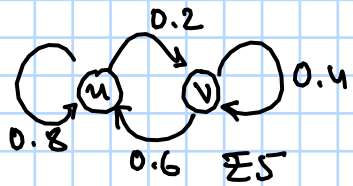
IE 620 Tutorials

Tutorial-1 :

1.



£10 £5

$$\text{now,} \quad \begin{matrix} & u & v \\ u \\ v \end{matrix} \begin{bmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{bmatrix} = \text{transition matrix}$$

$$\begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{bmatrix} = \begin{bmatrix} x & y \end{bmatrix}$$

↑ steady state

$$0.8x + 0.6y = x$$
$$0.2x + 0.4y = y$$

$$0.6y = 0.2x$$
$$x + y = 1$$

$$\Rightarrow 4y = 1 \quad x = 3/4$$
$$y = 1/4$$

so, average revenue $= 3/4 \times 10 - 1/4 \times 5$

$$= \frac{30}{4} - \frac{5}{4}$$
$$= \frac{25}{4}$$
$$= \frac{12.5}{2}$$
$$= 6.25$$

$$\begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 0.8 & 0.2 \\ 0.7 & 0.3 \end{bmatrix} = \begin{bmatrix} x & y \end{bmatrix}$$

$$x + y = 1$$

$$0.8x + 0.7y = x$$
$$0.7y = 0.2x$$
$$2x = 7y$$

$$2x + 2y = 2$$
$$9y = 2$$
$$y = 2/9$$
$$x = 7/9$$

$$\text{Avg} = 7/9 \times 10 - 2/9 \times 6$$
$$= \frac{70 - 12}{9}$$
$$= \frac{58}{9}$$

2. $\lambda = 0.7 \quad \lambda = 0.9$

$$V_u = \gamma_u + \lambda \left(P(u,u) V_u + P(v,u) V_v \right)$$

(on state u)

$$V_V = r_V + \lambda(P(U,V)V_U + P(V,V)V_V)$$

$\lambda = 0.7$: $V_U = 10 + 0.7(0.8V_U + 0.2V_V)$
$V_V = -5 + 0.7(0.6V_U + 0.4V_V)$

$\Rightarrow V_U = 17.14 \quad V_V = 3.09$

$\lambda = 0.9$:
$V_U = 10 + 0.9(0.8V_U + 0.2V_V)$
$V_V = -5 + 0.9(0.6V_U + 0.4V_V)$
$\Rightarrow V_U \approx 37.5$
$V_V \approx 28.36$

$\lambda = 0.7$: $V_U = 10 + 0.7(0.8V_U + 0.2V_V)$
$V_V = -6 + 0.7(0.7V_U + 0.3V_V)$
$\Rightarrow V_U \approx 17.82$
$V_V \approx 3.58$

$\lambda = 0.9$: $0.28V_U - 0.18V_V = 10$
$-0.63V_U + 0.73V_V = -6$
$V_U = 41.47$
$V_V = 30.08$

3. let $\pi^* = (a_t^*, \pi_{t+1}^*)$

if we know π^* is optimal for full we are done
if not
then $\exists \tilde{\pi}$ s.t it gives higher value at t
ie $V_t^{\tilde{\pi}}(x) > V_t^{\pi^*}(x)$

$$V_t^{\tilde{\pi}}(x) = r(x, \tilde{a}_t) + \mathbb{E}[V_{t+1}^{\tilde{\pi}}(X_{t+1})]$$
$$V_t^{\pi^*}(x) = r(x, \tilde{a}_t) + \mathbb{E}[V_{t+1}^{\pi^*}(X_{t+1})]$$
$$V_{t+1}^{\pi^*}(x) = \max_{\pi} V_{t+1}^{\pi}(x)$$
$$\Rightarrow V_{t+1}^{\tilde{\pi}}(x) \leq V_{t+1}^{\pi^*}(x) \quad \forall x$$
$$\Rightarrow r(x, \tilde{a}_t) + \mathbb{E}[V_{t+1}^{\tilde{\pi}}(X_{t+1})] \leq r(x, \tilde{a}_t) + \mathbb{E}[V_{t+1}^{\pi^*}]$$

as a_t^* maximises bellman equation

$$r(x, a_t^*) + \mathbb{E}[V_{t+1}^{\pi^*}] \geq r(x, \tilde{a}_t) + \mathbb{E}[V_{t+1}^{\pi^*}]$$

$\Rightarrow V_t^{\tilde{\pi}}(x) \leq V_t^{\pi^*}(t)$ this is a contradiction

bellman:

$$V(x) = r(x, a^*) + \sum P(y|x, a^*)V(y)$$

4. (a) $V^{\pi}(x) = \mathbb{E}_{\pi}\left[\sum_{t=1}^{\infty} \lambda^{t-1} r(X_t, A_t, X_{t+1}) \mid X_0 = x\right]$

bellman form:
$$V^{\pi}(x) = \sum_a \pi(a|x) \sum_y P(y|x,a)[r(x,a,y) + \lambda V^{\pi}(y)]$$

