)}_{1 \text{ period}}$$

16th Jan:

Finite horizon model:

① DP equation and optimality equation

② principles of optimality, cost-to-go fution

Infinite horizon discounted model:

If $|S| \times |A| < \infty$

$v_\alpha^\pi(\cdot)$ and v_α^* and $v_\alpha^{\pi^*} = v_\alpha^*$

discounted optimal equation

OE is $v_\alpha^\pi(s) = \min_a \{ r(s,a) + \lambda \sum_i p(i|s,a) v_\alpha^\pi(i) \}$ $\forall s$

↳ solution of this is the optimal π^*

solution exist, and have a v_α^* & unique
and ∃ optimal policy

policy class that is enough/adequate, we want to know how rich the class is

↳ Π^{SD}, Π^{MD}, Π^{HD}, Π^{SR}, Π^{MR}, Π^{HR}, k step memory can also be used

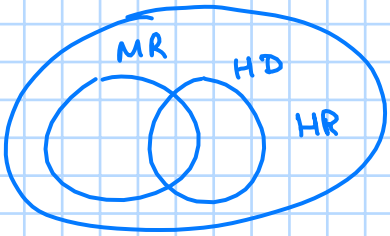
Eg: policy class ① all available history
② only current state / 'latest' state
③ any given k-step memory

one more way to characterise policies is by

- picking one action only — deterministic policy
- randomise any available actions — randomised

combinations of the two types of class classification gives us diff't class

markov deterministic → . . . . ← history dependent randomised

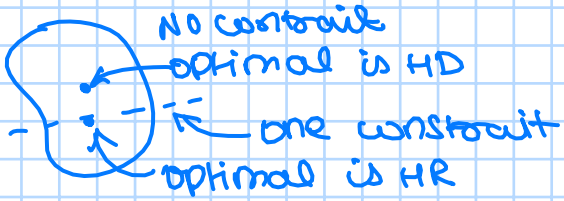


one does not contain another

for a given class, say H*, HR v/s HD ——— ①

↑ this is enough as

objective function and optimal policy HR exist



Constraints : ← intrinsic to the model

① constraint of action taken $\{A, R\}$, $\{R\}$

② attain some functional constraints, battery usage $\leq c$

now, we know if ② type of constraint not given, we will do Deterministic results

there is one more result which is

$$\sup_{\pi \in \pi^{HR}} v_\alpha^\pi(s) = \sup_{\pi \in \pi^{MD}} v_\alpha^\pi(s) \quad \forall s \in S$$

this result is due to Robert Strauch (PhD thesis with David Blackwell) that markov random policies are good enough ——— ②

from ①, ② we get markov decision is enough without loosing optimality

we took a history dependent policy, we generate markov randomised policies, the state action pair probabilites match, and so Expectations will be the same

$\pi \in \pi^{HR}$, $P^\pi(X_n = j, Y_n = a \mid X_1 = s_1)$, $\exists\, \pi' \in MR$ s.t

$$P^{\pi'}(X_n = j, Y_n = a \mid X_1 = s_1) = P^\pi(X_n = j, Y_n = a \mid X_1 = s_1)$$

if we have probability on X_1, we get $\forall \pi \in \pi^{HR}$, $\exists \pi' \in \pi^{MR}$ s.t

$$P^\pi(X_n = j, Y_n = a) = P^{\pi'}(X_n = j, Y_n = a)$$

now, for some unconstrained models, randomisation is not needed, deterministic processes are enough

so, enough to explore MD

Note: proof by Robert Strauch is a constructive proof to an Existance question

Computations and others :

in discounted model, there is
↪ value iteration (solve sequence of finite horizon model)
n period computation, take $n \to \infty$

Note: finite horizon models have something called, like in $|T|=90$ for a shoping sale when to announce the discount

day → only once

demand increase, profit go down if very early

and

demand decrease, profit go down as untold, very late

this kind of problems are called end of horizon models as well discount announcement, demand changes

above in value iteration, as $n \to \infty$ we don't get this effect of end of horizon

in VI, find v^* first, then recover policy by OE

↳ policy iteration (PI), we fix a policy, we tweak policy to get a better one and we keep doing that

↳ PE (policy evaluation)

PI (policy improve) using optimality equations

to avoid inf loop, we check if policy is optimal or not

now, $v_\alpha(\cdot) = \min_a \{ \quad \}$

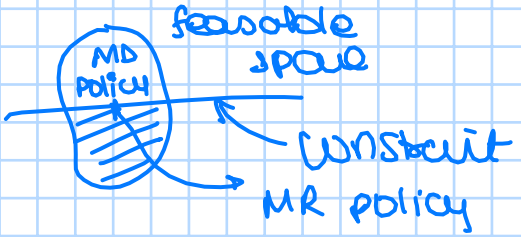
$v_\alpha(\cdot) \geq \{ \quad \} \forall a$, $\forall s$ we will have a bunch of linear inequalities

$\Rightarrow \min_{\alpha} \sum_{s \in S} \alpha(s) v(s)$ → returns α^*, v^* s.t

this is LP

then we can do dual also to get

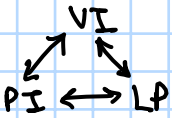
$\max_{(a,s)} \sum_{s,a} x(s,a) r(s,a)$ → returns x^*



Note: PI and LP are same, this is one-one correspondence, two diffent representations

21st Jan:

we saw approaches: ① VI
② PI (we also saw PI iff LP)



Total cost problems:

shortest path problems (SSP)

C = random time

reach end node / goal from start node

↑ there can also be congestion cost

$\pi \rightarrow$ policy, under this policy

$\pi^* = \underset{\pi}{\operatorname{argmin}} \mathbb{E}^{\pi}\left[\sum_{i=1}^{\tau} C_i\right]$ ($\tau \rightarrow$ random variable)

↑ optimal policy ↑ may be random variable

Note: we assume anything to be random but finite

Eg: path planning, board games, navigation

popular but not well studied

Note: learning version papers are very new, dp version is old

we want to know the learnability of an algorithm

MDP: $\{P_{ij}^{\pi}\}, \{\gamma^{\pi}(\cdot,\cdot)\}$

RL: ? → ? we have no probability / reward but we have data

(learning version)

so we have jumps and we want to know
① optimal policies
② we keep guessing algorithm and update it as time goes
we keep on updating policies

adaptive algorithms

recursive algorithms / adaptive / bootstraping:

$\pi_{t+1} = \pi_t + (\text{correction term})$ ← using data from period $t+1$

none generally, $\pi_{t+1} = f_t(\pi_t, \text{data from period } t+1)$

↑ some function

above is a recursion

f_t is a design question, key is to find such f_t

fdbk = feedback update

say, $s_t, a_t, r_t, s_{t+1}, a_{t+1}, \ldots$
→ data stream

SARSA and many other algorithms

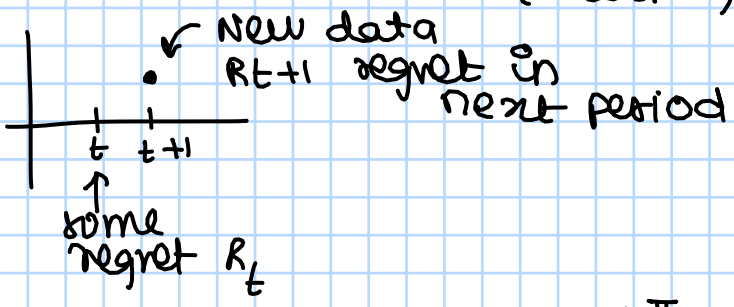
on policy / off policy
↑ ↖ data from diff. policy
adaptively (data generated by current algorithm)

reward is function of estimate → fdbk

Correctness/learnability:

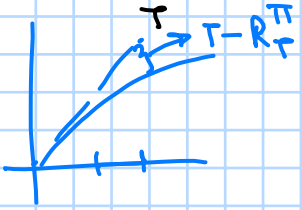
Regret of algo = (cost of algo) − (optimal cost) ≥ 0

≥ (optimal cost)



on planning horizon, R_T^π = end regret (output of π)

$\frac{R_T^\pi}{T} \to 0$ is aim or R_T^π is sublinear



the gap should widen as $t \to t+1$

i.e
$t - (R_t) < (t+1) - (R_{t+1})$
gap widens

Algos (learnability):

convergence, sublinear regret
↑ discount/avg cost ↑ total cost solution

Model free and model based:

$S, A, r(\cdot,\cdot), P(\cdot,\cdot)$ using data, we can estimate $r(\cdot,\cdot)$, $P(\cdot,\cdot)$

RL
MDP

model based approach

model based RL, Estimate r, P and 'use' MDP theory
(2 stage)

model free: they don't use MDP, direct estimate π and then keep updating hoping convergence

model free: estimate optimal policy (there is no model)

non-linear functional approximation: Deep RL (Neural network)

linear functional approximation

Q-learning:

It is an Off-policy ↖ suit of algorithms (Q-learning one of them)
a TD algorithms (temporal differences)

Q-factors

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \max Q(S_{t+1}, a) - Q(S, a) \right]$$

↑ update rule

$\gamma \in (0,1)$ ↑ discount parameter

$\alpha \in (0,1)$ ↑ learning rate

recall, value $V^{\pi}(s)$ $s \in S$

$$V^{\pi}(s) = \mathbb{E}^{\pi} \left[\sum_{1}^{\infty} \lambda^{n-1} r(s_n, a_n) \,\middle|\, s_1 = s \right]$$

↑ given $s_1 = s$

$$Q^{\pi}(s,a) = \mathbb{E}^{\pi} \left[\sum \lambda^{n-1} r(s_n, a_n) \,\middle|\, s_1 = s, a_1 = a \right]$$

↑ given $s_1 = s$, $a_1 = a$

23rd Jan:

MDP:

S = state space $\quad S = \{1, 2, \ldots n\}$

A = action space $= \{a_1, a_2 .. , a_m\}$

or A_S

$P(j|s,a)$ = prob of landing in j (if in s, choose action a)

$r(s,a)$, $r(s,a,j)$ then $\quad r(s,a) = \sum_{j \in S} P(j|s,a)\, r(j,s,a)$

Markovian randomised:

$$d_t : S \longrightarrow \Delta(A)$$

$$\pi = (d_1, \ldots)$$
$$d_2(0) = (\ )$$

markov deterministic:

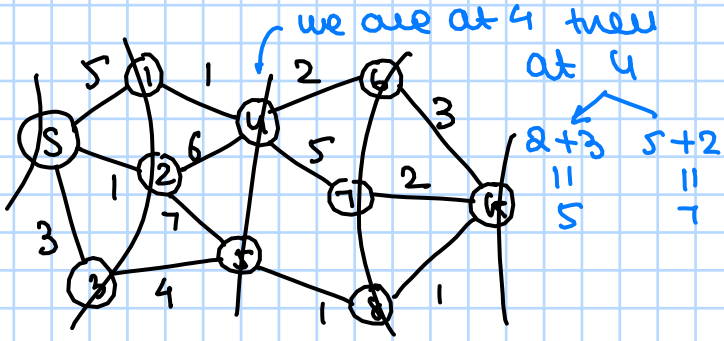
$$d_t : S \longrightarrow A$$

$$\pi(d_1, d_2 \ldots)$$

Transition prob matrix
TPM (markov chain)

$$P(j|s) = P(j|s, d(s))$$

shortest path:



we do backward induction at $t=3$

$6 \to 6 \quad 3$

$7 \to 6 \quad 2$

$8 \to 6 \quad 1$

$t=2$: $\quad 4 \to 6 \to 6 \quad 5$

$5 \to 8 \to 6 \quad 2$

$t=1$: $\quad 1 \to 4 \to 6 \to 6 \quad 6$

$2 \to 5 \to 8 \to 6 \quad 9$

⋮

stochastic shortest path:

now, $T=\{0,1\ldots N\}$ $T=N$ $h(s)\ \forall s$ → terminal reward function

$h(s)\ \forall s\in S$

$t=N-1,\quad U_t^*(s)=\max_{a\in A_s}\left\{r(s,a)+\sum_{j\in S}P(j|s,a)U_{t+1}^*(j)\right\}$

$$U_{N-1}^*(s)=\max_{a\in A_s}\left\{r(s,a)+\sum_{j\in S}P(j|s,a)h(j)\right\}$$

deterministic

$$\text{then}\quad d_{N-1}^*(s)=\operatorname{argmax}_{a\in A_s}\left\{r(s,a)+\sum_{j\in S}P(j|s,a)h(j)\right\}$$

secretary problem:

candidates $=\{1,2,\ldots N\}$

$S=\{0,1\}$ $A=\{C,Q\}$ $1\rightarrow$ Best so far, $0\rightarrow$ Not best so far

interviewed t so far, current is 0

at time t: $P_t(1|0)=\frac{1}{t+1}$

$P_t(1|1)=\frac{1}{t+1}$

$P(\text{best is in first } t)=\frac{\binom{N-1}{t-1}}{\binom{N}{t}}=\frac{t}{N}$

(1 is always with us; choose t sets)

$$P_t(0|1)=P_t(0|0)=\frac{t}{t+1}$$

$h_N(0)=0\quad h_N(1)=1$

$U_t^*(0)=\max_{C,Q}\left\{0,\ 0+\sum_{j=0,1}P(j|s)U_{t+1}^*(s)\right\}$

$=\max_{C,Q}\left\{0,\ \frac{1}{t+1}U_{t+1}^*(1)+\frac{t}{t+1}U_{t+1}^*(0)\right\}=\frac{t}{t+1}U_{t+1}^*(0)+\frac{1}{t+1}U_{t+1}^*(1)$

$U_t^*(1)=\max_{C,Q}\left\{\frac{t}{N},\ \frac{1}{1+U}U_{t+1}^*(1)+\frac{t}{t+1}U_{t+1}^*(0)\right\}$

↑ prob this is best

so, if $\exists \tau$ s.t $u_\tau^*(1) > \frac{\tau}{N}$ then $\forall t < \tau$, $u_t^*(1) > \frac{t}{N}$

as $u_\tau^*(1) = u_c^*(0)$

$u_{\tau-1}^*(0) = u_\tau^*(0)$

$\Rightarrow u_{\tau-1}^*(1) = \max\left\{\frac{\tau-1}{N}, \underbrace{u_\tau^*(0)}_{> \tau/N}\right\}$

$\Rightarrow u_{\tau-1}^*(1) = u_\tau^*(0)$

then it follows through induction

Policy Iteration algorithm:

step 1: choose a policy π

step 2: find the value of policy

$v^\pi = (I - \lambda P_d)^{-1} r_d$

there are iterative methods also

$d_n(s) = \underset{a \in A_s}{\operatorname{argmax}} \left\{ r(s,a) + \lambda \sum_{j \in S} P(j|s,a)\, v_n(j) \right\}$ ← iterate

$d_{n+1} = d_n$ then we are done, we repeate process till equal

$\pi = (d_{n+1}, d_{n+1}, d_{n+1} \ldots) \rightarrow$ optimal policy

28th Jan:

Q-Learning:

over episode, loop over
choose action as per ε-greedy (below)
note R, S'
$Q(S,a) \leftarrow Q(S,a) + \alpha[R + \gamma \max_a Q(S',a) - Q(S,a)]$
$S \leftarrow S'$

ε-greedy scheme:

$\varepsilon \in (0,1)$ is input parameter
non-greedy actions, choose with probability $\frac{\varepsilon}{|A(S)|} > 0$
and greedy action w.p $1-\varepsilon + \frac{\varepsilon}{|A(S)|} \in (0,1)$
also called ε-soft policy

Note: above is also called exploration-exploitation

from Q-learning → compute Q-table $Q(\cdot,\cdot)$
we can also put ε-greedy by ① each action $\frac{\varepsilon}{|A(S)|}$
② $1-\varepsilon$ rest all policy

ε-greedy (ε):
if unif$(0,1) < \varepsilon$:
randomly pick($A(S)$)
else:
$a = \arg\max_a (Q(S,a))$

algorithm parameters: step size $\alpha \in (0,1]$, small $\varepsilon > 0$
initialise $Q(S,a)$ $\forall S \in \mathcal{S}$, $a \in A(S)$ $Q(\text{terminal},\cdot) = 0$

loop for each episode:
initialise S
loop for each step of episode:
choose A from S using policy derived from Q (ε-greedy)
take action A, observe R, S'

$$Q(S,a) \leftarrow Q(S,a) + \alpha[R + \gamma \max_a Q(S',a) - Q(S,a)]$$

$S \leftarrow S'$
until S is terminated

↑ in optimality equations $Q^*(S,a) = R + \gamma \max_{a'} Q^*(S',a')$

so, $R + \gamma \max_a Q(S',a) - Q(S,a)$ is the deviation of current Q to optimal
or so if Q^* stabilises, $R + \gamma \max_{a'} Q(S',a) - Q(S,a) \to 0$ or no need to update

