Markov Decision Processes[1]

[1] Sections 2.1-2.3: Yang&Ying

Deterministic Policy Gradient Algorithms

David Silver DAVID@DEEPMIND.COM
DeepMind Technologies, London, UK
Guy Lever GUY.LEVER@UCL.AC.UK
University College London, UK
Nicolas Heess, Thomas Degris, Daan Wierstra, Martin Riedmiller *@DEEPMIND.COM
DeepMind Technologies, London, UK

2. Background

2.1. Preliminaries

We study reinforcement learning and control problems in which an agent acts in a stochastic environment by sequentially choosing actions over a sequence of time steps, in order to maximise a cumulative reward. We model the problem as a *Markov decision process* (MDP) which comprises: a *state space* $\mathcal{S}$, an *action space* $\mathcal{A}$, an *initial state distribution* with density $p_1(s_1)$, a stationary *transition dynamics distribution* with conditional density $p(s_{t+1}|s_t, a_t)$ satisfying the Markov property $p(s_{t+1}|s_1, a_1, ..., s_t, a_t) = p(s_{t+1}|s_t, a_t)$, for any trajectory $s_1, a_1, s_2, a_2, ..., s_T, a_T$ in state-action space, and a *reward function* $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. A *policy* is used to select actions in the MDP. In general the policy is stochastic and denoted by $\pi_\theta : \mathcal{S} \to \mathcal{P}(\mathcal{A})$, where $\mathcal{P}(\mathcal{A})$ is the set of probability measures on $\mathcal{A}$ and

the parameters θ of the policy in the direction of the performance gradient $\nabla_\theta J(\pi_\theta)$. The fundamental result underlying these algorithms is the *policy gradient theorem* (Sutton et al., 1999),

$$\nabla_\theta J(\pi_\theta) = \int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(a|s) Q^\pi(s, a) \mathrm{d}a\mathrm{d}s$$
$$= \mathbb{E}_{s\sim\rho^\pi, a\sim\pi_\theta} \left[\nabla_\theta \log \pi_\theta(a|s) Q^\pi(s, a)\right] \quad (2)$$

The policy gradient is surprisingly simple. In particular, despite the fact that the state distribution $\rho^\pi(s)$ depends on the policy parameters, the policy gradient does not depend on the gradient of the state distribution.

This theorem has important practical value, because it reduces the computation of the performance gradient to a simple expectation. The policy gradient theorem has been used to derive a variety of policy gradient algorithms (De-

Markov chains

X_m, X_k are dependent $m > k$

X_m and X_k are independent given X_ℓ

$m > \ell > k$

Markov chains

- Consider a random process $X = \{X_0, X_1, X_2, \ldots\}, X_k \in S$ and assume S is a finite set.
- X is a Markov chain if

$$\underline{P(X_k} = j | X_{k-1} = i, X_{k-2} = i_{k-2}, \ldots, X_0 = i_0) \\ = P(X_k = j | X_{k-1} = i) \quad \forall k, i, j, i_{k-2}, \ldots, i_0$$

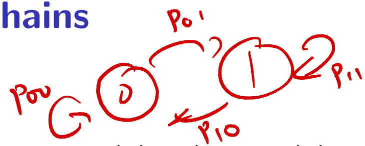
Time-Homogeneous Markov chains (MC):

- If $P(X_k = j | X_{k-1} = i)$ does not depend on k, X is called a time-homogeneous Markov chain
- Matrix $\underline{P}$ such that $P_{ij} = P(X_k = j | X_{k-1} = i)$ is called the transition probability matrix

next.

current

Markov chains



$$[p_0(k),\ p_1(k)]\begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix} = [p_0(k)p_{00} + p_1(k)p_{10},\ p_0(k)p_{01} + p_1(k)p_{11}]$$

$$p_0(k)p_{00} + p_1(k)p_{10} = p_0(k+1), \qquad p_0(k)p_{01} + p_1(k)p_{11} = p_1(k+1)$$

- Row vector $p(k) = (\cdots,\ p_i(k),\ \cdots)$
  $= (\cdots,\ P(X_k = i),\ \cdots)$

$$p(k+1) = p(k)P \implies p(k) = p(0)P^k$$

$$= p(k-1)PP = p(k-1)P^2$$

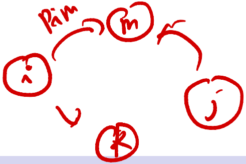
$p(k-1) = \pi$

$p(k) = p(k-1)P = \pi P = \pi$

Tow basic questions about Markov chains

- Does there exist a π so that $\pi = \pi\mathbf{P}$? If such a π exists, it is called a stationary distribution.
- If there exists a unique stationary distribution, does $\lim_{k\to\infty} p(k) = \pi$ for all $p(0)$?