(b) optimal: $V^*(x) = \max_\pi V^\pi(x) \quad \forall x \in S$

(c) $V^*(x) = \max_a \sum_y P(y|x,a)[r(x,a,y) + \lambda V^*(y)]$

satisfy bellman condition

(d) $V_{k+1}(x) = \max_a \sum_y P(y|x,a)[r(x,a,y) + \lambda V_k(y)]$

$V_k \to V^*$

(e) $Q^\pi(x,a) = E_\pi\left[\sum_{t=1}^{\infty} \lambda^{t-1} r(X_t, A_t, X_{t+1}) \,\middle|\, X_0 = x, A_0 = a\right]$

bellman form:

$Q^\pi(x,a) = \sum_y P(y|x,a)\left[r(x,a,y) + \lambda \sum_b \pi(b|y) Q^\pi(y,b)\right]$

(f) $Q^*(x,a) = \max_\pi Q^\pi(x,a)$

(g) $Q^*(x,a) = \sum_y P(y|x,a)[r(x,a,y) + \lambda \max_b Q^*(y,b)]$

(h) $Q_{k+1}(x,a) = \sum_y P(y|x,a)\left[r(x,a,y) + \lambda \max_b Q_k(y,b)\right]$

(i) $V^*(x) = \max_a Q^*(x,a)$

$Q^*(x,a) = \sum_y P(y|x,a)[r(x,a,y) + \lambda V^*(y)]$

(j) $Q_{t+1}(x_t, a_t) = Q_t(x_t, a_t) + \alpha[r_t + \gamma \underbrace{\max_a Q_t(x_{t+1}, a)}_{\text{step 1}} - Q_t(x_t, a_t)]$

step 1: $V_t(x_{t+1}) = \max_a Q_t(x_{t+1}, a)$

(estimating next step value)

step 2: updates $Q_{t+1}(x_t, a_t)$

by computing $r_t + \gamma V_t(x_{t+1})$ → value iteration update

9. (a) $H_q(x,a) = \sum_y P(y|x,a)\left[r(s,a,y) + \gamma \max_b q(y,b)\right]$

$$\|Hq_1 - Hq_2\|_\infty = \max_{(x,a)} \Big| \sum_y P(y|x,a) r(s,a,y) - \sum_y P(y|x,a) r(s,a,y) + \gamma \sum_y P(y|x,a)\left(\max_b q_1(y,b) - \max_{b'} q_2(y,b)\right) \Big|$$

$$\leq \max_{(x,a)} |\gamma \sum_y P(y|x,a)| \, \|q_1 - q_2\|_\infty$$

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

as $\gamma < 1 \Rightarrow H$ is a contraction map

i.e $\exists Q^*$ unique fixed point of H

Q^* is optimal value (convergence of Q-learning)

(b) $\sum_t \alpha_t(x,a) = \infty \quad \sum_t \alpha_t^2(x,a) < \infty$

$0 \le \alpha_t(x,a) < 1$

$\alpha_t = \frac{1}{t+1}$

(c) $\Delta_t(x,a) = Q_t(x,a) - Q^*(x,a)$

then $\Delta_{t+1}(x,a) = Q_{t+1}(x,a) - Q^*(x,a)$

$= Q_t(x,a) + \alpha_t(x,a)[r_t(x,a) + \gamma \max_b q_t(x_{t+1},b) - q_t(x_t,a_t)]$

$- Q^*(x,a) - \alpha_t Q^* + \alpha_t Q^*$

$= (1-\alpha_t)(Q_t - Q^*) + \alpha_t[r_t(x,a) + \gamma \max_b q_t(x_{t+1},b) - Q^*]$

$F_t(x,a)$

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

now, α_t satisfies (b)

$\mathbb{E}[F_t|\mathcal{P}_t] = HQ_t - Q^*$

$\Rightarrow \| \mathbb{E}[F_t|\mathcal{P}_t] \|_\infty \le \gamma \|\Delta_t\|_\infty$

$\mathrm{var}(F_t|\mathcal{P}_t) = \mathbb{E}[(r_t(x,a) + \gamma \max_b q_t(x_{t+1},b) - Hq_t)^2]$

