Lecture 2: Introduction to Markov Decision Process

Yasuyuki Sawada, Yaolang Zhong

University of Tokyo

October 8, 2025

The Markov Process

- Definition: A Markov chain (or Markov process) is a stochastic process $\{s_t\}_{t\geqslant 0} = (s_0, s_1, \dots) \in \mathcal{S}$ describing a sequence of random variables in which <u>the outcome in the next period depends only on the state in the current period</u>.
- (memorylessness) Mathematically, for all measurable subsets $\bar{\mathcal{S}} \subseteq \mathcal{S}$,

$$\Pr(s_{t+1} \in \bar{\mathcal{S}} \mid s_t, s_{t-1}, \dots) = \Pr(s_{t+1} \in \bar{\mathcal{S}} \mid s_t).$$

- Key Concepts:
  - Environment/Dynamic system: An exogenous system evolving in discrete time $t = 0, 1, 2, \dots$.
  - State $s_t \in S$: The minimal information from the past needed to predict the future.
  - Transition function/Transition kernel/Stochastic matrix: $P_t(s' \mid s) = \Pr(s_{t+1} = s' \mid s_t = s)$. In the time-homogeneous case, this simplifies to $P(s' \mid s)$.
  - Initial distribution: $\mu(s_0)$

Examples of the Markov Process

- Example 2.1: Days of the Week. $\mathcal{S} = \{\text{Mon}, \ldots, \text{Sun}\}$. The next day depends only on today, so $\{s_t\}$ is a *deterministic, time-homogeneous* Markov chain. It is periodic with period 7.
- Example 2.2: Weather States. $\mathcal{S} = \{\text{Sunny}, \text{Cloudy}, \text{Rainy}\}$. Suppose tomorrow's weather depends only on today's weather, with transition matrix

$$P = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.3 & 0.4 & 0.3 \\ 0.2 & 0.5 & 0.3 \end{bmatrix}, \quad P_{ij} = \Pr(s_{t+1} = j \mid s_t = i).$$

  Then $\{s_t\}$ is a *stochastic, finite-state, time-homogeneous* Markov chain.
- Example 2.3: AR(1) Process.

$$s_{t+1} = \rho s_t + \varepsilon_{t+1}, \qquad \varepsilon_{t+1} \sim \text{i.i.d. } (0, \sigma^2), \ |\rho| < 1.$$

  The next value depends only on the current state, so $\{s_t\}$ is a *stochastic, continuous-state* Markov process.

Examples of the Markov Process (continued)

Example 2.4: Employment and Unemployment Transitions (Sargent et al., 2020)

- A worker is either *unemployed* ($s_t = 0$) or *employed* ($s_t = 1$). Each month:
  - An unemployed worker finds a job with probability $\alpha \in (0, 1)$.
  - An employed worker loses a job with probability $\beta \in (0, 1)$.
- The Markov representation:

$$\mathcal{S} = \{0, 1\}, \quad P = \begin{pmatrix} 1 - \alpha & \alpha \\ \beta & 1 - \beta \end{pmatrix}.$$

- Once α and β are specified, we can ask:
  - What is the average duration of unemployment?
  - Over the long run, what fraction of time does a worker spend unemployed?
  - Conditional on employment, what is the probability of becoming unemployed at least once over the next 12 months?

Examples of the Markov Process (continued)

Example 2.5: U.S. Business Cycle Transitions (Hamilton, 2005)

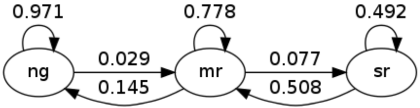
- Based on monthly U.S. unemployment data, Hamilton (2005) estimated the transition matrix

$$P = \begin{pmatrix} 0.971 & 0.029 & 0 \\ 0.145 & 0.778 & 0.077 \\ 0 & 0.508 & 0.492 \end{pmatrix},$$

  where the states represent:

  - *Normal growth (ng)*, *Mild recession (mr)*, *Severe recession (sr)*.

- Large diagonal entries indicate high persistence in the Markov process $\{s_t\}$, meaning each regime tends to last for several periods before switching.



Markov Processes: Fundamental Properties (all stochastic P)

- Finite-state representation: When $\mathcal{S} = \{1, \ldots, S\}$, the process is represented by a *row-stochastic matrix* P with elements $P_{ij} = \Pr(s_{t+1} = j \mid s_t = i)$ and $\sum_j P_{ij} = 1$.
- Expectation update: For any function $g : \mathcal{S} \to \mathbb{R}$,

$$\mathbb{E}[g(s_{t+1}) \mid s_t = i] = \sum_j P_{ij} g(j).$$

- n-step transitions: The probability of moving from i to j in n steps is $\Pr(s_{t+n} = j \mid s_t = i) = (P^n)_{ij}$.
- Chapman–Kolmogorov equation: For a time-homogeneous process, $P^{n+m} = P^n P^m$.

