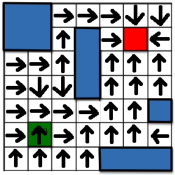
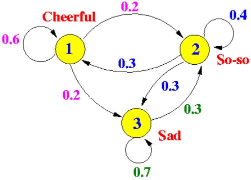
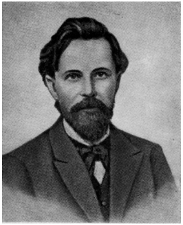
Reinforcement Learning

Markov Decision Processes



Marcello Restelli





Modelling the Environment

- **Deterministic vs Stochastic**
- **Finite** vs Continuous States
- **Finite** vs Continuous Actions
- **Discrete** vs Continuous Time
- **Fully** vs Partially Observable
- **Stationary** vs Non–Stationary

Introduction to MDPs

- **Markov Decision Processes** formally describe an environment for **reinforcement learning**
- The environment is **fully observable**
  - the **current state** completely characterizes the process
- Almost all RL problems can be formalized as MDPs
  - Optimal control primarily deals with **continuous** MDPs
  - **Partially observable** problems can be converted into MDPs
  - **Bandits** are MDPs with one state

Stochastic Processes

- A **stochastic process** is an indexed collection of **random variables** $\{X_t\}$
  - e.g., time series of weekly demands for a product
- **Discrete case**: At a particular time t, labeled by integers, system is found in exactly one of a finite number of mutually exclusive and exhaustive categories or states, labeled by integers, too
- Process could be **embedded** in that time points correspond to occurrence of specific events (or time may be evenly–spaced)
- Random variables may **depend on others**, e.g.,

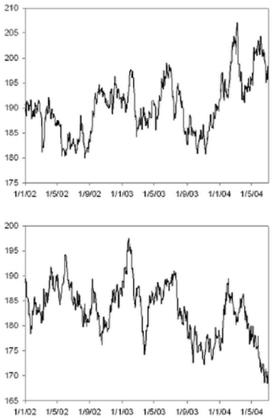
$$X_{t+1} = \begin{cases} \max\{(3 - D_{t+1}), 0\} & \text{if } X_t \leq 0 \\ \max\{(X_t - D_{t+1}), 0\} & \text{if } X_t \geq 0 \end{cases}$$

or

$$X_{t+1} = \sum_{k=0}^{K} \alpha_k X_{t-k} + \xi_t \text{ with } \xi_t \sim \mathcal{N}(\mu, \sigma^2)$$

Examples of Stochastic Processes

- Exchange rates
- Photon emission
- Epidemic models
- Earthquakes
- Budding yeast

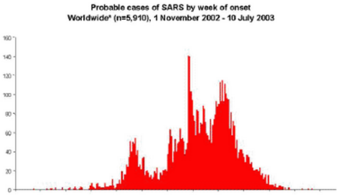


Examples of Stochastic Processes

- Exchange rates
- Photon emission
- Epidemic models
- Earthquakes
- Budding yeast

Examples of Stochastic Processes

- Exchange rates
- Photon emission
- Epidemic models
- Earthquakes
- Budding yeast

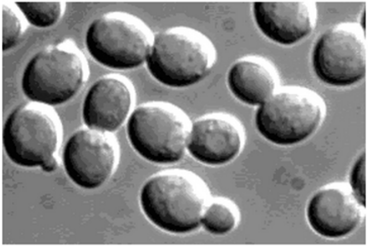


Examples of Stochastic Processes

- Exchange rates
- Photon emission
- Epidemic models
- Earthquakes
- Budding yeast

Examples of Stochastic Processes

- Exchange rates
- Photon emission
- Epidemic models
- Earthquakes
- Budding yeast



Markov Assumption

"The future is independent of the past given the present"

Definition

A stochastic process X_t is said to be **Markovian** if and only if

$$\mathbb{P}(X_{t+1} = j | X_t = i, X_{t-1} = k_{t-1}, \ldots, X_1 = k_1, X_0 = k_0) = \mathbb{P}(X_{t+1} = j | X_t = i)$$

- The state **captures all the information** from history
- Once the state is known, the history may be **thrown away**
- The state is a **sufficient statistic** for the future
- The conditional probabilities are **transition probabilities**
- If the probabilities are **stationary** (time invariant), we can write:

$$p_{ij} = \mathbb{P}(X_{t+1} = j | X_t = i) = \mathbb{P}(X_1 = j | X_0 = i)$$

Markov Processes

A Markov process (or Markov chain) is a **memoryless** stochastic process, i.e., a sequence of random states $s_1, s_2, \dots$ with the Markov property

Definition

A **Markov Process** is a tuple $\langle \mathcal{S}, P, \mu \rangle$

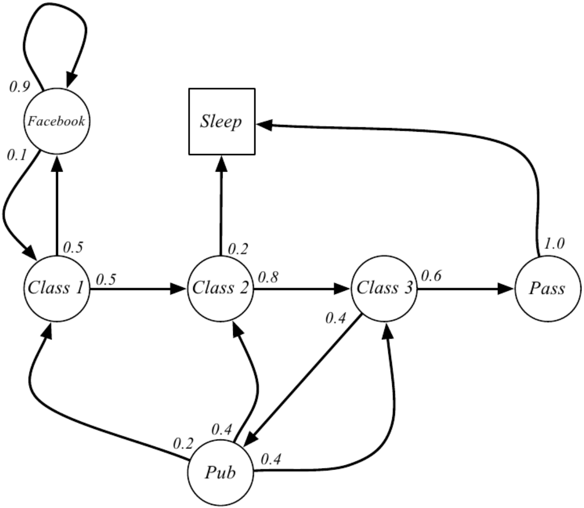
- $\mathcal{S}$ is a (finite) set of states
- P is a state transition probability matrix, $P_{ss'} = \mathbb{P}(s'|s)$
- a set of initial probabilities $\mu_i^0 = \mathbb{P}(X_0 = i)$ for all i