↑ Bounded

$\le C \le C + C\|\Delta_t\|^2$

some $C > 0$

then $\Delta_t \to 0$ wp 1 $\Rightarrow Q_t \to Q^*$

6. this follows from proof

7. $Hq(x,a) = \sum_y P(y|x,a)[r(x,a,y) + \gamma \max_b q(y,b)]$

by induction let's assume $q_i \le q_{i+1}$

then if we show

$q_{i+1} \le q_{i+2}$ we are done

$q_i \le q_{i+1} \Rightarrow \max_b q_i(y,b) \le \max_b q_{i+1}(y,b)$

$$\Rightarrow \gamma \max_b q_i^*(y,b) \le \gamma \max_b q_{i+1}^*(y,b)$$

$$\Rightarrow r(x,a,y) + \gamma \max_b q_i^*(y,b) \le r(x,a,y) + \gamma \max_b q_{i+1}^*(y,b)$$

(f_i under the first two terms; f_{i+1} under the last two terms)

$$\Rightarrow \sum_y P(y|x,a)[f_i] \le \sum_y P(y|x,a)[f_{i+1}]$$

$$\Rightarrow q_{i+1} \le q_{i+2}$$

8. (MC) methods

(a) it estimates value/action-value function

$$v^\pi(x) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid X_0 = x\right]$$

$$Q^\pi(x,a) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid X_0 = x, A_0 = a\right]$$

(b) data: full episode

$$(x_0, a_0, r_0, x_1, a_1, r_1 \ldots r_T)$$

(c) it is model free, does not require $P(y|x,a)$, $r(x,a,y)$

9. first-visit MC for $V \approx V_\pi$

given π

$V(x) \leftarrow 0$

$N(x) \leftarrow 0$

generate $(x_0, a_0, r_0, x_1, \ldots x_T)$

$\forall t, G_t = \sum_{k=t}^{T-1} \gamma^{k+1} r_k$

if x_t first visit in this episode

$N(x_t) \leftarrow N(x_t) + 1$

update: $V(x_t) \leftarrow V(x_t) + \frac{1}{N(x_t)}(G_t - V(x_t))$

in every visit:

$N(x_t) \leftarrow N(x_t) + 1$

$V(x_t) \leftarrow V(x_t) + \frac{1}{N(x_t)}(G_t - V(x_t))$

so in f_V, ε_V, we are estimating functions from experience

updates/episode	f_V	ε_V
var	1	multiple
Bias	>	<
	None	small

MLE: we estimate the model

i.e we estimate ① transition probabilities
② rewards
③ solve bellman equation

10. MC: MSE

$$\min_{V} \mathbb{E}\left[\left(G_t - V(x)\right)^2\right]$$

MSE: negative log-likelihood

$$\min_{p} -\sum \log P(y|x,a)$$

↑ choose transition probability to maximise probability of observed pairs

Tutorial - 2:

1. $v^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} r_t \mid s_0 = s\right]$ $\theta \in \mathbb{R}^d$ parameterise policy via

$\pi(a|s,\theta)$

$J(\theta) = \mathbb{E}[v^\pi(s_0)] = \langle \mu, v^\pi \rangle$ ← distribution of initial state

$d^\pi(s) = \sum_{t=0}^{\infty} P(s_t = s)$

$Q_w(s,a)$ $w \in \mathbb{R}^d$ approximation of $Q(s,a)$

value function $\rho(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^{t+1} r_t \mid \pi, s_0\right]$

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

Any policy parameterised by $\theta \in \mathbb{R}^d$

then

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

proof:

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

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

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

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

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

$$\frac{\partial \rho}{\partial \theta} = \frac{\partial}{\partial \theta} \mathbb{E}\left[\sum_1 \gamma^{t+1} r_t \mid s_0, \pi\right] = \frac{\partial}{\partial \theta} v^\pi(s_0) = \sum_s d^\pi(s) \sum_a \frac{d\pi(s,a)}{d\theta} Q^\pi(s,a)$$

(a) To prove: $d^\pi = \mu + P^\pi d^\pi$

proof:

$$d^\pi(s) = P(s_0 = s) + \sum_{t=1}^{\infty} P(s_t = s)$$

$P(s_0 = s) = \mu(s)$

$t \geq 1$: $P(s_t = s) = \sum_{s'} P(s_{t-1} = s') P^\pi(s|s')$

$$\sum_{t=1}^{\infty} P(s_t = s) = \sum_{s'} \left(\sum_{t=0}^{\infty} P(s_t = s')\right) P^\pi(s|s')$$

$$\Rightarrow d^\pi(s) = \mu(s) + \sum_{s'} d^\pi(s') P^\pi(s|s')$$

$$\Rightarrow d^\pi = \mu + P^\pi d^\pi$$

and so, $d^\pi = (I - P^\pi)^{-1}\mu$

(b) $J(\theta) = \sum_s \mu(s) V^\pi(s)$

$$\nabla_\theta J(\theta) = \sum_s \mu(s) \left[\nabla_\theta V^\pi(s)\right]$$

$$= \sum_s d^\pi(s) \sum_a \nabla_\theta \pi(a|s,\theta) Q^\pi(s,a)$$

(∵ follows similar to policy gradient proof)

