Risk-Aware Markov Decision Processes

Jefferson Huang, PhD

Assistant Professor
Department of Operations Research
Naval Postgraduate School

STOR-i Annual Conference

Lancaster University

Lancaster, UK
8 January, 2026

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 $A_t \in \mathcal{A}$.
- After the action A_t is taken:
  1. The system moves to a (possibly) new state S_{t+1} according to **transition probabilities** $p(S_{t+1}|S_t, A_t)$.
  2. A **one-step cost** $c(S_t, A_t, S_{t+1})$ is incurred.

Example: Network Component

A network component (e.g., a link in a fuel 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).

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

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

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

Example: Network-Level Maintenance

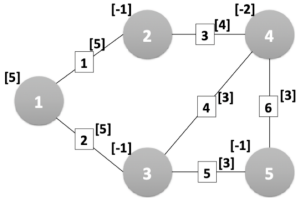


Figure 4.1. This represents a simple fuel network with a supply node (N1), four demand nodes (N2:N5), and six edges. Each of these edges represents a pipe, or component, in the fuel network and are labeled 1–6 in the boxes on each arc. The demand/supply is displayed in brackets above each node and the 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*: selection of which components to repair or replace.
- Network-level *one-step cost*: **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 maps each *history* $H_t = (S_0, \ldots, A_{t-1}, S_t)$ of the process to a recommended action.

- Let Π be the set of all (deterministic and history-dependent) policies.

Solving an MDP usually means finding a policy that minimizes the **expected value** of the total discounted* cost that will be incurred, for each possible starting state s_0:

$$\underset{\pi \in \Pi}{\text{minimize}} \quad E\left[\sum_{t=0}^{\infty} \beta^t c(S_t, \pi(H_t), S_{t+1}) \;\middle|\; S_0 = s_0 \right], \quad s_0 \in \mathcal{S}$$

- It has been known since the 1950s† that when $\mathcal{S}$ and $\mathcal{A}$ are finite, there exists a Markovian policy $\pi_* : \mathcal{S} \to \mathcal{A}$ that is **optimal** for all starting states that can be computed, e.g., 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).

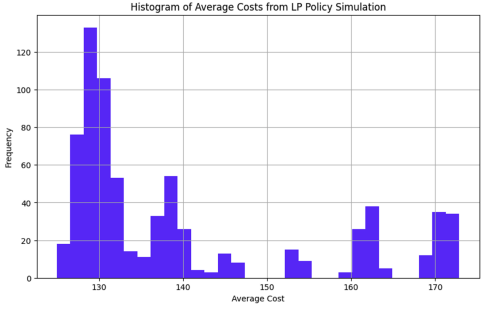
Accounting for Risk

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

$$C_\beta = \sum_{t=0}^{\infty} \beta^t c(S_t, \pi(H_t), S_{t+1})$$

The expected value of C_β does not capture the shape of the **distribution** of C_β.

Example: Estimated distribution of C_β under an optimal policy for a 6-component network considered in Stisser (2025):



Risk Measures

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

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

$$\mathrm{CVaR}_\alpha(C_\beta)$$

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

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

Conditional Value-at-Risk (CVaR)

The **conditional value-at-risk** of a random variable X at a risk level of $\alpha \in [0,1]$ is

$$\mathrm{CVaR}_\alpha(X) = \begin{cases} \operatorname{ess\,sup}(X) & \text{if } \alpha = 0 \quad (\text{“worst-case”}) \\ \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_q(X)\ dq & \text{if } 0 < \alpha < 1 \\ E(X) & \text{if } \alpha = 1 \quad (\text{“average-case”}) \end{cases}$$

where for $q \in (0,1)$, the corresponding **value-at-risk**

$$\mathrm{VaR}_q(X) = \inf\{x \in \mathbb{R} \mid P(X \leqslant x) \geqslant 1-q\}$$

is the $100(1-q)^{\text{th}}$ percentile of X.