Markov chains



Reachable

State j is called **reachable** from state i if there exists time T such that

$$P(X_T = j | X_0 = i) > 0$$

i.e. there exists a nonzero probability to go to state j from state i over a finite number of steps.

Irreducible

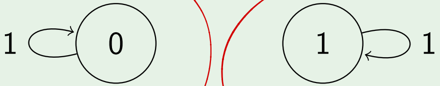
A Markov chain is **irreducible** if j is reachable from i $\quad \forall i, j$.

$$\left(\tfrac{1}{8}, \tfrac{7}{8}\right) I = \left(\tfrac{1}{8}, \tfrac{7}{8}\right)$$

Example

Consider a trivial two-state Markov chain with two states 0 and 1, and transition probability matrix

$$\mathbf{P} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$



Therefore, $\pi\mathbf{P} = \pi$ is true for any distribution π and the stationary distribution is not unique.

Markov chains

Theorem

A **finite-state, irreducible** MC has a unique stationary distribution π such that

$$\pi = \pi P.$$

- Does the distribution $p(k)$ converge to π as $k \to \infty$?

Markov chains

Theorem

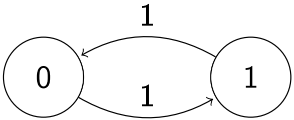
A **finite-state, irreducible** MC has a unique stationary distribution π such that

$$\pi = \pi P.$$

- Does the distribution $p(k)$ converge to π as $k \to \infty$?

Example:

$\pi = (\frac{1}{2}, \frac{1}{2})$



$p(0) = (1, 0)$

$p(1) = (0, 1)$

$p(2) = (1, 0)$

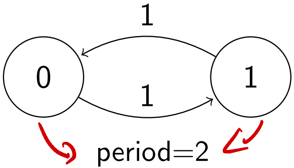
$p(3)$ ...

$\vdots$

Markov chains

- Period: state i is said to have a **period** k if the MC returns to state i in T steps only if T is a multiple of k.
- Aperiodic: a Markov chain is **aperiodic** if all states have period 1.

Examples:

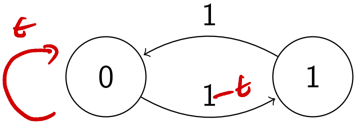


Markov chains

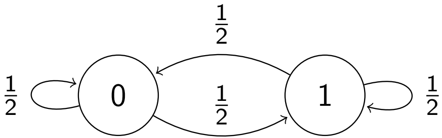
- Period: state i is said to have a **period** k if the MC returns to state i in T steps only if T is a multiple of k.
- Aperiodic: a Markov chain is **aperiodic** if all states have period 1.

Examples:

2, 3, 4, . . .

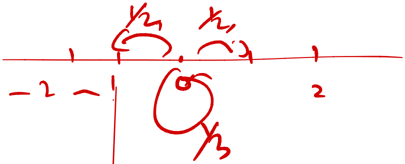


period=2



period=1

Markov chains



Theorem

A **finite-state, aperiodic, irreducible** MC has a unique stationary distribution π such that

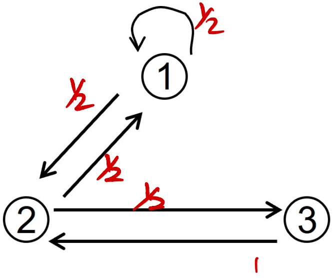
$$\pi = \pi P.$$

Furthermore,

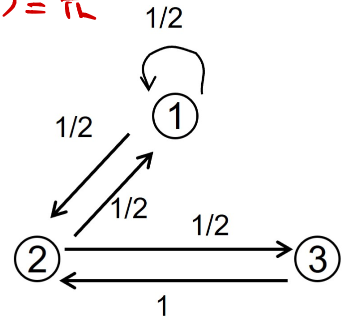
$$\lim_{k \to \infty} p(k) = \pi$$

Example: PageRank

- Imagine a random Web surfer.
- At any time t, surfer is on web i. At time $t+1$ the surfer picks a hyperlink uniformly at random and goes to the next web.
- $p_i(t)$: probability that the surfer is at web i at time t.



$r = p(\infty) = \pi$



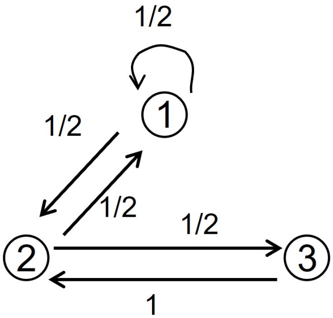
Example: PageRank

- Imagine a random Web surfer.
- At any time t, surfer is on web i. At time $t+1$ the surfer picks a hyperlink uniformly at random and goes to the next web.
- $p_i(t)$: probability that the surfer is at web i at time t.