(c) $\nabla_\theta J(\theta) = \sum_a \nabla_\theta \pi(a|s,\theta) Q^\pi(s,a)$ → summing over all action, but taking only gives one state actioned

(d) as $\nabla_\theta \pi(a|s) = \pi(a|s) \nabla_\theta \log \pi(a|s)$

$$\Rightarrow \nabla_\theta J(\theta) = \sum_s d^\pi(s) \sum_a \pi(a|s) Q^\pi(s,a) \nabla_\theta \log \pi(a|s)$$

$$= \mathbb{E}_\pi\left[\sum_{t=0} Q^\pi(s_t, a_t) \nabla_\theta \log \pi(a_t|s_t)\right]$$

(e) as no longer need to sum over all actions

(f) $Q^\pi(s,a)$ above is the value function, not parameterised

2. a = order quantity

D ~ F
↑
Random demand

selling price: p
cost: c

profit $= p \cdot \min(a, D) - c \cdot a$ (D ← Demand, a ← ordered)

$\max_a \mathbb{E}[\text{profit}]$ J = Objective

a* ← optimal

$$a^* = F^{-1}\left(\frac{p-c}{p}\right)$$

$r(a, D) = p \cdot \min(a, D) - c \cdot a$

$J(\theta) = \mathbb{E}_{a \sim \pi_\theta}[\mathbb{E}_D[r(a, D)]]$

$R(a) := \mathbb{E}_D[r(a, D)]$

$J(\theta) = \mathbb{E}_{a \sim \pi_\theta}[R(a)]$

$\nabla_\theta J(\theta) = \mathbb{E}_{a \sim \pi_\theta}[R(a) \nabla_\theta \log \pi_\theta(a)]$

$Q^\pi(s,a) = R(a)$

$$\Rightarrow \nabla_\theta J(\theta) = \int \pi_\theta(a) R(a) \nabla_\theta \log \pi_\theta(a) \, da$$

3. (a) compatible feature condition:

$$\nabla_w Q(s,a;w) = \nabla_\theta \log \pi(a|s;\theta)$$

in linear form:

$$Q_w(s,a) = w^T \nabla_\theta \log \pi_\theta(a|s)$$

(b) $w^* = \underset{w}{\operatorname{argmin}}\ \mathbb{E}_\pi\left[\left(Q^\pi(s,a) - Q(s,a;w)\right)^2\right]$

$$\mathbb{E}_\pi\left[\left(Q^\pi(s,a) - Q_w(s,a)\right)\nabla_w Q_w(s,a)\right] = 0$$

(c) $\nabla_w Q_w(s,a) = \nabla_\theta \log \pi_\theta(a|s)$

$$\Rightarrow \mathbb{E}_\pi\left[\left(Q^\pi(s,a) - Q_w(s,a)\right)\nabla_\theta \log \pi_\theta(a|s)\right] = 0$$

↑ as this is orthogonal, replacing $Q^\pi(s,a)$ with $Q_w(s,a)$ it does not bias gradient

(d) as $\nabla_\theta J(\theta) = \mathbb{E}_\pi\left[Q^\pi(s,a)\nabla_\theta \log \pi(a|s)\right]$

by orthogonality

$$= \mathbb{E}_\pi\left[Q_{w^*}(s,a)\nabla_\theta \log \pi(a|s)\right]$$

$$= \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} Q_{w^*}(s_t,a_t)\nabla_\theta \log \pi(a_t|s_t)\right]$$

(e) as we only need $(s_t,a_t) \to Q_{w^*}(s_t,a_t)$ estimate

$$\hat{\nabla} = \sum_t Q_{w^*}(s_t,a_t)\nabla_\theta \log \pi(a_t|s_t)$$

no need for summing over actions

4. They cover average reward and start-state ↑ both episodic & discounted

same theorem in both cases

5. Every actions

- d^π is discounted → finite → initial dist matters
- stationary distribution → infinite → initial dist not matter

6. $d^\pi(s)$ never differentiated critic → fast, actor → slow (DO LOOP), with compatibility we get exact gradient

7. (a) softmax:

$$\pi(a|s) = \frac{e^{\theta^T \phi(s,a)}}{\sum_{a'} e^{\theta^T \phi(s,a')}}$$

$$\nabla_\theta \log \pi_\theta(a|s) = \phi(s,a) - \sum_{a'} \pi_\theta(a'|s)\phi(s,a')$$

$$Q_w(s,a) = w^T \nabla_\theta \log \pi(a|s)$$
$$= w^T [\phi(s,a) - \sum_{a'} \pi_\theta(a'|s)\phi(s,a')]$$