Convergence analysis:

In finite $\mathcal{S}, \mathcal{A}$ MDPs, $Q(x, A_t) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R(X_t, A_t) \mid X_0 = x]$
and have optimal value $v^*(S) = \max_{A_t} Q(S, A_t)$

and in DP equation:

$$v^*(s) = \max_a \left\{ r(s,a) + \sum_{j \in S} P(j|s,a) v^*(j) \right\}$$

$$= \max_a \left\{ \sum_{j \in S} P_a(j|s) \left(r(s,a,j) + v^*(j) \right) \right\}$$

$$Q^*(s,a) = \sum_{y \in S} P_a(s,y) [r(s,a,y) + \gamma v^*(y)]$$

this is a fixed point equation of the nature $f(x) = x$

Existance, uniqueness, etc comes from fixed point equation

Theorem: $Q(\cdot,\cdot)$ is a fixed point of the contraction map/operator H

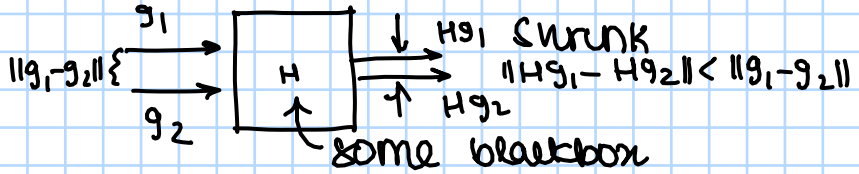
$$(HQ)(s,a) = \sum_{y \in S} P_a(s,y) \left[r(s,a,y) + \gamma \max_{b \in A} Q(y,b) \right]$$

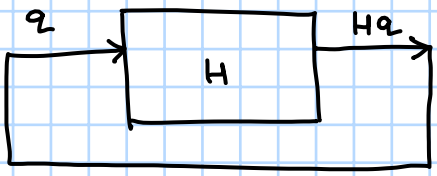
via sup-norm

proof:

if we show $\|Hq_1 - Hq_2\|_\infty \leq \gamma \|q_1 - q_2\| < \|q_1 - q_2\|$

$\gamma \in (0,1)$





$H^n q \to q^*$ as $n \to \infty$

$\{q, Hq, H^2q, H^3q, \ldots H^n q \ldots\} \to$ sequence $\{H^i q\}_{i=0}^{\infty}$

this sequence converges

$$\|Hq_1 - Hq_2\|_\infty = \max_{x,a} \left| \sum_{y \in S} P(x,y) \left[\gamma(x,a,y) + \gamma \max_b q_1(y,b) - \gamma(x,a,y) - \gamma \max_b q_2(y,b) \right] \right|$$

$$= \gamma \max_{x,a} \left| \sum_{y \in S} P(x,y) \left[\max_b q_1(y,b) - \max_b q_2(y,b) \right] \right|$$

$$\leq \gamma \max_{x,a} \left(\sum_{y \in S} P(x,y) \left| \max_b q_1(y,b) - \max_b q_2(y,b) \right| \right)$$

$(\because \max E \leq E \max$ and $|\ |$ is convex $)$

$$\leq \gamma \max_{x,a,y,b} |q_1 - q_2| \underbrace{\left(\sum_{y \in S} P(x,y) \right)}_{1} \quad (\because |\max(\,) - \max(\,)| \leq \max|(\,) - (\,)|)$$

$$\leq \gamma \|q_1 - q_2\|_\infty$$

$\mathcal{B}$, H is a contraction
by Banach fixed point theorem
$H^n g \longrightarrow Q^*$ and is unique

now, as arg of Q learning is $H \Rightarrow$ Q learning conv to Q^*

Note: they converge no matter what initial value of Q is, rate of convergence is geometric at rate of γ

Note: value iteration in MDP setup is similar to this

above as we have $\sum P(x,y)$ we used to visit all (S,a) pairs infinitely; which is ensured by ε-greedy policy

let π be a policy, $P_\pi(A_t = a \mid X_t = s) > 0$ } Behavior policy that generates samples

$\{x_t\}$, $\{a_t\}$, $\{r_t\}$ $r(\cdot,\cdot,\cdot)$ → data used for Q learning

$$Q(S_t, a_t) \leftarrow Q(S_t, a_t) + \alpha \left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, a_t) \right]$$

30th Jan :

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

$H_q(x,a) = \sum_{y \in S} P_a(x,y)[r(x,a,y) + \gamma \max_{b \in A} q(y,b)]$ is a contraction even if we do $q(s,a) = 0 \ \forall s \in S, a \in A$ ($\because$ initialisation)

we had proved, $\|Hq_2 - Hq_1\|_\infty \leq \gamma \|q_1 - q_2\|_\infty$

do this action

now, $\theta_{t+1}(x_t, a_t) = [1 - \alpha_t(x_t, a_t)]\,\theta_t(x_t, a_t)$

$+ \alpha_t(x_t, a_t)[r_t + \gamma \max_{b \in A} \theta_t(x_{t+1}, b)]$ ——— (1)

r_t ↑ get this reward; $\max_{b \in A}$: next state

Eg: $\alpha(x_t, a_t) = \frac{1}{n}$ $\quad \sum \frac{1}{n} \to \infty$ but $\sum \frac{1}{n^2} < \infty$

Theorem: given a finite MDP (S, A, P, γ) the Q learning algorithm with update rule (1) converges with probability 1 to Q^* under assumptions, $\sum \alpha_t(x_t, a_t) = \infty$, $\sum \alpha_t^2(x_t, a_t) < \infty$

Note: we need $\alpha_t(x_t, a_t)$ to visit every state, action infinity many times (step size)

if $\{X_t\}$ is stochastic process, $\mathcal{F}_t = \sigma(X_1, .. X_t)$

Theorem: The random process $\{\Delta_t\}$ taking values in $\mathbb{R}^n$

$\Delta_{t+1}(x) = [1 - \alpha_t(x)]\,\Delta_t(x) + \alpha_t(x) F_t(x)$ converg to 0 wrt $\|\cdot\|$, under assumptions

(a) $0 \leq \alpha_t \leq 1$, $\sum \alpha_t = \infty$, $\sum \alpha_t^2 < \infty$

(b) $\|\mathbb{E}[F_t(x) | \mathcal{F}_t]\|_\infty \leq \eta \|\Delta_t\|_\infty$ with $\eta \in (0,1)$

(c) $\mathrm{var}[F_t(x) | \mathcal{F}_t] \leq C[1 + \|\Delta_t\|_\infty^2]$

in Q-learning, $\theta^* - \theta_t \to$ i.e $\Delta_t = \theta_t - \theta^*$ is true seq Δ_t then $\Delta_t \to 0$ with prob = 1

in 1, subtract $\theta_t - \theta^*$ on both sides to get

$\theta_{t+1}(x_t, a_t) - \theta^*(x_t, a_t) = (1 - \alpha_t(x_t, a_t))(\theta_t(x_t, a_t) - \theta^*(x_t, a_t))$

$+ \alpha_t(x_t, a_t)[r_t + \gamma \max_{b \in A} \theta_t(x_t, b) - \theta^*(x_t, a_t)]$

$x = (x_t, a_t)$

$\Rightarrow \Delta_{t+1}(x) = (1 - \alpha_t(x))\Delta_t(x) + \alpha_t(x)[r_t + \gamma \max_{b \in A} \theta_t(x_t, b) - \theta^*(x)]$

$r_t + \gamma \max_{b \in A} \theta_t(x_t, b) - \theta^*(x) = F_t(x)$

$F_t(x_t, a_t) = r_t + \gamma \max_{b \in A} \theta_t(x_{t+1}, b) - \theta^*(x_t, a_t)$

r_t ↑ reward at time t $\quad r_t = r_t(x_t, a_t, x_{t+1})$ ↖ $X(x_t, a_t)$

↳ Random variable on state $t+1$

$$F_t(x,a) = r_t(x,a,X(x,a)) + \gamma \max_{b\in A} Q_t(X(x,a),b) - Q^*(x,a)$$

$$\mathbb{E}[F_t(x,a)\,|\,\mathcal{F}_t] = \sum_{y\in S} P(x,y)\left(r_t(x,a,y) + \gamma \max_{b\in A} Q_t(y,b) - Q^*(x,a)\right)$$

$$= HQ_t(x,a) - Q^*(x,a)$$

(Q^*: fixed point of H)

$$= HQ_t(x,a) - HQ^*(x,a)$$

$$\le \gamma \|Q_t(x,a) - HQ^*(x,a)\|_\infty$$

$$= \gamma \|\Delta_t\|_\infty$$

$$\Rightarrow \|\mathbb{E}[F_t(x,a)\,|\,\mathcal{F}_t]\|_\infty \le \gamma \|\Delta_t\|_\infty$$ → assumption (b) above

NOTE: the assumption (a) was already been taken

now, $\mathrm{var}[F_t(x,a)\,|\,\mathcal{F}_t]$

$$= \mathrm{var}[r(x,a,X(s,a)) + \gamma \max_{b\in A} Q_t(X(x,a),b) - Q^*(x,a)]$$

$$= \mathbb{E}[r(x,a,X(s,a)) + \gamma \max_{b\in A} Q_t(X(x,a),b) - Q^*(x,a) - (HQ_t(x,a) - Q^*(x,a))]^2$$

$$= \mathbb{E}[r(x,a,X(s,a)) + \gamma \max_{b\in A} Q_t(X(x,a),b) - HQ_t(x,a)]^2$$

$$= \mathrm{var}[r(x,a,X(s,a)) + \gamma \max_{b\in A} Q_t(X(x,a),b)]$$ ↳ also bounded

$$\le C \quad (\because \text{rewards are bounded})$$

$$\le C(1+\|Q_t\|^2)$$

$$\max r(x,a,X(s,a)) \le \max_y r(x,a,y)$$

$$Q_1 = (1-\alpha)Q_0 + \alpha(r + \gamma Q_0(\,)) \le C \Rightarrow Q_2 \text{ is bounded and so on}$$

(α: bounded, Q_0: bounded, C: bounded)

By some bound $\le M\left[\frac{1}{1-\gamma}\right]$ ↑ By r bound

THEOREM: given a finite MDP(S,A,P,γ) the Q learning algorithm with update rule (1) converges with probability 1 to Q^* under assumptions, $\sum \alpha_t(x_t,a_t) = \infty$, $\sum \alpha_t^2(x_t,a_t) < \infty$

proof: using the above stochastic convergence theorem (Δ_t) and steps done above on F_t, this follows trivially

4th Feb:

Our iterative scheme:

$$X_{n+1}(\omega) = (1-\alpha_n) X_n(\omega) + \alpha_n F_n(\omega) \quad \text{——— (1)}$$

$\{X_n\}$ converges almost surely to 0, under

(1) $\sum \alpha_n^2 < \infty$, $\sum \alpha_n = \infty$ $\alpha_n \in (0,1)$

(2) $\| \mathbb{E}[F_n(x) | \mathcal{F}_n] \|_w < \gamma \| X_n \|_w$ where $\forall \in (0,1)$

(3) $\text{var}(F_n | \mathcal{F}_n) \leq c(1 + \| X_n \|_w^2)$ for some $c > 0$

Note: convex combination of X_n and F_n, F_n such that its expectation bounded

it involves noise F_n

mean, var grow sublinearly

$$X_{n+1}(\omega) = (1-\alpha_n) X_n(\omega) + \alpha_n F_n(\omega)$$

new signal — signal — Noise

Noise should not affect signal so we need some properties

now, 'eliminated' by two signals in (1)

say $\alpha_n = \frac{1}{n}$, then $\sum_{n=1}^{\infty} \alpha_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots \to \infty$

$$\sum_{n=1}^{\infty} \alpha_n^2 = 1 + \frac{1}{2^2} + \frac{1}{3^2} \cdots < \infty$$

such a convex combination is true in like LLN

$\{X_n\}$ & $S_n = X_1 + \cdots + X_n$

$$\frac{S_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{S_{n-1} + X_n}{n}$$

$$= \frac{S_{n-1}}{n} + \frac{X_n}{n}$$

$$= \frac{(n-1)}{(n-1)} \frac{1}{n} S_{n-1} + \frac{X_n}{n}$$

$$= \left(1 - \frac{1}{n}\right)\left(\frac{S_{n-1}}{n-1}\right) + \frac{X_n}{n}$$

if $Y_n = \frac{S_n}{n}$

then $Y_n = \left(1 - \frac{1}{n}\right) Y_{n-1} + \left(\frac{1}{n}\right) X_n$

contributions and recent observations under above considerations, then SLLN, $Y_n \to \mathbb{E}[X_1]$ a.s

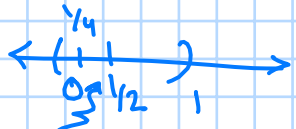
$$P\left(\omega : \frac{S_n(\omega)}{n} \to \mathbb{E}[X_1]\right) = 1 \quad (\because \text{SLLN})$$

Set on which this event (set of ω) occurs has probability 1

Eg: $X_i^\omega = \begin{cases} 1 & ; \text{ith coin is H} \\ 0 & ; \text{ith coin is T} \end{cases}$

$$\mathbb{E}[X_1] = \tfrac{1}{2}\times 1 + \tfrac{1}{2}\times 0 = \tfrac{1}{2}$$

$\frac{S_n(\omega)}{n} \to \frac{1}{2}$, set of trajectories is uncountable



$X_1 = H$ | $X_2 = T \ldots$

keep zooming in

course will also work in binary (Emily Borel construction)
$\forall$ point on $(0,1)$, $\exists \omega$ s.t

$$P\left(\omega : \frac{S_n(\omega)}{n} \to \mathbb{E}[X_1]\right) = 1$$

Note: HHTHHTHHT.... HHT.... they will not occur as $\mathbb{E}[X_i] = \frac{1}{2}$

$P(\omega = HHTHHT\ldots.) = 0$

deterministic sequences, not random

set of such is uncountable as $|2^{\mathbb{N}}|$

Note: $\Delta_t = \theta_t - \theta_t^*$ is error, and $\xrightarrow{a.s} 0$ $\sum a_n = \infty$, $\sum a_n^2 < \infty$ (say $\frac{1}{n}$)

SA:

stochastic approximation — SLLN — martingale theory — markov chains

martingale → generalisation of LLN

markov chains → $\frac{\sum f(X_n)}{n} \to \sum \pi_i f(i)$ a.s

martingale were real valued (not countable)

$$\mathbb{E}[X_{n+1} | \mathcal{F}_n] = X_n \quad a.s$$

$\{X_n\}$ converge almost surely (martingale convergence theorem)

$\mathbb{E}[X|Y]$ X → unknown (to observe), Y → observed

$x - g(Y) = \text{error}$

x ↑ target, $g(Y)$ ↑ Estimate

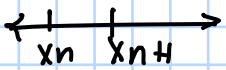
$\inf_{g(\cdot)} \mathbb{E}[(X - g(Y))^2]$ ← MSE, g^* exist and $g^* = \mathbb{E}[X|Y]$

Note: we needed Δ_t as it's dependent to Δ_{t-1}, so we cant apply LLN

sub martingale: $\mathbb{E}[X_{n+1} | \mathcal{F}_n] \leq X_n$ a.s

$\Rightarrow \mathbb{E}[X_{n+1}] \leq \mathbb{E}[X_n]$

ODE approach to SA :



α_n of $1/n$ nature

$$x_{n+1} = x_n + \alpha_n f(\ )$$
$$\Rightarrow \frac{x_{n+1} - x_n}{\alpha_n} = f(\ )$$

as $n \to \infty$, $\alpha_n \to 0$

↑ we will first study this ODE

so, $\frac{x_{n+1} - x_n}{\alpha_n}$ becomes a differential

the SA is borrowed from ODE convergence

Random process $\xrightarrow[\text{jump}]{}$ deterministic ODE

$\sum \frac{1}{n} \to \infty$ as we want to interpolate in $\mathbb{R}$ by ODE
we want to cover $\mathbb{R}$

convergence of random process related to fluctuations, var of a_i inflation is α_i^2

we need $\sum \alpha_i^2 < \infty$ for process to converge

11th Feb:

Q-learning:

this is via Stocastic approximation (a.s convergence), to complete this we need 2 things

the SA result relied on a paper by Tommi Takkola and a broder result relies on Aryeh Dvoretzky (on stocastic approximation)

theorm: (Dvoretzky) let $\alpha_n, \beta_n, \gamma_n \in [0,\infty)$ $\forall n \in \mathbb{N}$ s.t

① $\lim_{n\to\infty} \alpha_n = 0$ —— ①

② $\sum_{n\in\mathbb{N}} \beta_n < \infty$ —— ②

③ $\sum_{n\in\mathbb{N}} \gamma_n = \infty$ —— ③

let $\theta \in \mathbb{R}$, T_n be a measurable tranformation from $\mathbb{R}^n \to \mathbb{R}$

i.e $T_n^{-1}[a,\infty)$ is measurable in $\mathbb{R}^n$

s.t $|T_n(r_1, r_2, \ldots r_n) - \theta| \leq \max\{\alpha_n, (1+\beta_n)|r_n - \theta| - \gamma_n\}$ $\forall r_1, r_2, \ldots r_n \in \mathbb{R}$ —— ④

let X_n, Y_n be random variables s.t

$$X_{n+1}(\omega) = T_n[X_1(\omega), X_2(\omega), \ldots X_n(\omega)] + Y_n(\omega)$$ —— ⑤

$X = X(\omega)$ $\Omega = \{\omega\}$

then if the conditions: ① $\mathbb{E}[X_1^2] < \infty$ —— ⑥

② $\sum_{n\in\mathbb{N}} \mathbb{E}[Y_n^2] < \infty$ —— ⑦

③ $\mathbb{E}[Y_n | x_1, x_2 \ldots x_n] = 0$ —— ⑦ with prob 1 $\forall n$

then this implies: $\lim_{n\to\infty} \mathbb{E}[(X_n - \theta)^2] \to 0$ —— ⑧ and $\mathbb{P}[\lim_{n\to\infty} X_n = \theta] = 1$ —— ⑨

proof: wlog let $\theta = 0$ as otherise we first swift

let m be any $\mathbb{N}$, we will first show that for all $A \in \mathcal{F}$ which can be made in terms of $X_1, X_2, \ldots X_m$

then if $a > a_m$

$$\int_A [(|X_{m+1}| - a)^+]^2 d\mu \leq \int_A (\alpha\beta_m(1+\alpha\beta_m)[(|X_m| - a)^+]^2 + \gamma_m^2 + (1+\beta_m)^2(1+\alpha\beta_m)) d\mu$$

$r^+ = \frac{|r| + r}{2}$ → positive part of r

now this is true as:

$Z = Z(\omega)$ and $r \cdot \omega$ s.t $|Z| < a$ $\forall \omega$

then

$$\int_A [(X_{m+1}-a)^+]^2 d\mu \le \int_A [(X_{m+1}-Z)]^2 d\mu$$

← trivial after taking 2 cases

$$= \int_A [(T_m(X_m) - Z + Y_m)]^2 d\mu$$

t_1

$$= \int_A (T_m(X_m)-Z)^2 d\mu + \int_A Y_m^2 d\mu$$

$$+ 2 \int_A \int_{x_1 \dots x_n} (T_m(X_m)-Z)[Y_m | x_1 \dots x_m] d\mu$$

← Z goes out as of terms of A

$$= t_1 + 2 \int_A T_m(X_m) - Z \int_{x_1 \dots x_n} [Y_m | x_1 \dots x_m] d\mu$$

$\to 0$ ($\because$ ⑦)

$$= t_1$$

if Z defined in terms of $X_1, X_2, \dots X_n$

$$\text{so, } \int_A [(X_{m+1}-a)^+]^2 d\mu \le \int_A [T_m(X_m) - Z]^2 d\mu + \int_A Y_m^2 d\mu$$

$$\text{let } Z = \begin{cases} T_m(X_m) : |T_m(X_m)| \le a \\ a \quad ; T_m(X_m) > a \\ -a \quad ; T_m(X_m) < -a \end{cases}$$

$$\text{then } T_m(X_m) - Z = \begin{cases} 0 & ; |T_m(X_m)| \le a \\ T_m(X_m) - a & ; T_m(X_m) > a \\ T_m(X_m) + a & ; T_m(X_m) < -a \end{cases}$$

$$= \begin{cases} 0 & ; |T_m(X_m)| \le a \\ |T_m(X_m)| - a & ; |T_m(X_m)| > a \end{cases}$$

$$= (|T_m(X_m)| - a)^+$$

$$\text{so, } \int_A [T_m(X_m) - Z]^2 d\mu = \int_A [(|T_m(X_m)| - a)^+]^2 d\mu$$

now, a_1, a_m, then from ④:

case I: $|\gamma_m| \le a$ then $|T(\gamma_m)| \le \max\{\underbrace{\alpha_m}_{\le a}, \underbrace{(1+\beta_m)|\gamma_m|}_{\le (1+\beta_m)a} - \underbrace{\gamma_m}_{\le 0}\}$

$$|T(\gamma_m)| \le (1+\beta_m)a$$

$$\Rightarrow |T(\gamma_m)| - a \le (1+\beta_m)a - a = a\beta_m$$

case II: $|\gamma_m| > a$: then also from ④

$$|T_m(\gamma_m)| - a \le (1+\beta_m)|\gamma_m| - a = (1+\beta_m)(|\gamma_m| - a) + a\beta_m$$

so, from I, II : $0 \le (|T_m(\gamma_m)| - a)^+ \le (1+\beta_m)(|\gamma_m| - a)^+ + a\beta_m$

now $(u+v)^2 \le (1+v)u^2 + (1+v)v$ when $v \ge 0$

as $\frac{u^2+1}{2} \ge |u|$ ($\because AP \ge GP$)

$\Rightarrow u^2+1 \ge 2|u|$

$\Rightarrow (u^2+1)v \ge 2|uv|$

$\Rightarrow u^2+v^2+v(u^2+1) \ge (u+v)^2$

$\Rightarrow u^2(v+1) + v(v+1) \ge (u+v)^2$

so, $[(\tau_m(x_m)-a)^+]^2 \le [(1+\beta_m)(|x_m|-a)^+ + a\beta_m]^2$

$\le (1+\alpha\beta_m)(1+\beta_m)^2([|x_m|-a]^+)^2$

$+ \alpha\beta_m(1+\alpha\beta_m)$

$\Rightarrow \int_A [\tau_m(x_m) - z]^2 d\mu = \int_A [(\tau_m(x_m)-a)^+]^2 d\mu$

$\le \int_A (1+\alpha\beta_m)(1+\beta_m)^2([|x_m|-a]^+)^2 d\mu$

$+ \int_A \alpha\beta_m(1+\alpha\beta_m) d\mu$

$\Rightarrow \int_A [(|x_{m+1}|-a)^+]^2 d\mu \le \int_A (1+\alpha\beta_m)(1+\beta_m)^2([|x_m|-a]^+)^2$

$+ \alpha\beta_m(1+\alpha\beta_m) + \gamma_m^2 d\mu$

so our claim is done

now, $n > m$, $a \ge \max_{m \le j \le n} \alpha_j$, then what we did above gives us

$$\int_A [(|x_n|-a)^+]^2 d\mu \le u_{m,n} \int_A \{ v_{m,n} + \sum_{i=m}^{n-1} \gamma_i^2 + [(|x_m|-a)^+]^2 \} d\mu$$

$$u_{m,n} = \prod_{j=m}^{n-1} (1+\beta_j)^2(1+a\beta_j)$$

↑ this comes directly from above claim

$$v_{m,n} = a \sum_{j=1}^{n-1} \beta_j (1+a\beta_j)$$

then now, $\varepsilon > 0$ $a = a(\varepsilon)$ s.t $0 < a \le \sqrt{\frac{\varepsilon}{8}}$

choose $K = K(a,\varepsilon) > 1$ s.t $\max_{j \ge K-1} \alpha_j \le a$

($\because \lim_{j \to \infty} \alpha_j = 0$)

and $u_K \left(v_K + \sum_{j=K-1}^{\infty} \mathbb{E}[\gamma_j^2] \right) \le \frac{\varepsilon}{8}$ where

$$u_K = \prod_{j=K}^{\infty} (1+\beta_j)^2(1+a\beta_j) \qquad v_K = a \sum_{j=K}^{\infty} \beta_j(1+a\beta_j)$$

13th Feb:

theorm : (Dvoretzky) Let $\alpha_n, \beta_n, \gamma_n \in [0,\infty)$ $\forall n \in \mathbb{N}$ s.t

$$\textcircled{1} \lim_{n\to\infty} \alpha_n = 0 \quad \textcircled{1}$$

$$\textcircled{2} \sum_{n\in\mathbb{N}} \beta_n < \infty \quad \textcircled{2}$$

$$\textcircled{3} \sum_{n\in\mathbb{N}} \gamma_n = \infty \quad \textcircled{3}$$

Let $\Theta \in \mathbb{R}$, T_n be a measurable transformation from $\mathbb{R}^n \to \mathbb{R}$

i.e $T_n^{-1}[a,\infty)$ is measurable in $\mathbb{R}^n$

s.t $|T_n(x_1, x_2, \dots x_n) - \Theta| \le \max\{\alpha_n, (1+\beta_n)|x_n - \Theta| - \gamma_n\}$

$\forall x_1, x_2, \dots x_n \in \mathbb{R}$ $\quad \textcircled{4}$

Let X_n, Y_n be random variables s.t

$$X_{n+1}(\omega) = T_n[X_1(\omega), X_2(\omega), \dots X_n(\omega)] + Y_n(\omega) \quad \textcircled{5}$$

$X = X(\omega)$ $\Omega = \{\omega\}$

then if the conditions: $\textcircled{1}$ $\mathbb{E}[X_1^2] < \infty$ $\quad \textcircled{6}$

$\textcircled{2}$ $\sum_{n\in\mathbb{N}} \mathbb{E}[Y_n^2] < \infty$ $\quad \textcircled{7}$

$\textcircled{3}$ $\mathbb{E}[Y_n | x_1, x_2 \dots x_n] = 0$ $\quad \textcircled{8}$

with prob 1 $\forall n$

then this implies: $\lim_{n\to\infty} \mathbb{E}[(X_n - \theta)^2] \to 0$ $\quad \textcircled{8}$ and $\mathbb{P}[\lim_{n\to\infty} X_n = \theta] = 1$ $\quad \textcircled{9}$

proof:

$$\int_A [(|X_n| - a)^+]^2 \, d\mu \le u_{m,n} \int_A \{ v_{m,n} + \sum_{i=m}^{n-1} Y_i^2 + [(|X_m| - a)^+]^2 \} \, d\mu$$

$$u_{m,n} = \prod_{j=m}^{n-1} (1+\beta_j)^2 (1 + a\beta_j)$$

$$v_{m,n} = a \sum_{j=1}^{n-1} \beta_j^\circ (1 + a\beta_j^\circ)$$

then now, $\varepsilon > 0$ $a = a(\varepsilon)$ s.t $0 < a \le \sqrt{\frac{\varepsilon}{8}}$

choose $K = K(a, \varepsilon) > 1$ s.t $\max_{j \ge K-1} \alpha_j < a$

$(\because \lim_{j\to\infty} \alpha_j = 0)$

and $u_K \left(v_K + \sum_{j=K-1}^{\infty} \mathbb{E}[Y_j^2] \right) \le \frac{\varepsilon}{8}$ where

$$u_K = \prod_{j=K}^{\infty} (1+\beta_j^\circ)^2 (1 + a\beta_j^\circ) \qquad v_K = a \sum_{j=K}^{\infty} \beta_j^\circ (1 + a\beta_j^\circ)$$

ε exist as $u_K \to 0$, $v_K \to 0$ as $a \to 0$ $\sum_{j=K-1}^{\infty} \mathbb{E}[Y_j^2] < \infty$

now, $\forall j$, we let $s_j^\circ = s_j(\omega) = \operatorname{sgn}(T_j(x_j)) \to \{-1, 0, 1\}$

$j>1$, let $B_j' = \{\omega : \operatorname{sgn} X_j^{\circ} \neq s_{j-1}\}$ $\quad \operatorname{sgn} X_j = \operatorname{sgn}[T_{j-1}(X_{j-1}) + Y_{j-1}]$

$B_j'' = \{\omega : |T_{j-1}(X_{j-1})| \leq a\}$

let $B_j^{\circ} = B_j' \cup B_j''$ and $m > k$

$$A_m = B_m - \bigcup_{j=k}^{m-1} B_j$$

$$T_n = \bigcup_{m=k}^{n} A_m \qquad \Delta_n = \Omega - T_n$$

now, $\omega \in B_m'$ $X_m = T_{m-1}[X_1, X_2, \ldots X_{m-1}] + Y_{m-1}$

case I: $X_m > 0$:

$|X_m| \leq |Y_{m-1}|$

as $\operatorname{sgn}(X_m) \neq \operatorname{sgn}(T_{m-1}[X_1, \ldots X_{m-1}])$

so $X_m = \underset{+}{T_{m-1}}(X_{m-1}) + \underset{-}{Y_{m-1}}$

case II: $X_m \leq 0$:

$$\underset{-}{X_m} = \underset{+}{T_{m-1}}(X_{m-1}) + Y_{m-1}$$

$$-X_m \leq |Y_{m-1}|$$

$$\Rightarrow |X_m| \leq |Y_{m-1}|$$

so, from above, $|X_m| \leq |Y_{m-1}| \quad \forall \omega \in B_m'$

$\omega \in B_m''$, $|T_{m-1}(X_{m-1})| \leq a$

$$\Rightarrow |X_m| \leq a + |Y_{m-1}|$$

so, $\forall \omega \in A_m$ from $|X_m| \leq |Y_{m-1}|$
$|X_m| \leq a + |Y_{m-1}|$

we get

$$\int_{A_m} [(|X_n| - a)^+]^2 d\mu \leq u_{m,n} \int_{A_m} \left\{ v_{m,n} + \sum_{j=m}^{n-1} Y_i^2 + [(|X_m| - a)^+]^2 \right\} d\mu$$

$$u_{m,n} = \prod_{j=m}^{n-1} (1+B_j)^2 (1+aB_j) \qquad v_{m,n} = a \sum_{j=m}^{n-1} B_j (1+aB_j)$$

$\leq u_m$ over A_m $\leq v_m$

$$\Rightarrow \int_{A_m} [(|X_n| - a)^+]^2 d\mu \leq u_k \int_{A_m} v_k + \sum_{j=k-1}^{n-1} Y_j^2 \, d\mu$$

upper bound

now $n > m > k$:

$$\Rightarrow \int_{T_m} [(|X_n| - a)^+]^2 d\mu \leq u_k \int_{T_n} v_k + \sum_{j=k-1}^{\infty} Y_j^2 \, d\mu \leq \varepsilon/8$$

as all A_m disjoint, we make this integral !

$$|X_n| \leq (|X_n| - a)^+ + a$$

$$\Rightarrow \int_{T_m} X_n^2 d\mu \leq \varepsilon/8 + 2\int_{T_m} a[(|X_n| - a)^+] d\mu + \int_{T_m} a^2 d\mu$$

as, $2ab \le a^2+b^2$

$$2\int_{T_m} a[(|X_n|-a)^+]\,d\mu \le \varepsilon/8 + \int_{T_m} a^2\,d\mu$$

so, $$\int_{T_m} X_n^2\,d\mu \le \varepsilon/8+\varepsilon/8+2\underbrace{\int_{T_m} a^2\,d\mu}_{\le \varepsilon/8 \times 2}$$

$$\Rightarrow \int_{T_m} X_n\,d\mu \le \varepsilon/2$$

for $\forall n>k$, $\omega \in \Delta_n = \Omega - T_n$ outside B_j'

$$|X_j| = X_j \operatorname{sgn}(X_j) = X_j S_{j-1}$$
$$= S_{j-1}[T_{j-1}(X_{j-1})+Y_{j-1}]$$
$$= |T_{j-1}(X_{j-1})| + S_{j-1}Y_{j-1}$$

$\omega \notin B_j''$ $\quad |T_j(X_{j-1})| > a$

$$\Rightarrow |T_{j+1}(X_{j-1})| \le (1+\beta_{j-1})|X_{j-1}| - Y_{j-1} \text{ as } T(\cdot) \le \max(a,\cdot)$$