Markov Processes: Structural Properties (conditions on P)

- Irreducibility: Two states $s_i, s_j \in \mathcal{S}$ *communicate* if each can be reached from the other with positive probability in some finite number of steps:

$$\exists\, m, n \geqslant 1 \text{ such that } (P^m)_{ij} > 0 \quad \text{and} \quad (P^n)_{ji} > 0.$$

  The chain (or matrix P) is *irreducible* if all states communicate— meaning that from any starting point, it is possible (eventually) to reach any other state. Intuitively, there are no isolated groups of states.

- Aperiodicity: A Markov chain is called *periodic* if it moves through states in a fixed, repeating cycle, and *aperiodic* otherwise. More formally, the *period* of a state i is the greatest common divisor of all possible return times:

$$D(i) := \{\, n \geqslant 1 : (P^n)_{ii} > 0 \,\}, \qquad d(i) = \gcd(D(i)).$$

  For example, if $D(i) = \{3, 6, 9, \dots\}$, then $d(i) = 3$. A stochastic matrix is *aperiodic* if $d(i) = 1$ for all states, and *periodic* otherwise.

Markov Processes: Long-run Properties (conditions on P)

- Ergodicity (finite $\mathcal{S}$): When the transition matrix P is *irreducible* and *aperiodic*, the Markov chain admits a unique *stationary distribution* π satisfying

$$\pi = \pi P, \qquad \sum_i \pi_i = 1.$$

Under these conditions,

$$P^n \to \mathbf{1}\pi \quad \text{as } n \to \infty,$$

meaning that, regardless of the initial state, the distribution of the process converges to π.

Moreover, along any sufficiently long realization, the *time-average frequency* of visiting each state equals its stationary probability:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\{s_t = i\} = \pi_i \quad \text{with probability 1.}$$

Hence, long-run empirical frequencies coincide with theoretical steady-state probabilities, ensuring that time averages and cross-sectional distributions are consistent.

From Markov Process to Markov Decision Process (MDP)

- A *Markov Decision Process (MDP)* extends the Markov chain by allowing an **agent** to influence the evolution of the state through **actions**. At each period t, the agent observes the current state s_t, chooses an action a_t, receives a reward, and transitions probabilistically to a new state s_{t+1}.

- Formally, an MDP is defined by the tuple

$$(\mathcal{S},\, \mathcal{A},\, P,\, r,\, \beta)$$

  where:

  - State space $\mathcal{S}$: Possible system states.
  - Action space $\mathcal{A}(s)$: Feasible actions when in state s.
  - Transition kernel $P(s' \mid s, a)$: Probability of moving from s to s' given action a.
  - Reward function $r(s, a)$: Instantaneous payoff from taking action a in state s. (Sometimes expressed as a *cost* $-r(s, a)$.)
  - Discount factor $\beta \in (0, 1)$: Weights future rewards relative to current ones.

Agent–Environment Interaction (MDP loop)

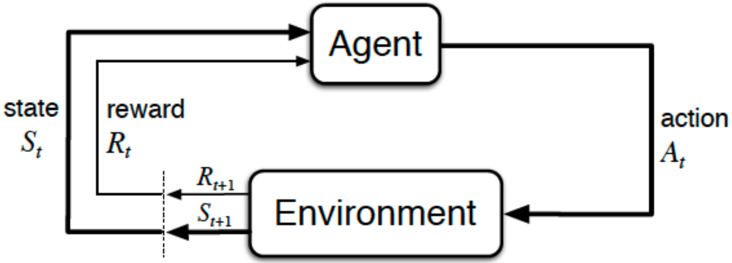


Figure 1: Agent–environment loop (adapted from Sutton and Barto (2018), Fig. 3.1).

MDP Taxonomy at a Glance

Axis	**Option**	**Notes**
Dynamics	Deterministic Stochastic	
Horizon	Finite Infinite	Episodes of length T: $t = 0, \ldots, T-1$ $T \to \infty$.
State space $\mathcal{S}$	Discrete Continuous	$\mathcal{S} = \{1, \ldots, S\}$ $\mathcal{S} \subseteq \mathbb{R}^n$
Action space $\mathcal{A}$	Discrete Continuous	$\mathcal{A}(S) \subseteq \{1, \ldots, A\}$ $\mathcal{A}(S) \subseteq \mathbb{R}^m$
Stationarity	Time-homogeneous Time-varying	$P(s' \mid s, a), \;\; r(s, a)$ $P_t(s' \mid s, a), \;\; r_t(s, a)$

Examples of Markov Decision Process (MDP)

Example 2.5: Maze Escape

- State: current position
- Action: Up, Low, Left, Right
- Reward: ?