(b) gaussian linear:

$$\pi_\theta(a|s) = N(\mu_\theta(s), \sigma^2)$$

$$\nabla_\theta \log \pi_\theta(a|s) = \frac{(a - \mu_\theta(s))}{\sigma^2} \nabla_\theta \mu_\theta(s)$$

$$\Rightarrow Q_w(s,a) = w^T \left[\frac{(a-\mu_\theta(s))}{\sigma^2} \nabla_\theta \mu_\theta(s) \right]$$

(c) simplex:

simplex is $\pi_\theta(a|s) = \theta_a(s)$ $\quad \sum_a \theta_a(s) = 1 \quad \forall s \in S$

$$\nabla_\theta \log \pi_\theta(a|s) = \frac{1}{\theta_a(s)} e_a$$
$a \in \{1, 2, \dots n\}$

then

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \cdots$$

$$\Rightarrow Q_w(s,a) = w^T \left[\frac{1}{\theta_a(s)} e_a \right]$$

8. gaussian compatible features
- Natural
- smooth gradients
- low var for σ closer 1
- sensitive to σ
- only linear
- requires differentiable mean function μ_θ

9. π, π' policies corresponding to θ, θ' episodic task

(a) as $v^\pi = r + P^\pi v^\pi$

$\Rightarrow (I - P^\pi) v^\pi = r$

$\Rightarrow v^\pi = (I - P^\pi)^{-1} r$ as $I - P^\pi$ invertible

(b) $V^{\pi'} - V^{\pi} = (A')^{-1}\gamma - A^{-1}\gamma$

$= (A')^{-1}\gamma - (A')^{-1}A'A^{-1}\gamma$

(c) $V^{\pi'} - V^{\pi} = (A')^{-1}[\gamma - A'A^{-1}\gamma]$

$= (A')^{-1}[AA^{-1}\gamma - A'A^{-1}\gamma]$

$= (A')^{-1}[(A - A')A^{-1}\gamma]$

$= (I - P^{\pi'})^{-1}(P^{\pi'} - P^{\pi})V^{\pi}$

(d) $J(\theta) = \mu^T V^{\pi}$

$J(\theta') - J(\theta) = \mu^T(V^{\pi'} - V^{\pi})$

and $\mu^T(I - P^{\pi'})^{-1} = d^{\pi'}$

$\Rightarrow J(\theta') - J(\theta) = \mu^T(I - P^{\pi'})^{-1}(P^{\pi'} - P^{\pi})V^{\pi}$

$= d^{\pi'}(P^{\pi'} - P^{\pi})V^{\pi}$

now $V^{\pi} - V^{\pi}P^{\pi} = \gamma$

$\Rightarrow J(\theta') - J(\theta) = d^{\pi'}(P^{\pi'}V^{\pi} - P^{\pi}V^{\pi})$

$= d^{\pi'}(P^{\pi'}V^{\pi} + \gamma - V^{\pi})$

(e) $Q^{\pi}(s,a) = r(s,a) + \sum_{s'} P(s'|s,a)V^{\pi}(s')$

$\Rightarrow \sum_a \pi'(a|s)Q^{\pi}(s,a) = r(s) + \sum_{s'} P^{\pi'}(s'|s)V^{\pi}(s')$

(f) $J(\theta') - J(\theta) = d^{\pi'}(\gamma + P^{\pi'}V^{\pi} - V^{\pi})$

$= \sum_s d^{\pi'}(s)\left[\sum_a \pi'(a|s)Q^{\pi}(s,a) - V^{\pi}(s)\right]$

$= \sum_s d^{\pi'}(s) \sum_a \pi'(a|s)\left[Q^{\pi}(s,a) - V^{\pi}(s)\right]$

(g) $Q(s,a) \geq V(s)$ θ' improves θ

$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s)$

$\Rightarrow J(\theta') - J(\theta) = \mathbb{E}_{s\sim d^{\pi'},\, a\sim\pi'}[A^{\pi}(s,a)]$

(h) as we don't know $d^{\pi'}$, π' unknown

(i) we used $\nabla_\theta J(\theta) = \mathbb{E}_{s\sim d^{\pi},\, a\sim\pi}[Q^{\pi}(s,a)\nabla_\theta \log \pi_\theta(a|s)]$

(j) we want to update

"action better than expected"

as $\mathbb{E}[V^{\pi}(s)\nabla_\theta \log \pi(a|s)] = 0$

$\Rightarrow \nabla_\theta J(\theta) = \mathbb{E}_{s\sim d^{\pi},\, a\sim\pi}[A^{\pi}(s,a)\nabla_\theta \log \pi_\theta(a|s)]$

