DPRL-Assignment 2

Federico Donati, Filippo Donghi

November 2025

1 Problem description and modelling assumptions

We consider a system with two components. Each component has a deterioration level

$$x, y \in \{1, \dots, 10\},$$

where 1 is new and 10 is failure. At every time step, and only when not being repaired, each component independently deteriorates to the next level with probability $p = 0.1$ (capped at 10). A component in state 10 fails and is immediately sent for corrective repair.

The system produces one unit of reward in any time step where both components are functioning ($x < 10, y < 10$) and no repair is in progress.

Two repair types are available:

- *Corrective repair*: required when a component is in state 10; cost 25 per failed component.
- *Preventive repair*: optional when $1 \le x, y \le 9$; cost 5 per repaired component.

Any repair takes one full period. Repaired components return to state 1; non-repaired components do not deteriorate during that period. Repairing both components simultaneously still consumes a single time step.

The system is modelled as an MDP $(\mathcal{X}, \mathcal{A}, P, r)$.

$$\mathcal{X} = \{1, \dots, 10\}^2, \qquad s = (x, y).$$

In functioning states ($x < 10, y < 10$) the action set is

$$a^0 = \text{run}, \quad a^1 = \text{repair component 1}, \quad a^2 = \text{repair component 2}, \quad a^3 = \text{repair both}.$$

In failure states, corrective repair of the failed component(s) is mandatory; preventive repair of the healthy component may or may not be allowed depending on the assignment part.

Transitions under a^0 are

$$(x, y) \to (x + i, y + j), \qquad i, j \in \{0, 1\},$$

with probabilities $(1-p)^{2-i-j} p^{i+j}$, clipped at 10. Repair actions deterministically move repaired components to 1 in the next step.

Rewards are

$$r(s, a^0) = 1 \text{ if } x < 10, y < 10, \qquad r(s, a) = -(\text{repair cost}) \text{ otherwise}.$$

The objective in all experiments is to maximise the long-run average reward

$$\liminf_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T-1} r(X_t, A_t)\right].$$

Implementation details

In the implementation we keep the state space equal to $\{1, \dots, 10\}^2$. A repair action (corrective or preventive) consumes exactly one time step and immediately moves each repaired component to state 1; components that are not repaired in that step keep their current state and do not deteriorate during that period. Deterioration under action a^0 is simulated independently for the two components using Bernoulli(p) trials, which induces the transition probabilities above. In all dynamic programming algorithms we work with the relative value function and anchor it by setting $V(1, 1) = 0$ after each iteration. RVI is terminated once $\|V_{\text{new}} - V\|_\infty < 10^{-8}$ or after 10^4 iterations.

2 Problem formulation

Let $V(x, y)$ denote the relative value function and let ϕ denote the optimal long–run average reward. For every state (x, y) we have the Bellman optimality equation

$$V(x, y) + \phi = \max_{a \in \mathcal{A}(x,y)} \Big\{ r(x, y, a) + \sum_{(x', y')} P\big((x', y') \mid (x, y), a\big)\, V(x', y') \Big\}.$$

The optimal action in each state is any maximiser of the expression on the right. The three model variations in parts (b)–(d) differ only in their admissible action sets $\mathcal{A}(x, y)$.

(b) System without preventive repair

(b.1) Simulation. With no preventive repair, the system simply runs whenever $x < 10, y < 10$ and performs corrective repair on any component in state 10. This induces a time-homogeneous Markov chain, which we simulate to estimate ϕ.

(b.2) Stationary distribution. The transition matrix of the induced Markov chain on $\{1, \ldots, 10\}^2$ is constructed explicitly. Let π denote its stationary distribution. Since the chain is unichain under this fixed policy, the long–run average reward is

$$\phi = \sum_{x,y} \pi(x, y)\, r(x, y, a^0).$$

(b.3) Relative value iteration. For the same fixed policy, the Poisson equation is solved using relative value iteration (RVI). At each iteration we update the value function, extract the implied gain

$$\phi = V_{\text{new}}(1, 1) - V(1, 1),$$

and anchor the value function by setting $V(1, 1) = 0$.