$\omega \notin B_j' \cup B_j''$

$$|X_j| \le (1+\beta_{j-1})|X_{j-1}| - Y_{j-1} + S_{j-1}Y_{j-1}$$

as $\Delta_n \subset B_j^c \quad \forall k \le j \le n$

$$\Rightarrow |X_n| \le \omega_{k,n}|X_k| - \sum_{m=k}^{n-1}\omega_{m,n}Y_m + \sum_{m=k}^{n} S_m \omega_{m,n} Y_m$$

$$\omega_{m,n} = \prod_{j=m}^{n}(1+\beta_j)$$ ← this follows from above inequality

well, $Z_n = \omega_{k,n}|X_k| - \sum_{m=k}^{n}\omega_{m,n}Y_m$

$$Z_n' = \sum_{m=k}^{n} S_m \omega_{m,n} Y_m$$

$$\Rightarrow \int_{\Delta_n} X_n^2\,d\mu \le \int_{\Delta_n}[(Z_n+Z_n')^+]^2\,d\mu$$

as $|X_n| \le Z_n + Z_n' \le [Z_n+Z_n']^+$

$$\Rightarrow \int_{\Delta_n} X_n^2\,d\mu \le \int_{\Delta_n}(Z_n^+ + |Z_n'|)^2\,d\mu$$

$$\le 2\int_{\Delta_n}(Z_n^+)^2\,d\mu + 2\int_{\Delta_n}|Z_n'|^2\,d\mu$$

as cross terms $\mathbb{E}[Y_m | x_1 \dots x_m] = 0$

now, $$\int_{\Delta_n}|Z_n'|^2\,d\mu \le \int_{\Omega}|Z_n'|^2\,d\mu = \sum_{m=k}^{n}\omega_{m,n}^2\,\mathbb{E}[Y_m^2]$$
$$\le \prod_{i=k}^{\infty}(1+\beta_j)^2 \sum_{m=k}^{\infty}\mathbb{E}[Y_m^2] \le \frac{\varepsilon}{8}$$

recall if Z is r.v with $\mathbb{E}[Z^2] < \infty$

$$\mathbb{E}[(Z-b)^+]^2 \xrightarrow{\text{then}} 0 \text{ as } b \to 0$$

$m=1, n=k$:

X_k $Z = |X_k| \prod_{j=k}^{\infty}(1+\beta_j)$ → finite second moment

$$Z_n^+ \leq Z - \sum_{m=k}^{n} \gamma_m$$

$$\Rightarrow \int_{\Delta_n} (Z_n^+)^2 d\mu \leq \mathbb{E}[(Z_n^+)^2] < \varepsilon/8 \text{ as } \to 0 \text{ as } \sum_{m=k}^{n} \gamma_m \to \infty$$

$\therefore$ $\int_{\Delta_n} (X_n)^2 d\mu \leq (\varepsilon/8 + \varepsilon/8) = \varepsilon/2$

$$\Rightarrow \int_{T_m} (X_n)^2 d\mu + \int_{\Delta_n} (X_n)^2 d\mu \leq \varepsilon/2 + \varepsilon/2 = \varepsilon$$

$$\Rightarrow \int_{\Omega} (X_n)^2 d\mu < \varepsilon$$

$$\Rightarrow \mathbb{E}[(X_n)^2] \to 0 \text{ as } n \to \infty$$

18th Feb:

theorem: (Dvoretzky) Let $\alpha_n, \beta_n, \gamma_n \in [0,\infty)$ $\forall n \in \mathbb{N}$ s.t

$$\textcircled{1} \lim_{n\to\infty} \alpha_n = 0 \quad \textcircled{1}$$

$$\textcircled{2} \sum_{n\in\mathbb{N}} \beta_n < \infty \quad \textcircled{2}$$

$$\textcircled{3} \sum_{n\in\mathbb{N}} \gamma_n = \infty \quad \textcircled{3}$$

Let $\Theta \in \mathbb{R}$, T_n be a measurable transformation from $\mathbb{R}^n \to \mathbb{R}$

i.e $T_n^{-1}[a,\infty)$ is measurable in $\mathbb{R}^n$

s.t $|T_n(\gamma_1, \gamma_2, \ldots \gamma_n) - \Theta| \le \max\{\alpha_n, (1+\beta_n)|\gamma_n - \Theta| - \gamma_n\}$

$\forall\, \gamma_1, \gamma_2, \ldots \gamma_n \in \mathbb{R}$ $\quad \textcircled{4}$

Let X_n, Y_n be random variables s.t

$$X_{n+1}(\omega) = T_n[X_1(\omega), X_2(\omega), \ldots X_n(\omega)] + Y_n(\omega) \quad \textcircled{5}$$

$X = X(\omega)$ $\Omega = \{\omega\}$

then if the conditions: $\textcircled{1}$ $\mathbb{E}[X_1^2] < \infty$ $\quad \textcircled{6}$

$\textcircled{2}$ $\sum_{n\in\mathbb{N}} \mathbb{E}[Y_n^2] < \infty$ $\quad \textcircled{7}$

$\textcircled{3}$ $\mathbb{E}[Y_n | x_1, x_2 \ldots x_n] = 0$ $\quad \textcircled{8}$ with proof $\forall n$

then this implies: $\lim_{n\to\infty} \mathbb{E}[(X_n - \Theta)^2] = 0$ $\quad \textcircled{8}$ and $\mathbb{P}[\lim_{n\to\infty} X_n = \Theta] = 1$ $\quad \textcircled{9}$

proof: we have already seen that

if $a > \max_{m \le j \le n} a_j$ then for $A \in \mathcal{F}_{n-1}$

$$\int_A [(|X_n| - a]^+)^2 d\mu \le \mu_{m,n} \int_A \left\{ \nu_{m,n} + \sum_{j=1}^{n-1} Y_j^2 + [(|X_m| - a)^+]^2 \right\} d\mu$$

now, $\mathbb{E}[X_n^2]$ for $A = \Omega$ $\forall n > m$:

$$\mathbb{E}[X_n^2] < H\left(\max_{j \ge m} a_j, \sum_{j=m}^{\infty} \beta_j, \mathbb{E}[X_m^2] + \sum_{j=m}^{\infty} \mathbb{E}[Y_j^2]\right)$$

as $\mathbb{E}[X_n^2] \to 0$

H is an increasing function of those 3 variables, and $H \downarrow 0$ as those variables $\downarrow 0$

now given any $\varepsilon, \delta > 0$, we have $\exists\, \eta(\varepsilon, \delta) > 0$ s.t

if all those arguments $< \eta$

then

$$H < \delta^2 \varepsilon$$

$$\Rightarrow H\left(\max_{j \ge m} a_j, \sum_{j=m}^{\infty} \beta_j, \mathbb{E}[X_m^2] + \sum_{j=m}^{\infty} \mathbb{E}[Y_j^2]\right) < \delta\varepsilon^2 \quad \text{for } m = m(\varepsilon, \delta)$$

Let T_n', X_n', Y_n' for $m = m(\varepsilon, \delta)$ $\quad j < m$

$$T_j' = T_j \quad X_j' = X_j \quad Y_j' = Y_j$$

for $j > m$:

$$T_j'(r_1, r_2, \ldots r_j) = \begin{cases} T_j(r_1, r_2, \ldots r_j) & ; |r_j| < \delta \\ r_j & ; |r_j| \geq \delta \end{cases}$$

$$Y_j' = \begin{cases} Y_j & ; |X_j| < \delta \\ 0 & ; |X_j| \geq \delta \end{cases}$$

$$X_{j+1}' = T_j'(X_j') + Y_j' \quad \text{i.e } Y_j', X_{j+1}' \text{ are defined recursively}$$

with $r_n = 0$ we have

$$|T_n'(r_1, r_2, \ldots r_n)| \leq \max(a_n, (1+\beta_n)|r_n|)$$

for $n < m$ trivial and
for $n \geq m$

as $|r_n| < \max(a_n, (1+\beta_n)|r_n|)$
we get T_n' satisfies the conditions

$\mathbb{E}[Y_n' | X_1' \ldots X_n'] = 0$ trivial for $n \geq m$
for $n < m$ again trivial

$\mathbb{E}[Y_n'^2] \leq \mathbb{E}[Y_n^2] \Rightarrow \sum_{n=1}^{\infty} \mathbb{E}[Y_n'^2] < \infty$

$\Rightarrow \mathbb{E}[X_n'^2] < \delta^2 \varepsilon$ as $X_m' = X_m$
$(\because \text{above } H < \delta^2 \varepsilon)$ ← this comes from before

so, $\delta^2 P[|X_n'| > \delta] \leq \mathbb{E}[X_n'^2]$ ($\because$ markov inequality)

$\Rightarrow \delta^2 P[|X_n'| > \delta] < \delta^2 \varepsilon$

$\Rightarrow P[|X_n'| > \delta] < \varepsilon$

for $j > m$: if $|X_j| \geq \delta$ some $j \geq m$ then

$T_j'(r_1 \ldots r_j) = r_j$ as $|r_j| > \delta$

$Y_j' = 0 \quad X_{j+1} = T_n(X_1 \ldots X_n) + Y_n$
$= r_j$

$\Rightarrow |X_{j+1}| \geq \delta$

and so, $|X_{j+1}'| = |T_j'(X_j') + Y_j'|$
$\geq \delta$ ($Y_j' \to 0$)

$\Rightarrow |X_j'| > \delta \quad \forall j > m$

$\Rightarrow P[\max_{m < j \leq n} |X_j'| \geq \delta] \leq P[|X_n'| > \delta] < \varepsilon$

$\Rightarrow P[\lim_{n \to \infty} X_n = 0] = 1$ as $\varepsilon \to 0$

monte-carlo is a simulation based methods

Q-learning – tabular method (many deep methods exist)
– it is off policy

SARSA:

initialised 0 value, it is an on-policy algorithm

$\ldots S_t, a_t, r_t, S_{t+1}, a_{t+1} \rightarrow$ on policy

S A R S A

we saw the state so on-policy

we have to do some action here

this $a_{t+1} = \pi(S_{t+1}, Q_t)$ → depends on (S_{t+1}, Q_t)

and update Q table

$$Q_{t+1}(S_t, a_t) = Q_t(S_t, a_t) + \alpha_t(S_t, a_t)\{r_t + \gamma Q_t(S_{t+1}, a_{t+1}) - Q_t(S_t, a_t)\}$$

$\pi(\,)$: greedy is one step ahead → NO exploration

$\pi(\,)$: uniform prob on all actions → exploration

Eg: ε-greedy randomly with $\frac{\varepsilon}{(|\mathcal{A}|)}$

else max with $1-\varepsilon$

$\varepsilon_n = \frac{1}{|\mathcal{A}|}$

softmax $\frac{e^{\beta_t Q(s,a)}}{\sum_{a'} e^{\beta_t Q(s,a')}}$ ← time dependent = prob of moving a / $\pi(s,a)$

$\{\beta_t\} \rightarrow$ {schedules}

$\beta_t = 0 \ \forall t \Rightarrow$ uniform

$\beta_t \uparrow \infty \Rightarrow$ max/greedy algorithm

eventually becoming greedy
but at start we are doing exploring / exploiting

so, actions which are good are used infinitely often: 'reinforced'

theorem: (SARSA converges) under ε-greedy policies, SARSA converges almost surely

proof.

$$(T^{*\varepsilon}_{q_s})(s,a) := r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[\varepsilon/|\mathcal{A}| \sum_{b} q_s(s',b') + (1-\varepsilon)\max_{a'} q_s(s',a')\right]$$

then from below exercise $T^{*\varepsilon}_{q}$ converge almost surely, say $Q^{*,\varepsilon}$

then now, if $Q^{*,\varepsilon} \rightarrow Q^*$ we will be done

$T^* Q^* = Q^*$ (Bellman equation)

and $Q^{*,\varepsilon}(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\varepsilon/|\mathcal{A}| \sum_{b'} q_s(s',b') + (1-\varepsilon) \max_{a'} q_s(s',a') \right]$

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

then $\| Q^* - Q^{*,\varepsilon} \|_{\max} \to 0$ from below exercise

$\Delta^\varepsilon = \| Q^{*,\varepsilon} - Q \|_\infty$

$\vdots$

$\| Q^{*,\varepsilon} - Q^* \|_{\max} \to 0$

Exo: show that $T^{*\varepsilon}_q$ is a contraction operator → done down

Exo: show that $\| Q^* - Q^{*,\varepsilon} \|_{\max} \to 0$ → done down

Exo: show that $T^{*\varepsilon}_q$ is a contraction operator

Ans:

$$(T^{*\varepsilon}_{q_s})(s,a) := r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\varepsilon/|\mathcal{A}| \sum_b q_s(s',b') + (1-\varepsilon) \max_{a'} q_s(s',a') \right]$$

$$(T^\pi_q)(s,a) = r(s,a) + \gamma \left[\sum_{y \in S} \left(P(y|s,a) \times \sum_{b \in A} \pi(b|y) q(y,b) \right) \right]$$

($\pi(b|y)$: probability of taking this action)

$$\| T^\pi_{q_1} - T^\pi_{q_2} \|_\infty = \max_{s,a} \left| \gamma \sum_y \left(P(y|s,a) \times \sum_{b \in A} \pi(b|y) (q_1(y,b) - q_2(y,b)) \right) \right|$$

$$\leq \gamma \max_{s,a} \sum_y \left(P(y|s,a) \times \sum_{b \in A} \pi(b|y) \right) \| q_1 - q_2 \|_\infty$$

$$\leq \gamma \| q_1 - q_2 \|_\infty$$

so T^π_q is a contraction map

Exo: show that $\| Q^* - Q^{*,\varepsilon} \|_{\max} \to 0$

Ans: as T^π_q is a contraction map

$\exists\, Q^\pi$ s.t. $T^\pi Q^\pi = Q^\pi$

$$Q_{t+1}(x_t, a_t) = Q_t(x_t, a_t) + \alpha_t (r_t + \gamma Q_b(x_{t+1}, a_{t+1}) - Q_t(x_t, a_t))$$

$$= (1 - \alpha_t) Q_t + \alpha_t (r_t + \gamma Q_b(x_{t+1}, a_{t+1}))$$

let $\Delta_t = Q_t - Q^\pi$ (Q^π: fixed point of T^π contraction)

then $Q_{t+1} - Q^\pi = \Delta_{t+1} = (1 - \alpha_t) Q_t + \alpha_t (r_t + \gamma Q_t(x_{t+1}, a_{t+1})) - Q^\pi$

$$= (1-\alpha_t)\theta_t + \alpha_t(T^\pi\theta_t) - \theta^\pi - \alpha_t\theta^\pi + \alpha_t\theta^\pi$$

$$= (1-\alpha_t)(\theta_t - \theta^\pi) + \alpha_t(T^\pi\theta_t - T^\pi\theta^\pi)$$

$$= (1-\alpha_t)\Delta_t + \alpha_t F_t$$

$$F_t = T^\pi\theta_t - T^\pi\theta^\pi$$

now the theorem $\{\Delta_t\}$ in $\mathbb{R}^n$

$\Delta_{t+1} = (1-\alpha_t)\Delta_t + \alpha_t F_t$

converg to zero w.p.1 under following assumptions:

given to us $\rightarrow 0 \le \alpha_t \le 1$ $\sum_t \alpha_t(x) = \infty$ $\sum_t \alpha_t^2 < \infty$

$$\mathbb{E}[F_t|\mathcal{F}_t] = (T^\pi\theta_t) - \theta^\pi$$

$$= T^\pi\theta_t - T^\pi\theta^\pi$$

$$\|\mathbb{E}[F_t|\mathcal{F}_t]\|_\infty \le \gamma\|\Delta_t\|_\infty \quad \gamma<1$$

as F_t is bounded ($\because \theta_0 - \theta^\pi$ bounded and contraction)

$$\text{var}(F_t|\mathcal{F}_t) = \mathbb{E}\left[\left(r(x,a,X(x,a)) + \gamma\left[\theta_t(X(s,a),\pi(X(s,a))\right] - T^\pi\theta^\pi - (T^\pi\theta_t + T^\pi\theta^\pi)\right)^2\right]$$

r is bounded

$$\le C \le C(1+\|\Delta_t\|_\infty^2)$$

so, $\Delta_t \to 0$ wp1

$\Rightarrow \theta_t \to \theta^\pi$ wp 1

20th Feb:

we first showed contraction, then used SA converges to show it

Note: there is a contraction and then AS converges

implicitly SARSA has same flavour, even though it is an on-policy algorithm unlike Q-learning

these two are examples of monte-carlo method, other MC algos include ① first visit
② every visit

in same episode only one entry for first visit
- lot of data gets wasted

Exe: why do we need contraction and almost sure convergence → done down

there are other algos like TD(0), TD(λ), TD(1)
← extremes → extremes

n-step estimator

in Q-learning we had only one entry updation

but in functional approximation methods this is not true
we update whole vector

Exe: correct Q-learning proof by SA-way to finish it → done down

to port Q-learning with SA-algorithm, we will make 2 tables for variables
and do auto on each variable

