REINFORCEMENT LEARNING

7/31/2023

THE GOAL IS TO CONTROL AN AGENT (NOT JUST MAKING PREDS) IN AN ENVIRONMENT. EACH EXAMPLE INFLUENCES EACH OTHER, SEQUENTIAL DECISION MAKING

MARKOV - DECISION PROCESS

TUPLE $(S, A, \{P_{SA}\}, \gamma, R)$

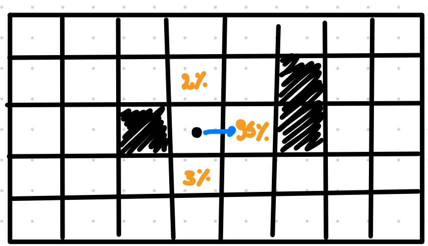
S - SET OF STATES EX $(\underbrace{x, y, z}_{\text{POSITION}}, \underbrace{v_x, v_y, v_z}_{\text{VELOCITY}}, \underbrace{\theta, \phi, \alpha}_{\text{DIRECTIONS}}, \dots) \rightarrow$ HELICOPTER

WE WILL 1^{ST} CONSIDER IT DISCRETE

A - SET OF ACTIONS EX ('MOVE LEFT', 'MOVE RIGHT')

P_{SA} - STATE TRANSITION PROBABILITIES

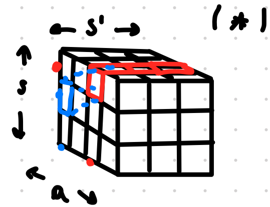
EX



RANDOMNESS IN THE ENVIRONMENT

DIFFERENT DISTRIBUTION FOR EACH (s, a)

WE CAN THINK OF $\{P_{SA}\}$ AS A 3D STRUCTURE (TENSOR)



— P_{SA} (PMF)

γ – DISCOUNT FACTOR, $\gamma \in [0, 1)$ HOW MUCH DO WE DESIRE TO GET THE R VS LATER

R – REWARD, OF BEING IN A STATE, $R(s) \in \mathbb{R}$

Ex

$$S_0 \xrightarrow{a_1} S_1 \xrightarrow{a_2} S_2 \xrightarrow{a_3} \cdots$$

$R(S_0)$ $R(S_1)$ $R(S_2)$

WE WANT TO MAXIMIZE THE EXPECTATION OF (*)

$$\max \mathbb{E}\left[\underbrace{R(S_0) + \gamma R(S_1) + \gamma^2 R(S_2) + \cdots + \gamma^d R(S_{d+1})}_{(*)}\right]$$

WILL ALWAYS CONVERGE!

WE NEED TO DEFINE

- POLICY π : FUNCTION THAT MAPS CURRENT STATE TO ACTION (AN AGENT EXECTUTES POLICY)

$$\pi : S \longmapsto A$$

- VALUE $V_\pi(s)$: FUNCTION THAT GIVES EXPECTED VALUE FOR BEING IN S AND FOLLOWING π

$$V_\pi(s) = \mathbb{E}\left[R(s_0) + \gamma R(s_1) + \dots + \gamma^{k-1} R(s_k)\right]$$

$$V_\pi(s) = \mathbb{E}\left[R(s_0) + \dots + \gamma^{k-1} R(s_k) \mid \pi\right]$$

ONCE REACHED K, π IS GIVEN

$$V_\pi(s) = R(s_0) + \mathbb{E}\left[\gamma R(s_1) + \dots + \gamma^{k-1} R(s_k) \mid \pi\right]$$

AND, SINCE THIS HAS A RECURSIVE STRUCTURE

BELLMAN EQUATION

$$V_\pi(s) = R(s) + \gamma \sum_{s' \in S} P_{s\pi(s)}(s') V_\pi(s')$$

$$V_\pi(s) = R(s) + \mathbb{E}_{s \sim P_{s\pi(s)}}\left[V_\pi(s')\right]$$

POLICY π $\xrightarrow{V=(I-\gamma P_{s\pi(s)})^{-1}R}$ VALUE f^n V

$\pi(s) = \arg\max_{a\in A} \sum_{s'\in S} P_{sa}(s')\, V(s')$

REPEATING THIS CONVERGES TO π^*

$$V(s) = [V(s_0), V(s_1), V(s_2), \dots]$$

$$P_{s\pi(s)} = \begin{bmatrix} \text{———} \\ \text{———} \\ (*) \end{bmatrix}$$

$$R(s) = [R(s_0), \dots]$$

$$\gamma \in \mathbb{R}$$

SYSTEM OF K LINEAR EQUATIONS (K VARS)

