Blackwell Optimality in Robust MDPs

Niels van Duijl

Vineet Goyal and Julien Grand-Clément, Robust Markov Decision Process: Beyond Rectangularity, 2023

Julien Grand-Clement, Marek Petrik, Nicolas Vieille, Beyond discounted returns: Robust Markov decision processes with average and Blackwell optimality, 2024

Contents

- Introduction
- Main Paper: Robust Markov Decision Process: Beyond Rectangularity
- Second Paper: Beyond discounted returns: Robust Markov decision processes with average and Blackwell optimality

Introduction

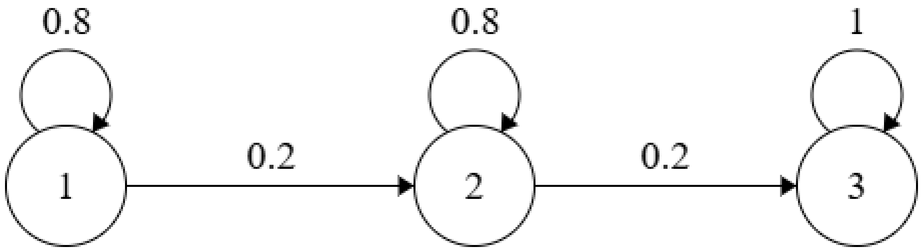
Markov Chains

- Discrete-time stochastic process
- State space ***S***, time horizon ***T***

$$\{X_t, t = 0, 1, 2, \ldots, T\} \quad X_t \in S, \forall t$$

- Transition probabilities ***P***
- **Markov Chain** when P only depends on the current state.

$$P\{X_{t+1} = j \mid X_t = i, X_{t-1} = i_{t-1}, \ldots, X_1 = i_1, X_0 = i_0\} = P_{ij}$$



Markov Decision Process (MDP)

- Markov Chain, but with an action space ***A***
- P_{ij} now depends on the chosen action as well
- Each state-action pair has a reward $r_{s,a}$
- Choose actions based on a policy π
- Stationary policy:

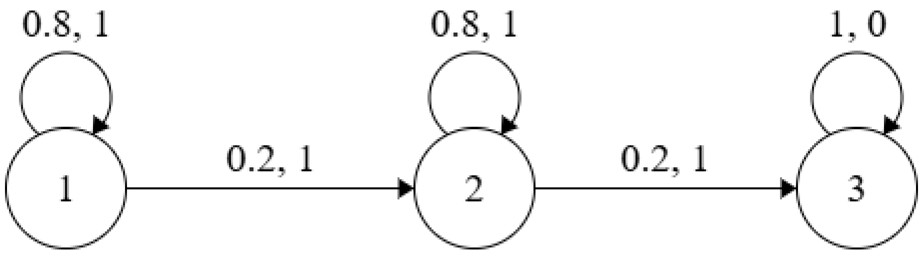
$$\pi = \{\pi_s(a), a \in A, s \in S\}$$

- Deterministic when $\pi_s(a) \in \{0, 1\}$

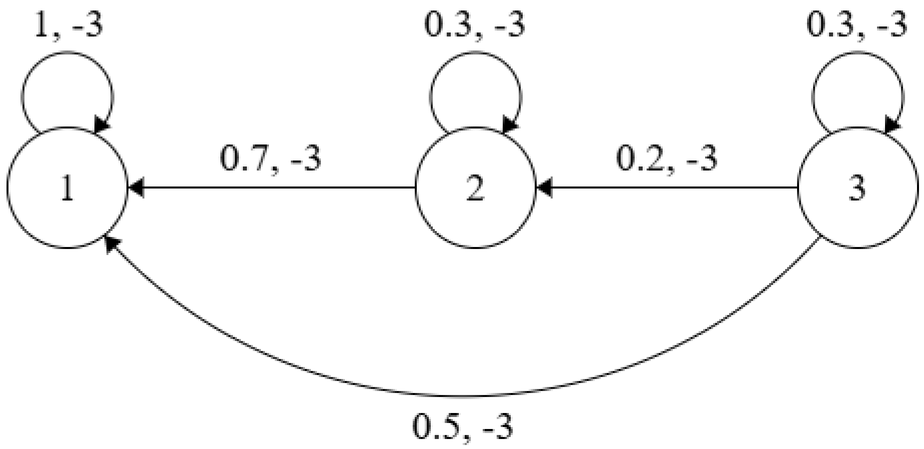
MDP Example

- We now have rewards (second label on edge)
- We have *wait* and *repair* as actions

