Sequential decisions under uncertainty

KTH/EES PhD course

Lecture 6

Lecture 6

- MDP with average reward criterion
  - Finite Markov chain
  - Optimality

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:

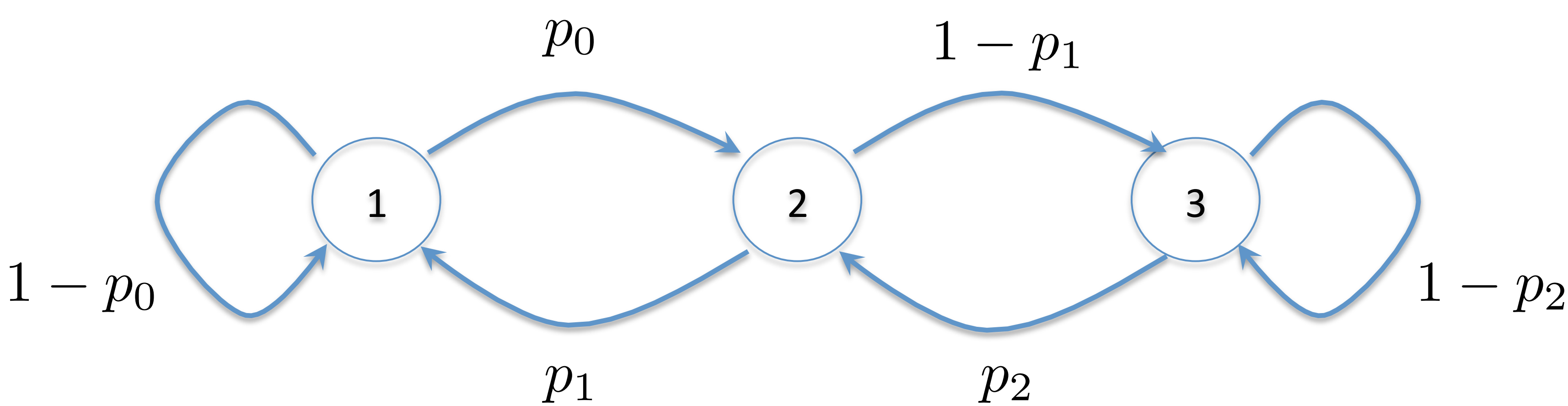
$$i \to j \iff \exists m : p^m(i,j) > 0$$

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

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

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

Limiting matrix

- Definition: $P^\star = \lim_{n \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} P^k$

$$p^\star(i,j) = \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} p^k(i,j)$$

- Properties:

$$PP^\star = P^\star P = P^\star$$

$$H_P = (I - P + P^\star)^{-1}(I - P^\star) = \lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} \sum_{k=0}^{n-1} (P^k - P^\star)$$

Fundamental matrix

$$H_P = \lim_{N \to \infty} \sum_{k=0}^{N-1} (P^k - P^\star)$$ for aperiodic chains

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

Average reward MDP: model

- Stationary reward and transitions: $r(s, a)$

  $p(j|s, a)$

- Bounded reward: $|r(s, a)| \leq M, \quad \forall s, a$

- Finite state space: S

Average reward MDP: model

- HR policies: $\pi = (\pi_1, \pi_2, \ldots)$

$$\pi_t : H_t \to \mathcal{P}(\mathcal{A})$$

- Value / gain of a policy: $g^\pi(s) = \lim_{N \to \infty} \frac{1}{N} v^\pi_{N+1}(s)$

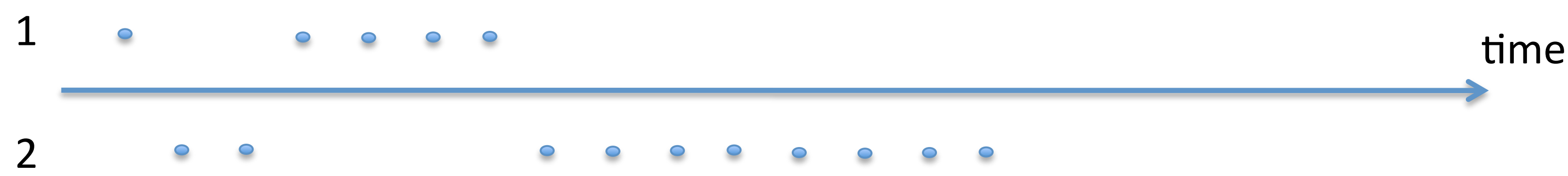
$$v^\pi_{N+1}(s) = E^\pi_s [\sum_{t=1}^{N} r(X_t, Y_t)]$$

