Risk-Aware Markov Decision Processes

Jefferson Huang, PhD
Assistant Professor
Department of Operations Research
Naval Postgraduate School

This work is sponsored by the Office of Naval Research.
(Grant N0001425GI01179)

MORS Emerging Techniques Forum

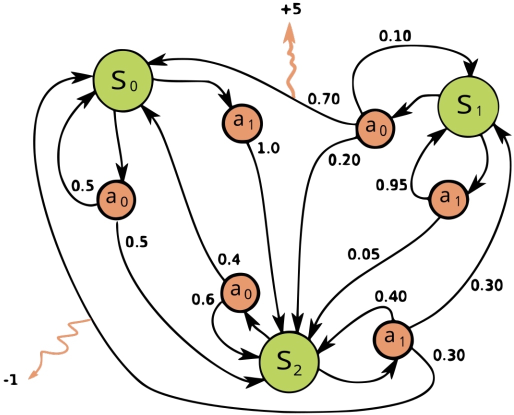
Alexandria, VA
4 December, 2025

Markov Decision Processes (MDPs)

A **Markov decision process (MDP)** models a *sequential decision-making* problem that is subject to *uncertainty*.

- A decision is made at each time step $t = 0, 1, \ldots$
- On each time step t, the decision-maker (DM):
  1. **Observes** the current state $S_t \in \mathcal{S}$ (e.g., of the system being controlled).
  2. **Selects** an action $A_t \in \mathcal{A}$.
- As a result of the action A_t being taken when the state is S_t,
  1. a one-step **cost** $c(S_t, A_t)$ is incurred, and
  2. the system moves to a (possibly) new state S_{t+1} according to **transition probabilities** $p(S_{t+1}|S_t, A_t)$.

Example: Notional MDP



Source: Wikipedia

Example: Network Component

A network component (e.g., a pipe segment in a water distribution network) is either New, Old, or Failed.

After observing the component's state, the DM may Do Nothing, Repair it, or Replace it.

While the component is operational (i.e., not Failed), an **operating cost** c_o is incurred. While the component has Failed, a **failure cost** $c_f > c_o$ is incurred.

Depending on the component's state, it is subject to certain deterioration/failure rates.

(See the NPS Master's thesis by Stisser (2025).)

$\mathcal{S} = \{\text{New, Old, Fail}\}$

$\mathcal{A} = \{\text{DN, Repa, Repl}\}$

S_t	A_t	$c(S_t, A_t)$	S_{t+1}	$p(S_{t+1} \vert S_t, A_t)$
New	DN	c_o	New	$1 - p_2 - q_2$
New	DN	c_o	Old	p_2
New	DN	c_o	Fail	q_2
New	Repa	$c_o + c_1$	New	1
New	Repl	$c_o + c_2$	New	1
Old	DN	c_o	Old	$1 - q_1$
Old	DN	c_o	Fail	q_1
Old	Repa	$c_o + c_1$	Old	1
Old	Repl	$c_o + c_2$	New	1
Fail	DN	c_f	Fail	1
Fail	Repa	$c_f + c_1$	Old	1
Fail	Repl	$c_f + c_2$	New	1

Example: Network-Level Maintenance

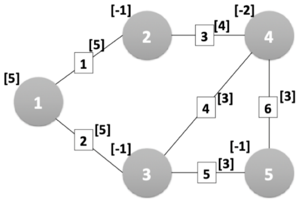


Figure 4.1. This represents a simple fuel network with a supply node (N1), four demand nodes (N2:N5), and six edges. Each of these edges represents a pipe, or component, in the fuel network and are labeled 1–6 in the boxes on each arc. The demand/supply is displayed in brackets above each node and the arc capacity in brackets above each arc. The flow cost is one unit per unit of flow, and the unmet demand penalty is ten per unit of unmet demand.

Source: Stisser (2025)

Each link in the network is a “component”.

- Network-level action is selection of which components to repair or replace.
- Network-level cost is the **min-cost flow** cost from source(s) to sink(s).
- Each component’s state evolves independently according to its own deterioration/failure rates.

