P01

Lecture 2 : Tabular RL

Last time: Value and Policy Iteration

Value Iteration

$Q_0 = 0$.

For $t = 0, 1, \ldots$

$$Q_{t+1} = TQ_t$$

Policy Iteration

Initial policy π_0

For $t = 0, 1, \ldots$

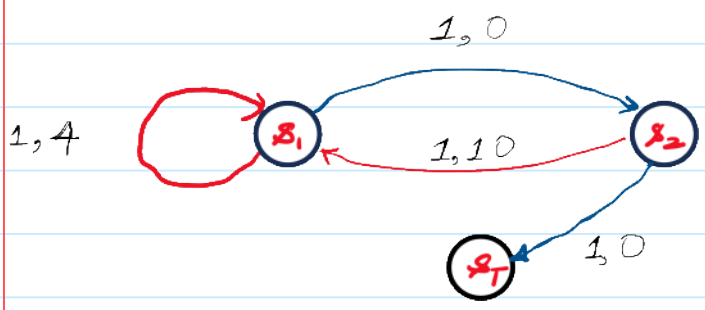
Policy Evaluation

Find Q_{π_t}

Policy Update

Set $\pi_{t+1} = \text{greedy}(Q_{\pi_t})$

Illustrative Example



Let $Q_0 = [0, 0, 0, 0, 0]^T$.
& $r = 0$.g.

Then, $Q_1 = TQ_0$

$\Rightarrow Q_1(s_1, L)$

$= R(s_1, L) + \gamma \max_{a' \in \{L,R\}} Q_0(s_1, a')$

$= 4.$

Similarly,

$Q_1(s_1, R) = 0$

$+ \gamma \max_{a' \in \{L,R\}} Q(s_2, a')$

$= 0.$

Continuing this way, we get

$Q_1(s_2, L) = 10$

$Q_1(s_2, R) = 0$

& $Q_2(s_T, a) = 0 \quad \forall a \in \{L, R\}$

Observe that the greedy policy at s_1 & s_2 suggest picking the action L at both places

Let us compute Q_2 now

$Q_2(s_1, L) = 4 +$

$\gamma \max\{Q_1(s_1, L), Q_1(s_1, R)\}$

$= 4 + 0.9 \times 4$

$= 7.6$

$Q_2(s_1, R) = 0 +$

$\gamma \max\{Q_1(s_2, L), Q_1(s_2, R)\}$

$= 0.9 \times 10 = 9.$

$Q_2(s_2, L)$

$= 10 + \gamma \max\{Q_1(s_1, L), Q_1(s_1, R)\}$

$= 10 + 0.9 \max\{4, 0\}$

$= 13.6$

$Q_2(s_2, R) = 0$.

The behaviour of the Policy Iteration method can similarly be studied for the above example.

We now look at model-free variants of the above algorithm.

These algorithms do not need knowledge of the transition probabilities, instead they learn using sampled trajectories.

To begin with, we first look at the problem of policy evaluation.

Here, given a policy π, the goal is to find

$$Q_\pi.$$

Recall that, for any s, a,

$$Q_\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t r(s_t, a_t) \,\middle|\, \begin{matrix} s_0 = s, \\ a_0 = a \end{matrix}\right]$$

and the expectation is with respect to

$$s_{t+1} \sim p(\cdot \mid s_t, a_t)$$

and

$$a_{t+1} \sim \pi(\cdot \mid s_{t+1})$$

for all $t \geq 0$.

Temporal-Difference Learning for Policy Evaluation.

We presume $\pi(a|s) > 0 \quad \forall s, a$.

Initialize Q_0 and initial state-action pair s_0, a_0

For $t \geq 0$,

Observe $r(s_t, a_t)$

Observe $s_{t+1} \sim p(\cdot|s_t, a_t)$

sample $a_{t+1} \sim \pi(\cdot|s_{t+1})$

Update $Q_{t+1}(s, a)$

$$= \begin{cases} Q_t(s, a), & \text{if } (s, a) \neq (s_t, a_t) \\ Q_t(s_t, a_t), & \text{otherwise} \end{cases}$$

$$+ \alpha_t \Big[r(s_t, a_t) + \gamma Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t) \Big]$$

where (α_t) is some step-size sequence.

Remark: This algorithm is known as TD(0) and it updates Q_n as soon as one transition occurs.

Intuition behind algorithm design

Let

$$f(Q) = \frac{1}{2} \|Q - Q_\psi\|_{D_\psi}^2$$

where D_ψ is a $SA \times SA$ diagonal matrix whose (s,a)-th diagonal entry is

$$d_\psi(s)\, \psi(a|s)$$

and d_ψ is the stationary distribution of the Markov chain induced by ψ.

