Sequential decisions under uncertainty

KTH/EES PhD course

Lecture 2

Lecture 2

- A few words on probability theory
- Finite-horizon Markov Decision Processes

Probability theory

Probability space

- The goal is to formally model "random" experiments (e.g. coin tossing)
- Samples: all information you need in understanding an experiment is contained in a sample randomly selected by nature
- Set of samples: Ω
- Example 1: throwing a dice, $\Omega = \{1, 2, 3, 4, 5, 6\}$
- Example 2: select a real number uniformly at random between 0 and 1, $\Omega = [0, 1]$

σ-algebras

- A σ-algebra is a subset of sets of the sample set such that:

  1. $\Omega \in \mathcal{F}$
  2. $F \in \mathcal{F} \Rightarrow {}^{c}F \in \mathcal{F}$
  3. If $F_n \in \mathcal{F}$ for all $n \in \mathbb{N}$, then $\cup_{n\in\mathbb{N}} F_n \in \mathcal{F}$

- σ-algebra generated by a set G of subsets is the smallest σ-algebra containing the subsets of G
- <u>Example 1:</u> throwing a dice, a natural σ-algebra is the set of all subsets of the sample space

σ-algebras

- Example 2: select a real number uniformly at random between 0 and 1, the Borel algebra is that generated by the open sets of [0,1].

  Notation: $\mathcal{F} = \mathcal{B}([0,1])$

Probability measures

- Measurable space: $(\Omega, \mathcal{F})$
- A probability measure is $P : \mathcal{F} \to [0, 1]$ such that:

1. $P(\emptyset) = 0, P(\Omega) = 1$
2. If $F_n \in \mathcal{F}$ for all $n \in \mathbb{N}$, and $F_n \cap F_m = \emptyset$, for all n, m, then

$$P(\cup_{n \in \mathbb{N}} F_n) = \sum_{n \in \mathbb{N}} P(F_n)$$

- <u>Example 1:</u> throwing a dice, $P(\omega) = 1/6, \quad \forall \omega \in \Omega$

Probability measures

- Example 2: select a real number uniformly at random between 0 and 1, $\mathcal{F} = \mathcal{B}([0,1])$

  Lebesgue measure: $P([0,x)) = x, \quad \forall x \in [0,1]$

- Terminology:
  - $(\Omega, \mathcal{F}, P)$ is a probability space
  - $F \in \mathcal{F}$ is an *event*

Random variables

- Measurable space: $(\Omega, \mathcal{F})$
- A random variable is a measurable function $X : \Omega \to \mathbb{R}$

$$\forall A \in \mathcal{B}(\mathbb{R}), X^{-1}(A) \in \mathcal{F}$$

- Example 1: throw a dice

$$\forall \omega \in \Omega, X(\omega) = \begin{cases} 1, & \text{if } \omega \text{ is even,} \\ 0, & \text{otherwise.} \end{cases}$$

- Interpretation: we run an experiment, and observe the value of a random variable. It provides *partial* information about the sample selected by nature.

σ-algebras generated by random variables

- Family of random variables on $(\Omega, \mathcal{F})$: $(X_\gamma, \gamma \in G)$
- The σ-algebra generated by $(X_\gamma, \gamma \in G)$ is the smallest algebra $\mathcal{G} \subset \mathcal{F}$ such that for all $\gamma \in G, X_\gamma$ is $\mathcal{G}$-measurable
- Notation: $\mathcal{G} = \sigma(X_\gamma, \gamma \in G)$
- Interpretation: We run an experiment with $(\Omega, \mathcal{F}, P)$. Nature selects a sample ω. We observe the values $X_\gamma(\omega)$. The algebra $\mathcal{G} = \sigma(X_\gamma, \gamma \in G)$ consists of those events *F* for which for all sample, you are able to decide whether *F* occurred or not observing $X_\gamma(\omega)$
- <u>Example 1.</u> (cf. previous slide) .