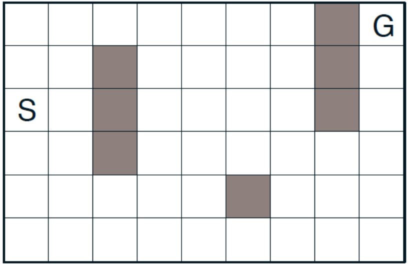


Figure 2: An Maze Problem

Examples of Markov Decision Process (MDP)

Example 2.6: The Cart-Pole game GIF

- State:
  - Cart Position: [-4.8, 4.8]
  - Cart Velocity: [-Inf, Inf]
  - Pole Angle: [-24°, 24°]
  - Pole Angular Velocity: [-Inf, Inf]
- Action: 0 (Left) or 1 (Right)
- Reward: $+1$ for every step unless failed
- Horizon: infinite but set $T = 200$ in practice

Examples of Markov Decision Process (MDP)

Example 2.7: The Pendulum GIF

- State: continuous, $s_t = [\cos\theta,\, \sin\theta,\, \dot{\theta}]$
- Action: continuous torque $a_t \in [-2,\, 2]$
- Reward: negative cost of energy and deviation from upright:

$$r_t = -(\theta^2 + 0.1\,\dot{\theta}^2 + 0.001\,a_t^2)$$

- Goal: swing up and balance the pole in the upright position ($\theta = 0$)
- Horizon: typically finite (e.g., $T = 200$)

Examples of Markov Decision Process (MDP)

Example 2.8: McCall Job Search (Discrete Action, Infinite Horizon)

- State: employment status and current offer $S_t = (E_t, W_t)$
  - $E_t \in \{\text{Unemployed}, \text{Employed}\}$
  - If unemployed, wage offer $W_t \sim F$ arrives i.i.d.
- Action: $A_t \in \{\text{Accept}, \text{Reject}\}$ (only if unemployed)
- Reward:
  - Reject $\rightarrow R_t = b$ (benefit)
  - Accept at $w \rightarrow R_t = w$ each future period
- Transition:
  - Reject $\rightarrow$ stay unemployed, draw new $W_{t+1} \sim F$
  - Accept $\rightarrow$ employed at fixed wage w (absorbing)
- Spaces: $\mathcal{S} = \{\text{U}, \text{E}(w)\} \times \text{supp}(F), \quad \mathcal{A}(\text{U}) = \{\text{Accept}, \text{Reject}\}$

Examples of Markov Decision Process (MDP)

Example 2.9: Cake-Eating / Consumption–Savings (Continuous State & Action)

- State: cake/asset stock $S_t = K_t \in [0, \bar{K}]$
- Action: consumption $A_t = C_t \in [0, K_t]$ (continuous)
- Reward: instantaneous utility from consumption, e.g. $R_t = u(C_t)$ (commonly $u(c) = \log c$ or CRRA)
- Transition (no production, no shocks): $K_{t+1} = K_t - C_t \quad$ (resource constraint)
- Horizon: infinite; discount $\beta \in (0, 1)$
- State space: $\mathcal{S} = [0, \bar{K}]$ (continuous)
- Action space: $\mathcal{A}(K) = [0, K]$ (continuous, state-dependent feasible set)

Policy of the Agent / Decision Maker

- In an MDP, the policy (or *decision rule*) specifies how the agent chooses actions based on the current state.
- A (deterministic) policy is a mapping:
$$\sigma : \mathcal{S} \to \mathcal{A}, \qquad A_t = \sigma(S_t),$$
where $\sigma(s)$ gives the action taken when the system is in state s.
- A stochastic policy assigns probabilities to actions:
$$\sigma(a \mid s) = \Pr(A_t = a \mid S_t = s),$$
meaning the agent randomizes its choice in state s.
- A stationary policy does not depend on time t:
$$\sigma_t = \sigma \quad \forall t.$$
By contrast, a *nonstationary policy* $\sigma_t(s)$ may vary with time.

References & Further Reading

- Dimitri P. Bertsekas, **Reinforcement learning and optimal control (2025 Spring course at ASU)**, Lecture 1 and 2: https://web.mit.edu/dimitrib/www/RLbook.html
- Sutton and Barto, **Reinforcement Learning: An Introduction**, Chapter 1, 2, 3: https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf

Hamilton, James D., *Regime-Switching Models in Economics and Finance*, Elsevier, 2005. Originally published work on Markov-switching models in macroeconomics.

Sargent, Thomas J., John Stachurski, and the QuantEcon team, "QuantEcon: Lectures and Code for Economics, Finance, and Data Science," https://quantecon.org/ 2020. Open-source project for quantitative economics education.

Sutton, Richard S. and Andrew G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed., MIT Press, 2018.