- Looking forward in time, n–step transition probabilities

$$p_{i,j}^{(n)} = \mathbb{P}(X_{t+n} = j | X_t = i) = \mathbb{P}(X_n = j | X_0 = i)$$

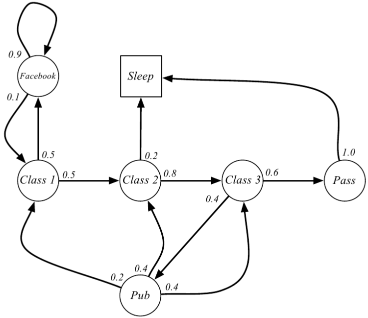
Markov Process Example 1

Student process



Markov Process Example 1

Student process

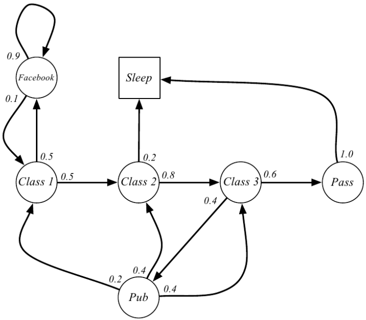


Sample paths

- C1 C2 C3 Pass Sleep
- C1 FB FB C1 C2 Sleep
- C1 C2 C3 Pub C2 C3 Pass Sleep
- C1 FB FB C1 C2 C3 Pub C1 FB FB FB C1 C2 Sleep

Markov Process Example 1

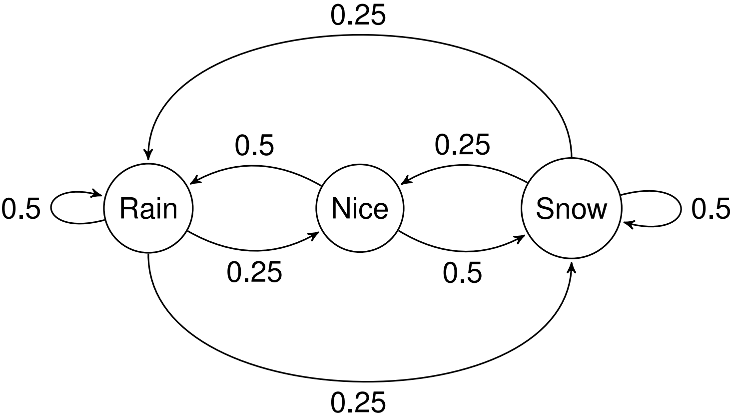
Student process



$$P = \begin{array}{l} FB \\ C1 \\ C2 \\ C3 \\ Pub \\ Pass \\ Sleep \end{array} \begin{bmatrix} 0.9 & 0.1 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.8 & 0 & 0 & 0.2 \\ 0 & 0 & 0 & 0 & 0.4 & 0.6 & 0 \\ 0 & 0.2 & 0.4 & 0.4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Markov Process Example 2

Land of Oz



$$P = \begin{array}{l} \textit{Rain} \\ \textit{Nice} \\ \textit{Snow} \end{array} \begin{bmatrix} 0.5 & 0.25 & 0.25 \\ 0.5 & 0 & 0.5 \\ 0.25 & 0.25 & 0.5 \end{bmatrix}$$

Chapman–Kolmogorov

- n–step transition probabilities can be obtained from 1–step transition probabilities **recursively**

$$p_{ij}^{(n)} = \sum_{k=0}^{M} p_{ik}^{(v)} p_{kj}^{(n-v)}, \quad \forall i, j, n; \quad 0 \leq v \leq n$$

- We can get this via the **matrix** too

$$P^{(n)} = \underbrace{P \cdots P}_{n \text{ times}} = P^n = PP^{n-1} = P^{n-1}P$$

- Given the transition matrix P and the **starting distribution** represented by the probability vector μ, the probability distribution over the state space after n steps is:

$$\mu^{(n)} = \mu P^n$$

Chapman–Kolmogorov for Student Process

$$P = \begin{array}{l} FB \\ C1 \\ C2 \\ C3 \\ Pub \\ Pass \\ Sleep \end{array} \begin{bmatrix} 0.90 & 0.10 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.50 & 0.00 & 0.50 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.80 & 0.00 & 0.00 & 0.20 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.40 & 0.60 & 0.00 \\ 0.00 & 0.20 & 0.40 & 0.40 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix}$$

$$P^2 = \begin{array}{l} FB \\ C1 \\ C2 \\ C3 \\ Pub \\ Pass \\ Sleep \end{array} \begin{bmatrix} 0.86 & 0.09 & 0.05 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.45 & 0.05 & 0.00 & 0.40 & 0.00 & 0.00 & 0.10 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.32 & 0.48 & 0.20 \\ 0.00 & 0.08 & 0.16 & 0.16 & 0.00 & 0.00 & 0.60 \\ 0.10 & 0.00 & 0.10 & 0.32 & 0.16 & 0.24 & 0.08 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix}$$

$$P^{10} = \begin{array}{l} FB \\ C1 \\ C2 \\ C3 \\ Pub \\ Pass \\ Sleep \end{array} \begin{bmatrix} 0.59 & 0.07 & 0.04 & 0.04 & 0.02 & 0.02 & 0.22 \\ 0.33 & 0.04 & 0.02 & 0.03 & 0.01 & 0.02 & 0.55 \\ 0.03 & 0.00 & 0.00 & 0.01 & 0.01 & 0.01 & 0.94 \\ 0.04 & 0.01 & 0.00 & 0.01 & 0.00 & 0.01 & 0.93 \\ 0.10 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.85 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix}$$

$$P^{100} = \begin{array}{l} FB \\ C1 \\ C2 \\ C3 \\ Pub \\ Pass \\ Sleep \end{array} \begin{bmatrix} 0.01 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.99 \\ 0.01 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.99 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix}$$