Function approximation:

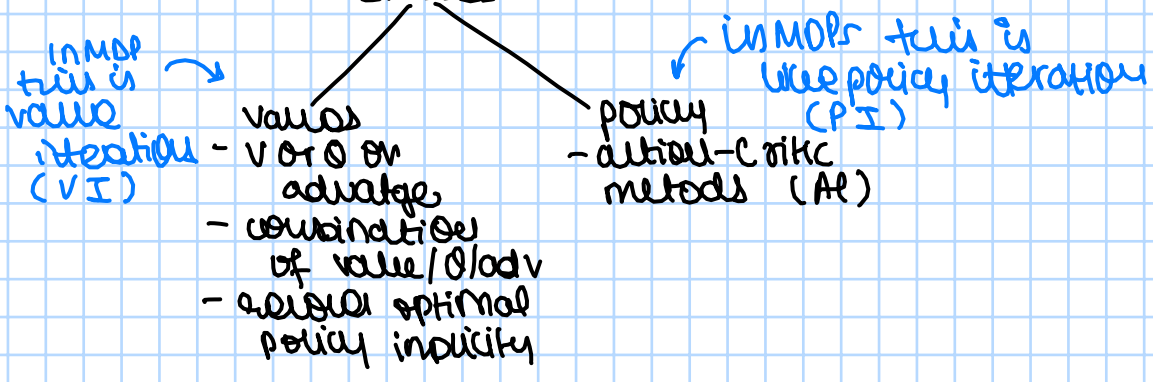
Linear: LFA- linear function approximation
Non-linear: DRL (Deep-RL)
- DRL is data hungry
- we don't know if it converges or not
- there is a counter example on 2 state neural network

Note: All negative examples are very sparse

Eg: Autonomous vehicle is not trained on accident data

Single agent:

Similar to MDPs, in value space we can operate



discount factor, Avg cost, Total cost, critic alone, Action only, Action-critic

value $V(\cdot)$ $\theta \in \mathbb{R}^q$ → parameter

$V_\theta(x) = \langle \theta, \phi(x) \rangle$ ← $\phi(x)^T\theta$ $\phi(x) \in \mathbb{R}^n$, $\theta \in \mathbb{R}^n$

x ↑ state, ϕ → quantization

$Q_\theta(\cdot,\cdot) = \langle \theta, \phi(x,a) \rangle$

π policy is parameterized by $\gamma \in \mathbb{R}^q$

$\pi_\theta(x, a)$

critic ↔ policy evaluation
action ↔ policy update

Temporal difference:

$x_k, a_k, r_{k+1}, x_{k+1}$

$\delta_k = r_k + \gamma V_\theta(x_{k+1}) - V_\theta(x_k)$ ← discounted model

$\theta_{k+1} = \theta_k + \alpha_{c,k} \nabla_\theta V_{\theta_k}(x_k)$ → gradient update ← $\min \|V_\theta - V^*\|^2$ (TD error)

for LFA's $\theta_{k+1} = \theta_k + \alpha_{c,k} \delta_k \phi(x_k)$ TD(0)

$\theta_{k+1} = \theta_k + \alpha(G_t - V_\theta(S_t))\phi(S_t)$ $G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k}$ TD(1)

Exe: why do we need contraction and almost sure convergence
Ans: contraction gives us a unique fixed point

then we use other SA result to show
$\Delta_t \to 0$
i.e. $Q_t \to Q^\pi$
↑ fixed point of given map

Exe: correct Q-learning proof by SA-ODE to finish it

proof:

$\Delta_{t+1} = (1-\alpha_t)\Delta_t + \alpha_t F_t$

$0 \le \alpha_t \le 1$ $\sum \alpha_t = \infty$ $\sum \alpha_t^2 < \infty$

$\| \mathbb{E}[F_t | \mathcal{F}_t] \|_\infty \le \gamma \| \Delta_t \|_\infty$ $\gamma < 1$

$\mathrm{var}(F_t | \mathcal{F}_t) \le C(1 + \|\Delta_t\|_\infty^2)$ $C > 0$

then $\Delta_{t+1} = \Delta_t + \alpha_t [\mathbb{E}[F_t | \mathcal{F}_t] - \Delta_t) + \alpha_t (F_t - \mathbb{E}[F_t | \mathcal{F}_t])$

let $T_t(\Delta_t) = \Delta_t + \alpha(\mathbb{E}[F_t | \mathcal{F}_t] - \Delta_t)$

$Y_t = \alpha_t (F_t - \mathbb{E}[F_t | \mathcal{F}_t])$

$$\Delta_{t+1} = T_t(\Delta_t) + Y_t$$

now, $\mathbb{E}[Y_t | \mathcal{F}_t] = \mathbb{E}[\alpha_t (\mathbb{E}[F_t | \mathcal{F}_t] - \mathbb{E}[F_t | \mathcal{F}_t])]$

$= 0$

$$\mathbb{E}[Y_t^2] = \alpha_t^2 \operatorname{Var}(F_t | \mathcal{F}_t)$$

$$\leq \alpha_t^2 C \left[1 + \|\Delta_t\|_\infty^2\right]$$

$$\Rightarrow \sum_{t=1}^{\infty} \mathbb{E}[Y_t^2] \leq \left(\sum_{t=1}^{\infty} \alpha_t^2\right) C \left(1 + \|\Delta_t\|_\infty^2\right)$$

$< \infty$ as $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$

now $\|T_t(\Delta_t)\|_\infty = \|(1-\alpha)\Delta_t + \alpha \mathbb{E}[F_t | \mathcal{F}_t]\|$

$\leq (1-\alpha)\|\Delta_t\|_\infty + \alpha\varphi\|\Delta_t\|_\infty$

$\leq \|\Delta_t\|_\infty - \alpha(1-\varphi)\|\Delta_t\|_\infty$

let $\varphi_n = \alpha(1-\varphi)$

then $\sum_{n=1}^{\infty} \varphi_n = \infty$ as $\sum \alpha = \infty$

so, then this means $\mathbb{P}[\lim_{n \to \infty} \Delta_t = 0] = 1$

4th March:

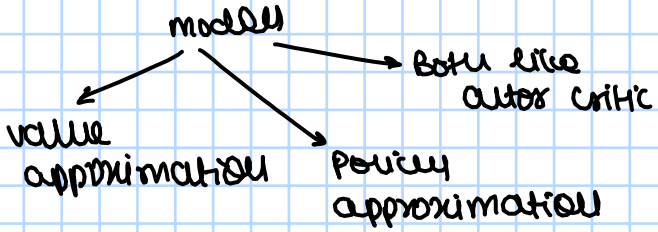
Fution approximation:

actor-critic methods → policy gradient

Q-learning, SARSA } are known to converge (a.s)

Note: we could convert the convergence of Q-learning\SARSA on fullional approximation (under some assumption)

Note: Q-learning with FA is critic only (as it is off-policy) and SARSA is also critic-only as only state-action pair



Note: AC is in lines of policy iteration in MDP scheme

we will follow the survey of actor-critic in RL

$J(\pi) = \mathbb{E}[g(r_1, r_2, \dots) | \pi]$ g = discounted / avg sum

$\pi : X \times U \to [0, \infty)$

discounted reward: $J(\pi) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{k+1} \mid d_0, \pi\right]$

$x_0 \in X$ initial state $x_0 \sim d_0(\cdot)$ distribution

one period reward

$$= \int_X d_\gamma^\pi(x) \int_U \pi(x,u) \int_X f(x,u,x') \rho(x,u,x') \, dx' du dx$$

$\sum_{k=0}^{\infty} \gamma^k p(x_k = x | d_0, \pi)$

(discounted state distribution under policy π)

$f : X \times U \times X \to [0, \infty)$ is the state transition prob density (kernel)

$$p(x_{k+1} \in X_{k+1} | x_k, u_k) = \int_{X_{k+1}} f(x_k, u_k, x') dx'$$

$$V^\pi(x) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{k+1} \middle| x_0 = x, \pi\right]$$

$$Q^\pi(x,u) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{k+1} | x_0 = x, u_0 = u, \pi\right]$$

$V^\pi(x) = \mathbb{E}[\rho(x,u,x') + \gamma V^\pi(x')]$

$Q^\pi(x,u) = \mathbb{E}[\rho(x,u,x') + \gamma Q^\pi(x',u')]$

$V^*(x) = \max_u \mathbb{E}[\rho(x,u,x') + \gamma V^*(x')]$

$Q^*(x,u) = \mathbb{E}[\rho(x,u,x') + \gamma \max_{u'} Q^*(x',u')]$

Actor-only:

Let $\nu \in \mathbb{R}^p$ (parameter) $\leftrightarrow$ policy π

then, let π_ν be the policy

$$\nabla_\nu J = \frac{\partial J}{\partial \pi_\nu} \times \frac{\partial \pi_\nu}{\partial \nu} \quad (\because \text{chain rule})$$

$$\nu_{k+1} = \nu_k + \underbrace{\alpha_{a,k}}_{>0 \text{ small enough}} \nabla_\nu J_k$$

learning rate of actor at time k

a = actor
c = critic

$\sum_{k=0}^{\infty} \alpha_{a,k} = \infty \qquad \sum_{k=0}^{\infty} \alpha_{a,k}^2 < \infty$ imposes small enough α

$\Rightarrow$ convergence

Critic-only:

$$\pi(x) = \arg\max_u Q(x,u)$$

Now, we will do AC methods from policy gradients methods for RL with FA
(SP Singh, R Sutton)

Actor-Critic:

$|S| \times |A|$ = no of policies needed in policy gradient

policy θ

$$\Delta\theta = \alpha \frac{\partial \rho}{\partial \theta}$$

θ = vector of policy parameters
ρ = performance (avg sum / discounted sum)

$t = \{0, 1, \ldots\}$

$s_t \in S$, $a_t \in A$, $r_t \in \mathbb{R}$

$$P_{ss'}^a = P_r[s_{t+1} = s' \mid s_t = s, a_t = a]$$

$$R_s^a = \mathbb{E}[r_{t+1} \mid s_t = s, a_t = a]$$

$$\pi(s, a, \theta) = P(a_t = a \mid s_t = s, \theta)$$

$\theta \in \mathbb{R}^l$, $l << |S|$, π is differentiable wrt parameters

$$\rho(\pi) = \lim_{n \to \infty} \frac{1}{n} \mathbb{E}[r_1 + r_2 + \ldots + r_n \mid \pi]$$

$$= \sum_s d^\pi(s) \sum_a \pi(s,a) R_s^a$$

$\hookrightarrow d^\pi(s) := \lim_{t \to \infty} P(s_t = s \mid s_0, \pi)$

$$\theta^\pi(s,a) = \sum_{t=1}^{\infty} \mathbb{E}[r_t - \rho(\pi) \mid s_0 = s, a_0 = a, \pi]$$ value of state action pair

Note: $Q^\pi(s,a)$ defined as bias from MDP setting, so for future we can piggyback on MDP terms

bias of a policy $h(s)$

gain of a policy $g(s_0) = \sum_s d^\pi(s) \sum_a \pi(s,a) R_s^\pi$

↑ will be independent of s_0

$\Rightarrow g = \sum_s d^\pi(s) \sum_a \pi(s,a) R_s^\pi = \rho(\pi)$

$\Rightarrow g = \rho(\pi)$

all states are reachable (assumption under policy π)

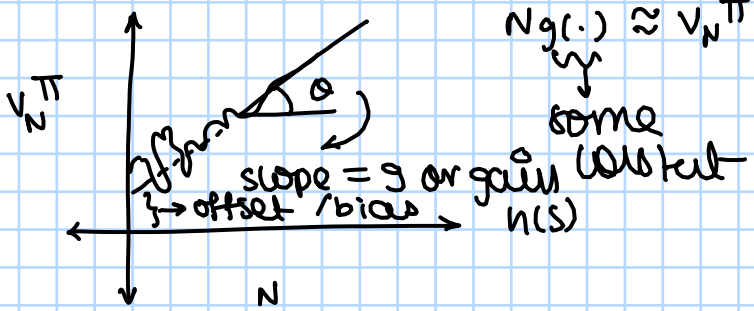
true convergence

$$\rho_s(\pi) = g^\pi(s) = \lim_{n\to\infty} \frac{1}{n} \mathbb{E}\left[\sum_{i=1}^{n} r_i \mid x_0 = s, \pi\right]$$

$$= \lim_{n\to\infty} \frac{1}{n} V_n^\pi(s)$$

$N g(\cdot) \approx V_N^\pi$ for large N

$\approx$ some constant



Note: h is unique upto additive constant, so bias is also called relative values

↳ fix one component to get

variables = states + 1 extra → stable

(∵ fixing one component of h)

6th March:

Start-state model ↔ discounted model

average reward and some related background

$$Q^{\pi}(s,a) := \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid s_0, \pi\right]$$

for avg: $V_{N+1}^{\pi}(s) = \mathbb{E}_s^{\pi}\left[\sum_{t=1}^{N} r_t(X_t, Y_t)\right]$

↳ total reward upto N periods

gain ↙

$$g^{\pi}(s) = \lim_{N\to\infty} \frac{1}{N} V_{N+1}^{\pi}(s)$$

d → decision rule, policy $\pi = (d_1, d_2, \dots)$ for stationary policy

$d^{\infty} \in SR$

$P_d(t \times M)$

$P_d^* = \text{cesaro sum} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} P^{i-1}$

then gain of policy $g^{\pi}(s) = \lim_{N\to\infty} \frac{1}{N} V_{N+1}(s) = P_d^* r_d(s)$

average optimal gain $g^*(\cdot) \geq g^{\pi}(\cdot) \ \forall \ \pi \in SR$

and r (bounded)

so, $g(s) = P^* r$

$g^{\pi}(j) = g^{\pi}(k)$ for j, k in same irreducible chain

MDP is irreducible / unichain then $g^{\pi}(\cdot)$ is constant vector

gain and bias:

$|S| < \infty \quad h = H_P r \quad H_P = (I - P + P^*)^{-1}(I - P^*)$

$P P^* = P^*$

for aperiodic model

$$h = \sum_{t=0}^{\infty} (P^t - P^*) r_t$$

$$= \sum_{t=0}^{\infty} P^t (r_t - g)$$

$$h(s) = \mathbb{E}_s\left[\sum_{t=0}^{\infty} (r(X_t) - g(X_t))\right]$$

↑ total deviation of g to true rewards from stationary rand

$$h = \sum_{t=0}^{\infty} P^t (r - g)$$

$g(\cdot)$ is true average of rewards (per period)

Aperiodic MC:

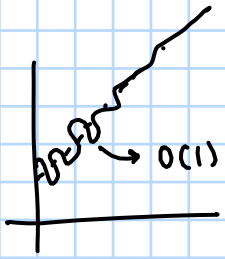
Bias is small, mostly in first few periods (transient quality)

$$V_{N+1} = \sum_{1}^{N} p_{t-1}\gamma \Rightarrow h = \sum_{1}^{N} p_{t+1}\gamma - Ng + \underbrace{\sum_{N+1}^{\infty}(p_{t+1}-p^*)\gamma}_{\to 0 \text{ can be shown as } h \to \text{converges}}$$

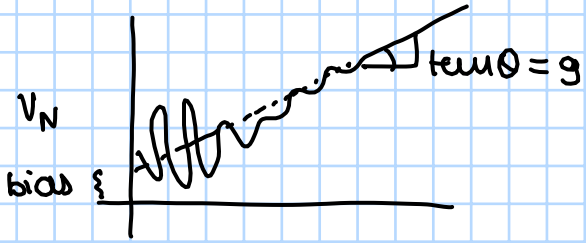
$$\lim_{N\to\infty} \sum_{N+1}^{\infty}(p_{t+1}-p^*)\gamma = 0$$

i.e $V_{N+1} = Ng + h + o(1)$ → with some oscillation

V_{N+1}: linearly growing



for aperiodic model



$$V(s) - V(\tilde{s}) = h(s) - h(\tilde{s})$$

Relative value / difference of starting in two distinct states

(g, h) → Role in optimality equations

Now, back to policy-gradient theorem (PGT)

$$\rho(\pi) = \lim_{n\to\infty} \frac{1}{n} \mathbb{E}[r_1 + r_2 + \cdots r_n \mid \pi] = \sum_s d^{\pi}(s) \sum_a \pi(s,a) R_s^{\pi}$$

$$\text{or } \lim_{n\to\infty} \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid s_0, \pi\right]$$

$$\theta_{\pi}(s,a) = \sum_{t=1}^{\infty} \mathbb{E}[r_t - \rho(\pi) \mid s_0 = s, a_0 = a, \pi] = h(\cdot) \quad \text{(bias)}$$

$$\text{or } \sum_{t=1}^{\infty} \mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k} \mid s_t = s, a_t = a, \pi\right]$$

$$d^{\pi}(s) = \lim_{n\to\infty} \text{Prob}(s_n = s, \pi)$$

Theorm: (Policy gradient theorm) For any MDP, in either average or state-stable formulation

$$\frac{\partial \rho}{\partial \theta} = \sum_s d^{\pi}(s) \sum_a \frac{\partial \pi(s,a)}{\partial \theta} \theta^{\pi}(s,a) \qquad \theta := \text{policy parameters}$$

now, in RL we don't know this, so we estimate

Note: Above we parametrised policy alone

$\pi(\cdot,\cdot)$ and $Q(\cdot,\cdot)$ was same state space, so we used comparable features

