10. Reinforcement learning

Based on dynamic programming.

Transition function: $x' = \delta(x, a)$

Reward function: $r' = \varrho(x, a)$

we are searching for an optimal action sequence $a_0, a_1, a_2, \ldots = \{a_k\}_{k=0}^{\infty}$ and and control policy $a_k = \mu(x_k)$ leading to an optimal infinite-horizon discounted value function

$$J^*(x_0) = \max_{\{a_k\}_{k=0}^{\infty}} \sum_{k=0}^{\infty} \gamma^k \varrho(x_k, a_k)$$

where $\gamma < 1$ is a discounting factor which determines how far ahead rewards should influence a control policy.

$\gamma < 1 \Rightarrow J^*$ is finite

$$J^*(x_0) = \max_{a_0}\Big[\varrho(x_0, a_0) +$$
$$\gamma \max_{\{a_{k+1}\}_{k=0}^{\infty}} \sum_{k=0}^{\infty} \gamma^k \varrho(x_{k+1}, a_{k+1})\Big]$$

$$= \max_{a_0} \left[\rho(x_0, a_0) + \gamma J^*(x_1) \right] =$$

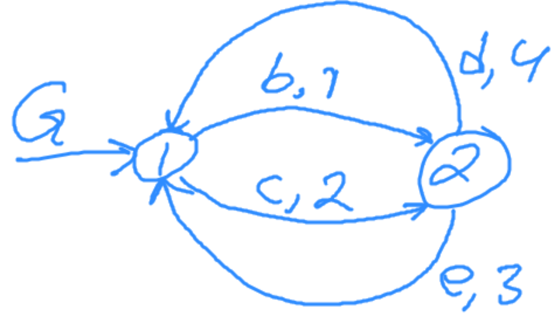
$$= \max_{a_0} \left[\rho(x_0, a_0) + \gamma J^*(\delta(x_0, a_0)) \right]$$

This expression can be applied to any state x $\Rightarrow$

$$\boxed{J^*(x) = \max_{a \in \Sigma(x)} \left[\rho(x,a) + \gamma J^*(\delta(x,a)) \right]}$$

This famous equation is called Bellman's equation which is the core of dynamic programming.

Ex



$$J^*(1) = \max\{\rho(1,b), \rho(1,c)\} + \gamma J^*(2)$$

$$= \max\{1, 2\} + \gamma J^*(2) =$$

$$= 2 + \gamma J^*(2)$$

$$J^*(2) = \max\{\underbrace{\rho(2,d)}_{4}, \underbrace{\rho(2,e)}_{3}\} + \gamma J^*(1) = 4 + \gamma J^*(1)$$

Fixed point

$$J^*(1) = 2 + \gamma J^*(2) = 2 + \gamma(4 + \gamma J^*(1))$$

$$\Rightarrow J^*(1) = \frac{2+4\gamma}{1-\gamma^2} \qquad J^*(2) = \frac{4+2\gamma}{1-\gamma^2}$$

Q-function

$$Q(x,a) = g(x,a) + \gamma J^*(\delta(x,a))$$

state action pair: (x,a)

$$J^*(x) = \max_{a \in \Sigma(x)} Q(x,a)$$

Bellman's equation is now expressed in terms of the Q-function.

$$J^*(\underbrace{\delta(x,a)}_{x'}) = \max_{b \in \Sigma(\delta(x,a))} Q(\delta(x,a), b)$$

$$\boxed{Q(x,a) = g(x,a) + \gamma \max_{b \in \Sigma(\delta(x,a))} Q(\delta(x,a), b)}$$

For all possible actions we are looking for the optimal one, which determines the optimal control action

$$\mu(x) = \arg \max_{a \in \Sigma(x)} Q(x,a) = a^*$$

Ex

$$Q(x,a) = g(x,a) + \gamma J^*(\delta(x,a))$$

$$Q(1,b) = 1 + \gamma J^*(2)$$

$$Q(1,c) = 2 + \gamma J^*(2)$$

$$Q(2,d) = 4 + \gamma J^*(1)$$

$$Q(2,e) = 3 + \gamma J^*(1)$$

Optimal policy

$$\mu(1) = \max_{a \in \{b,c\}} Q(1,a) = c$$

$$\mu(2) = \max_{a \in \{d,e\}} Q(2,a) = d$$

Model-free Q-function iteration

Replace $\delta(x,a)$ with x'
$\rho(x,a)$ with r'

Send an action to the plant an wait for the next x' and the transition reward r'

The Q-function is then updated without any model but instead by feedback from the plant.

$$Q(x,a) = r' + \gamma \max_{b \in \Sigma(x')} Q(x',b)$$

Problem: We don't know the Q-function.

Q-learning

Based on (x',r') an estimate $\hat{Q}_k(x,a)$ is updated

$$\boxed{\hat{Q}_{k+1}(x,a) = (1-\alpha_k)\hat{Q}_k(x,a) + \alpha_k\left(r' + \gamma \max_{b\in\Sigma(x')} \hat{Q}_k(x',b)\right)}$$

α_k = learning factor that is reduced when time k is increasing.