- Transition probability matrix: $\begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \end{pmatrix}$

Example: PageRank

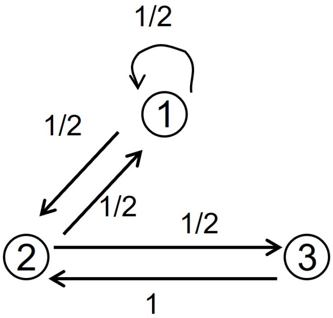
- Imagine a random Web surfer.
- At any time t, surfer is on web i. At time $t+1$ the surfer picks a hyperlink uniformly at random and goes to the next web.
- $p_i(t)$: probability that the surfer is at web i at time t.



- Transition probability matrix: $\begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \end{pmatrix}$
- Irreducible and aperiodic

Example: PageRank

- Imagine a random Web surfer.
- At any time t, surfer is on web i. At time $t+1$ the surfer picks a hyperlink uniformly at random and goes to the next web.
- $p_i(t)$: probability that the surfer is at web i at time t.



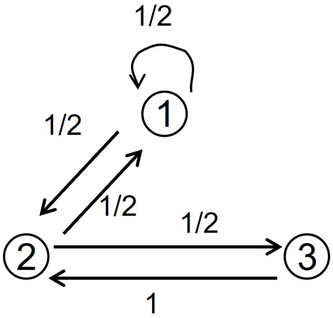
- Transition probability matrix: $\begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \end{pmatrix}$
- Irreducible and aperiodic
- Initial distribution: $p(0) = \left(\frac{1}{3} \quad \frac{1}{3} \quad \frac{1}{3}\right)$

$p(1) = p(0)\,P$

$p(2) = p(0)\,P^2$

Example: PageRank

- Imagine a random Web surfer.
- At any time t, surfer is on web i. At time $t+1$ the surfer picks a hyperlink uniformly at random and goes to the next web.
- $p_i(t)$: probability that the surfer is at web i at time t.



- Change of the distribution over time:

$$p(t+1) = p(t) \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \end{pmatrix}$$

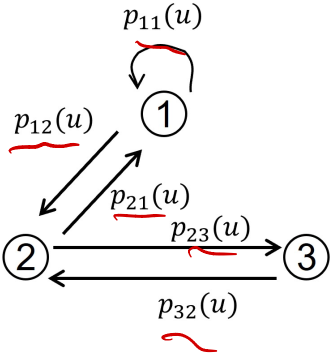
$p(\infty)$

- Stationary distribution: $\pi = \left(\frac{2}{5}, \frac{2}{5}, \frac{1}{5}\right)$.

Markov Decision Processes

Controlled Markov chain

state-transition probabilities can be controlled. In particular, under action u, transition probability from i to j is $P_{ij}(u)$.



Markov Decision Processes

- $u \in A$: for simplicity, assume A is a finite set
- $r(x, u)$: reward of taking action u at state x. $r(x, u)$ can be a random variable.

  Assume $r(x, u) \geq 0$ and takes finite number of values.
- MDP (Markov decision process) with discounted cost:

$$\lim_{T \to \infty} E\left[\sum_{k=0}^{T} \alpha^k r(x_k, u_k)\right], \quad 0 < \alpha < 1.$$

Markov Policy and Stationary Policy

- At time k, we have access to $\{x_0, \ldots, x_k\}$, $\{u_0, \ldots, u_{k-1}\}$ and $\{r(x_0, u_0), \ldots, r(x_{k-1}, u_{k-1})\}$, i.e. the past history of states, actions and rewards, and the current state.
- Define μ_k, a function (policy) which decides the action to be taken at time k.

$$\implies u_k = \mu_k(x_0, \ldots, x_k, u_0, \ldots, u_{k-1}, r_0, \ldots, r_{k-1})$$

a function of past history and current state.

μ_k can be a random function (random policy)

Markov Policy and Stationary Policy

For discounted MDP, it is sufficient to consider policies depending only on the current state x_k, i.e.

$$u_k = \mu_k(x_k) \quad \text{(Markov policy)}$$

If a policy does not explicitly depend on k, i.e. $u_k = \mu(x_k)$, then it is called a stationary policy.

Assume one of the optimal policies is stationary for the MDPs we consider.

Reference

- Chapter 3.3 of R. Srikant and Lei Ying, *Communication Networks: An Optimization, Control and Stochastic Networks Perspective*, Cambridge University Press, 2014.