$f_w: S \times A \to \mathbb{R}$ for Q^π (in addition to π parametrisation)
(approximation to Q^π with parameter w)

Objective is to learn f_w

$$\Delta w_t \propto \frac{\partial}{\partial w}\left[\hat{Q}^\pi(s_t,a_t) - f_w(s_t,a_t)\right]^2$$

($\hat{Q}^\pi$: unbiased estimator for $Q^\pi(s_t,a_t)$; one can tune it to used all rewards)

differentiate this; mean square error

minimising the mean square error

$$\Rightarrow \Delta w_t \propto \left[\hat{Q}^\pi(s_t,a_t) - f_w(s_t,a_t)\right] \frac{\partial f_w(s_t,a_t)}{\partial w}$$

Suppose the above converges (to a local optimum)

$$\Rightarrow \sum_s d^\pi(s) \sum_a \pi(s,a)\left[Q^\pi(s,a) - f_w(s,a)\right] \frac{\partial f_w(s,a)}{\partial w} = 0$$

Theorem: (Policy gradient with function approximation) If f_w satisfy

$$\sum_s d^\pi(s) \sum_a \pi(s,a)\left[Q^\pi(s,a) - f_w(s,a)\right] \frac{\partial f_w(s,a)}{\partial w} = 0$$

and it is compatible with policy parametrisation in sense that

$$\frac{\partial f_w(s,a)}{\partial w} = \frac{\partial \pi(s,a)}{\partial \theta} \frac{1}{\pi(s,a)}$$

then, $\frac{\partial \rho}{\partial \theta} = \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s,a)}{\partial \theta} f_w(s,a)$

proof:

as $\sum_s d^\pi(s) \sum_a \pi(s,a)\left[Q^\pi(s,a) - f_w(s,a)\right] \frac{\partial f_w(s,a)}{\partial w} = 0$

$\pi(s,a) \to \frac{1}{\frac{\partial f_w}{\partial w}(s,a)} \frac{\partial \pi(s,a)}{\partial \theta}$

$$\Rightarrow \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s,a)}{\partial \theta}\left[Q^\pi(s,a) - f_w(s,a)\right] = 0 \quad \text{—①}$$

this approximate error is $\perp$ to policy parametrisations

substituting ① to $\frac{\partial \rho}{\partial \theta}$ in policy theorem

$$\frac{\partial \rho}{\partial \theta} - 0 = \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s,a)}{\partial \theta} f_w(s,a)$$

now, value parametrization attain with compatibility condition

say $\pi(s,a) = \frac{e^{\langle \theta, \phi_{s,a} \rangle}}{\sum_{b \in s} e^{\langle \theta, \phi_{s,b} \rangle}}$ ← softmax

features given $\phi_{s,a} \in \mathbb{R}^d \quad \forall s,a \in S \times A$

$\theta \in \mathbb{R}^d$ (a parameter)

$\theta \in \mathbb{R}^d$

$$\frac{\partial f_w(s,a)}{\partial w} = \frac{\partial \pi(s,a)}{\partial \theta} \frac{1}{\pi(s,a)} = \phi_{s,a} - \sum_b \pi(s,b) \phi_{s,b}$$

$\Rightarrow f_w$ is linear with w

$\Rightarrow f_w = \langle w, \phi_{s,a} - \sum_{b \in s} \pi(s,b) \phi_{s,b} \rangle$

parameter $w \in \mathbb{R}^d$

$\phi_{s,a}$ are features which are given to us, we need to calculate

$$\pi(s,a) = \frac{e^{\langle \theta, \phi_{s,a} \rangle}}{\sum_b e^{\langle \theta, \phi_{s,b} \rangle}}$$

$$f_w(s,a) = \langle w, \phi_{s,a} - \sum_b \pi(s,b) \phi(s,b) \rangle$$

we also know $\frac{d\rho}{d\theta} = g(\theta, w)$ ← we can tune these parameters

and $\frac{df_w}{dw} = \frac{\partial \pi}{\partial \theta} \times \frac{1}{\pi}$

11th March:

Theorem: (policy gradient) for any MDP, in either average-reward or start-state

$$\frac{\partial \rho}{\partial \theta} = \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a)$$

we also saw that under functional approximation

$$\frac{\partial f_w(s,a)}{\partial w} = \frac{\partial \pi(s,a)}{\partial \theta} \frac{1}{\pi(s,a)}$$

then $\frac{\partial \rho}{\partial \theta} = \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s,a)}{\partial \theta} f_w(s,a)$

we, also saw $\pi(s,a) = \frac{e^{\theta^T \phi_{s,a}}}{\sum_b e^{\theta^T \phi_{s,b}}} \Rightarrow f_w(s,a) = w^T [\phi_{s,a} - \sum_b \pi(s,b) \phi_{s,b}]$

Theorem: (policy iteration with fution Approximation) let π and f_w be any differiable fution approximation for policy and value fution that satisfy compatibility fution

and $\max_{\theta, s, a, i, j} \left| \frac{\partial^2 \pi(s,a)}{\partial \theta_i \partial \theta_j} \right| < B < \infty$ some $B > 0$

$\{\alpha_k\}_{k=0}^\infty$ be any step-size sequence s.t $\lim_{k \to \infty} \alpha_k = 0$ and $\sum_k \alpha_k = \infty$

then, for any MDP with bounded rewards, $\{\rho(\pi_k)\}_{k=0}^\infty$ sequence defined by any θ_0, $\pi_k = \pi(\cdot, \cdot, \theta_k)$

some $\theta_0 \in \mathbb{R}^d$

and, $w_k = w$ s.t $\sum_s d^{\pi_k}(s) \sum_a \pi_k(s,a) [Q^{\pi_k}(s,a) - f_w(s,a)] \frac{\partial f_w(s,a)}{\partial w} = 0$

$$\theta_{k+1} = \theta_k + \alpha_k \sum_s d^{\pi_k}(s) \sum_a \frac{\partial \pi_k(s,a)}{\partial \theta} f_{w_k}(s,a)$$

converges s.t $\lim_{k \to \infty} \frac{\partial \rho(\pi_k)}{\partial \theta} = 0$

proof: form policy gradient theorm θ_k update in the direction of gradient (it is in correct direction)

bcoz on $\frac{\partial^2 \pi(s,a)}{\partial \theta_i \partial \theta_j}$ and MDP rewards together assure that $\frac{\partial^2 \rho}{\partial \theta_i \partial \theta_j}$ bounded, so by earlier anya Borkar argument, $\{\theta_k\}$ iterates converge almost surely by below exercise

Exe: put above algorithm in anya-diwedi framework → done down

Theorm: for any MDP, in either average-reward or start-state

$$\frac{\partial \rho}{\partial \theta} = \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a)$$

proof:

$$\frac{\partial V^\pi(s)}{\partial \theta} \overset{\text{def}}{=} \frac{\partial}{\partial \theta} \sum_a \pi(s,a) Q^\pi(s,a) \quad \forall s \in S$$

$$= \sum_a \left[\frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \pi(s,a) \frac{\partial}{\partial \theta} Q^\pi(s,a) \right]$$

for average-reward:

$$Q^\pi(s,a) = R_s^a - \rho(\pi) + \sum_{s'} P_{ss'}^a V^\pi(s')$$

$$Q^\pi(s,a) = \mathbb{E}_s \left[\sum_0^\infty r_t - \rho(\pi) \right]$$

$$= \mathbb{E}_s \left[r_1 - \rho(\pi) + \sum_1^\infty (r_{t+1} - \rho(\pi)) \right]$$

$$= R_s^a - \rho(\pi) + \sum_{s'} P_{ss'}^a \left(\sum_b \pi(s',b) Q(s',b) \right)$$

$$= R_s^a - \rho(\pi) + \sum_{s'} P_{ss'}^a V^\pi(s')$$

so, $$\frac{\partial V^\pi(s)}{\partial \theta} = \sum_a \left[\frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \pi(s,a) \frac{\partial}{\partial \theta} \left(R_s^a - \rho(\pi) + \sum_{s'} P_{ss'}^a V^\pi(s') \right) \right]$$

$$= \sum_a \left[\frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \pi(s,a) \left(-\frac{\partial \rho}{\partial \theta} + \sum_{s'} P_{ss'}^a \frac{\partial V^\pi(s')}{\partial \theta} \right) \right]$$

$$\sum_a \pi(s,a) \frac{\partial \rho}{\partial \theta} = \sum_a \left[\frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \pi(s,a) \sum_{s'} P_{ss'}^a \frac{\partial V^\pi(s')}{\partial \theta} \right] - \frac{\partial V^\pi(s)}{\partial \theta}$$

$$\Rightarrow \frac{\partial \rho}{\partial \theta} = \sum_a \left[\frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \pi(s,a) \sum_{s'} P_{ss'}^a \frac{\partial V^\pi(s')}{\partial \theta} \right] - \frac{\partial V^\pi(s)}{\partial \theta}$$

summing both sides over stationary distribution d^π

$$\sum_s d^\pi(s) \frac{\partial \rho}{\partial \theta} = \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a)$$

$$+ \sum_s d^\pi(s) \sum_a \pi(s,a) \sum_{s'} P_{ss'}^a \frac{\partial V^\pi(s')}{\partial \theta}$$

$$- \sum_s d^\pi(s) \frac{\partial V^\pi(s)}{\partial \theta} \rightarrow \sum_{s'} d^\pi(s') \frac{\partial V^\pi(s')}{\partial \theta}$$

(as d^π is stationary distribution, no need to calculate 2 step)

$$\Rightarrow \frac{\partial \rho}{\partial \theta} = \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a)$$

start-state:

$$\frac{\partial V^\pi(s)}{\partial \theta} \overset{\text{def}}{=} \frac{\partial}{\partial \theta} \sum_a \pi(s,a) Q^\pi(s,a) \quad \forall s \in S$$

$$= \sum_a \left[\frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \pi(s,a) \frac{\partial}{\partial \theta} Q^\pi(s,a) \right]$$

$$= \sum_a \left[\frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \pi(s,a) \frac{\partial}{\partial \theta} \left[R_s^a + \sum_{s'} \gamma P_{ss'}^a V^\pi(s') \right] \right]$$

$$= \sum_a \left[\frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \pi(s,a) \sum_{s'} \gamma P_{ss'}^a \frac{\partial}{\partial \theta} V^\pi(s') \right]$$

$$= \sum_s \sum_{k=0}^{\infty} \gamma^k \Pr(s \to s, k, \pi) \sum_a \frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a)$$

$$\Rightarrow \frac{\partial J}{\partial \theta} = \frac{\partial}{\partial \theta} \mathbb{E}\left[\sum_1^\infty \gamma^{t-1} r_t \,\middle|\, s_0, \pi\right] = \frac{\partial}{\partial \theta} V^\pi(s_0)$$

$$= \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a)$$

Note: our algorithm above is actor-critic algorithm / method

actor updates the policy

critic gives critism on policy

fixed π, first update w parameter

$$w_k = w \text{ s.t } \sum_s d^{\pi_k}(s) \sum_a \pi_k(s,a)\left[Q^{\pi_k}(s,a) - f_w(s,a)\right] \frac{\partial f_w(s,a)}{\partial w} = 0$$

$$\theta_{k+1} = \theta_k + \alpha_k \sum_s d^{\pi_k}(s) \sum_a \frac{\partial \pi_k(s,a)}{\partial \theta} f_{w_k}(s,a)$$

w should converge first
(best w approximation
in Q mean square approximation)

Note: we need a do loop for w_k approximation to w

we need a faster $w_k \to w$ $\{\beta_k\}$ $\sum \beta_k = \infty$, $\sum \beta_k^2 < \infty$

critic is giving feedback

and this $w_k \to w$ goes into θ_k to see the actor, it acts on θ_k to get θ_{k+1}

if we don't need a do loop to get $w_k \to w$, we can have

$\frac{\alpha_k}{\beta_k} \to 0$ i.e β_k is faster, then in one loop we can
allow for fixed α_k, $w_k \to w$
then
$\theta_k \leftarrow$ update

bigger stepsizes of β_k, as $k \to \infty$ will make w_k converge faster
than θ_k

Exe: put above algorithm in actor-critic framework

Ans:

$$\theta_{k+1} = \theta_k + \alpha_k \underbrace{\sum_s d^{\pi_k}(s) \sum_a \frac{\partial \pi_k(s,a)}{\partial \theta} f_{w_k}(s,a)}_{g(\theta_k, w_k)}$$

then as

$$w_k: \sum_s d^{\pi_k}(s) \sum_a \pi_k(s,a)\left(Q^{\pi_k}(s,a) - f_w(s,a)\right) \frac{\partial f_w}{\partial w} = 0$$

let $h(\theta) = \nabla \rho(\theta)$

then

$g(\theta_k, w_k) = h(\theta_k) + \varepsilon_k$ → approximation noise due to imperfect f_w

$$\theta_{k+1} = \theta_k + \alpha(h(\theta_k) + \varepsilon_k)$$

$$X_{n+1} = T_n(X_n) + Y_n$$

$$T_k(\theta) = \theta + \alpha_k h(\theta)$$

$$Y_k = \alpha_k \varepsilon_k$$

$$\theta_{k+1} = T_k(\theta_k) + Y_k$$

as $\alpha_k \to 0$, $\sum_k \alpha_k = \infty$, $\mathbb{E}[Y_k \mid \theta_1, \ldots, \theta_k] = 0$

as $\mathbb{E}[\varepsilon_k \mid \theta_k] = 0$

↓

unbiased gradient estimate

as $\sum Y_k = \sum \alpha_k \varepsilon_k \Rightarrow \sum \mathbb{E}[Y_k^2] = \sum O(\alpha_k^2) < \infty$

as, $\max_{\theta, s, a, i, j} \left| \frac{\partial^2 \pi(s,a)}{\partial \theta_i \partial \theta_j} \right| \leq B < \infty$ some $B > 0$

if $|T_n(x) - T_n(y)| \leq (1 + \beta_n)|x - y|$

then $|T_n(x) - \theta^*| \leq \alpha_n(1 + \beta_n)|x - \theta^*|$

now, $\frac{\partial T_k}{\partial \theta_j} = I + \alpha_k \frac{\partial}{\partial \theta_j} \left(\sum_s d^{\pi_k}(s) \sum_a \frac{\partial \pi(s,a)}{\partial \theta} f_{w_k}(s,a) \right)$

as $\max_{\theta, s, a, i, j} \left| \frac{\partial^2 \pi(s,a)}{\partial \theta_i \partial \theta_j} \right| \leq B$

$\Rightarrow \left| \frac{\partial}{\partial \theta_j} \left(\frac{\partial \pi(s,a)}{\partial \theta_i} \right) \right| \leq B$

$\Rightarrow \left| \frac{\partial T_k}{\partial \theta} \right| \leq 1 + \alpha_k C \quad C \in \mathbb{R}$

then, $|T_k(\theta) - T_k(\theta')| \leq (1 + \alpha_k C)|\theta - \theta'|$

$\Rightarrow |T_k(\theta) - \theta^*| \leq (1 + \alpha_k C)|\theta - \theta^*|$

as $\theta_{k+1} = \theta_k + \alpha_k \nabla \rho(\theta_k)$

as $\theta_{k+1} \to \theta$

$\Rightarrow \theta = \theta + \lim_{n \to \infty} \alpha_k \nabla \rho(\theta)$

$\Rightarrow \nabla \rho(\theta) = h(\theta) = 0$

13th March :

we saw a DO LOOP algorithm to first find w_k and then θ_k

we can convert this to a single loop algorithm by taking two stochastic approximates s.t