- P_{ij} when choosing *wait:*



- P_{ij} when choosing *repair:*

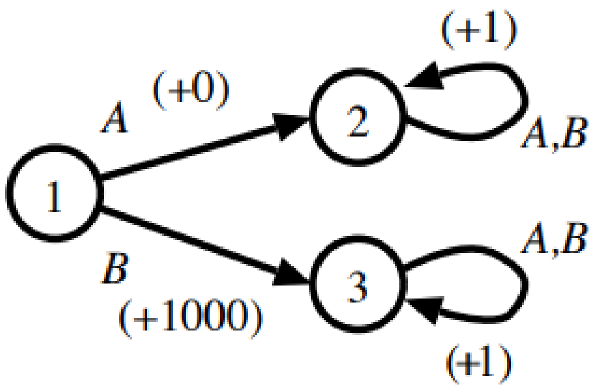


Bellman Equations

- Computing the optimal policy
- How to deal with infinite time horizon?
- Discounted vs average reward:

$$v(\pi^*, s) = \max_{a \in A}\{r_{s,a} + \lambda \sum_{s' \in S} P(s' \mid s, a)\, v(\pi^*, s')\} \quad \lambda \in (0, 1).$$

$$v(\pi^*, s) = \max_{a \in A}\{r_{s,a} + \sum_{s' \in S} P(s' \mid s, a)\, v(\pi^*, s')\} - g(\pi^*).$$

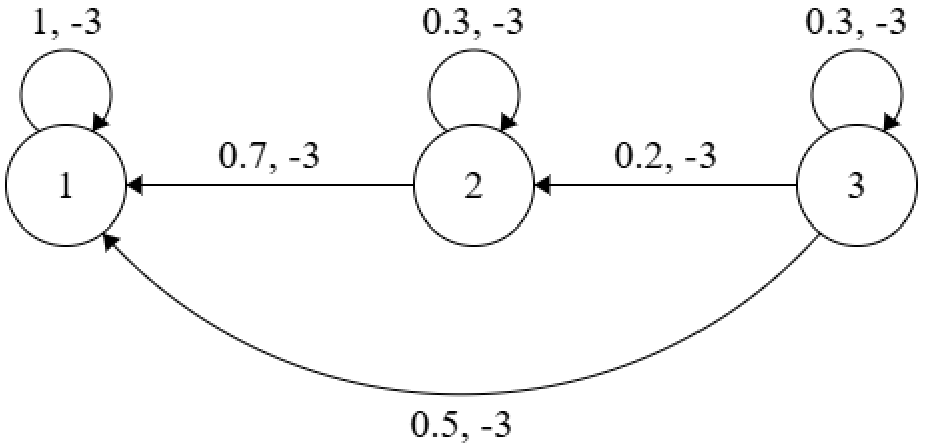


- A policy **π** is **Blackwell optimal** if it is optimal for all **λ** close enough to 1
- These are also optimal policies for the average reward counterpart.
- We will use discounted reward for the rest of the talk

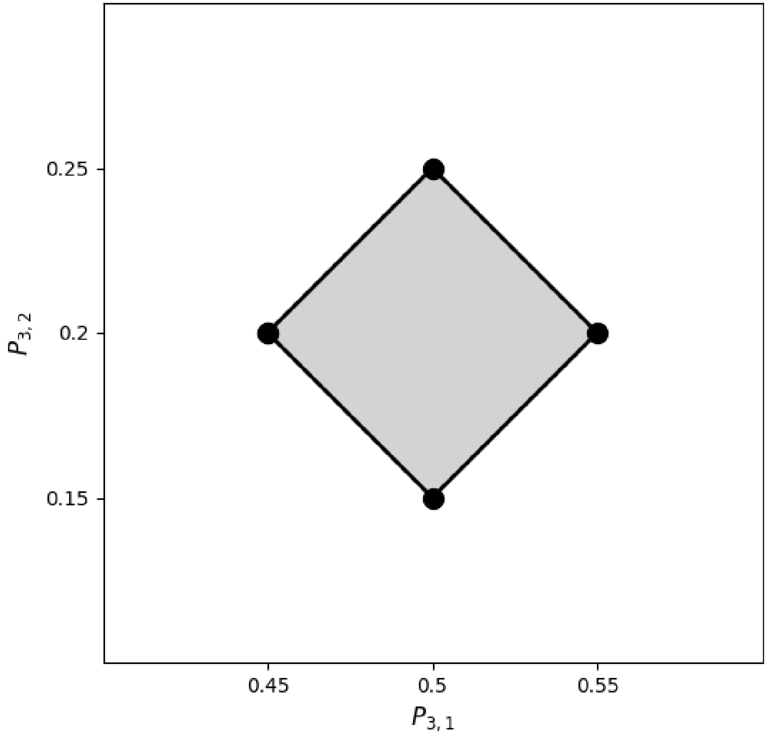
Graph Source: Schwartz, A. (1993). *A Reinforcement Learning Method for Maximizing Undiscounted Rewards*. 298–305. https://doi.org/10.1016/B978-1-55860-307-3.50045-9

