Sequential decisions under uncertainty

KTH/EES PhD course

Lecture 5

Lecture 5

- Q-learning
  - A deterministic example
  - Proof of convergence

- MDP with average reward criterion
  - Finite Markov chain
  - Optimality (lecture 6)

Q-learning

Q-values

- The Q-value: the maximum expected rewards starting from a given state and selecting a given action

$$q(s,a) = r(s,a) + \lambda \sum_j p(j|s,a) v(j)$$

- Q-values vs. value function: $v(s) = \max_{a \in A_s} q(s,a)$

$$q(s,a) = r(s,a) + \lambda \sum_j p(j|s,a) \max_{b \in A_s} q(j,b)$$

- Q-value iteration:

$$q_{n+1}(s,a) = r(s,a) + \lambda \sum_j p(j|s,a) \max_{b \in A_j} q_n(j,b)$$

Q-values

- The Q-values: the unique fixed point of *F*

$$q(s,a) = r(s,a) + \lambda \sum_j p(j|s,a) \max_{b \in A_s} q(j,b)$$

$$F : \mathbb{R}^{S \times A} \to \mathbb{R}^{S \times A}$$

$$F(q)_{sa} = r(s,a) + \lambda E_{p(.|s,a)} \left[\max_{b \in A_J} q(J,b) \right]$$

Q-learning, bandit version

- Bandit: in a given state, you shave to select an action and you may observe the corresponding reward and new state

$$s \xrightarrow{a} S(s,a) \sim p(\cdot|s,a)$$

- Choose a stationary randomized policy arbitrarily such that:

$$P^\pi(Y_t = a | X_t = s) > 0, \quad \forall s, a$$

- In a given state, each action is explored an infinite number of times

Q-learning, bandit version

- Algorithm: Initialize $q_0 \in \mathbb{R}^{S\times A}, \quad s_0$

$$q_{n+1}(s_n, a_n) = q_n(s_n, a_n)$$

$$+ \alpha_n(s_n, a_n)\left[r(s_n, a_n) + \lambda \max_b q_n(S'(s_n, a_n), b) - q_n(s_n, a_n)\right]$$

$$s_{n+1} = S'(s_n, a_n)$$

- Convergence: $\forall s, a,$

$$\sum_n \alpha_n(s,a) = \infty, \quad \sum_n \alpha_n^2(s,a) < \infty$$

The algorithm approximates ODE: $\dot{q} = F(q) - q$

Q-learning
A deterministic example

Deterministic Q-learning

- Deterministic model:

Current state	Action	Next state
s	$\xrightarrow{a}$	$s'(s,a)$

- Q-value: fixed point

$$q(s,a) = r(s,a) + \lambda \max_{b \in A} q\left(s'(s,a), b\right)$$

Deterministic Q-learning

- Algorithm:

$$q_{n+1}(s_n, a_n) = q_n(s_n, a_n)$$
$$+ \alpha_n(s_n, a_n)\left[r(s_n, a_n) + \lambda \max_b q_n(s'(s_n, a_n), b) - q_n(s_n, a_n)\right]$$

Randomized stationary policy:

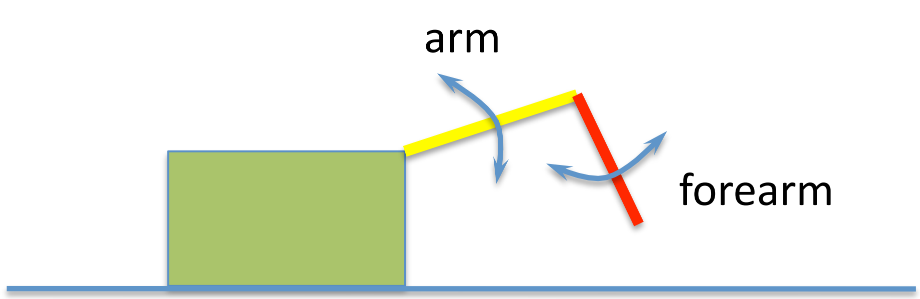
$w.p.\ 1-\epsilon, \quad a_n \in \arg\max_a q_n(s_n, a)$ (exploitation)

$w.p.\ \epsilon, \quad a_n$ random (exploration)

A robot learning to walk

- Example by **Frank Vanden Berghen**
  http://www.applied-mathematics.net/qlearning/

- A one-arm robot:



- Actions: move the arm or the forearm up or down
- States: angles of the arm and forearm
- Goal: maximize the discounted distance (going to the right)

Convergence proof’s ingredients

Doob convergence results

Theorem If $\sup_n E[\|X_n\|] < \infty,$

then $X_\infty = \lim_{n\to\infty} X_n,$ almost surely, and X_∞ is finite.