In most cases the action is selected to maximize $\hat{Q}_k(x,a)$. But to explore the whole state space it is also necessary to take actions which do not maximize $\hat{Q}$.

An alternative arbitrary action $a \in \Sigma(x)$ is then taken with equal (uniform) but reduced probability

$$p_k(x) = \frac{1}{k|\Sigma(x)|}$$

as time k increases.

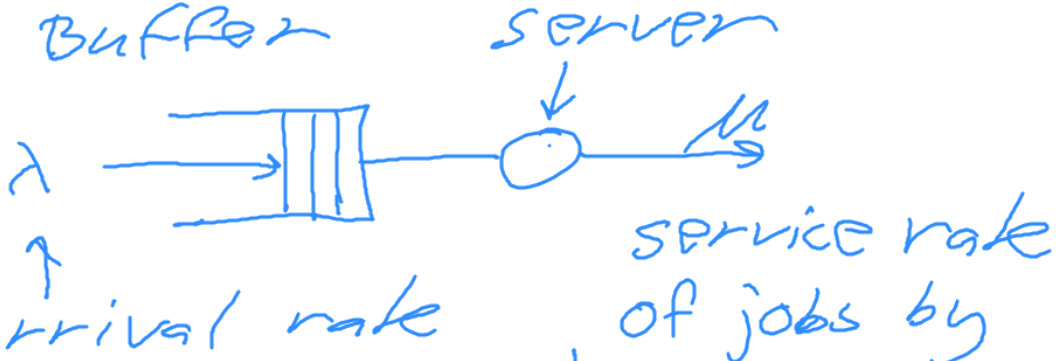
This done for all $a \in \Sigma(x)$ and for all states x. The optimal (greedy) solution is then taken with prob. $1 - p_k(x) \to 1$ $k \to \infty$ (*)

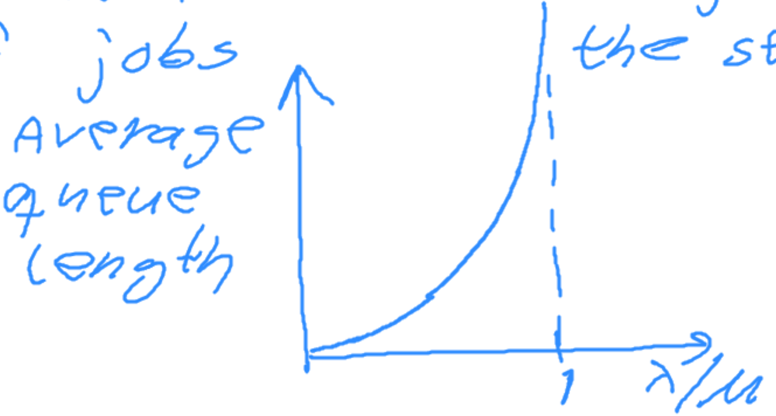
(*) This strategy is called ε-greedy.

8. Markov processes (last part of ch8)

Probabilistic models

Motivating example





8. Cont. Probabilistic models

Markov chains

stochastic process: $\{x(t) : t \in T\}$

$x(t)$ = random variable for each $t \in T$

T = countable set of time instances
$= \{t_0, t_1, t_2, \dots\}$

For this discrete-time set T we get a discrete-time stochastic process.

If instead $t \in \mathbb{R}^+ \Rightarrow \{x(t)\}$ is a continuous stochastic process.

Markov chains

State space of $\{x(t)\}$ is a discrete set Q

$x(t)$ takes values $q_i \in Q$

State probability:

$p_i(t_k) = P(x(t_k) = q_i)$

This stochastic process is called a Markov chain if the next state conditional probability only depends on the current state

$$P(x(t_{k+1})=q(t_{k+1}) \mid x(t_k)=q(t_k) \wedge x(t_{k-1})=q(t_{k-1}) \wedge \ldots \wedge x(t_0)=q(t_0)) =$$

$$= P(x(t_{k+1})=q(t_{k+1}) \mid x(t_k)=q(t_k))$$

Transition probability:

$$p_{ij} = P(x(t_{k+1})=q_j \mid x(t_k)=q_i)$$

$$P\{B \mid A\} = \frac{P[A \cap B]}{P[A]} = \left[\begin{array}{l}\text{probability of} \\ B \text{ when } A \text{ has} \\ \text{already occured}\end{array}\right]$$

$$A_i = x(t_k)=q_i \qquad B = x(t_{k+1})=q_j$$

Total probability

$$P\{B\} = \sum_{i=1}^{n} P\{A_i \cap B\} = \sum_{i=1}^{n} P[B \mid A_i] P[A_i]$$