(c) Corrective repair with optional preventive repair at failure

In working states ($x < 10, y < 10$) the dynamics are identical to part (b) and only action a^0 is available. When exactly one component fails, one can choose between: (i) repairing only the failed component, or (ii) repairing both (the healthy component preventively). Thus, for $y < 10$,

$$V(10, y) + \phi = \max\bigl\{-25 + V(1, y),\ -30 + V(1, 1)\bigr\},$$

and symmetrically for $(x, 10)$ when $x < 10$. If both components have failed, both are repaired:

$$V(10, 10) + \phi = -50 + V(1, 1).$$

(d) Full preventive maintenance

In this case, the full action set

$$\mathcal{A}(x, y) = \{a^0, a^1, a^2, a^3\}$$

is available in all working states ($x < 10, y < 10$), allowing preventive repair of either or both components at any time. The Bellman equation therefore becomes:

$$\begin{aligned} V(x, y) + \phi = \max\Bigl\{ & 1 + (1-p)^2 V(x, y) + p(1-p)V(x+1, y) \\ & + (1-p)pV(x, y+1) + p^2 V(x+1, y+1), \\ & -5 + V(1, y), \\ & -5 + V(x, 1), \\ & -10 + V(1, 1)\Bigr\}, \qquad x < 10,\ y < 10, \end{aligned}$$

with $x + 1$ and $y + 1$ clipped at 10. Failure states are treated as in part (c), since corrective repair is mandatory for any component in state 10.

3 Results

Table 1 reports the long–run average rewards for all model variants. Parts B.1–B.3 agree to numerical precision, confirming the consistency of the simulation, stationary distribution, and RVI implementations. For the dynamic programming cases we also report the number of relative value iteration (RVI) updates required for convergence.

Method		ϕ^*	**RVI iters**
B.1	Simulation (corrective only)	0.435183	–
B.2	Stationary distribution	0.434835	–
B.3	Value iteration (corrective only)	0.434835	925
C	Limited preventive repair	0.585685	998
D	Full preventive repair	0.854172	651

Table 1: Computed long–run average rewards and convergence statistics.

4 Discussion

The three methods in part B return nearly identical average rewards, showing that the simulation, stationary distribution computation, and RVI solver are consistent and correctly model the corrective-only system. Allowing limited preventive repair in part C yields a clear improvement, since repairing the healthy component during a corrective intervention lowers the probability of near-term repeated failures. Full preventive maintenance in part D delivers the highest performance: repairing components before they approach failure avoids the high corrective cost and almost eliminates downtime. The ordering $\phi_B < \phi_C < \phi_D$ therefore matches the expected intuition that early, inexpensive prevention is preferable to corrective action.