$w_k \to w$ faster than $\theta_k \to \theta$

we can do this by $\{\beta_k\}_{k=1}^{\infty}$, $\{\alpha_k\}_{k=1}^{\infty}$ $\beta_k > \alpha_k$

Algorithm :

update θ → policy parameter → policy $\pi_\theta(s,a)$

update w ↑ $\theta(\cdot)$ fn approximation $\langle w, f_w(\cdot) \rangle$

↕

compatable features

$f_w : \nabla \log \pi_\theta$

$\nabla_w f_w = \nabla_\theta \pi \frac{1}{\pi} \Leftrightarrow$ features are compatible

so, $\nabla_w f_w = \nabla_\theta \pi$ is sufficient condition for compatable features

Average Reward :

$$J(\pi) = \lim_{n\to\infty} \frac{1}{n} \mathbb{E}\left[\sum_{k=0}^{n-1} r_{k+1} \,\middle|\, \pi\right]$$

$$= \int_X d^\pi(x) \int_U \pi(x,u) \int_X f(x,u,x')\, p(x,u,x')\, dx'\, du\, dx$$

$$V^\pi(x) = \mathbb{E}\left[\sum_{k=0}^{\infty} (r_{k+1} - J(\pi)) \,\middle|\, x_0 = x, \pi\right]$$

$$Q^\pi(x,u) = \mathbb{E}\left[\sum_{k=0}^{\infty} (r_{k+1} - J(\pi)) \,\middle|\, x_0 = x, u_0 = u, \pi\right]$$

Exe: show that $V^\pi(x) + J(\pi) = \mathbb{E}[p(x,u,x') + V^\pi(x')]$

$Q^\pi(x,u) + J(\pi) = \mathbb{E}[p(x,u,x') + Q^\pi(x',u')]$

$V^\pi(x) = \sum_a \pi(x,u) Q^\pi(x,a)$ → done down

Bellman equations :

$V^\pi(x) + J(\pi) = \mathbb{E}[p(x,u,x') + V^\pi(x')]$

$Q^\pi(x,u) + J(\pi) = \mathbb{E}[p(x,u,x') + Q^\pi(x',u')]$

Optimality equations :

$V^*(x) + J^* = \max_u \mathbb{E}[p(x,u,x') + V^*(x')]$

$Q^*(x,u) + J^* = \mathbb{E}[p(x,u,x') + \max_{u'} Q^*(x',u')]$

J^* := optimal reward for π^* optimal policy

now for $\theta \in \mathbb{R}^q$ $V_\theta(x)$ or $Q_\theta(x,u)$

$$V_\theta(x) = \langle \theta, \phi(x) \rangle$$

$$Q_\theta(x,u) = \langle \theta, \phi(x,u) \rangle$$

Critic: evaluate actor's policy

Actor: update accordingly

policy π is parametrized by $\vartheta \in \mathbb{R}^p$

$$\pi_\vartheta(x,u)$$

$\delta_k = r_{k+1} + \gamma V_{\theta_k}(x_{k+1}) - V_{\theta_k}(x_k)$ $(x_k, u_k, r_{k+1}, x_{k+1})$ transition

TD error / temporal difference

given δ_k error we want to update critic

$$\theta_{k+1} = \theta_k + \alpha_{c,k}\, \delta_k \nabla_\theta V_{\theta_k}(x_k)$$

↑ this δ_k is surprising but present as δ_k = TD error

also $\nabla_\theta V_{\theta_k}(x_k) = \phi(x_k)$

(∵ linear function approximation)

$$\theta_{k+1} = \theta_k + \alpha_{c,k}\, \delta_k \phi(x_k)$$

↓

TD(0) method

Eligibility trace TD(λ):

$z_k \in \mathbb{R}^q$ $z_k = \lambda\gamma z_{k-1} + \nabla_\theta V_{\theta_k}(x_k)$

eligibility trace vector

$\lambda \in (0,1)$

$\lambda\gamma \in (0,1)$

new $\theta_{k+1} = \theta_k + \alpha_{c,k}\delta_k z_k$

Discounted setup:

$$\delta_k = r_{k+1} + \gamma V_{\theta_k}(x_{k+1}) - V_{\theta_k}(x_k)$$

$$z_k = \lambda\gamma z_{k-1} + \nabla_\theta V_{\theta_k}(x_k)$$

$$\theta_{k+1} = \theta_k + \alpha_{c,k}\delta_k z_k$$

$$\vartheta_{k+1} = \vartheta_k + \alpha_{a,k} \nabla_\vartheta J_k$$

Average criterion:

Here TD error $\delta_k = r_{k+1} - \hat{J}_k + V_{\theta_k}(x_{k+1}) - V_{\theta_k}(x_k)$

$$\hat{J}_k = \hat{J}_{k-1} + \alpha_{J,k}\,(r_{k+1} - \hat{J}_{k-1})$$

$\in (0,1)$

$$z_k = \lambda z_{k-1} + \nabla_\theta V_{\theta_k}(x_k)$$

$$\theta_{k+1} = \theta_k + \alpha_{c,k}\,\delta_k\,z_k$$

$$\gamma_{k+1} = \gamma_k + \alpha_{a,k}\,\nabla_\gamma J_k$$

Convergence:

$$\sum_k \alpha_{a,k} = \infty \qquad \sum_k \alpha_{c,k} = \infty \qquad \sum_k \alpha_{a,k}^2 < \infty \qquad \sum_k \alpha_{c,k}^2 < \infty$$

$$\sum_k \left(\frac{\alpha_{a,k}}{\alpha_{c,k}}\right)^d < \infty \quad \text{for some } d > 0$$

$$\alpha_{J,k} = \alpha_{c,k}$$

Exe: show that $V^\pi(x) + J(\pi) = \mathbb{E}[\rho(x,u,x') + V^\pi(x')]$

$$Q^\pi(x,u) + J(\pi) = \mathbb{E}[\rho(x,u,x') + Q^\pi(x',u')]$$

$$V^\pi(x) = \sum_a \pi(x,u)\, Q^\pi(x,a)$$

Ans: as $V^\pi(x) = \mathbb{E}\left[\sum_{k=0}^{\infty}(r_{k+1} - J(\pi)) \,\middle|\, x_0 = x, \pi\right]$

$\rho : X \times U \times X \to \mathbb{R}$ is reward fuction

$$V^\pi(x) = \mathbb{E}\left[r_1 - J(\pi) + \sum_{k=1}^{\infty}(r_{k+1} - J(\pi)) \,\middle|\, x_0 = x, \pi\right]$$

$$= 0 - J(\pi) + \mathbb{E}[r_1 + V^\pi(x')]$$

(x, u, x') — x': next

$$V^\pi(x) + J(\pi) = \mathbb{E}[\rho(x,u,x') + V^\pi(x')]$$

same for $Q^\pi(x,u) + J(\pi) = \mathbb{E}[\rho(x,u,x') + Q^\pi(x',u')]$

$$\text{now, } V^\pi(x) = \sum_a \pi(x,a)\,\mathbb{E}\left[\sum_{k=0}^{\infty}(r_{k+1} - J(\pi) \,\middle|\, x_0 = x, u_0 = a, \pi\right]$$

$$= \sum_a \pi(x,a)\, Q^\pi(x,a)$$

18th March :

proximal policy optimization dada (PPO) and direct policy optimization (DPO)

decentralised MARL :

① Average cost (Replicated model)
② Total cost model ↔ stochastic shortest path problem

Multi-agent system MAS :

- swarm of drones/robots
- vehicles on network/ graph
- game like board game and other field games

$\underset{\sim}{S_t}$ → (a vector) state, $((s_1^1, s_1^2, \ldots s_1^n))$
one for each agent

$\underset{\sim}{a_t}$ → (a vector) action, $((a_1^1, a_1^2, \ldots a_1^n))$

$r^i(s_t, a_t)$
i^{th} agent reward depends on other actions as well

$s_t \in (s_t^1, s_t^2, \ldots s_t^n)$

$a_t \in \mathcal{A} \times \mathcal{A} \times \cdots \times \mathcal{A}$

Reward criterion :

→ average
→ discounted
→ total
:
} for each agent

if agents are antagonistic ← just a variant of MDP with multidimensional markov chain

MDP setup, multiple agent → antagonist (not cooperative) is 'Multiagent MDP'

these are also called stochastic games : Lloyd Shapley 1970

if $n=2$ then policy pair equilibrium exist or not : John Nash

$r^1(f,g)$ $f = \pi^1 = (d_1^1, d_2^1, \ldots d_n^1, \ldots) \in$ Markov randomized class
$r^2(f,g)$ $g = \pi^2 = (d_1^2, d_2^2, \ldots d_n^2, \ldots)$

(f^*, g^*), (f_a^*, g_a^*), (f_d^*, g_d^*)
(f_a^*, g_a^*) ↓ for average
(f_d^*, g_d^*) → for discounted

if $|\mathcal{S}| < \infty$, $|\mathcal{A}| < \infty$, then above $\exists (f^*, g^*)$ (equilibrium exist)

say average reward criterion (for each player), say $I \leq P_5$ (all costs)

and now consider each has further constraints to meet, like from common budget

↳ service loads/latency → constrained stochastic games

Mild conditions:

under constrained stochastic game, $\exists (f^*, g^*)$ it in MR

Here we need randomised as every agent can't cater agent

constrained stochastic games:

we can know an optimisation problem (OP) which has a

global minima (x^*)

$OP(x) \geq 0 \Rightarrow OP(x^*) \geq 0$

we need objective function and feasible space to be non-convex

Note: $\{(f_a^*, g_a^*) \in MR \times MR\} \Leftrightarrow \{\text{global minima } OP(x^*)_\varepsilon = 0\}$

they have a one-one mapping

to get above result, we will have (p^i, d^i) → primal-dual pair $\forall i = 1, 2, \ldots n$
and by rearraging them we get above

23rd March:

Multi agent RL:

we have seen stochastic games, non-coop games now we found the nash equilibrium

now, decentralised learning:

agents act locally / individually

private actions

$r_i(\cdot,\cdot)$ are private information } privacy of agent i preserved

$f = \sum_{i \in N} f_i \rightsquigarrow \{f_t^i\}$ state

now, $\pi^i : \mathcal{S} \times \mathcal{A}^i \to [0,1]$

joint policy: $\pi : \mathcal{S} \times \mathcal{A} \to [0,1]$

policy $\pi(s,a) = \prod_{i \in N} \pi^i(s,a^i)$

$\pi^i_{\theta^i}$, $\theta^i \in \Theta^i \subseteq \mathbb{R}^{M_i}$ is a compact set

$\pi_\theta \to$ vector of vectors

Assume

$\pi^i(s,a^i) \to$ differentiable

'Exploration' > 0

$P_\theta(s'|s) = \sum_{a \in A} \pi_\theta(s,a) P(s'|s,a)$

↑ transition probability matrix of underlying

Note: DTMC (Discrete time markov chain is aperiodic and ergodic

now, objective r^i_{t+1}

average criterion:

$$\max_\theta J(\theta) = \max_\theta \lim_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T-1} \frac{1}{N} \sum_{i \in N} r^i_{t+1}\right]$$

$$= \sum_s \sum_a d_\theta(s) \pi_\theta(s,a) \bar{R}(s,a)$$

$\bar{R}(s,a) = \frac{1}{N} \sum_{i \in N} R^i(s,a)$ } → globally averaged reward

let $\bar{r}_t = \frac{1}{N} \sum_{i \in N} r^i_t$

then $\bar{R}(s,a) = \mathbb{E}[\bar{r}_{t+1} \mid s_t = s, a_t = a]$

$$\Theta_\theta(s,a) = \sum_{t=0}^{\infty} \mathbb{E}[\bar{r}_{t+1} - J(\theta) \mid s_0 = s, a_0 = a, \pi_\theta]$$

$$V_\theta(s) = \sum_{a \in A} \pi_\theta(s,a) Q_\theta(s,a)$$

$$A_\theta(s,a) = Q_\theta(s,a) - V_\theta(s)$$

AC methods for networked MDPs:

Note: policy gradient theorem for MARL holds

local advantage function

$$A^i(s,a) = Q_\theta(s,a) - \underbrace{\tilde{V}^i_\theta(s,a^{-i})}_{\sum_{a^{-i}} \pi^{-i}_{\theta^{-i}}(s,a^{-i}) Q_\theta(s,a^i,a^{-i})}$$

Note: consensus algos need $V^i_t \to V^i$ and $V^i = V \ \forall i \in N$

parameterised rewards/cost:

$\forall$ agent communication network graph N $\{c_{ij}(t)\}$ $i \leftrightarrow j$

gossip algorithms: social network analysis - information of every agent needs to be diffused among all agents

Fully decentralised MARL with network agents:

ICML Kaiqing Zhang

network multi-agent MDP

$$(\mathcal{S}, \{\mathcal{A}^i\}_{i \in N}, P, \{R^i\}_{i \in N})$$

$$\{s_t\}_{t \geq 0} \leftrightarrow \{a_t\}_{t \geq 0}$$

$$R^i : \mathcal{S} \times \mathcal{A}^i \to \mathbb{R} \quad \forall i \in N$$

$$P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$$

$$\nabla_{\theta^i} J(\theta) = \mathbb{E}_{s \sim d_\theta, a \sim \pi_\theta}[\nabla_{\theta^i} \log \pi^i_{\theta^i}(s,a^i) A_\theta(s,a)]$$

$$= \mathbb{E}_{s \sim d_\theta, a \sim \pi_\theta}[\nabla_{\theta^i} \log \pi^i_{\theta^i}(s,a^i) \times \underbrace{A^i_\theta(s,a)}_{\text{local } ?}]$$

empirical average

parametrise private rewards

$$\bar{r}(s,a) \quad \bar{r}(s,a;w) \ w \in \mathbb{R}^K, \ K \ll |\mathcal{S}||\mathcal{A}|$$

w to use via MLE

$$\min_w \mathbb{E}_{s,a}[\bar{r}(s,a) - \bar{r}(s,a;w)] \quad \text{—— ①}$$

this ① is equivalently characterized by

$$\min_{w} \sum_{i} \mathbb{E}_{s,a} \left[\left(r^{\theta}(s,a) - \bar{r}(s,a;w) \right)^2 \right]$$

25th March :

Decentralized Models :

We gossip models / networks consensus algorithm is given
we saw that policy gradient theorem holds here

Algorithm:

$Q(s,a;w): \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ $\quad w^i \leftrightarrow i^{th}$ agent component

↑ vector of vectors

$Q(s,a;w^i) \leftrightarrow$ local estimate of Q_θ ← via PLT for MARL

↑ Let each agent i have local parameter w^i with regards to reach consensus

(relative value for average reward MDP)

in critic step, the update resembles the action-value TD-learning for policy evaluation followed by linear combination of its neighbours parameter

↳ parameter sharing (consensus update) involves

$C_t = [C_t(i,j)]_{N \times N}$ matrix

$C_t(i,j)$ = weight on message transmitted from i to j at time t

Critic: $\mu^i_{t+1} = (1-\beta_{w,t})\,\mu^i_t + \beta_{w,t} r^i_{t+1}$ → μ^i_t tracks long-term returns of agent i, $\beta_{w,t} > 0$

$$\tilde{w}^i_t = w^i_t + \beta_{w,t} \cdot \delta^i_t \cdot \nabla_w Q_\theta(w^i_t)$$

$$w^i_{t+1} = \sum_{j \in N} C_t(i,j)\, \tilde{w}^j_t$$

$$\delta^i_t = r^i_{t+1} - \mu^i_t + Q_{t+1}(w^i_t) - Q_t(w^i_t)$$ ← TD-error

↑ Action-valued TD-error update

Actor: $\theta^i_{t+1} = \theta^i_t + \beta_{\theta,t} A^i_t \psi^i_t$

$$A^i_t = Q_t(w^i_t) - \sum_{a' \in A^i} \pi_{\theta^i_t}(s_t, a^i) \cdot Q(s_t, a^i, a^i; w^i)$$

$$\psi^i_t = \nabla_{\theta^i} \log \pi_{\theta^i_t}(s_t, a^i_t)$$

↑ this is how we go around compatible features

also the estimator we used is unbiased estimator

$$\mathbb{E}\left[\underbrace{r_{t+1} - J(\theta) + V_\theta(s_{t+1}) - V_\theta(s_t)}_{\delta_t} \,\middle|\, s_t = s, a_t = a, \pi_\theta\right] = A_\theta(s,a) \quad \forall s,a$$

ex: show that above is an unbiased estimator, i.e → done down

$$\mathbb{E}[\bar{\delta}_t \mid s_t=s, a_t=a, \pi_\theta] = A_\theta(s,a) \quad \forall s,a$$