Thus,

$$f(Q) = \frac{1}{2} \sum_{s,a} d_\psi(s)\, \psi(a|s) \times \left(Q(s,a) - Q_\psi(s,a) \right)^2$$

$$= \frac{1}{2} (Q - Q_\psi)^T D_\psi (Q - Q_\psi)$$

Suppose $\psi(a|s) > 0 \ \forall s,a$ & the Markov chain induced by ψ is ergodic so that a unique stationary distribution d_ψ exists & $d_\psi(s) > 0 \ \forall s$.

Then, D_ψ is positive definite &

$$\nabla f(Q) = D_\psi (Q - Q_\psi)$$

Hence, if we would have known D_ψ & Q_ψ,

then one could have used the update rule

$$Q_{n+1} = Q_n + \alpha_n D_\psi \left[Q_\psi - Q_n \right]$$

which is the gradient descent method for the optimization problem

$$\min_{\theta} f(\theta).$$

Since θ_y is unknown, we now use its properties to modify the update rule.

Recall

$$T_y x = r + \gamma P_y x,$$

where P_y is $SA \times SA$ of

$$P_y(s', a' | s, a)$$

$$= P(s' | s, a)\, \pi(a' | s').$$

Moreover,

$$T_y \theta_y = \theta_y.$$

Then, $T_y \theta_y = r + \gamma P_y \theta_y$

$$= \theta_y.$$

In particular,

$$D_\psi Q_\psi$$

$$= \sum_{s,a} e_{s,a}\, d_\psi(s)\, \mu(a/s)\, Q_\psi(s,a)$$

$$= \sum_{s,a,s',a'} e_{s,a})\, d_\psi(s)\, \mu(a/s)$$

$$p(s'/s,a)\, \mu(a'/s')$$

$$\left[r(s,a) + \gamma\, Q_\psi(s',a') \right]$$

$$\propto \sum_{s,a,s',a'} e_{s,a}\, d_\psi(s)\, \mu(a/s)$$

$$p(s'/s,a)\, \mu(a'/s')$$

$$\left[r(s,a) + \gamma\, Q_\theta(s',a') \right]$$

$$= \mathbb{E}\left[e_{s,a} \left(r(s,a) + \gamma\, Q_\theta(s',a') \right) \,\middle|\, \mathcal{F}_n \right],$$

where $\mathcal{F}_n = \sigma(Q_0, s_0, a_0, s_1, a_1, \ldots, s_{n-1}, a_{n-1}, s_n, a_n)$

& the expectation is presuming

$$s \sim d_\gamma(\cdot)$$

$$a \sim \mu(\cdot|s)$$

$$s' \sim p(\cdot|s,a)$$

$$a' \sim \mu(\cdot|s').$$

Hence, the SGD can be approximated by

$$Q_{n+1} = Q_n + \alpha_n \left[r(s_n, a_n) + \gamma Q_n(s_{n+1}, a_{n+1}) - Q_n(s_n, a_n) \right] e_{s_n, a_n}.$$

which is exactly the TD(0) method.

Observe that

$$Q_{n+1} = Q_n + \alpha_n D_y \left[T_y Q_n - Q_n \right] + \alpha_n M_{n+1},$$

where

$$M_{n+1} = \delta_n e_{s_n, a_n} - D_y \left[T_y Q_n - Q_n \right]$$

and

$$\delta_n = r(s_n, a_n) + \gamma Q_n(s_{n+1}, a_{n+1}) - Q_n(s_n, a_n).$$

Using stochastic approximation theory, it can be shown that the (Q_n) iterates track the solution trajectories of the ODE

$$\dot{Q}(t) = D_y \left[T_y Q_n - Q_n \right]$$

Additionally, by using the fact that T_y is a contraction,

it can be shown that

Q_y is a globally asymptotically stable equilibrium for this limiting ODE.

Using this, it can be concluded that

$$Q_t \xrightarrow{a.s.} Q_y$$

as $t \to \infty$

Optimal Control:

Using a similar recipe, one can come up with the following algorithm to find Q^*:

$$Q_{t+1} = Q_t + \alpha_t \left[r(s_t, a_t) + \max_{a'} Q_t(s_t', a') - Q_t(s_t, a_t) \right] e_{s_t, a_t}$$

At time t, a_t can be chosen in multiple ways:

① fixed behaviour policy

$$a_t \sim \mu(\cdot | s_t)$$

where μ is some fixed behaviour policy

② ϵ-greedy behaviour policy:

$$a_t = \begin{cases} \text{arbitrary action} & \text{w.p. } \epsilon \\ \underset{a'}{\text{argmax}}\, Q_t(s_t, a') & \text{w.p. } (1-\epsilon) \end{cases}$$

In either case,

$$Q_t \xrightarrow{a.s.} Q^*$$

as $t \to \infty$.