$$\begin{bmatrix} V(s_0) \\ V(s_1) \\ \vdots \end{bmatrix} = \begin{bmatrix} R(s_0) \\ R(s_1) \\ \vdots \end{bmatrix} + \gamma \begin{bmatrix} P_{s\pi(s)} \end{bmatrix} \begin{bmatrix} V(s'_0) \\ V(s'_1) \end{bmatrix}$$

$$V = R + \gamma P_{s\pi(s)} V \;\Big|\; (I - \gamma P_{s\pi(s)}) V = R \;/\; V = (I - P_{s\pi(s)})^{-1} R$$

THEN WE EXPRESS THE RATIONAL POLICY

$$\pi(s) = \arg\max_{a \in A} \sum_{s' \in S} P_{sa}(s') V(s')$$

AND THEN FIND V W/ THE GREEDY POLICY AND BELLMAN EQUATION

OPTIMAL VALUE FUNCTION

$$V^*(s) = \max_{\pi} V_{\pi}(s)$$

$$V^*(s) = R(s) + \max_{a \in A} \sum_{s' \in S} P_{sA}(s') V^*(s)$$

WE ARE NO MORE FOLLOWING A POLICY, BUT DIRECTLY MAXIMIZING THROUGH ACTIONS
SO WE CAN DEFINE AN OPTIMAL POLICY π^*

$$\pi^*(s) = \arg\max_{a \in A} \sum_{s' \in S} P_{sA}(s') V^*(s')$$

ALL POSSIBLE POLICIES' DOMAIN IS $|A|^{|S|}$, SO A LOT

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

ALGORITHMS

1. VALUE ITERATION
2. POLICY ITERATION

VALUE ITERATION

I. SET $V(s) = 0 \quad \forall s$

II. REPEAT UNTIL CONVERGENCE {

$$\forall s \quad V(s) := R(s) + \max_{a \in A} \gamma \sum_{s'} P_{sa}(s') V(s')$$

}

EX ON 1^{ST} ITER. THIS IS R(S)

EX $V^{(0)}(s) = 0 \rightarrow V^{(1)}(s) = R(s) \longrightarrow V^{(2)}(s) = R(s) + \gamma R(s_1) + \dots \longrightarrow V^{(T)}(s) = V^{*}(s)$

$V^{(t)}(s) \longrightarrow V^{(t+1)}(s)$

BELLMAN BACKUP OPERATOR

WILL ALWAYS CONVERGE TO V^{*}, π^{*}

POLICY ITERATION

I. INIT π RANDOMLY

II. REPEAT UNTIL CONVERGENCE {

(a) LET $V = V^{\pi}$

(b) FOR EVERY s, LET

$$\pi(s) = \arg\max_{a \in A} \sum_{s'} P_{sa}(s') V(s')$$

}

ALWAYS CONVERGE TO V^*, π^*

LEARNING THE MDP

UNTIL NOW WE KNEW R AND P_{SA}

IF WE WERE GIVEN ONLY A SUBSET OF THE MDP

$MDP = (S, A, \{P_{SA}\}, \gamma, R)$ → WE DON'T KNOW THE RULES OF THE GAME, SO WE NEED TO LEAR THE GAME WHILE GETTING GOOD AT IT

EX

EPISODES

$t_1: \quad S_0^{(1)} \xrightarrow{a_0^{(1)}} S_1^{(1)} \xrightarrow{a_1^{(1)}} S_2^{(1)} \longrightarrow \dots$

$t_2: \quad S_0^{(2)} \xrightarrow{a_0^{(2)}} S_1^{(2)} \longrightarrow \dots$

RUN TRIALS OVER AND OVER (EXPLORING STATE), THEN WE CAN DEFINE

$$P_{SA}(s') = \frac{\#\text{TIMES WE TOOK } a \text{ IN } S \text{ AND GOT IN } S'}{\#\text{TIME WE TOOK } a \text{ IN } S}$$