Chapman–Kolmogorov for Land of Oz

$$P = \begin{bmatrix} 0.500 & 0.250 & 0.250 \\ 0.500 & 0.000 & 0.500 \\ 0.250 & 0.250 & 0.500 \end{bmatrix}$$

$$P^2 = \begin{bmatrix} 0.438 & 0.188 & 0.375 \\ 0.375 & 0.250 & 0.375 \\ 0.375 & 0.188 & 0.438 \end{bmatrix}$$

$$P^3 = \begin{bmatrix} 0.406 & 0.203 & 0.391 \\ 0.406 & 0.188 & 0.406 \\ 0.391 & 0.203 & 0.406 \end{bmatrix}$$

$$P^4 = \begin{bmatrix} 0.402 & 0.199 & 0.398 \\ 0.398 & 0.203 & 0.398 \\ 0.398 & 0.199 & 0.402 \end{bmatrix}$$

$$P^5 = \begin{bmatrix} 0.400 & 0.200 & 0.399 \\ 0.400 & 0.199 & 0.400 \\ 0.399 & 0.200 & 0.400 \end{bmatrix}$$

$$P^6 = \begin{bmatrix} 0.400 & 0.200 & 0.400 \\ 0.400 & 0.200 & 0.400 \\ 0.400 & 0.200 & 0.400 \end{bmatrix}$$

First Passage Times

- Number of transitions to go from i to j for the first time
  - if $i = j$, this is the **recurrence time**
  - First Passage Times are random variables
- n–step recursive relationship for first passage probability

$$
\begin{aligned}
f_{ij}^{(1)} &= p_{ij}^{(1)} = p_{ij} \\
f_{ij}^{(2)} &= p_{ij}^{(2)} - f_{ij}^{(1)} p_{jj} \\
&\vdots \\
f_{ij}^{(n)} &= p_{ij}^{(n)} - f_{ij}^{(1)} p_{jj}^{(n-1)} - f_{ij}^{(2)} p_{jj}^{(n-2)} - \dots - f_{ij}^{(n-1)} p_{jj}
\end{aligned}
$$

- For fixed i and j, these $f_{ij}^{(n)}$ are non–negative numbers so that $\sum_{n=1}^{\infty} f_{ij}^{(n)} \leq 1$
- If $\sum_{n=1}^{\infty} f_{ii}^{(n)} = 1$ that state is a **recurrent** state

Classification of States

- State j is **accessible** from i if $p_{ij}^{(n)} > 0$ (for some $n \geq 0$)
- If state j is accessible from i and vice versa, the two states are said to **communicate**
- As a result of communication, one may **partition** the general Markov process into states in disjoint classes
- **Positive recurrent**: a state that is recurrent and has a finite expected return time
- If the Markov process can only visit the state at integer multiples of t, we call it **periodic**
- Positive recurrent states that are aperiodic are called **ergodic** states
- A state j is said to be an **absorbing** state if $p_{jj} = 1$
- A state which is not absorbing is called **transient**

Classification of Markov Processes

Absorbing

A Markov process is called **absorbing** if it has **at least one** absorbing state and if that state can be **reached from every** other state (not necessarily in one step)

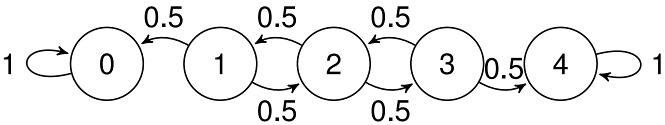
Ergodic

A Markov process is called **ergodic** (or **irreducible**) if it is possible to go from every state to every state (not necessarily in one move)

Regular

A Markov process is called **regular** if some power of the transition matrix has **only positive elements**

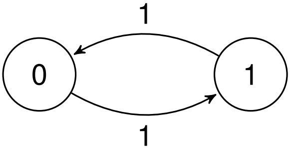
Drunkard's Walk



$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Examples

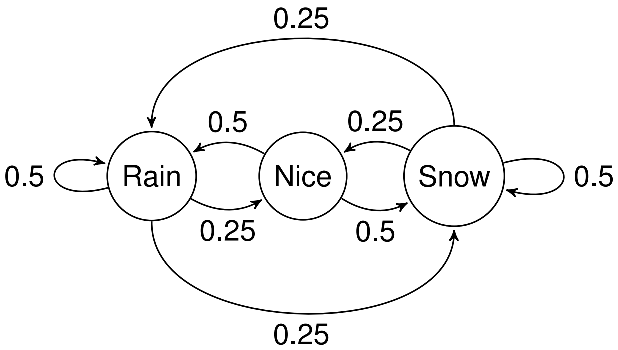
Switching Process



$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Examples

Land of Oz



$$P = \begin{bmatrix} 0.500 & 0.250 & 0.250 \\ 0.500 & 0.000 & 0.500 \\ 0.250 & 0.250 & 0.500 \end{bmatrix}$$

Stationary Distribution

- Consider Markov process with transition probability matrix P. A probability vector $\mathbf{p}$ is called a **stationary distribution** of the Markov process if:

$$\mathbf{p}P = \mathbf{p}$$

Theorem

Let P be the transition matrix for a ***regular*** *chain. Then, as $n \to \infty$, the powers P^n approach a limiting matrix W with all rows the same vector $\mathbf{w}$. The vector $\mathbf{w}$ is a strictly positive probability vector (i.e., the components are all positive and they sum to one).*

Stationary Distribution

Theorem

Let P be a ***regular*** *transition matrix,* $\mathbf{w}$ *be the common row of W, and let* $\mathbf{c}$ *be the column vector all of whose components are 1. Then*