10. discounted setting:

$$J^\pi = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1}\right]$$

$$V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \,\middle|\, s_0 = s\right] \qquad V^\pi(s) = \sum_a \pi(a|s)\, Q^\pi(s,a)$$

$$Q^\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \,\middle|\, s_0 = s,\ a_0 = a\right] \qquad Q^\pi(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) V^\pi(s')$$

$$A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$$

Average cost:

$$p^\pi = J^\pi = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} r_t\right]$$

$$g^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} (r_{t+1} - p^\pi) \,\middle|\, s_0 = s\right]$$

$$h^\pi(s) = g^\pi(s)$$

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

$$Q^\pi(s,a) = r(s,a) - p^\pi + \sum_{s'} p(s'|s,a)\, g^\pi(s')$$

$$g^\pi(s) = \sum_a \pi(a|s)\, Q^\pi(s,a)$$

$$A^\pi(s,a) = Q^\pi(s,a) - g^\pi(s)$$

Tutorial-3:

1.(a) $M = \langle \mathcal{S}, \mathcal{A}, P, R \rangle$

$\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2 \times \cdots \times \mathcal{S}_n$

$\mathcal{S}_i = \{A, B\}$

$\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \mathcal{A}_n$

$\mathcal{A}_i = \{\text{Jump}, \text{Not Jump}\}$

$= \{0, 1\} \quad a = (a_1 \ldots a_n) \in \mathcal{A} = \{0,1\}^N$

$P(s'|s,a)$

$R_i(s,a)$ each agent

$$\bar{R}(s,a) = \frac{1}{N} \sum_{i=1}^{N} R_i(s,a)$$

policy: $\pi(a|s)$

$$J(\pi) = \lim_{T \to \infty} \frac{1}{T} \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} \bar{r}_{t+1} \right]$$

$$\bar{r}_{t+1} = \frac{1}{N} \sum_{i=1}^{N} r_{t+1}^i$$

$$J(\pi) = \sum_{s \in \mathcal{S}} d_\pi(s) \sum_{a \in \mathcal{A}} \pi(a|s) \bar{R}(s,a)$$

$$Q^\pi(s,a) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} (r_{t+1} - J(\pi)) \,\middle|\, s_0 = s, a_0 = a \right]$$

$$V^\pi(s) = \sum_a \pi(a|s) Q^\pi(s,a)$$

$$\pi^* = \arg\max_\pi J(\pi)$$

(b) $M = (\mathcal{S}, \{\mathcal{A}_i\}_{i=1}^N, P, \{R_i\}_{i=1}^N, \{G_t\}_{t \geq 0})$

$\mathcal{S} = \{A, B\}^N$

$\mathcal{A}_i = \{0, 1\}$

$a = (a_1 \ldots a_n) \in \mathcal{A} = \prod_{i=1}^{N} \mathcal{A}_i$

$P(s'|s,a)$

$r_{t+1}^i \sim R_i(s_t, a_t)$ agent i do not observe

$G_t = (N, E_t)$ Nodes = agents

Edge $(i,j) \in E_t \rightarrow$ agent communicate at t

↓ time varying graph (communication network)

each agent $\pi^i_{\theta_i}(a^i|s)$

$$\pi_\theta(a|s) = \prod_{i=1}^{N} \pi^i_{\theta_i}(a^i|s)$$

$$J(\theta) = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T-1} \frac{1}{N} \sum_{i=1}^{N} r^i_{t+1}\right]$$

objective

$$= \sum_s d_\theta(s) \sum_a \pi_\theta(s,a) \bar{R}(s,a)$$

$$\bar{R}(s,a) = \frac{1}{N} \sum_{i=1}^{N} R^i(s,a)$$

each agent i has also its own observation

global state s_t

own reward r^i_{t+1}

neighbors message via C_t

$$\max_{\theta=(\theta_1 \dots \theta_N)} J(\theta)$$

$$Q_\theta(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} (r_{t+1} - J(\theta)) \,\middle|\, s_0 = s, a_0 = a\right]$$

$$V_\theta(s) = \sum_a \pi_\theta(s,a)\, Q_\theta(s,a)$$

$Q(\cdot, w)$ or $V(\cdot, w)$ → each agent maintains estimates

$$w^i \leftarrow \sum_j C_t(i,j) w^j$$

↑ spread information across networks

(c) communication weights $C_t = [C_t(i,j)] \in \mathbb{R}^{N\times N}$

$C_t e = e$ ← Row stochasticity

i.e $\sum_{j=1}^{N} C_t(i,j) = 1 \quad \forall\, i$

Connectivity / mixing

$$\rho\left(\mathbb{E}[C_t^T (I - \frac{1}{N} e e^T) C_t]\right) < 1$$

i.e $x - \bar{x} e \to 0$ all agents estimates converge to common value

graph connectivity $C_t(i,j) > 0 \Rightarrow (i,j) \in E_t$

all agents converge in estimate $w_1 \approx w_2 \cdots \approx w_N$

(d) policy: $\pi_{\theta_p^i}(a_p^i|s) > 0$ (actor)