$$\sigma(X) = \{\emptyset, \Omega, \{1,3,5\}, \{2,4,6\}\}$$

Expectation

- Restrict attention to countable sample sets
- Probability space: $(\Omega, \mathcal{F}, P)$
- Random variable: $X : \Omega \to \mathbb{R}$
- Define $A = \{X(\omega), \omega \in \Omega\}, \quad F_a^X = X^{-1}(\{a\}), \forall a \in A$
- Expectation (if it exists):

$$E[X] = \sum_{a \in A} aP(F_a^X) = \sum_{a \in A} aP[X = a]$$

Conditional expectation

- Restrict attention to countable sample sets
- Probability space: $(\Omega, \mathcal{F}, P)$
- Conditional probability: $F, G \in \mathcal{F}$

$$P(F|G) = P(F \cap G)/P(G)$$

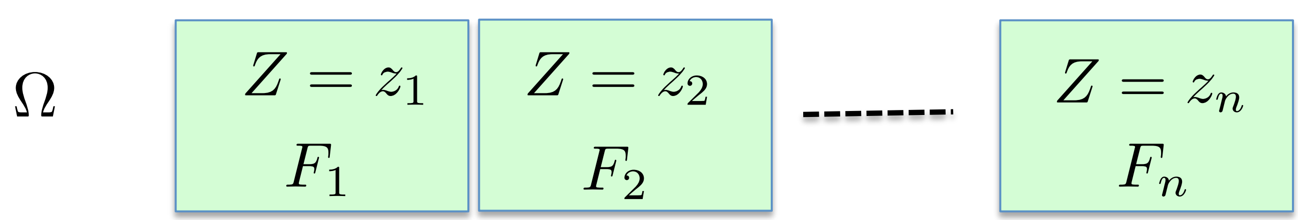
- Two random variables *X* and *Z* with respective values $(x_1, \ldots, x_m), (z_1, \ldots, z_n)$

$$E[X|Z = z_j] = \sum_{i=1}^{n} x_i P[X = x_i | Z = z_j]$$

Conditional expectation

- Random variable $Y = E[X|Z]$

if $Z(\omega) = z_j, \quad Y(\omega) = E[X|Z = z_j]$



Y is constant over $F_j \iff Y \sigma(Z)-$measurable

Conditional expectation

- Interpretation: An experiment has been performed. The available information is $Z(\omega)$. $Y(\omega)$ is the expectation of *X* given that information.

Properties

- For any pair of r.v. $X, Z, \quad E[X] = E[E[X|Z]]$
- If *X* is $\sigma(Z)-$measurable, $X = E[X|Z]$
- Tower property: two algebras $\mathcal{H} \subset \mathcal{G}$