- $\mathbf{w}P = \mathbf{w}$*, and any row vector* $\mathbf{v}$ *such that* $\mathbf{v}P = \mathbf{v}$ *is a constant multiple of* $\mathbf{w}$
- $P\mathbf{c} = \mathbf{c}$*, and any column vector* $\mathbf{x}$ *such that* $P\mathbf{x} = \mathbf{x}$ *is a multiple of* $\mathbf{c}$*.*

Theorem

Let P be the transition matrix for a ***regular*** *chain and* $\mathbf{v}$ *an arbitrary probability vector. Then*

$$\lim_{n \to \infty} \mathbf{v}P^n = \mathbf{w},$$

where $\mathbf{w}$ *is the* ***stationary distribution*** *of the chain.*

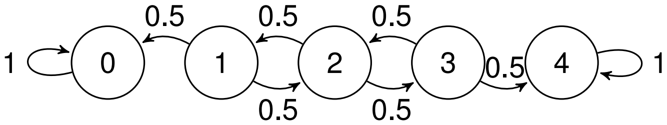
- **Canonical form**: renumber the states so that **transient** states come first:

$$P = \begin{bmatrix} Q & R \\ 0 & I \end{bmatrix}$$

- $Q^n \xrightarrow{n \to \infty} 0$

Theorem

*For an absorbing Markov process the matrix $I - Q$ has an inverse N (called **fundamental matrix**) and $N = \sum_{i=0}^{\infty} Q^i$. The ij–entry n_{ij} of the matrix N is the **expected number of times** the chain is in state s_j, given that it starts in state s_i. The initial state is counted if $i = j$.*



Canonical form

$$P = \begin{bmatrix} Q & R \\ 0 & I \end{bmatrix} = \begin{array}{c} 1 \\ 2 \\ 3 \\ 0 \\ 4 \end{array} \left[\begin{array}{ccc|cc} 0 & 0.5 & 0 & 0.5 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0.5 \\ \hline 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right]$$

$$N = (I - Q)^{-1} = \begin{array}{c} 1 \\ 2 \\ 3 \end{array} \begin{bmatrix} 1.5 & 1 & 0.5 \\ 1 & 2 & 1 \\ 0.5 & 1 & 1.5 \end{bmatrix}$$

Fundamental Matrix

Ergodic Markov Processes

Proposition

Let P be the transition matrix of an ergodic chain, and let W be the matrix all of whose rows are the fixed probability row vector for P. Then the matrix $I - P + W$ has an inverse Z that is called **fundamental matrix**.

Theorem

The mean first passage matrix M for an ergodic chain is determined from the fundamental matrix Z and the fixed row probability vector $\mathbf{w}$ *by* $m_{ij} = \frac{z_{jj} - z_{ij}}{w_j}$.

Marcello Restelli

$$I - P + W = \begin{bmatrix} 0.9 & -0.05 & 0.15 \\ -0.1 & 1.2 & -0.1 \\ 0.15 & -0.05 & 0.9 \end{bmatrix}$$

$$Z = \begin{bmatrix} 1.147 & 0.04 & -0.187 \\ 0.08 & 0.84 & 0.08 \\ -0.187 & 0.04 & 1.147 \end{bmatrix}$$

$$w = [0.4 \quad 0.2 \quad 0.4]$$

$$m_{12} = \frac{z_{22} - z_{12}}{w_2} = \frac{0.84 - 0.04}{0.2} = 4$$

Mixing Rate

Definition

The **total variation distance** between two probability distributions μ and ν on Ω is

$$\|\mu - \nu\|_{TV} = \frac{1}{2}\sum_{x \in \Omega} |\mu(x) - \nu(x)|$$

Definition

Given ϵ, the **mixing time** is

$$\tau(\epsilon) = \min_{t}\left\{\max_{x \in \Omega} \|P^{t'}(x, \cdot) - w\|_{TV} < \epsilon, \quad \forall t' \geq t\right\}$$

Definition

A Markov chain is **rapidly mixing** if $\tau(\epsilon)$ is $O(\text{poly}(log(\frac{|\mathcal{S}|}{\epsilon})))$

Spectral gap

Definition

A Markov process is **time–reversible** if $w_i P_{ij} = w_j P_{ji}$.

The **eigenvalues** of P are
$1 = \lambda_1 > \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_{|\mathcal{S}|} \geq -1$

Definition

The **spectral gap** β of a Markov process defined by transition matrix P is $1 - \max(|\lambda_2|, |\lambda_{|\mathcal{S}|}|)$

Theorem (Alon, Sinclair)

$$\frac{1-\beta}{\beta} log\left(\frac{1}{2\epsilon}\right) \leq \tau(\epsilon) \leq \frac{1}{\beta} log\left(\frac{1}{\epsilon \min_i w_i}\right)$$

Markov Reward Processes

A Markov reward process is a Markov process with **values.**

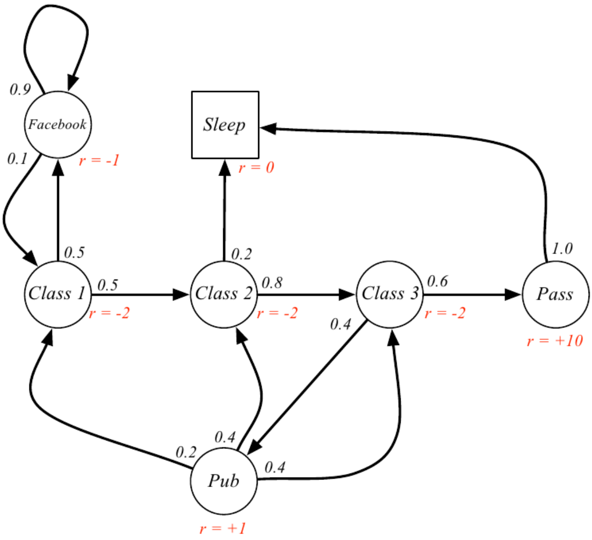
Definition

A **Markov Reward Process** is a tuple $\langle \mathcal{S}, P, R, \gamma, \mu \rangle$