$\pi_\theta(s,a) = \prod_{i=1}^{N} \pi_{\theta_i^i}(a_i^i|s)$

diff wrt θ_i^i

critic (value fution)

$Q(s,a|\omega) = \omega^T \phi(s,a)$ } ①

$\phi(s,a) \in \mathbb{R}^K$

or $V(s|v) = v^T \varphi(s)$ $\quad R(s,a|\lambda) = \lambda^T f(s,a)$ } ②

now objetive:

$J(\theta) = \sum_s d_\theta(s) \sum_a \pi_\theta(s,a) \bar{R}(s,a)$

$\max_\theta J(\theta)$

$\nabla_{\theta_p} J(\theta) = \mathbb{E}_{s,a}[\nabla_{\theta_i^i} \log \pi_{\theta_p^i}(s,a_i^i) A_\theta(s,a)]$

(∵ policy gradient)

algorithm: Actor-critic

TD error: $\delta_t^i = r_{t+1}^i - \mu_t^i + Q(s_{t+1}, a_{t+1}|\omega_t^i) - Q(s_t,a_t|\omega_t^i)$

update: $\omega_t^i \leftarrow \omega_t^i + \beta_{\omega,t} \delta_t^i \nabla_\omega Q(s_t,a_t|\omega_t^i)$

consensus: $\omega_t^i \leftarrow \sum_{j=1}^{N} c_t(i,j) \omega_t^j$

avg reward: $\mu_{t+1}^i = (1-\beta_{\omega,t}) \mu_t^i + \beta_{\omega,t} r_{t+1}^i$

actor: slow time scale

$A_t^i = Q(s_t,a_t|\omega_t^i) - \sum_{a_i^i} \pi_{\theta_p^i}(s_t,a_t^i) Q(s_t,(a_t^i, a_t^{-i}), \omega_t^i)$

score fution: $\psi_t^i = \nabla_{\theta_p} \log \pi_{\theta_p^i}(s_t,a_t^i)$

policy update: $\theta_{t+1}^i = \theta_t^i + \beta_{\theta,t} A_t^i \psi_t^i$

assumption: MDP ergodicity, Policy assumptions, communication matrix

↑ irreducible & aperiodic

↑ $\pi_{\theta_i^i}(s,a_i^i) > 0$

$\sum_t \beta_{\omega,t} = \infty$, $\sum_t \beta_{\theta,t} = \infty$, $\sum_t (\beta_{\omega,t}^2 + \beta_{\theta,t}^2) < \infty$

$\beta_{\theta,t} = o(\beta_{\omega,t})$, $|r_t^i| \le C$ bounded rewards

convergence : $\omega_t^i \to \omega_\theta$

$\omega^1 = \cdots = \omega^N$

$\nabla_{\theta_i^i} J(\theta) \approx 0$

(e) $\theta = (\theta_1 \dots \theta_N) \in \Theta = \prod_{i=1}^{N} \theta_i^i$

$$\pi_\theta(s,a) = \prod_{i=1}^{N} \pi_{\theta_i^i}^i(s,a_i)$$

$$J(\theta) = \sum_{s \in S} d_\theta(s) \sum_{a \in A} \pi_\theta(s,a) \bar{R}(s,a)$$

$$= \sum_{s \in S} d_\theta(s) \sum_{a \in A} \left(\prod_{i=1}^{N} \pi_{\theta_i}^i(s,a_i) \right) \left(\frac{1}{N} \sum_{i=1}^{N} R_i^i(s,a) \right)$$

$$J(\theta) = \mathbb{E}_{s \sim d_\theta, a \sim \pi_\theta} \left[\frac{1}{N} \sum_{i=1}^{N} R_i^i(s,a) \right]$$

(f) $Q_\theta(s,a)$, $V_\theta(s)$

$$Q_\theta(s,a) = \mathbb{E}_\theta \left[\sum_{t=0}^{\infty} (\bar{r}_{t+1} - J(\theta)) \,\middle|\, s_0 = s, a_0 = a \right]$$

$$V_\theta(s) = \sum_{a \in A} \pi_\theta(s,a) Q_\theta(s,a)$$

(g) $A_\theta(s,a) = Q_\theta(s,a) - V_\theta(s)$

$a = (a_i^i, a^{-i})$

$$V_\theta^i(s, a^{-i}) = \sum_{a_i^i \in A_i^i} \pi_{\theta_i^i}^i(s, a_i^i) Q_\theta(s, (a_i^i, a^{-i}))$$

$$A_\theta^i(s,a) = Q_\theta(s,a) - V_\theta^i(s, a^{-i})$$

$$\mathbb{E}_{a_i^i \sim \pi_i^i} [A_\theta^i(s,a)] = 0$$

(h) $\nabla_{\theta_i^i} J(\theta) = \mathbb{E}_{s \sim d_\theta, a \sim \pi_\theta} [\nabla_{\theta_i^i} \log \pi_{\theta_i^i}^i(s,a_i) \cdot \hat{A}_\theta(s,a)]$