- When X is a *continuous* random variable,

$$\mathrm{CVaR}_\alpha(X) = E\left[X \mid X \geqslant \mathrm{VaR}_\alpha(X)\right]$$

Optimization Representations of CVaR

- Rockafellar & Uryasev (2000) showed that for $\alpha \in (0,1)$, the conditional value-at-risk $\mathrm{CVaR}_\alpha(X)$ has a "**primal** representation" as the solution to an optimization problem:

$$\mathrm{CVaR}_\alpha(X) = \inf_{w \in \mathbb{R}} \left\{ w + \frac{1}{\alpha} E(|X - w|^+) \right\}$$

- When X is square-integrable, there is a corresponding "**dual** representation"

$$\mathrm{CVaR}_\alpha(X) = \sup_{B \in \mathcal{U}(\alpha)} E(BX),$$

where the optimization is over the "**risk envelope**"

$$\mathcal{U}(\alpha) = \{\text{random variables } B \mid \alpha B \in [0,1] \text{ almost surely, and } E(B) = 1\}.$$

Replacing Expectation with CVaR in MDPs

Question: Given a risk level $\alpha \in (0, 1)$, how do we find a policy that minimizes the conditional value-at-risk

$$\mathrm{CVaR}_\alpha(C_\beta)$$

of the total discounted cost C_β?

- The "primal representation" of $\mathrm{CVaR}_\alpha(C_\beta)$ leads to a (difficult) parametric optimization problem; see e.g., Bäuerle & Jaśkiewicz (2024).
- The "dual representation" motivates a **Markov game** formulation.
  - Chow, Tamar, Mannor, & Pavone (2015) used a "temporal decomposition"[§] of $\mathrm{CVaR}_\alpha(C_\beta)$ to formulate a Markov game where the decision-maker plays against an adversary that sequentially "allocates risk" to the states of the MDP.
  - Hau, Delage, Ghavamzadeh, & Petrik (2023) showed that the value of this Markov game is not necessarily the minimum achievable $\mathrm{CVaR}_\alpha(C_\beta)$.
  - The discrepancy is due to the fact that in the optimality equation for the game, "min-max does not equal max-min".

[§] The decomposition is due to Pflug & Pichler (2016).

The Markov Game

The **optimality equation** of the Markov game (MG) is

$$v(s,y) = \min_{a\in\mathcal{A}} \max_{\mathbf{b}\in\mathcal{U}(s,y,a)} \left\{ \sum_{s'\in\mathcal{S}} [c(s,a,s') + \beta v(s', y\mathbf{b}_{s'})]\, \mathbf{b}_{s'}\, p(s'|s,a) \right\} \qquad \text{for } \; s\in\mathcal{S},\; y\in[0,1].$$

- States of the MG are $(s,y)\in\mathcal{S}\times[0,1]$, where y is the "**current risk level**".
- The decision-maker has same available actions $a\in\mathcal{A}$ as in the original MDP.
- If the current state is $(s,y)\in\mathcal{S}\times[0,1]$ and the decision-maker takes action $a\in\mathcal{A}$, then the adversary's set of available actions is the "risk envelope"

$$\mathcal{U}(s,y,a) := \left\{ \mathbf{b}\in\mathbb{R}^{|\mathcal{S}|} \;\middle|\; y\mathbf{b}_{s'}\in[0,1]\ \forall s'\in\mathcal{S},\ \text{and} \sum_{s'\in\mathcal{S}} \mathbf{b}_{s'}\, p(s'|s,a) = 1 \right\}.$$

Some Open(?) Questions

- How do we deal with the fact that after time $t = 0$, the "current risk levels" are **un-observable** in practice?[¶]
- How well can we do if we can only use (deterministic and stationary) Markovian policies?[‖]
- How well does the value of the Markov game work as a "dual bound" for the purpose of obtaining performance bounds for sub-optimal polices?[**]

[¶] Feinberg & Ding (2025) use properties of the finite-horizon value functions to construct optimal history-dependent policies that do not depend on the current risk level.

[‖] Feinberg & Ding (2025) showed that it suffices to consider deterministic *history-dependent* policies.

[**] The value function of the game is a lower bound on $\mathrm{CVaR}_{\alpha}(C_{\beta})$.