- $\mathcal{S}$ is a (finite) set of states
- P is a state transition probability matrix, $P_{ss'} = \mathbb{P}(s'|s)$
- R is a reward function, $R_s = \mathbb{E}[r|s]$
- γ is a discount factor, $\gamma \in [0, 1]$
- a set of initial probabilities $\mu_i^0 = P(X_0 = i)$ for all i

Example

Student MRP



Return

- Time horizon
  - **finite**: finite and fixed number of steps
  - **indefinite**: until some stopping criteria is met (**absorbing** states)
  - **infinite**: forever
- Cumulative reward
  - total reward:
$$V = \sum_{i=1}^{\infty} r_i$$
  - average reward:
$$V = \lim_{n \to \infty} \frac{r_1 + \cdots + r_n}{n}$$
  - discounted reward:
$$V = \sum_{i=1}^{\infty} \gamma^{i-1} r_i$$
  - mean–variance reward

Infinite–horizon Discounted Return

Definition

The **return** v_t is the total discounted reward from time–step t.

$$v_t = r_{t+1} + \gamma r_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

- The discount $\gamma \in [0, 1)$ is the present value of future rewards
- The value of receiving reward r after $k+1$ time–steps is $\gamma^k r$
- **Immediate** reward vs **delayed** reward
  - γ close to 0 leads to "**myopic**" evaluation
  - γ close to 1 leads to "**far–sighted**" evaluation
- γ can be also interpreted as the **probability** that the process will **go on**

Why discount?

Most Markov reward (and decision) processes are discounted, why?

- **Mathematically** convenient to discount rewards
- **Avoids infinite returns** in cyclic Markov processes
- **Uncertainty** about the future may not be fully represented
- If the reward is **financial**, immediate rewards may earn more interest than delayed rewards
- **Animal/human behavior** shows preference for immediate reward
- It is sometimes possible to use **undiscounted** Markov reward processes (i.e. $\gamma = 1$), e.g. if all sequences **terminate**

Value Function

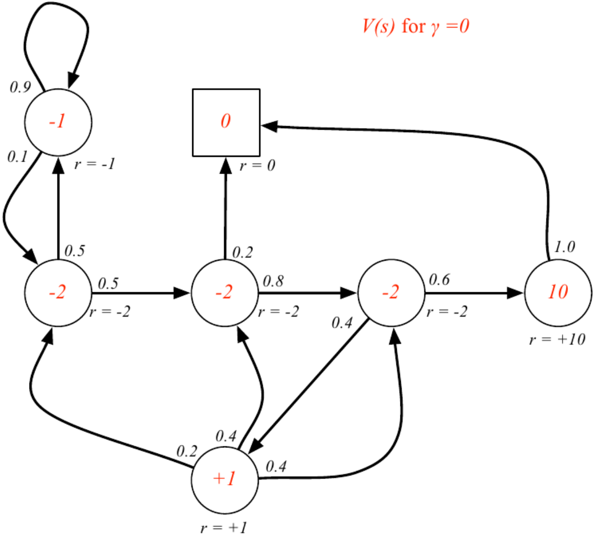
The value function $V(s)$ gives the long–term value of state s

Definition

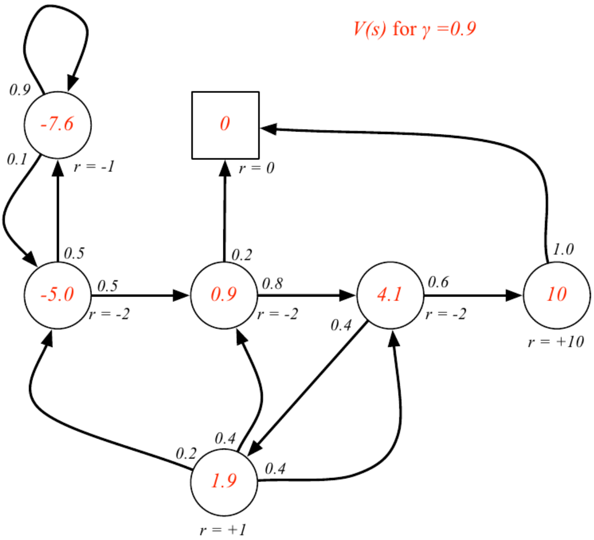
The **state value function** $V(s)$ of an MRP is the **expected** return starting from state s

$$V(s) = \mathbb{E}[v_t | s_t = s]$$

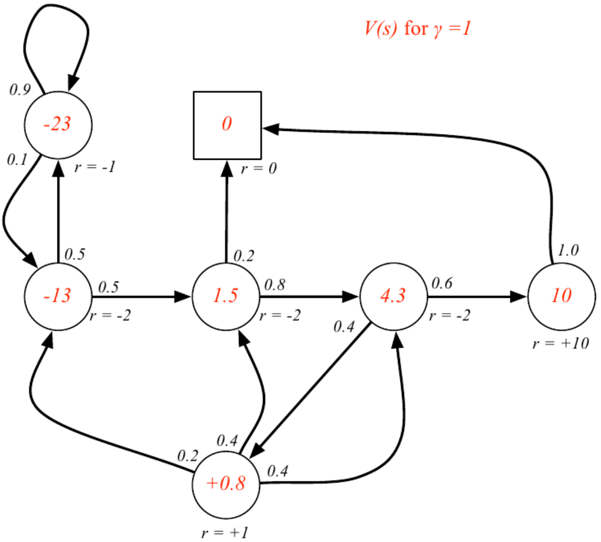
Example


V(s) for γ =0
0.9
-1
0.1
r = -1
0
r = 0
0.5
0.2
1.0
-2
0.5
r = -2
-2
0.8
r = -2
-2
0.6
r = -2
10
r = +10
0.4
0.4
0.2
+1
0.4
r = +1