AND W/ GOOD COVERAGE WE CAN GET GOOD ESTIMATES

AT THE BEGINNING WE ASSIGN $P_{SA}(s) = 0{,}1$ (FOR EXAMPLE), BECAUSE IT WOULD BE $\frac{0}{0}$

AND REWARDS CAN BE DEFINED AS THE MEAN FROM THE TRIALS

$$R(s) = \frac{1}{n} \sum_{i=1}^{n} R^{(i)}$$

ALGORITHM

I. INIT π RANDOMLY

II. REPEAT {

(a) EXECUTE π IN MDP FOR SOME TRIALS

(b) USE ACCUMULATED EXPERIENCE AND UPDATE P_{SA} AND R

→ LEARN THE GAME

(c) APPLY VALUE - ITERATION W/ EST. P_{SA}, R

(d) UPDATE π TO BE GREEDY W.R.T. V_{π}

→ BECOME GOOD AT IT

OPTIMIZE

WHEN INIT V IN V-I WE INIT V AS THE OBTAINED R (?)

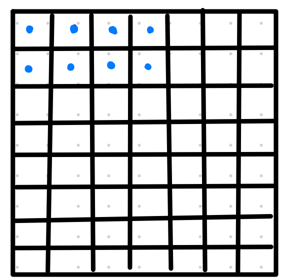
AS LONG AS R AND P_{SA} CONVERGE TO TRUE VALUES, THE ALGORITHM WILL CONVERGE

RL IN CONTINUOUS SPACES

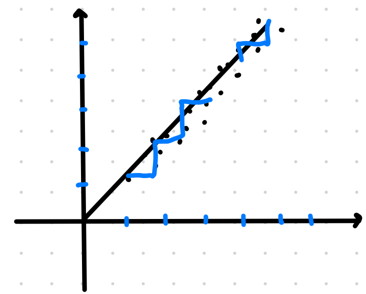
<u>EX</u>

$$s = (x, y, \theta, \dot{x}, \dot{y}, \dot{\theta})$$ IS A CONTINUOUS STATE SPACE

WE CAN DISCRETIZE!



STATE SPACE



TYPICALLY WE HAVE A SIMULATOR

$$s_t \longrightarrow \boxed{\text{SIMULATOR}} \longrightarrow s_{t+1} \sim P_{s_t a_t}$$

$$\uparrow$$

$$a_t$$

WE RUN AN n NUMBER OF TRIALS

$$
\begin{aligned}
&s_0^{(1)} \xrightarrow{a_0^{(1)}} s_1^{(1)} \xrightarrow{a_1^{(1)}} s_2^{(1)} \xrightarrow{a_2^{(1)}} \cdots \xrightarrow{a_{T-1}^{(1)}} s_T^{(1)} \\
&s_0^{(2)} \xrightarrow{a_0^{(2)}} s_1^{(2)} \xrightarrow{a_1^{(2)}} s_2^{(2)} \xrightarrow{a_2^{(2)}} \cdots \xrightarrow{a_{T-1}^{(2)}} s_T^{(2)} \\
&\cdots \\
&s_0^{(n)} \xrightarrow{a_0^{(n)}} s_1^{(n)} \xrightarrow{a_1^{(n)}} s_2^{(n)} \xrightarrow{a_2^{(n)}} \cdots \xrightarrow{a_{T-1}^{(n)}} s_T^{(n)}
\end{aligned}
$$

WE APPLY A LEARNING ALGORITHM TO PREDICT s_{t+1} GIVEN s_t, a_t
TIPICALLY A LINEAR ONE S.T.

$$s_{t+1} = f(s_t, a_t)$$

$$s_{t+1} = A s_t + B a_t$$

HERE THE PARAMETERS OF THE MODEL ARE A, B AND ESTIMATE THEM W/ DATA FROM n TRIALS

$$\arg\min_{A,B} \sum_{i=1}^{n} \sum_{t=0}^{T-1} \| s_{t+1}^{(i)} - (A s_t^{(i)} + B a_t^{(i)}) \|_2^2$$