Algorithm-2: a v based algorithm

$$V_t(v) = V(s_t; v) \quad v \in \mathbb{R}^L \quad L \ll |\mathcal{S}|$$

$$\tilde{\mu}_t^i = (1-\beta_{v,t})\,\mu_t^i + \beta_{v,t}\cdot r_{t+1}^i$$

$$\mu_{t+1}^i = \sum_{j\in N} c_t(i,j)\,\tilde{\mu}_t^j$$

this depend on (s,a) but to be unbiased estimator we need $\bar{V}_{t+1}$

$$\delta_t^i = r_{t+1}^i - \mu_t^i + V_{t+1}(v_t^i) - V_t(v_t^i) \quad \rightarrow \text{TD-error}$$

$$\tilde{v}_t^i = v_t^i + \beta_{v,t}\,\delta_t^i\,\nabla_v V_t(v_t^i)$$

$$v_{t+1}^i = \sum_{j\in N} c_t(i,j)\,\tilde{v}_t^i$$

so, we will estimate global avg reward $\bar{R}$ in critic step to get δ_t^i to be unbiased estimator

from last class we know that

$$\bar{R}(s,a;\lambda): \mathcal{S}\times\mathcal{A} \to \mathbb{R}, \quad \lambda\in\mathbb{R}^M, \; M \le |\mathcal{S}|\times|\mathcal{A}|$$

we minimize weighted mean-square error

$$\min_\lambda \sum_{\substack{s\in\mathcal{S}\\ a\in\mathcal{A}}} d_\beta(s)\,\pi_\theta(s,a)\,[\bar{R}(s,a;\lambda) - \bar{R}(s,a)]^2$$

so, $\tilde{\lambda}_t^i = \lambda_t^i + \beta_{v,t}\cdot[r_{t+1}^i - \bar{R}_t(\lambda_t^i)]\cdot\nabla_\lambda \bar{R}_t(\lambda_t^i)$

$$\lambda_{t+1}^i = \sum_{j\in N} c_t(i,j)\cdot\tilde{\lambda}_t^i$$

so, above is our critic step

actor: $\theta_{t+1}^i = \theta_t^i + \beta_{\theta,t}\,\tilde{\delta}_t^i\cdot\psi_t^i$

$$\tilde{\delta}_t^i = \bar{R}_t(\lambda_t^i) - \mu_t^i + V_{t+1}(v_t^i) - V_t(v_t^i)$$

↑ this is what now becomes an unbiased estimator

$$\psi_t^i = \nabla_\theta \log \pi_{\theta_t^i}(s_t^i, a_t^i)$$

where $\sum\beta_{w,t} = \sum\beta_{v,t} = \sum\beta_{\theta,t} = \infty$

$\sum\beta_{w,t}^2, \sum\beta_{v,t}^2, \sum\beta_{\theta,t}^2 < \infty$, $\beta_{\theta,t} = o(\beta_{w,t}) = o(\beta_{v,t})$

and $\frac{\beta_{w,t+1}}{\beta_{w,t}} \to 1$ & $\frac{\beta_{v,t+1}}{\beta_{v,t}} \to 1$

Convergence:

In algo 2, with 2-time scale setup

find θ_* and so π_{θ_*}

$$\left.\begin{aligned} \lim_{t\to\infty} \mu_t^i &= J(\theta_*) \\ \lim_{t\to\infty} \lambda_t^i &= \lambda_{\theta_*} \\ \lim_{t\to\infty} v_t^i &= v_{\theta_*} \end{aligned}\right\} \text{a.s}$$

and $\exists$ unique solution to

$F^T D_{\theta_*}^{sig}[\bar{R} - F\lambda_{\theta_*}] = 0$ ← state, action-features matrix

$\bar{\Phi}^T D_{\theta_*}^{s}[T_{\theta_*}^{v}(\bar{\Phi} v_{\theta_*}) - \bar{\Phi} v_{\theta_*}] = 0$ ← state-features matrix

Ex: show that above θ_* is an unbiased estimator, i.e

$$\mathbb{E}[\bar{\delta}_t \mid s_t = s, a_t = a, \pi_{\theta_*}] = A_{\theta_*}(s,a) \quad \forall s,a$$

Ans:

we want to show

$$\mathbb{E}[r_{t+1} - J(\theta_*) + V_{\theta_*}(s_{t+1}) - V_{\theta_*}(s_t) \mid s_t = s, a_t = a, \pi_{\theta_*}] = A_{\theta_*}(s,a)$$

now, $\mathbb{E}[r_{t+1} \mid s,a] = R(s,a)$

$$\mathbb{E}[V_{\theta_*}(s_{t+1}) \mid s,a] = \sum_{s'} P(s' \mid s,a) V_{\theta_*}(s')$$

now, $R(s,a) - J(\theta_*) + \sum_{s'} P(s' \mid s,a) V_{\theta_*}(s') - V_{\theta_*}(s) = \mathbb{E}[\bar{\delta}_t \mid s,a,\pi_{\theta_*}]$

also by bellman equation

$$Q_{\theta_*}(s,a) = R(s,a) - J(\theta_*) + \sum_{s'} P(s' \mid s,a) V_{\theta_*}(s')$$

$$\Rightarrow R(s,a) - J(\theta_*) + \sum_{s'} P(s' \mid s,a) V_{\theta_*}(s') - V_{\theta_*}(s)$$

$$= Q_{\theta_*}(s,a) - V_{\theta_*}(s)$$

$$= A_{\theta_*}(s,a)$$

27th march :

we had seen, $F^T D_\theta^{s,a}[\bar{R} - F\lambda_\theta] = 0$ ——①

$\phi^T D_\theta^s [T_\theta^V(\phi v_\theta) - \phi v_\theta] = 0$ ——②

similar to TD(0) for single-agent

recall that for fixed θ (Policy gradient algorithm) the 3 critic updates converge and their limits are as per ① ②

$$\left.\begin{aligned} \lim_{t\to\infty} w_t^i &= J(\theta) \\ \lim_{t\to\infty} \lambda_t^i &= \lambda_\theta \\ \lim_{t\to\infty} v_t^i &= v_\theta \end{aligned}\right\} \text{a.s}$$

$V(s;v) = v^T\varphi(s)$ $\quad \varphi(s) = [\varphi_1(s), \ldots \varphi_K(s)]^T \in \mathbb{R}^K$ ← features associated with states

$\bar{R}(s,a;\lambda) = \lambda^T f(s,a)$ $\quad f(s,a) = [f_1(s,a), \ldots f_M(s,a)]^T \in \mathbb{R}^M$ ← features associated with (s,a)

furthermore, $\phi \in \mathbb{R}^{|S| \times K}$ have $[\varphi_l(s), s \in S]^T$ as l^{th} column $l \in [K]$

$F \in \mathbb{R}^{|S| \cdot |A| \times M}$ have $[f_m(s,a), s \in S, a \in A]^T$ as m^{th} column, $m \in [M]$

full-rank as we are choosing features

now, $P_\theta[s'|s] = \sum_{a \in A} P[s'|s,a]\pi_\theta[s,a]$

$D_\theta^s = \text{diag}[d_\theta(s), s \in S]$

$\bar{R}_\theta = [\bar{R}_\theta(s), s \in S]^T \in \mathbb{R}^{|S|}$

$\bar{R}_\theta(s) = \sum_a \pi_\theta(s,a)\bar{R}(s,a)$

$T_\theta^V : \mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$ $\quad v \in \mathbb{R}^{|S|}$ as

$T_\theta^V(v) = \bar{R}_\theta - J(\theta)\cdot \mathbb{1} + P_\theta v$ → optimality equation for average

$\mathbb{1}$ ← vector of all ones

$F^T D_\theta^{s,a}[\bar{R} - F\lambda_\theta] = 0$ → MSE approximation

$\Phi^T D_\theta^s [T_\theta^V(\Phi v_\theta) - \Phi v_\theta] = 0$ → projection

recall, $\delta_{t,\theta}^i = f_t^T \lambda_\theta - J(\theta) + \varphi_{t+1}^T v_\theta - \varphi_t^T v_\theta$

$$\psi_{t,\theta}^i = \nabla_{\theta^i} \log \pi_{\theta^i}^i(s_t, a_t^i)$$

then we have following : conv of actor

$\{\theta_t^i\}$ converges a.s to the asymptotically stable equilibria

$$\dot{\theta}^{i\circ} = \hat{\Gamma}^{i\circ}\left[\mathbb{E}_{s_t \sim d_\theta, a_t \sim \pi_\theta}\left(\delta_{t,\theta}^{\gamma^i} \cdot \psi_{t,\theta}^{i}\right)\right] \quad \forall i \in N$$

↑ projection operator

Assumptions:

$\pi_\theta^i(\,) > 0$ as differentiable

also, $\theta_t^{i\circ}$ needs a local projection operator

$$\Gamma^{i\circ}: \mathbb{R}^{m_i} \to \Theta^{i\circ} \subset \mathbb{R}^{m_i}$$

that projects any θ_t^i onto compact set $\theta^{i\circ}$

also, $\theta = \prod_{i=1}^{N} \theta^{i\circ}$ is large enough to include atleast one local minimum of $J(\theta)$

ODE approach:

stochastic approximation (SA) iterates

$x_{n+1} = x_n + a_n(\,)$
x_n: config
↑ Random variable

- martingale
- Arya Akhavan
- ODE

1st April:

we have seen actor-critic method, now for example we see stochastic ODE approach

Eg: $S_n = \sum_{i=1}^{n} X_i$ $\quad \frac{S_{n+1}}{n+1} = \frac{\sum_{1}^{n+1} X_t}{n+1} = \frac{X_{n+1} + S_n}{n+1}$

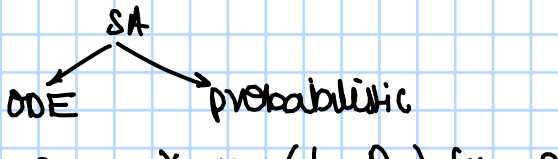
$$= \frac{1}{n+1} X_{n+1} + \frac{S_n}{n+1}$$

$$= \frac{1}{n+1} \theta_{n+1} + \left(1 - \frac{n}{n+1}\right) S_n$$

↓ set of limit points (long a.s)

a.s non similar to Q-learning martingale proof

SA-iteration:



$$\frac{S_{n+1} - S_n}{n+1} = \frac{X_{n+1}}{n+1} + \left(1 - \frac{n}{n+1}\right) S_n - \frac{S_n}{n+1}$$

$$= \frac{X_{n+1}}{n+1} + \left(1 + \left(-\frac{n-1}{n+1}\right)\right) S_n$$

$$= \frac{X_{n+1}}{n+1} + O(\,) \approx \frac{\Delta S_n}{n+1}$$

for large n: $\frac{1}{n+1} \approx \frac{1}{f(n+1)}$ then $s(t) \approx f(S(t))$ → ode

15th April:

we have seen the paper on decentralized MARL Algorithms

via Empirical framework
parametrize reward
& leverage average of those parameters

algos obs for $\sum_{i=1}^{n} f_i(\cdot)$ → one way of giving information

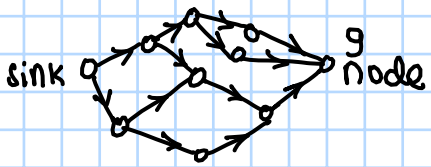
f_i ↦ avg reward of i^{th} agent

$$f_i = \lim_{T\to\infty} \sum_{i=1}^{T} \mathbb{E}[r_i(\,)]$$

Other sum models:

Total cost model, Stochastic Shortest path (SSP$_s$) model

Network ↔ graph



source node → sink } links ↔ costs
sink node → goal } Paths ↔ sum links/costs

cost of a path from sink to g → learn (mean/avg) cost expectation

planning Horizon:

planning horizon, random but finite say $\mathbb{E}[\sum_{i=1}^{\tau_p} r_i]$ τ_p: horizon of path p

$\min_p \mathbb{E}_p[\sum_{i=1}^{\tau_p} r_i]$ → Total cost MDP

DE of total cost model (DP setup) = $\min_{\text{link } i} \{(\text{cost } l_i) + (\text{min of cost to go})\}$

MDP: 15 yrs old approach

learning version (RL version, cost realization)

Learnability/convergence not known/complex

~5 yrs RL fun (single agent) SSP

~5 papers

episodic task ↔ sublinear regret (# of episodes, K)

pioneering work ⇔ single agent in a network

rarely, a single agent uses a network, multiagent version → leads to notions of congestion

(Z, E) = Network E = Edge = $\{(i,j), i,j \in Z\}$

$z = (\text{sink}_1, \ldots, \text{goal})$

↑ initial nodes

$N = \{1, \ldots n\}$ = set of agents

common objective :

all agents need to travel through the network

sink → goal

$\min_i \{ \mathbb{E}[\text{cost all agent reaching to goal node}] \}$

c_p^i

$i \rightarrow$!

$j \rightarrow$!

$\rightarrow$!

$\{N, \{V \setminus \{\text{sink}\}, g, \{A^i\}_{i \in N}, \{c^i\}_{i \in p}, \mathbb{P}(e_t)_{t \geq 0}\}$

f_i = expected travel cost of agent i

$f = \sum_i f_i \quad s = (s^1, s^2, \ldots s^n), \; s^i \in \mathbb{Z}$

$A^i(s)$, cost $c^i(s, a)$, $\bar{c}^i(s, a) = \underbrace{k^i(s,a)}_{\text{private}} \underbrace{\sum_{j \in N} \mathbb{I}\{a^i(s) = a^j(s)\}}_{\text{congestion public info}}$

real say fuel efficiency

$c^i(\,) \geq c_{min} > 0$

$V^\pi(s) = \lim_{T \to \infty} \mathbb{E}_\pi \left[\sum_{i=1}^{T} \bar{c}(s_t, \pi(s_t)) \middle| s_t = s \right]$

where $\bar{c}(s_t, \pi(s_t)) = \frac{1}{n} \sum_{i=1}^{n} c^i(s_t, \pi(s_t))$

for $\delta \; \exists \pi^* = \underset{\pi \in \Pi_p}{\text{argmin}} V^\pi(s)$

$\Rightarrow V^* = V^{\pi^*}$

$\mathbb{P}(V(s,a)) = \sum_{s'} \mathbb{P}(s'|s,a) V(s')$

$P(s'|s,a) = \langle \phi(s'|s,a), \theta^* \rangle$

16th April:

Assume that $\exists \theta^* \in Ad(\|\theta^*\|_2 \le \sqrt{n}\xi)$ so that

$$P(s'|s,a) = \langle \phi(s'|s,a), \theta^* \rangle$$

↖ from the beast

$$\phi_V(s,a) = \sum_{s'} \phi(s'|s,a)\, V(s')$$

for some $V : \mathcal{S} \to [0, B]$

$$\mathcal{L}V(s) = \min_a \{ \bar{c}(s,a) + PV(s,a) \}$$

$$B^* = \max_s V^*(s)$$

use $\bar{c}(\,,\,)$ and parametrize by w

$$\bar{c}(\,,\,,w) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$$
$$w \in \mathbb{R}^K, \quad K << |\mathcal{S} \times \mathcal{A}|$$

$$\min_w \; E_{s,a}[\bar{c}(s,a) - \bar{c}(s,a,w)]^2$$

$$\tilde{w}_t^i \leftarrow w_t^i + \gamma_t [c_t^i(\,,\,) - \bar{c}(\,,\,,w_t^i)] \times \nabla_w \bar{c}(\,,\,,w_t]$$

$$w_t^i = \sum_{j \in N} \ell_t(i,j)\, \tilde{w}_t^j$$

↑ communication matrix entries

$$Q^{i\pi}(s,a) = \bar{c}(s,a,w^*) + PV^*(s,a)$$

$$V^*(s) = \min_a Q^{i*}(s,a)$$

Feasability criteria:

K-episode

$$R_K = \sum_{j=1}^{K} \sum_{\ell=1}^{I_j} \frac{1}{n} \sum_{i \in N} (\bar{c}(s_{i,\ell}, a_{j,\ell}, w_{j,\ell}^i) - K V^{i*}(s_{init}))$$

I_j length of episode j period

$V^{i*} = V^*$ (Th 2) w_j^i's converge to w^*