Introduction
Robust MDPs

- The probabilities are now uncertain, picked from an uncertainty set $\mathbb{P}$
- Knowing exact probabilities is hard when deriving from data
- ($\boldsymbol{P_{3,3}} = 1 - \boldsymbol{P_{3,2}} - \boldsymbol{P_{3,1}}$)
- $\boldsymbol{P_{ij}}$ when choosing *repair:*



- $\boldsymbol{P_3}$ when choosing *repair:*



Robust MDPs

- The probabilities are now uncertain, picked from an uncertainty set $\mathbb{P}$
- Knowing exact probabilities is hard when deriving from data
- We consider the worst-case probabilities

$$\max_{\pi \in \Pi^G} \min_{\boldsymbol{P} \in \mathbb{P}} R(\pi, \boldsymbol{P}).$$

- Adversary MDP: Second player that controls the factor matrix that tries to minimize our reward
- Two-player game of (normal) MDPs.

Rectangularity

- Generally, it is NP-hard to compute optimal policy
- Independence assumptions are needed: this is rectangularity
- Neglecting constraints only makes the worst-case worse.

(s,a)-rectangularity, s-rectangularity

- (s,a)-rect: Each ***P*** can be chosen from the uncertainty set independently of others
- Optimal policy that is stationary and deterministic, and can be computed efficiently

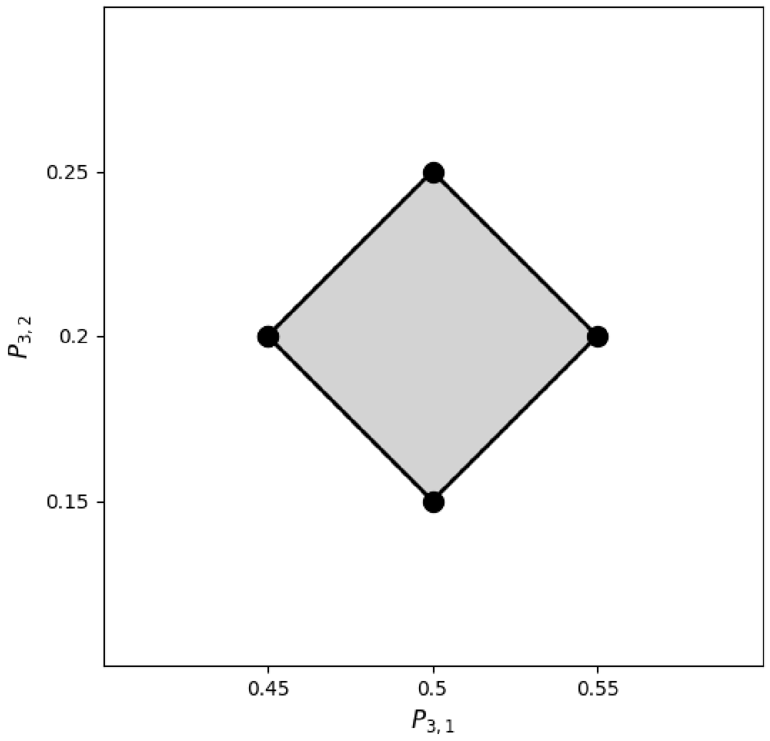
$$\mathbb{P} = \underset{(s,a)\in\mathbb{S}\times\mathbb{A}}{\times} \mathbb{P}_{sa}, \qquad \mathbb{P}_{sa} \subseteq \mathbb{R}_+^{|\mathbb{S}|}.$$