Example


V(s) for γ =0.9
0.9
-7.6
0.1
r = -1
0
r = 0
0.5
-5.0
0.5
r = -2
0.2
0.9
0.8
r = -2
4.1
0.6
r = -2
0.4
1.0
10
r = +10
0.4
0.2
1.9
0.4
r = +1

Example


V(s) for γ =1
-23
0
-13
1.5
4.3
10
+0.8
0.9
0.1
r = -1
r = 0
0.5
0.5
r = -2
0.2
0.8
r = -2
0.6
r = -2
0.4
1.0
r = +10
0.2
0.4
0.4
r = +1

Bellman Equation for MRPs

The value function can be decomposed into **two** parts:

- **immediate** reward r
- discounted value of **successor** state $\gamma V(s')$

$$
\begin{aligned}
V(s) &= \mathbb{E}[v_t | s_t = s] \\
&= \mathbb{E}[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots | s_t = s] \\
&= \mathbb{E}[r_{t+1} + \gamma(r_{t+2} + \gamma r_{t+3} + \ldots) | s_t = s] \\
&= \mathbb{E}[r_{t+1} + \gamma v_{t+1} | s_t = s] \\
&= \mathbb{E}[r_{t+1} + \gamma V(s_{t+1}) | s_t = s]
\end{aligned}
$$

Bellman Equation

$$V(s) = \mathbb{E}[r + \gamma V(s') | s] = R_s + \gamma \sum_{s' \in S} P_{ss'} V(s')$$

Bellman Equation in Matrix Form

The Bellman equation can be expressed **concisely** using matrices

$$V = R + \gamma P V$$

where V and R are column vectors with one entry per state and P is the state transition matrix.

$$\begin{bmatrix} V(1) \\ \vdots \\ V(n) \end{bmatrix} = \begin{bmatrix} R(1) \\ \vdots \\ R(n) \end{bmatrix} + \gamma \begin{bmatrix} P_{11} & \cdots & P_{1n} \\ \vdots & \ddots & \vdots \\ P_{n1} & \cdots & P_{nn} \end{bmatrix} \begin{bmatrix} V(1) \\ \vdots \\ V(n) \end{bmatrix}$$

Solving the Bellman Equation

- The Bellman equation is a **linear** equation
- It can be solved **directly**

$$
\begin{aligned}
V &= R + \gamma P V \\
(I - \gamma P) V &= R \\
V &= (I - \gamma P)^{-1} R
\end{aligned}
$$

- Computational **complexity** is $O(n^3)$ for n states
- Direct solution only possible for **small** MRPs
- There are many **iterative methods** for large MRPs, e.g.,
  - Dynamic programming
  - Monte–Carlo evaluation
  - Temporal–Difference learning

Discrete–time Finite Markov Decision Processes

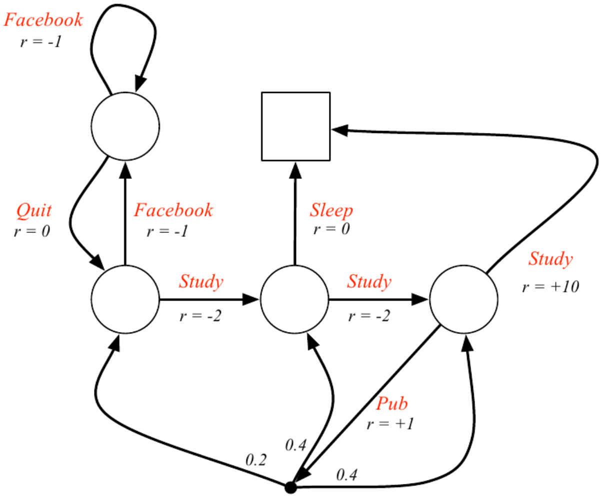
A Markov decision process (MDP) is Markov reward process with **decisions.** It models an environment in which all states are Markov and time is divided into **stages.**

Definition

A **Markov Process** is a tuple $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma, \mu \rangle$

- $\mathcal{S}$ is a (finite) set of states
- $\mathcal{A}$ is a (finite) set of actions
- P is a state transition probability matrix, $P(s'|s, a)$
- R is a reward function, $R(s, a) = \mathbb{E}[r|s, a]$
- γ is a discount factor, $\gamma \in [0, 1]$
- a set of initial probabilities $\mu_i^0 = P(X_0 = i)$ for all i

Example: Student MDP



Goals and Rewards

- Is a scalar **reward** an adequate notion of a **goal**?
  - **Sutton hypothesis**: That all of what we mean by goals and purposes can be well thought of as the maximization of the cumulative sum of a received scalar signal (reward)
  - Probably ultimately wrong, but so **simple** and **flexible** we have to disprove it before considering anything more complicated
- A goal should specify **what** we want to achieve, not **how** we want to achieve it
- The same goal can be specified by (infinite) **different reward functions**
- A goal must be outside the agent's direct control – thus outside the agent
- The agent must be able to measure success:
  - **explicitly**
  - **frequently** during her lifespan

Policies

- A policy, at any given point in time, **decides** which action the agent selects
- A policy fully defines the **behavior** of an agent
- Policies can be:
  - Markovian $\subseteq$ History–dependent
  - Deterministic $\subseteq$ Stochastic
  - Stationary $\subseteq$ Non–stationary