$$E[E[X|\mathcal{G}]|\mathcal{H}] = E[X|\mathcal{H}]$$

References

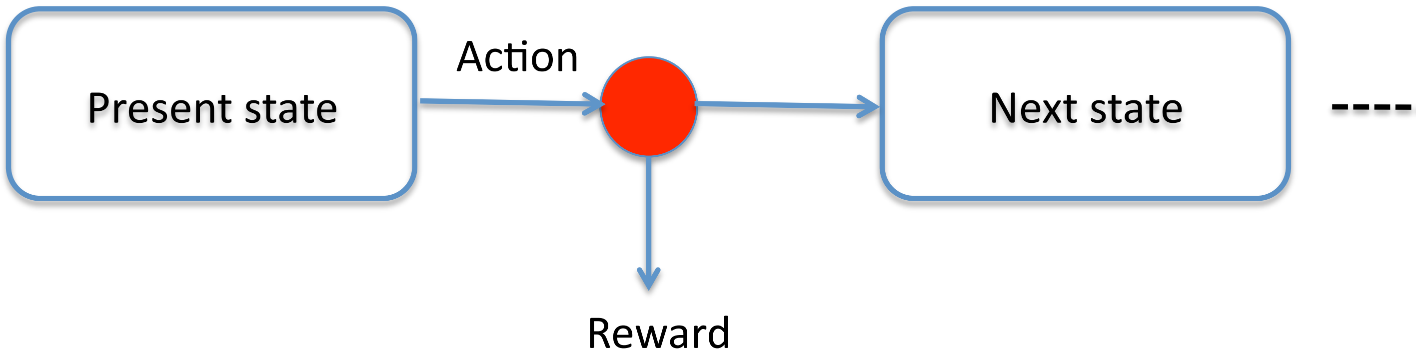
- See Chapters 2 – 9 in

  Probability with Martingales, David Williams

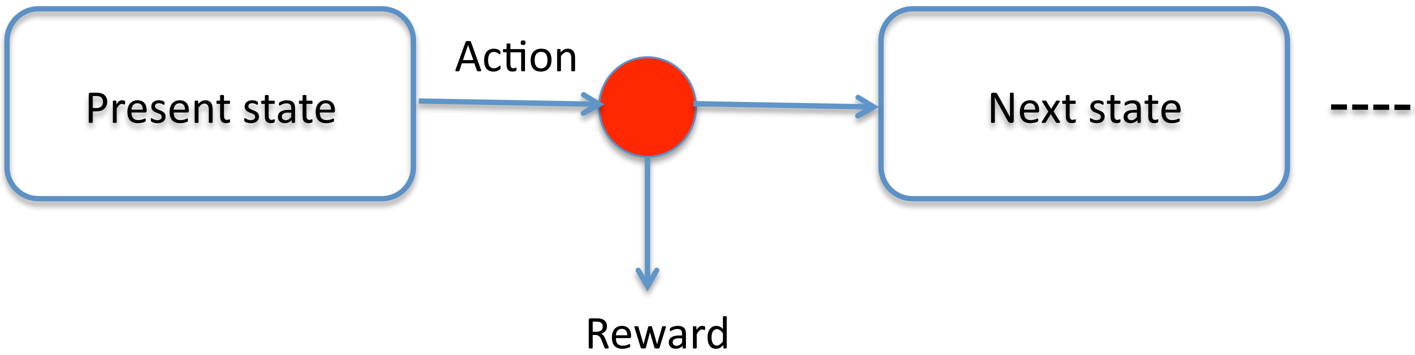
  Cambridge University Press

Finite-horizon Markov Decision Processes

Model

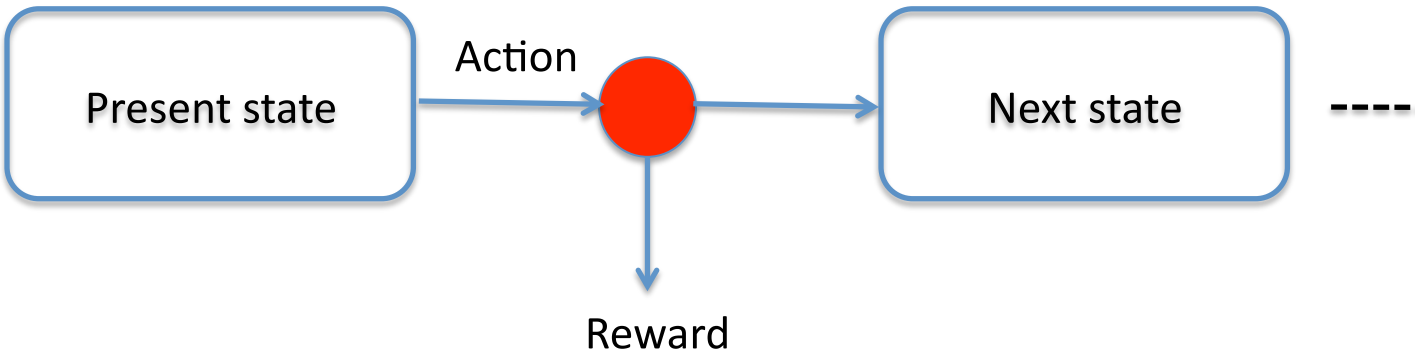


States, actions, time horizon



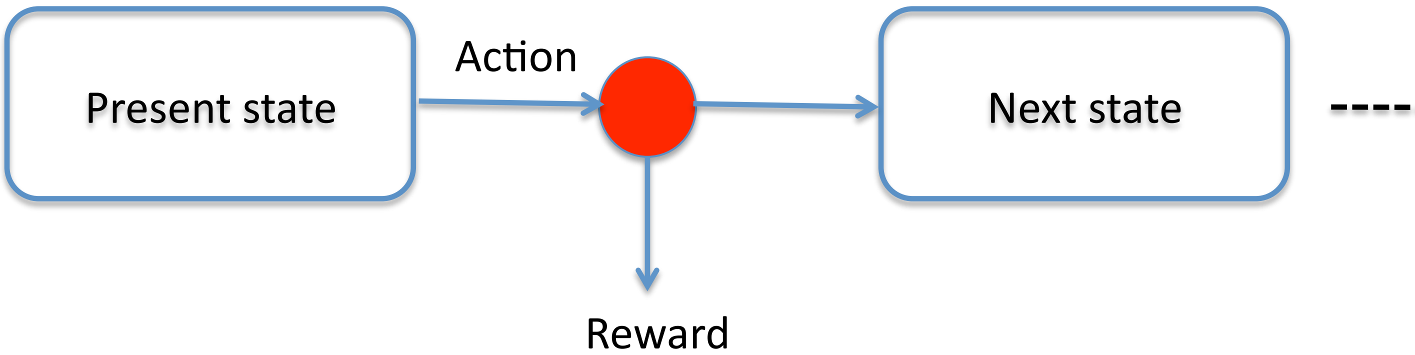
- Set of states: S
- Set of actions available in state *s*: $A_s, \quad A = \cup_{s \in S} A_s$
- These sets are finite, countably infinite, or compacts subsets of a Euclidian space (finite dimension)
- Time horizon *N*: $t \in \{1, \ldots, N\}$

Rewards and transitions



- Reward when selecting at time *t* action *a* in state *s*: $r_t(s, a)$
  It could also depend on the next state: $r_t(s, a, s')$
- Reward at time *N*: $r_N(s)$
- Probability to move from state *s* to *s'* when selecting at time *t* action *a*: $p_t(s'|s, a)$

Decision rules, policies



- History up to time *t*: $h_t = (s_1, a_1, \ldots, s_{t-1}, a_{t-1}, s_t)$
- Set of possible histories: $(S \times A)^{t-1} \times S$
- We distinguish different types of policies:
  - History dependent Randomized: HR