- Optimal value: $g^\star(s) = \sup_{\pi \in HR} g^\pi(s)$

- … the limit may not exist

Average reward MDP: model

- Example: two states 1 and 2 with respective rewards 1 and 2



$$\lim_{N\to\infty} \sup \frac{1}{N} v^{\pi}_{N+1}(s) > \lim_{N\to\infty} \inf \frac{1}{N} v^{\pi}_{N+1}(s)$$

Stationary policies

- Stationary policy: $\pi = (d, d, \ldots)$

$$g^{\pi}(s) = \lim_{N \to \infty} \frac{1}{N} v^{\pi}_{N+1}(s) = P^{\star}_{d} r_{d}(s)$$

- Notation: $r_d(s) = r(s, d(s))$

$$(P_d v)(s) = \sum_{j \in S} p(j|s, d(s)) v(j)$$

HR vs. MR policies

- Markovian policies are good enough: define for all $\pi \in HR$

$$g_{-}^{\pi}(s) = \lim \inf_{N \to \infty} \frac{1}{N} v_{N+1}^{\pi}(s)$$

$$g_{+}^{\pi}(s) = \lim \sup_{N \to \infty} \frac{1}{N} v_{N+1}^{\pi}(s)$$

For each $\pi \in HR$, there exists $\pi' \in MR$ such that

$$g_{+}^{\pi} = g_{+}^{\pi'}$$

$$g_{-}^{\pi} = g_{-}^{\pi'}$$

Evaluating stationary policies

- Stationary policy: $\pi = (d, d, \ldots)$

$$g^{\pi}(s) = \lim_{N \to \infty} \frac{1}{N} v_{N+1}^{\pi}(s) = P_d^{\star} r_d(s)$$

- Under a stationary policy, the state and action evolves as an homogenous Markov chain, and the reward starting at a given state is the steady-state reward (see ergodic theorem)

- $g^{\pi}(\cdot)$ is constant over communication classes

Evaluating stationary policies

- Bias: $h^{\pi} = H_{P_d} r_d \in \mathbb{R}^S$
- Difference between total reward and stationary reward
- Aperiodic chain:

$$h^{\pi} = \sum_{t=0}^{\infty} (P^t - P^{\star}) r_d = \sum_{t=0}^{\infty} P^t (r_d - g^{\pi})$$

$$h^{\pi}(s) = E_s[\sum_{t=1}^{\infty} (r_d(X_t) - g^{\pi}(X_t))]$$

- Periodic chain: expand the expressions to Cesaro-limits

Evaluating stationary policies

- Aperiodic chain:

$$v_{N+1} = \sum_{t=1}^{N} P_d^{t-1} r_d$$

$$h^\pi = \sum_{t=1}^{N} P_d^{t-1} r_d - N g^\pi + \sum_{t=N+1}^{\infty} (P_d^{t-1} - P_d^\star) r_d$$

$$\implies v_{N+1}^\pi = N g^\pi + h^\pi + o(1)$$

Evaluation equations

Theorem We have:

(i) $(I - P_d)g^\pi = 0$

(ii) $g^\pi + (I - P_d)h^\pi = r_d$

Unichain MDPs

- Unichain MDP: the Markov chain for every stationary policy is unichain (irreducible)
- Multichain MDPs (See Puterman chapter 9)

Optimality equation

- Unichain MDP (aperiodic case)
- Optimal expected total gain: $v_N^\star = (N-1)g^\star 1 + h + o(1)$

$$v_{N+1}^\star(s) = \max_a \left[r(s,a) + \sum_j p(j|s,a) v^\star(j) \right]$$

$$v_{N+1}^\star = \max_{d:S \to A} \left[r_d + P_d v^\star \right]$$

- Hence:

$$0 = \max_a \left[r(s,a) - g^\star + \sum_j p(j|s,a) h(j) - h(s) \right]$$

Optimality equation

- Unichain MDP (aperiodic case)

$$0 = \max_{a} \left[r(s,a) - g^\star + \sum_j p(j|s,a)h(j) - h(s) \right]$$

$$0 = \max_{d:S\to A} \left[r_d - g^\star 1 + (P_d - I)h \right]$$

Optimality

Theorem If there exists a scalar g and a vector h satisfying the optimality equation, then:

$$g1 = g_+^\star = g_-^\star$$

Theorem If the action space is finite, then optimality equations have a solution.

Optimality policies

- First method: let the discount factor tend to 1 …
- Second method: *h*-improving policies

$$r_{d_h} + P_{d_h} h = \max_{d:S \to A} (r_d + P_d h)$$

Theorem If there exists a scalar g and a vector h satisfying the optimality equation, then, *h*-improving policies are optimal.