Stationary Stochastic Markovian Policies

Definition

A policy π is a **distribution over actions** given the state:

$$\pi(a|s) = \mathbb{P}[a|s]$$

- MDP policies depend on the **current state** (not the history)
- i.e., Policies are **stationary** (time–independent)
- Given an MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, R, \gamma, \mu \rangle$ and a policy π
  - The state sequence $s_1, s_2, \ldots$ is a **Markov process** $\langle \mathcal{S}, P^{\pi}, \mu \rangle$
  - The state and reward sequence $s_1, r_2, s_2, \ldots$ is a **Markov reward process** $\langle \mathcal{S}, P^{\pi}, R^{\pi}, \gamma, \mu \rangle$, where

$$P^{\pi}(s'|s) = \sum_{a \in \mathcal{A}} \pi(a|s) P(s'|s, a) \quad R^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a|s) R(s, a)$$

Value Functions

Given a policy π, it is possible to define the **utility** of each state: **Policy Evaluation**

Definition

The state–value function $V^\pi(s)$ of an MDP is the expected return starting from state s, and then following policy π

$$V^\pi(s) = \mathbb{E}_\pi[v_t | s_t = s]$$

For **control purposes**, rather than the value of each state, it is easier to consider **the value of each action** in each state

Definition

The action–value function $Q^\pi(s, a)$ is the expected return starting from state s, taking action a, and then following policy π

$$Q^\pi(s, a) = \mathbb{E}_\pi[v_t | s_t = s, a_t = a]$$

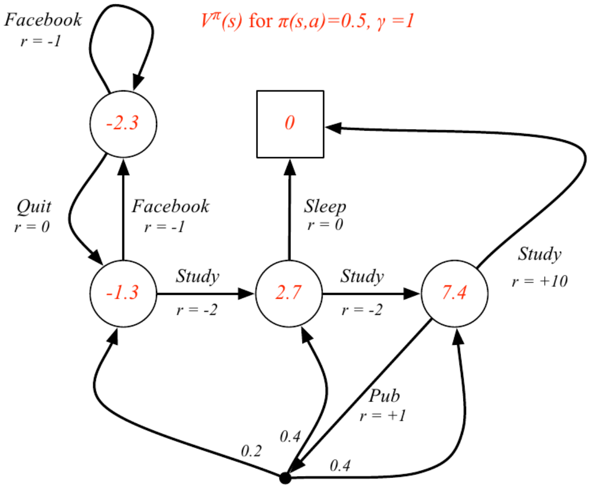
Example: Value Function of Student MDP

Marcello Restelli

Markov Processes

Markov Reward Processes

Markov Decision Processes



Bellman Expectation Equation

The state–value function can again be **decomposed** into immediate reward plus discounted value of successor state,

$$\begin{aligned} V^\pi(s) &= \mathbb{E}_\pi[r_{t+1} + \gamma V^\pi(s_{t+1}) | s_t = s] \\ &= \sum_{a \in A} \pi(a|s) \left(R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^\pi(s') \right) \end{aligned}$$

The action-value function can similarly be decomposed

$$\begin{aligned} Q^\pi(s,a) &= \mathbb{E}_\pi[r_{t+1} + \gamma Q^\pi(s_{t+1}, a_{t+1}) | s_t = s, a_t = a] \\ &= R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^\pi(s') \\ &= R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) \sum_{a' \in A} \pi(a'|s') Q^\pi(s',a') \end{aligned}$$

Bellman Expectation Equation (Matrix Form)

The Bellman expectation equation can be expressed **concisely** using the induced MRP

$$V^{\pi} = R^{\pi} + \gamma P^{\pi} V^{\pi}$$

with **direct solution**

$$V^{\pi} = (I - \gamma P^{\pi})^{-1} R^{\pi}$$

Bellman operators for V^π

Definition

The Bellman operator for V^π is defined as $T^\pi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ (maps value functions to value functions):

$$(T^\pi V^\pi)(s) = \sum_{a \in A} \pi(a|s) \left(R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^\pi(s') \right)$$

- Using Bellman operator, Bellman expectation equation can be **compactly** written as:

$$T^\pi V^\pi = V^\pi$$

- V^π is a **fixed point** of the Bellman operator T^π
- This is a **linear equation** in V^π and T^π
- If $0 < \gamma < 1$ then T^π is a **contraction** w.r.t. the maximum norm

Bellman operators for Q^π

Definition

The Bellman operator for Q^π is defined as $T^\pi : \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ (maps action–value functions to action–value functions):

$$(T^\pi Q^\pi)(s,a) = R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) \sum_{a' \in A} \pi(a'|s') \left(Q^\pi(s',a')\right)$$

- Using Bellman operator, Bellman expectation equation can be compactly written as:

$$T^\pi Q^\pi = Q^\pi$$

- Q^π is a fixed point of the Bellman operator T^π
- This is a linear equation in Q^π and T^π
- If $0 < \gamma < 1$ then T^π is a contraction w.r.t. the maximum norm

Optimal Value Function

Definition

The **optimal state–value function** $V^*(s)$ is the maximum value function over all policies

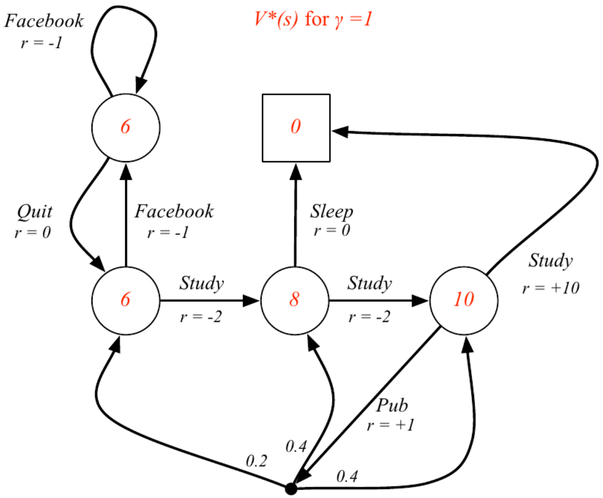
$$V^*(s) = \max_{\pi} V^{\pi}(s)$$

The **optimal action–value function** $Q^*(s, a)$ is the maximum action–value function over all policies

$$Q^*(s, a) = \max_{\pi} Q^{\pi}(s, a)$$

- The optimal value function specifies the **best** possible performance in the MDP
- An MDP is "**solved**" when we know the optimal value function