  - History dependent Deterministic: HD
  - Markov Randomized: MR
  - Markov Deterministic: MD

Decision rules, policies

- HR: $\pi = (\pi_1, \ldots, \pi_{N-1})$

  $\pi_t : (S \times A)^{t-1} \times S \to \mathcal{P}(A_{s_t})$

  $q_{\pi_t(h_t)}(a)$: probability to select action *a*

- HD: $\pi_t : (S \times A)^{t-1} \times S \to A_{s_t}$

  $\pi_t(h_t)$: selected action

- MR: $\pi_t : S \to \mathcal{P}(A_{s_t})$

  $q_{\pi_t(s_t)}(a)$: probability to select action *a*

Decision rules, policies

- MD: $\pi_t : S \to A_{s_t}$

  $\pi_t(s_t)$: selected action

- Note that: $MD \subset MR \subset HR$

  $MD \subset HD \subset HR$

- We will provide conditions under which MD are as good as HR policies

Induced probability space

- Restrict attention to discrete states and actions
- Probability space: $\Omega = (S \times A)^{N-1} \times S$
- Sample path: $\omega = (s_1, a_1, \ldots, s_{t-1}, a_{t-1}, s_t)$
- Algebra: all possible subsets of sample paths
- Random variables: $X_t(\omega) = s_t, \quad Y_t(\omega) = a_t, \quad Z_t(\omega) = h_t$
- A policy induces a probability measure
- When starting at $s_1 = s$
- For $\pi \in HR \quad P^\pi[X_1 = s] = 1$