↓ know — agent i has unbiased estimate of Q_θ or A_θ

$$J(\theta) = \sum_s d_\theta(s) \sum_a \pi_\theta(s,a) \bar{R}(s,a)$$

$\nabla_\theta J(\theta) = \mathbb{E}_{s,a} [\nabla_\theta \log \pi_\theta(s,a) Q_\theta(s,a)]$ from policy gradient

$$\pi_\theta(s,a) = \prod_{j=1}^{N} \pi_{\theta_j}^j(s,a_j)$$

$$\log \pi_\theta(s,a) = \sum_{j=1}^{N} \log \pi_{\theta_j}^j(s,a_j)$$

$$\Rightarrow \nabla_{\theta_j} \log \pi_\theta(s,a) = \nabla_{\theta_j} \log \pi^j_{\theta_j}(s,a_j)$$

$$\Rightarrow \nabla_{\theta_i} J(\theta) = \mathbb{E}_{s,a}\left[\nabla_{\theta_i} \log \pi^i_{\theta_i}(s,a_i)\, Q_\theta(s,a)\right]$$

for any $b(s,a^{-i})$ independent of a_i

$$\mathbb{E}_{a_i \sim \pi^i}\left[\nabla_\theta \log \pi^i(s,a_i)\, b(s,a^{-i})\right] = 0$$

as $\sum_{a_i} \pi^i(s,a_i) \nabla_{\theta_i} \log \pi^i(s,a_i) = \nabla_{\theta_i} \sum_{a_i} \pi^i(s,a_i)$

$$= \nabla_{\theta_i}(1)$$

$$= 0$$

let $b(s,a^{-i}) = V^i_\theta(s,a^{-i})$

$$\Rightarrow \nabla_{\theta_i} J(\theta) = \mathbb{E}\left[\nabla_{\theta_i} \log \pi^i_{\theta_i}(s,a_i)\left(Q_\theta(s,a) - V^i_\theta(s,a^{-i})\right)\right]$$

$$\Rightarrow \nabla_{\theta_i} J(\theta) = \mathbb{E}\left[\nabla_{\theta_i} \log \pi^i_{\theta_i}(s,a_i) A^i_\theta(s,a)\right]$$

then estimate $\hat{\theta}$, $\hat{A}$

$$\mathbb{E}[\hat{A}(s,a)] = A_\theta(s,a)$$

$$\Rightarrow \mathbb{E}\left[\nabla_{\theta_i} \log \pi^i \cdot \hat{A}\right] = \nabla_{\theta_i} J(\theta)$$

(i) $\min_\lambda \sum_{s,a} d_\theta(s,a)\left(R(s,a|\lambda) - R(s,a)\right)^2$

$\bar{d}_\theta(s,a) = d_\theta(s)\pi_\theta(s,a)$

save stationary point as

$$\min \sum_{i=1}^{N} \sum_{s,a} \bar{d}_\theta(s,a)\left(R(s,a|\lambda) - R_i(s,a)\right)^2$$

as $\bar{R} = \frac{1}{N}\sum_i R_i$

$$\sum_i (x - R_i)^2 = N(x - \bar{R})^2 + \text{constant}$$

(ii) $R(s,a|\lambda) = \lambda^T f(s,a)$

$\lambda_i \rightarrow$ local

$$\nabla_\lambda = \left(r^i_{t+1} - R(s_t,a_t|\lambda_i)\right) f(s_t,a_t)$$

$$\lambda^{t+\frac{1}{2}}_i = \lambda_i^t + \beta_t\left(r^i_{t+1} - R(s_t,a_t|\lambda^i_t)\right)\nabla_\lambda R(s_t,a_t|\lambda^t_i)$$

$$\lambda_i^{t} = \sum_{j=1}^{N} c_t(i,j)\, \lambda_j^{t+\frac{1}{2}}$$

actor critic:

$$\delta_t^i = R(s_t, a_t | \lambda_i^t) - \mu_i^t + V(s_{t+1} | v_i^t) - V(s_t | v_i^t)$$

actor: $\theta_i^{t+1} = \theta_i^t + \beta_{\theta,t}\, \delta_t^i \nabla_\theta \log \pi_{\theta_i}^i (s_t, a_t^i)$

$$v_i \leftarrow v_i + \beta_t \delta_t^i \nabla_v V$$

2. (a) in SSP per period cost of each agent capture local contribution of agent to global objective
↓ $c_t(s,a)$

(b) local learning + consensus + fictional approximation

(c) in MARL env is non-stationary from agent's prespective