Making Decisions via Policies

A **policy** π provides, for each state $s \in \mathcal{S}$, a recommended action $a = \pi(s) \in \mathcal{A}$.

- Let Π be the set of all policies.

Solving an MDP typically means finding a policy that minimizes the **expected value** of the total discounted* cost that will be incurred for each possible starting state:

$$\underset{\pi \in \Pi}{\text{minimize}} \quad \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t c(S_t, \pi(S_t)) \,\middle|\, S_0 = s \right], \quad s \in \mathcal{S}$$

- It has been known since the 1950s† that when $\mathcal{S}$ and $\mathcal{A}$ are finite, there exists a policy π_* that is **optimal** for all starting states simultaneously.
- An optimal policy can be computed by solving a linear program.

*The **discount factor** $\gamma \in [0, 1)$ captures the extent to which near-term costs are more important than longer-term costs.
†A standard reference on MDP theory is Puterman (2005).

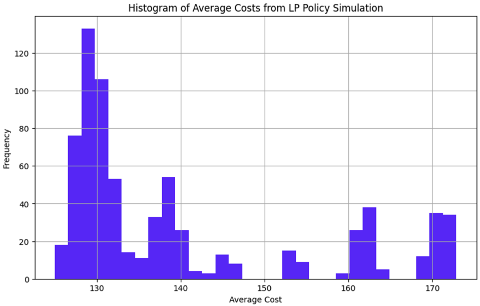
Accounting for Risk

The total discounted cost that will be incurred is a random variable C:

$$C = \sum_{t=0}^{\infty} \gamma^t c(S_t, \pi(S_t))$$

The expected value of C does not capture the shape of the distribution of C.

- The histogram below is the estimated distribution of C under an optimal policy for a 6-component network considered in Stisser (2025).



Risk Measures

Idea: Replace the expected value with something else that accounts for what's going on in the "tail" of the distribution of C.

- In finance, that "something else" is often the **conditional value-at-risk (CVaR)**.
- Given $\alpha \in (0, 1]$, the conditional value-at-risk

$$\mathrm{CVaR}_\alpha(C)$$

  can be interpreted as the conditional expectation of C, given that it exceeds the $100 \cdot (1 - \alpha)^{\text{th}}$ percentile of the distribution of C.
  - decreasing $\alpha \implies$ more risk-averse.
- CVaR is an example of a "coherent" risk measure.[‡]

[‡]A standard reference on risk measures is Chapter 4 of Föllmer & Scheid (2002).

Replacing Expectation with CVaR in MDPs

Problem: Standard techniques for analyzing and solving MDPs no longer apply.

- The "dynamic programming principle" may fail to hold.[§]
- Unclear whether it's possible compute an optimal policy via linear programming.

There are two main approaches so far:

- Formulate and solve a (difficult) parametric optimization problem.[¶]
- Work with a related **two-player zero-sum Markov game** against "nature". (Our project.)

[§] See Hau et al. (2023), who provide a counterexample.
[¶] See Section 5 of Bäuerle & Jaśkiewicz (2024).

The Related Markov Game

The DM plays against nature, who "allocates risk" over the states $s \in \mathcal{S}$ of the original decision process given the DM's action.

- The "state" of the game is a pair (s, y), where $y \in [0, 1]$ is the DM's current "risk level".
- This game was originally proposed by Chow et al. (2015), who (incorrectly) claimed that a solution of the game can be used to construct a CVaR-optimal policy.
- There are (at least) two issues:
  - The solution of the game does not necessarily correspond to a CVaR-optimal policy.
  - In the original decision process, the DM can see the initial risk level α, but cannot see the evolving risk levels y.
- Our project is focused on addressing these issues.[‖]

[‖] Some details can be found in Feinberg & Ding (2025).

Some Results (So Far)

- It is possible to convert a solution for the DM in the game to a corresponding policy that can be implemented in the original MDP.
- The value of the game provides a lower bound on the minimum achievable CVaR under any policy.
  - Can **bound** the sub-optimality of any policy.
- Much more to follow . . .