Lecture 3: Markov Chains and MDP

Course: Reinforcement Learning Theory
Instructor: Lei Ying
Department of EECS
University of Michigan, Ann Arbor

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 $\theta \in \mathbb{R}^n$ is a vector of n parameters, and $\pi_\theta(a_t|s_t)$ is the conditional probability density at a_t associated with the policy. The agent uses its policy to interact with the MDP to give a trajectory of states, actions and rewards, $h_{1:T} = s_1, a_1, r_1 ..., s_T, a_T, r_T$ over $\mathcal{S} \times \mathcal{A} \times \mathbb{R}$. The return r_t^γ is the total discounted reward from time-step t 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),

$$\begin{aligned}\nabla_\theta J(\pi_\theta) &= \int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(a|s) Q^\pi(s,a) da ds \\ &= \mathbb{E}_{s\sim\rho^\pi, a\sim\pi_\theta} [\nabla_\theta \log \pi_\theta(a|s) Q^\pi(s,a)] \quad (2)\end{aligned}$$

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 (Degris et al., 2012a), by forming a sample-based estimate of this expectation. One issue that these algorithms must address is how to estimate the action-value function $Q^\pi(s,a)$. Perhaps the simplest approach is to use a sample return r_t^γ to estimate the value of $Q^\pi(s_t, a_t)$, which leads to a variant of the REINFORCE algorithm (Williams, 1992).

Markov chains

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

$$\begin{aligned} &P(X_k = j | X_{k-1} = i, X_{k-2} = i_{k-2}, \dots, X_0 = i_0) \\ &= P(X_k = j | X_{k-1} = i) \quad \forall k, i, j, i_{k-2}, \dots, i_0 \end{aligned}$$

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 P such that $P_{ij} = P(X_k = j | X_{k-1} = i)$ is called the probability transition matrix

Markov chains

More notations about Markov chains (MC):

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

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

- 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 $i \quad \forall i, j$.

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

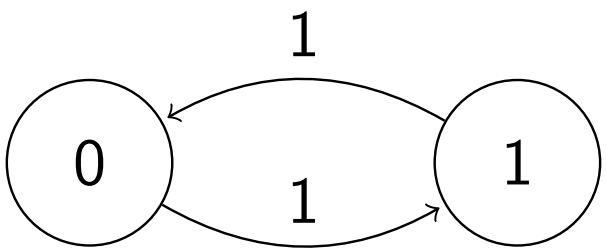
Theorem

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 $k \to \infty$?

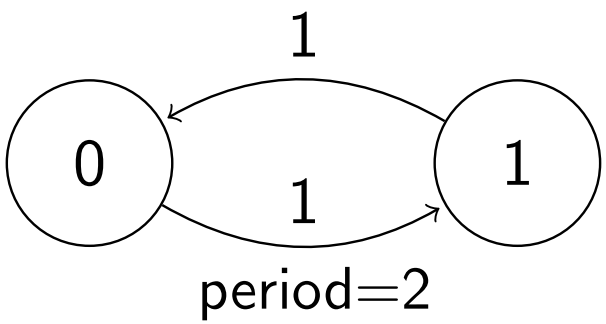
Example:



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.

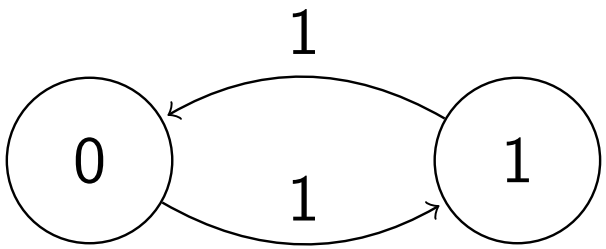
Examples:



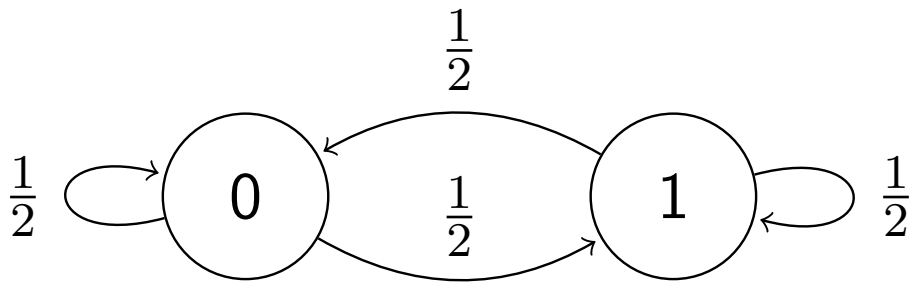
Markov chains

- 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:



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,

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

Markov chains

More definitions about Markov chains (MC):

- Controlled Markov chain: state-transition probabilities can be controlled. In particular, under action u, transition probability from i to j is $P_{ij}(u)$
- $u \in U$: for simplicity, assume U 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:

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

Action versus Policy:

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

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

a function of past history and current state.

μ_k can be a random function (random policy)

Action versus Policy:

- Claim: 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.
- Note: Assume optimal policies are 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.

Acknowledgements: I would like to thank Alex Zhao for helping prepare the slides, and Honghao Wei and Zixian Yang for correcting typos/mistakes.