Example: Optimal Value Function of Student MDP



Optimal Policy

Value functions define a partial ordering over policies

$$\pi \geq \pi' \text{ if } V^{\pi}(s) \geq V^{\pi'}(s), \quad \forall s \in \mathcal{S}$$

Theorem

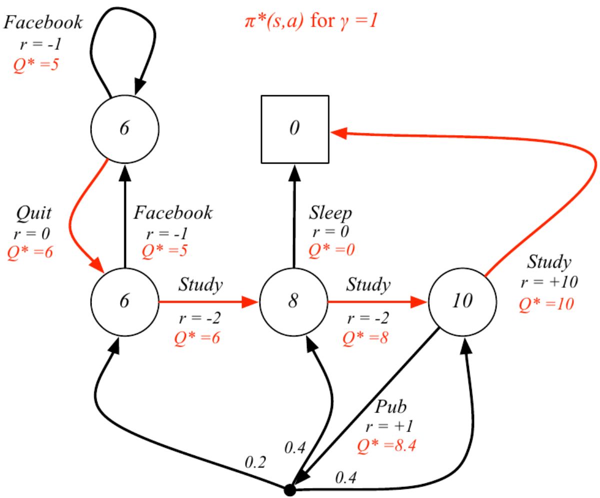
For any Markov Decision Process

- *There exists* ***an optimal policy*** π^* *that is better than or equal to all other policies* $\pi^* \geq \pi, \quad \forall \pi$
- ***All*** *optimal policies achieve the* ***optimal value function****,* $V^{\pi^*}(s) = V^*(s)$
- ***All*** *optimal policies achieve the* ***optimal action–value function****,* $Q^{\pi^*}(s,a) = Q^*(s,a)$
- *There is always a* ***deterministic optimal policy*** *for any MDP*

A deterministic optimal policy can be found by maximizing over $Q^*(s,a)$

$$\pi^*(a|s) = \begin{cases} 1 & \text{if } a = arg\max_{a \in \mathcal{A}} Q^*(s,a) \\ 0 & \text{otherwise} \end{cases}$$

Example: Optimal Policy for Student MDP



Bellman Optimality Equation

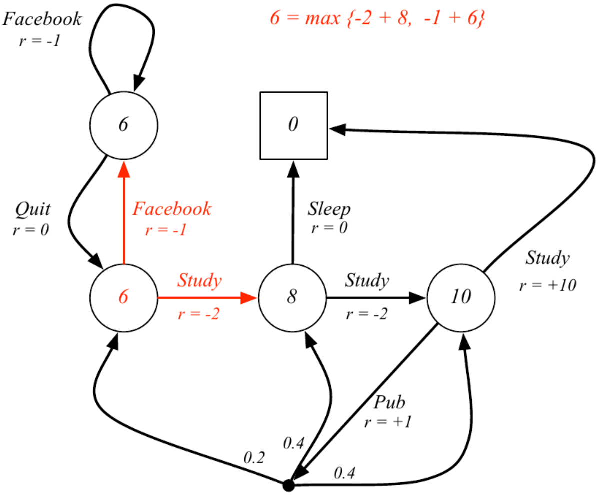
Bellman Optimality Equation for $V*$

$$
\begin{aligned}
V^*(s) &= \max_a Q^*(s,a) \\
&= \max_a \left\{ R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V^*(s') \right\}
\end{aligned}
$$

Bellman Optimality Equation for $Q*$

$$
\begin{aligned}
Q^*(s,a) &= R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V^*(s') \\
&= R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) \max_{a'} Q^*(s',a')
\end{aligned}
$$

Example: Bellman Optimality Equation in Student MDP



Bellman Optimality Operator

Definition

The Bellman optimality operator for V^* is defined as $T^* : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ (maps value functions to value functions):

$$(T^* V^*)(s) = \max_{a \in A} \left(R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^*(s') \right)$$

Definition

The Bellman optimality operator for Q^* is defined as $T^* : \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$ (maps action–value functions to action–value functions):

$$(T^* Q^*)(s,a) = R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) \max_{a'} Q^*(s',a')$$

Properties of Bellman Operators

- **Monotonicity**: if $f_1 \leq f_2$ component–wise

$$T^\pi f_1 \leq T^\pi f_2 \quad , \quad T^* f_1 \leq T^* f_2$$

- **Max–Norm Contraction**: for two vectors f_1 and f_2

$$\| T^\pi f_1 - T^\pi f_2 \|_\infty \leq \gamma \| f_1 - f_2 \|_\infty$$

$$\| T^* f_1 - T^* f_2 \|_\infty \leq \gamma \| f_1 - f_2 \|_\infty$$

- V^π is the **unique fixed point** of T^π
- V^* is the **unique fixed point** of T^*
- For any vector $f \in \mathbb{R}^{|\mathcal{S}|}$ and any policy π, we have

$$\lim_{k \to \infty} (T^\pi)^k f = V^\pi \quad , \quad \lim_{k \to \infty} (T^*)^k f = V^*$$

Solving the Bellman Optimality Equation

- Bellman optimality equation is **non–linear**
- **No closed form** solution for the general case
- Many **iterative** solution methods
  - Dynamic Programming
    - Value Iteration
    - Policy Iteration
  - Linear Programming
  - Reinforcement Learning
    - Q–learning
    - SARSA

Extensions to MDP

- Undiscounted, average reward MDPs
- Infinite and continuous MDPs
- Partially observable MDPs
- Semi-MDPs
- Non–stationary MDPs, Markov games