Appendix: Optimal Policies

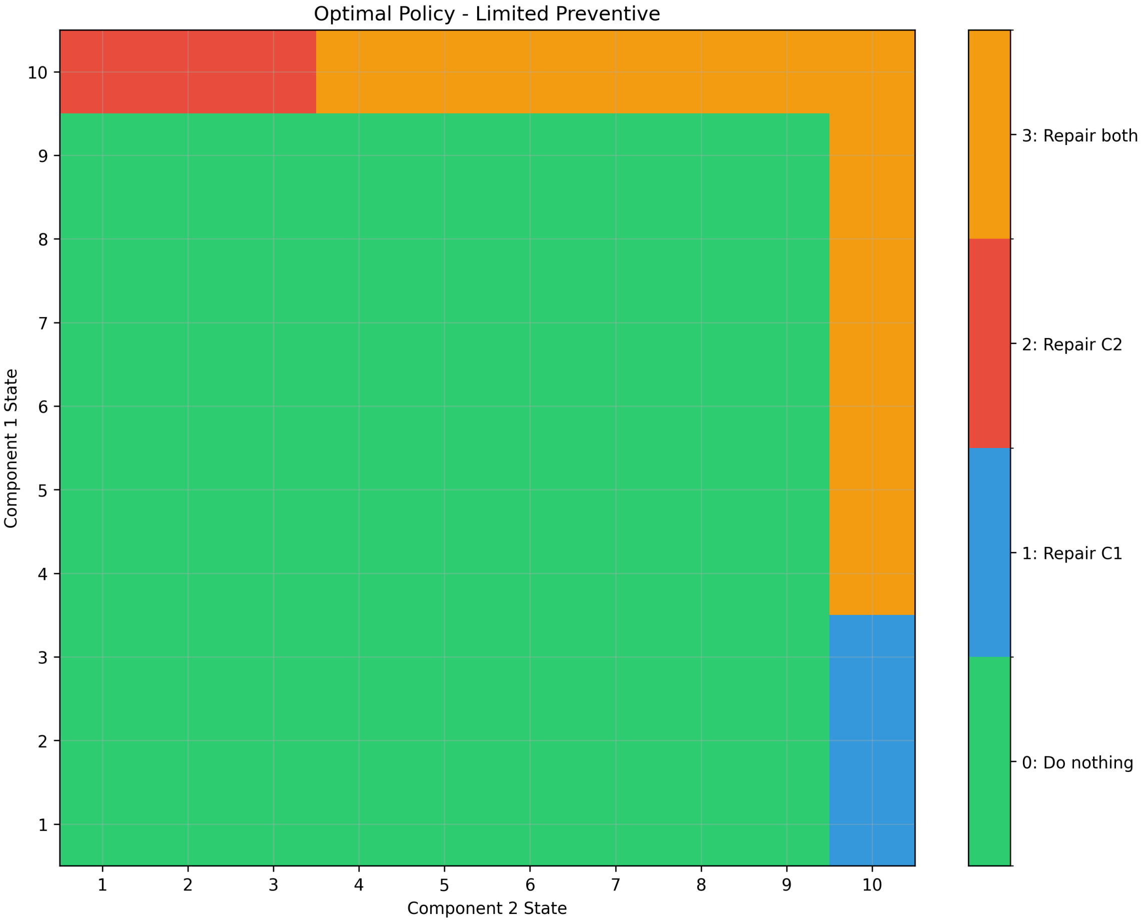

Figure 1: Optimal policy for the limited preventive maintenance model (Part C).

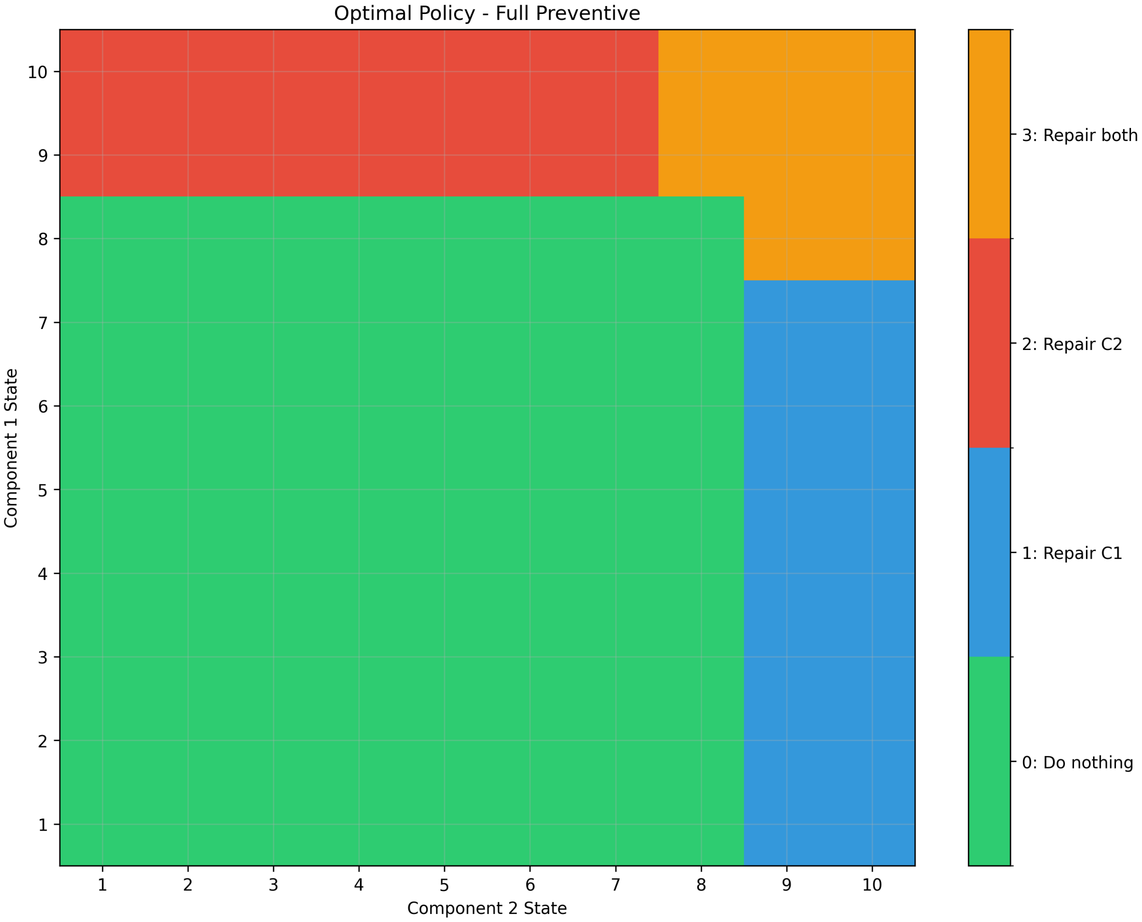

Figure 2: Optimal policy for the full preventive maintenance model (Part D).