Theorem If $E[\|X_n\|^2] < \infty, \forall n,$

and if $\sum_n E[\|X_n - X_{n-1}\|^2] < \infty,$

then $X_\infty = \lim_{n\to\infty} X_n,$ almost surely.

Gronwall lemmas

Lemma (Continuous) u, v positive continuous functions

$$u(t) \leq C + K \int_0^t u(s)v(s)ds, \quad \forall t \in [0,T]$$

$$\Longrightarrow \quad u(t) \leq C \exp(K \int_0^t v(s)ds), \quad \forall t \in [0,T]$$

Lemma (Continuous) x_n, a_n positive sequences

$$x_{n+1} \leq C + L \sum_{m=0}^{n} a_m x_m$$

$$\Longrightarrow \quad x_{n+1} \leq C \exp(L \sum_{m=0}^{n} a_m)$$

Stochastic Approximation

- Algorithm: $x_{n+1} = x_n + a_n \times (h(x_n) + \xi_{n+1}), \quad \forall n.$

- Assumptions: $E[\xi_{n+1} | \mathcal{F}_n] = 0, \quad a.s., \forall n$

  h *L*-Lipschitz

  $$\sum_n a_n = \infty, \quad \sum_n a_n^2 < \infty,$$

  $$E[\|\xi_{n+1}\|^2 | \mathcal{F}_n] \leq K(1 + \|x_n\|^2), \quad a.s., \forall n$$

  $$\sup_n \|x_n\| < \infty, a.s.$$

ODE method

- Time: $t(0) = 0, \quad t(n) = \sum_{k=0}^{n-1} a_k, \forall n \geq 1$

  $$\lim_{n \to \infty} t(n) = \infty$$

- Continuous piece-wise linear interpolation: $\bar{x}(t)$

  $$\bar{x}(0) = 0$$

  $$\bar{x}(t) = x_n + (x_{n+1} - x_n) \times \frac{t - t(n)}{t(n+1) - t(n)},$$

  $$\forall t \in [t(n), t(n+1))$$

ODE method

- Approximate ODE: $x^s(s) = \bar{x}(s)$

$$\dot{x}^s(t) = h(x^s(t)), \quad \forall t \geq s$$

- The interpolated algorithm trajectory is well approximated by the ODE:

Theorem For any $T > 0$,

$$\lim_{s \to \infty} \sup_{t \in [s, s+T]} \|\bar{x}(t) - x^s(t)\| = 0, a.s.$$

ODE method

Corollary If *h* has a unique globally asymptotically stable point $x^\star$ then $\lim_{n\to\infty} x_n = x^\star$.

MDP with average reward criterion

Finite Markov chain

- Probability space: $(\Omega, \mathcal{F}, P)$
- Definition:
  - Finite state space: S
  - A sequence of r.v. $(X_n, n \in \mathbb{N})$ with values in *S* is a Markov chain iff

$$\forall n \geq 0, s \in S, \quad P(X_{n+1} = s | X_0, \ldots, X_n) = P(X_{n+1} = s | X_n)$$

- Transition matrix for homogenous Markov chain

$$P = (p(i,j))_{i,j \in S}$$
$$p(i,j) = P(X_{n+1} = j | X_n = i)$$

Kolmogorov equations

- Distribution at time *n*: row vector μ_n

$$\mu_{n+1} = P\mu_n$$

- *m* steps transitions: $\mu_{n+m} = P^m \mu_n$

$$P^m = (p^m(i,j)_{i,j\in S}$$

$$p^m(i,j) = P(X_{n+m} = j | X_n = i)$$

- Accessibility, communication:

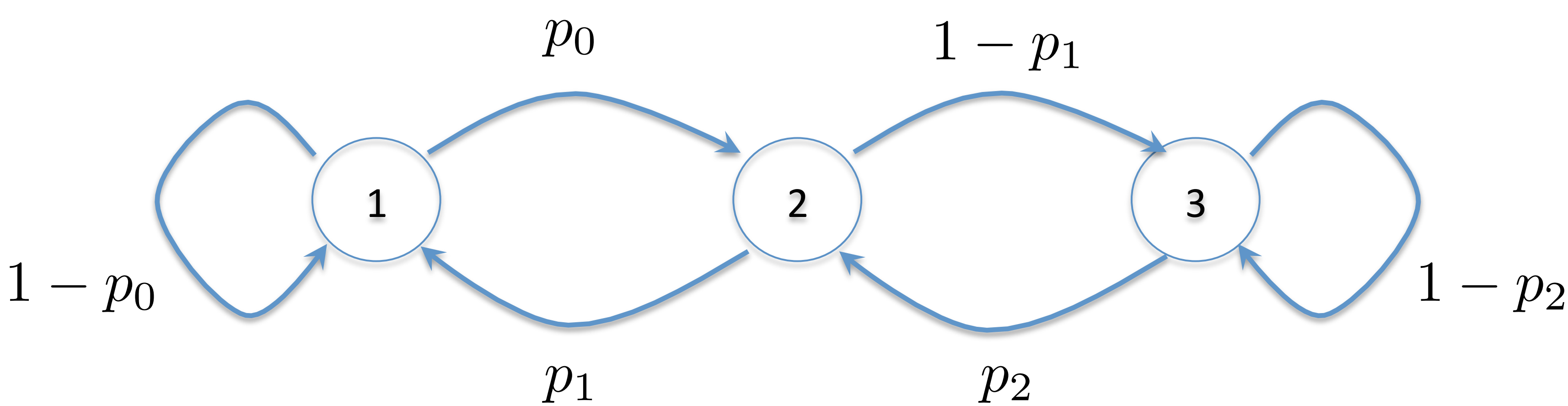
$$i \to j \iff \exists m : p^m(i,j) > 0$$

$$i \leftrightarrow j \iff (i \to j, j \to i)$$

Communication classes, Irreducibility

- By definition: each state communicates with itself
- Communication is an equivalence class
- Two communicating states are said to belong to the same communication class
- A finite Markov chain is irreducible iff there is a unique communication class

Transition graph



State classification

- Time to reach *i*: $\tau_i = \inf(n \geq 1 : X_n = i)$
- Recurrent state: $P_i(\tau_i < \infty) = 1$
- Positive recurrent state: $E_i(\tau_i) < \infty$
- Transient state: $P_i(\tau_i < \infty) < 1$
- Recurrence is a class property:

  $i \leftrightarrow j \Longrightarrow i, j$ are both recurrent or both transient
- Number of visits: $N_i = \sum_{n \geq 1} 1_{X_n = i}$

  $$P_i(\tau_i < \infty) = 1 \Longleftrightarrow P_i[N_i = \infty] = 1$$

Irreducibility and recurrence

- In an irreducible finite Markov chain, all states are positive recurrent

Periodicity

- The period of state *i* is the largest integer *d* satisfying:

$$(p^n(i,i) > 0 \Longrightarrow n \in d\mathbb{N})$$

- A state is aperiodic if its period is equal to 1
- In an irreducible Markov, all states have the same period
- An irreducible Markov chain with period *d* has a cyclic structure

$$\exists S_0, ..., S_{d-1} : \cup_l S_l = S, \;\; S_d = S_0$$

$$\forall i \in S_k, \sum_{j \in S_{k+1}} p(i,j) = 1$$

Periodicity

- An irreducible Markov chain with period *d* has a cyclic structure

$$P = \begin{pmatrix} 0 & A_0 & 0 & 0 \\ 0 & 0 & A_1 & 0 \\ 0 & 0 & 0 & A_2 \\ A_3 & 0 & 0 & 0 \end{pmatrix}$$

Stationary probability

- A distribution is stationary if: $\pi = \pi P$
- Global balance equations:

$$\forall i, \pi(i) = \sum_j \pi(j) p(j,i)$$

- A finite irreducible Markov chain has a stationary distribution

$$\forall i, \pi(i) = \frac{E_0[\sum_{n \geq 1} 1_{X_n = i} 1_{n \leq \tau_0}]}{E_0[\tau_0]}$$

$$\pi(i) = \frac{1}{E_i[\tau_i]}$$

Ergodic theorem

- For a finite irreducible Markov chain:

$$\forall f : S \to \mathbb{R}$$

$$\sum_i |f(i)|\pi(i) < \infty,$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) = \sum_{i\in S} f(i)\pi(i), \quad a.s.$$

References

- Markov chains, Pierre Bremaud, Springer, 1999
- Finite Markov chains and Algorithmic applications, Olle Haggstrom, Cambridge Univ. Press, 2002
- Markov chains and Mixing times, D. Levin, Y. Peres, E. Wilmer, AMS 2009
- Markov chains and Stochastic stability, S. Meyn and L. Tweedie, Cambridge Univ. Press, 1993
- Network Performance Analysis (Chapters 1-6), T. Bonald, M. Feuillet, Wiley, 2011