HERE WE CAN CHOOSE TO MAKE THE MODEL

- DETERMINISTIC
- STOCHASTIC

$$s_{t+1} = A s_t + B a_t + \varepsilon_t \quad (**)$$

$$\varepsilon \sim N(0, \Sigma)$$

$$s_{t+1} = A\,\phi_s(s_t) + B\,\phi_a(a_t)$$

$$s_{t+1} \sim N(A\,\phi_s(s_t), B\,\phi_a(a_t), \Sigma)$$

ϕ_s, ϕ_A NON LINEAR FEATURE MAPPINGS OF S AND a

WE CAN NOW DECRIBE FITTED VALUE ITERATION, THAT APPROX. V OF A CONTINUOUS STATE MDP

$$V(s) := R(s) + \gamma \max_a \int_{s'} P_{sa}(s') V(s')\, ds'$$

$$= R(s) + \gamma \max_a \mathbb{E}_{s' \sim P_{sa}} \left[V(s') \right]$$

$$= R(s) + \gamma \max_a \underbrace{\frac{1}{K} \sum_{i=1}^{K}}_{\text{MONTE-CARLO EST. FROM } \mathbb{E}} V(s_i) \qquad s_i \sim P_{sa}$$

$$V(s) = \theta^T \phi(s) \quad (*)$$

∞ NUMBER OF STATES, WE CANNOT STORE THEM. SO WE FIT A MODEL THAT APPROX THE RIGHT VALUE TO THIS (*) MODELING (PARAMETRIC) ASSUMPTION

GIVEN THE ASSUMPTION, WE CAN DEFINE THIS NEW ALGORITHM

ALG

I. SAMPLE n STATES $s^{(1)}, s^{(2)}, \ldots, s^{(n)}$

II. INIT $\theta = 0$ $\quad (V(s) = \theta^T \phi(s))$

III. REPEAT FOR $i = 1, \ldots, n$ {

FOR EACH $a \in A$ {

SAMPLE $s_1', \ldots, s_K' \underset{\text{FROM}}{\sim} P_{sa}$ → LEARNT IN PHASE 1 (**)

SET $q(a) = \frac{1}{K} \sum^{K} R(s^{(i)}) + \gamma V(s^{(i)})$ $\qquad$ $q(a)$ EMPIRICAL AVERAGE

}

SET $y^{(i)} = \max_a q(a)$

}

$y^{(i)}$ IS THE O: OF BELLMAN OPERATOR (EST. OF $R(s^{(i)}) + \gamma \max_a \mathbb{E}_{s' \sim P_{sa}} [V(s')]$)

$(s^{(i)}, y^{(i)})_{i=1}^{n}$ $\quad V^{(t+1)}(s^{(i)}) = y^{(i)}$ (*)

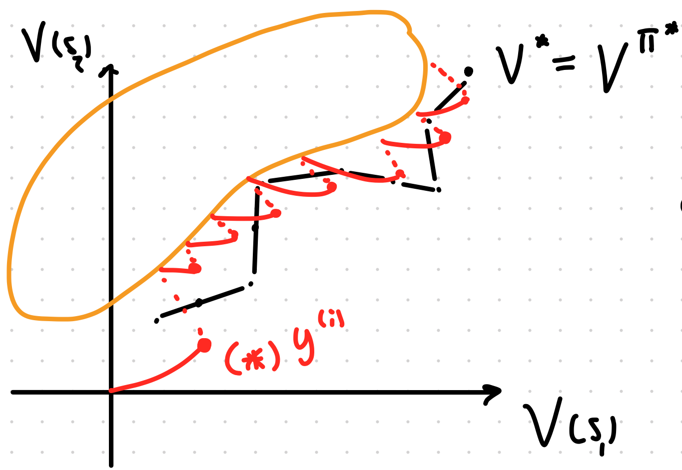
SET $\theta = \arg\min_{\theta} \sum_{i=1}^{n} (\theta^T \phi(s^{(i)}) - y^{(i)})^2$

BASICALLY, BY APPLYING THE BELLMANN BACKUP OPERATOR, WE WILL CONVERGE TO V^*

$V(s) = \theta^T(\phi(s))$

WE KNOW HOW TO REPRESENT THIS SUBSPACE (AND ITS f)

THIS SPACE DEPENDS ON THE FEATURE MAP



∞ - DIMENSIONAL SPACE (CONTINUOUS)

WE PROJECT $y^{(1)}$ ONTO THE SUBSPACE WE KNOW HOW TO REPRESENT
AT SOME POINT WE OBTAIN THE APPROX VALUE $V^{(T)}$