Total probability = state probability

$$p_j(t_{k+1}) = P\{\underbrace{x(t_{k+1})=q_j}_{B}\} =$$

$$= \sum_{i=1}^{n} P[\underbrace{x(t_{k+1})=q_j}_{B} \mid \underbrace{x(t_k)=q_i}_{A_i}] P[\underbrace{x(t_k)=q_i}_{A_i}]$$

$$= \sum_{i=1}^{n} p_{ij} \cdot p_i(t_k) =$$

$$= \underbrace{[p_1(t_k) \ \ldots \ p_n(t_k)]}_{p(t_k)} \begin{bmatrix} p_{1j} \\ \vdots \\ p_{nj} \end{bmatrix}$$

For $j = 1, \ldots, n$:

$$\underbrace{[p_1(t_{k+1}) \ldots p_n(t_{k+1})]}_{p(t_{k+1})} = p(t_k) \underbrace{\begin{bmatrix} p_{11} & p_{12} & \ldots & p_{1n} \\ \vdots & & & \vdots \\ p_{n1} & p_{n2} & \ldots & p_{nn} \end{bmatrix}}_{P}$$

$$\boxed{p(t_{k+1}) = p(t_k)\,\mathbb{P}}$$

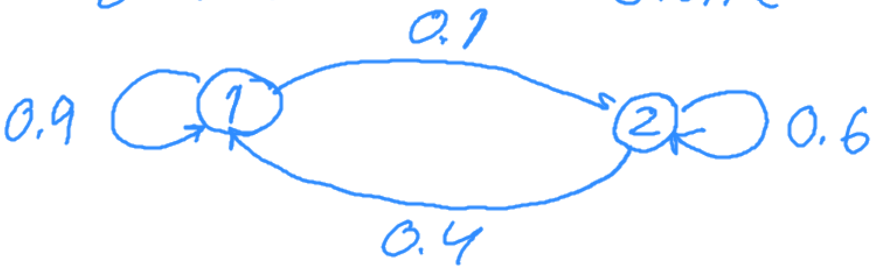
Important property:

$$\sum_{j=1}^{n} p_{ij} = \sum_{j=1}^{n} P\{x(t_{k+1}) = q_j \mid x(t_k) = q_i\} = 1$$

Ex Machine with failure

working state

failure state



$$\mathbb{P} = \begin{bmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{bmatrix}$$

$p(0) = [1 \quad 0]$ (initial state = working state)

$$p(t_1) = p(0)\,\mathbb{P} = [1 \quad 0]\begin{bmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{bmatrix} =$$

$$= [0.9 \quad 0.1]$$

$$p(t_2) = [0.9 \quad 0.1]\begin{bmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{bmatrix} = [0.85 \quad 0.15]$$

stationary solution

$$\bar{p} = \lim_{k\to\infty} p(t_k) \qquad \bar{p} = \bar{p}\,\mathbb{P}$$

$$\bar{p} = [p \quad 1-p] = [p \quad 1-p]\begin{bmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{bmatrix}$$

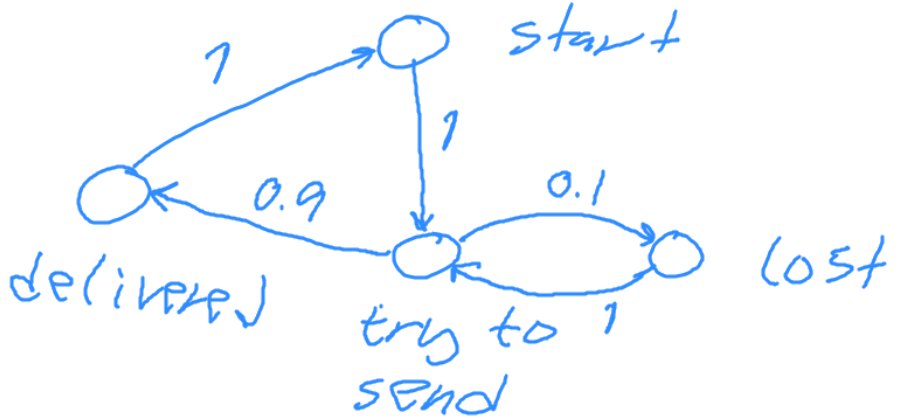
one unknown variable $p \Rightarrow$
enough to solve one equation

$$p = 0.9p + 0.4(1-p) = 0.4 + 0.5p$$

$$0.5p = 0.4 \Rightarrow p = 0.8$$

$$\bar{p} = [0.8 \quad 0.2]$$

Ex communication protocol



Markov processes

Continuous-time stochastic process with the Markov conditional probability property

$t_{k+1} - t_k = \Delta t$ where $\Delta t \to 0$

Assume a given transition probability $p_{ij} = a_{ij}\,\Delta t$

Here a_{ij} = transition rate

The row sum in $\mathbb{P}$

$$\sum_{l=1}^{n} p_{jl} = \overbrace{\underbrace{p_{j1} + \dots + p_{jj-1} + p_{jj+1} \dots + p_{jn}}_{\sum_{l\neq j}^{n} p_{jl}} + p_{jj}}^{=1}$$