$$P^\pi[Y_t = a | Z_t = h_t] = q_{\pi_t(h_t)}(a)$$

$$P^\pi[X_{t+1} = s | Z_t = h_t, Y_t = a_t] = p_t(s | s_t, a_t)$$

Induced probability space

- Sample path probability:

$$P^\pi[s,a_1,s_2,\ldots,s_t] = q_{\pi_1(s)}(a_1)p_1(s_2|s,a_1)$$
$$q_{\pi_2(h_1)}(a_2)\ldots q_{\pi_{t-1}(h_{t-1})}(a_{t-1})p_{t-1}(s_t|s_{t-1},a_{t-1})$$

- Conditional probability:

$$P^\pi[a_t,s_{t+1},\ldots,s_N|s,a_1,\ldots,s_t] = \frac{P^\pi[s,a_1,...,s_N]}{P^\pi[s,a_1,...,s_t]}$$

$$P^\pi[a_t,s_{t+1},\ldots,s_N|s,a_1,\ldots,s_t] = q_{\pi_t(h_t)}(a_t)p_t(s_{t+1}|s_t,a_t)$$
$$\ldots q_{\pi_{N-1}(h_{N-1})}(a_{N-1})p_{N-1}(s_N|s_{N-1},a_{N-1})$$

Induced probability space

- Reward from time *t*:

$$R_t(s_t, a_t, ..., s_N) = \sum_{u=t}^{N-1} r_u(s_u, a_u) + r_N(s_N)$$

- Given that the history is h_t :

$$E^{\pi}_{h_t}[R_t] = \sum_{(a_t, s_{t+1}, ..., s_N)} R_t(s_t, a_t, ..., s_N) \times P^{\pi}[a_t, s_{t+1}, ..., s_N | s, a_1, ..., s_t]$$

Value function

- Defined as: $v_N^\star(s) = \sup_{\pi \in HR} v_N^\pi(s)$

  with
  $$v_N^\pi(s) = E_s^\pi\left[\sum_{t=1}^{N-1} r_t(X_t, Y_t) + r_N(X_N)\right]$$

- Optimal policies may not exist (countably infinite actions), in which case we look for $\pi^\star : v_N^{\pi^\star}(s) \geq v_N^\star(s) - \epsilon$

Computing rewards

- Let $\pi \in HR$
- Define the reward from time *t* given history h_t

$$u_t^\pi(h_t) = E_{h_t}^\pi[\sum_{u=t}^{N-1} r_u(X_u, Y_u) + r_N(X_N)]$$

- Note that $v_N^\pi(s) = u_1^\pi(s)$
- We compute rewards using a backward induction

Algorithm

1. For $t = N, \quad \forall h_N, u_N^\pi(h_N) = r_N(s_N)$
2. Until $t = 1$

   $t - 1 \rightarrow t$

   $\forall h_t$:

$$u_t^\pi(h_t) = \sum_{a \in A_{s_t}} q_{\pi_t(h_t)}(a) \left[r_t(s_t, a) + \sum_{j \in S} p(j|s_t, a) u_{t+1}^\pi(h_t, a, j) \right]$$

Principle of optimality

- We construct optimal policies using backward induction
- i.e., we compute the optimal reward from time *t* given history h_t

$$u_t^\star(h_t) = \sup_{\pi \in HR} u_t^\pi(h_t)$$

- Optimality equations

$$u_N(h_N) = r_N(s_N)$$

$$u_t(h_t) = \sup_{a \in A_{s_t}} \left[r_t(s_t, a) + \sum_{j \in S} p_t(j|s_t, a) u_{t+1}(h_t, a, j) \right]$$

Principle of optimality

- Optimality equations

$$u_t(h_t) = \sup_{a \in A_{s_t}} \left[r_t(s_t, a) + \sum_{j \in S} p_t(j|s_t, a) u_{t+1}(h_t, a, j) \right]$$

- Case 1: the sup is always reached
- Case 2: the sup is not always reached

Principle of optimality

Theorem We have:

(i) $u_t(h_t) = u_t^\star(h_t), \quad \forall t = 1, \ldots N-1, \forall h_t$

(ii) $u_1(s) = v_N^\star(s)$

- In other words, we have identified the value function, i.e., the optimal reward

Optimal policy in HD

Theorem Let $\pi^\star \in HD$

Assume that for all $t = 1, \ldots, N-1, h_t$:

$$r_t(s_t, \pi_t^\star(h_t)) + \sum_{j \in S} p_t(j|s_t, \pi_t^\star(h_t)) u_{t+1}(h_t, \pi_t^\star(h_t), j)$$

$$= \max_{a \in A_{s_t}} \left[r_t(s_t, a) + \sum_{j \in S} p_t(j|s_t, a) u_{t+1}(h_t, a, j) \right]$$

Then for all $t = 1, \ldots, N, h_t$: $u_t^{\pi^\star}(h_t) = u_t(h_t)$

and $\pi^\star$ is optimal: $v_N^{\pi^\star}(s) = v_N^\star(s)$

ε-optimal policy in HD

Theorem Let $\epsilon > 0$, $\pi^\epsilon \in HD$

Assume that for all $t = 1, \ldots, N-1, h_t$:

$$r_t(s_t, \pi_t^\epsilon(h_t)) + \sum_{j \in S} p_t(j|s_t, \pi_t^\epsilon(h_t)) u_{t+1}(h_t, \pi_t^\epsilon(h_t), j)$$

$$\geq \sup_{a \in A_{s_t}} \left[r_t(s_t, a) + \sum_{j \in S} p_t(j|s_t, a) u_{t+1}(h_t, a, j) \right] - \frac{\epsilon}{N-1}$$

Then for all $t = 1, \ldots, N, h_t$: $u_t^{\pi^\epsilon}(h_t) \geq u_t(h_t) - \dfrac{(N-t)\epsilon}{N-1}$

and π^ϵ is ε-optimal: $v_N^{\pi^\epsilon}(s) \geq v_N^\star(s) - \epsilon$

HD optimality

Corollary

(a) For all ε>0, there exists an ε-optimal policy in HD;

(b) Assume that for all $t = 1, \ldots, N-1, h_t :$ there exists an action a':

$$r_t(s_t, a') + \sum_{j \in S} p_t(j|s_t, a') u_{t+1}(h_t, a', j)$$

$$= \sup_{a \in A_{s_t}} \left[r_t(s_t, a) + \sum_{j \in S} p_t(j|s_t, a) u_{t+1}(h_t, a, j) \right]$$

then there exists an optimal policy in HD.

Optimality of Markov Deterministic policies

- It would greatly simplify the analysis, and reduce the computational complexity of identifying optimal policies

MD optimality

Theorem

For any $t = 1, \ldots, N, \; u_t(h_t)$ depends on h_t only through s_t

(a) For all ε>0, there exists an ε-optimal policy in MD;

(b) Assume that for all $t = 1, \ldots, N-1, h_t :$ there exists an action a':

$$r_t(s_t, a') + \sum_{j \in S} p_t(j|s_t, a') u_{t+1}(h_t, a', j)$$

$$= \sup_{a \in A_{s_t}} \left[r_t(s_t, a) + \sum_{j \in S} p_t(j|s_t, a) u_{t+1}(h_t, a, j) \right]$$

then there exists an optimal policy in MD.

Algorithm: Optimal MD policy

1. For $t = N, \quad u_N(s) = r_N(s), \forall s \in S$
2. Until $t = 1$

$$t - 1 \rightarrow t$$

$$\forall s_t \in S :$$

$$u_t(s_t) = \max_{a \in A_{s_t}} \left[r_t(s_t, a) + \sum_{j \in S} p_t(j|s_t, a) u_{t+1}(j) \right]$$

$$A^{\star}_{s_t,t} = \arg\max_{a \in A_{s_t}} \left[r_t(s_t, a) + \sum_{j \in S} p_t(j|s_t, a) u_{t+1}(j) \right]$$