- s-rect: Probs may be dependent on probs for different actions in the same state. Still independent between states.
- Here the optimal policy is stationary but may not be deterministic

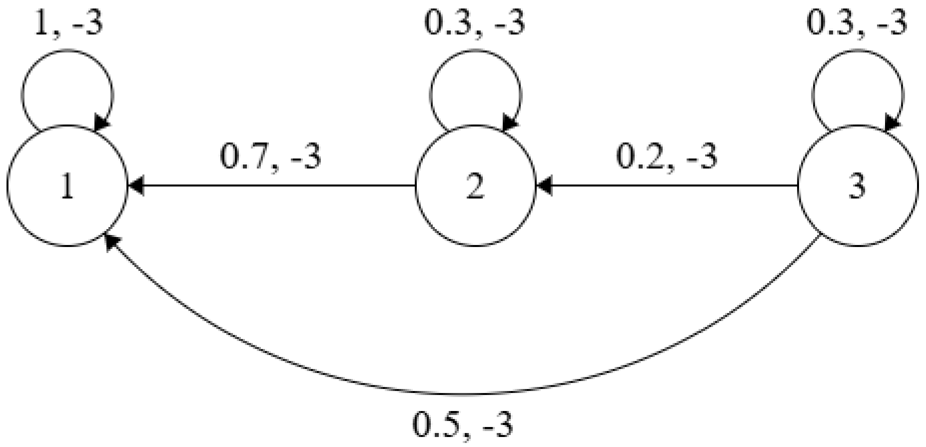
$$\mathbb{P} = \underset{s\in\mathbb{S}}{\times} \mathbb{P}_s, \qquad \mathbb{P}_s \subseteq \mathbb{R}_+^{|\mathbb{S}|\times|\mathbb{A}|}.$$

(s,a)-rectangularity, s-rectangularity

- (s,a)-rect: Each ***P*** can be chosen from the uncertainty set independently of others
- s-rect: Probs may be dependent on probs for different actions in the same state.
- $\boldsymbol{P_3}$ when choosing *repair:*



- $\boldsymbol{P_{ij}}$ when choosing *repair:*



Paper 1: Main Contributions

- **Introduce a new type of rectangularity**
- **Min-max duality:**

$$\max_{\pi \in \Pi} \min_{\boldsymbol{P} \in \mathbb{P}} R(\pi, \boldsymbol{P}) = \min_{\boldsymbol{P} \in \mathbb{P}} \max_{\pi \in \Pi} R(\pi, \boldsymbol{P}).$$

- **Algorithm to compute the optimal policy**
- **Blackwell optimality**

Paper 1: R-rectangularity

A new type of uncertainty set

- Idea: common underlying factors (healthcare)
- Fixed Coefficients ***u***, factors ***w*** themselves are uncertain
- Each factor is a probability distribution over the next state
- ***r*** is not the reward!

$$\mathbb{P} = \left\{ \left(\sum_{i=1}^{r} u_{sa}^{i} w_{i,s'} \right)_{sas'} \,\middle|\, \boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_r) \in \mathcal{W} \subseteq \mathbb{R}^{S \times r} \right\}$$

$$\sum_{i=1}^{r} u_{sa}^{i} = 1,\ \forall\, (s, a) \in \mathbb{S} \times \mathbb{A}, \sum_{s'=1}^{S} w_{i,s'} = 1,\ \forall\, i \in [r],$$

A new type of rectangularity

- r-rectangularity: when the factors are independent

$$\mathcal{W} = \mathcal{W}^1 \times \ldots \times \mathcal{W}^r, \; \textit{where}\; \mathcal{W}^1, \ldots, \mathcal{W}^r \subset \mathbb{R}^S_+ .$$

- (s,a)-rect → r-rect
- s-rect and r-rect not related

$$\mathbb{P} = \left\{ \left(\sum_{i=1}^{r} u^i_{sa} w_{i,s'} \right)_{sas'} \middle| \; \boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_r) \in \mathcal{W} \subseteq \mathbb{R}^{S \times r} \right\}$$

Assumption 2.4

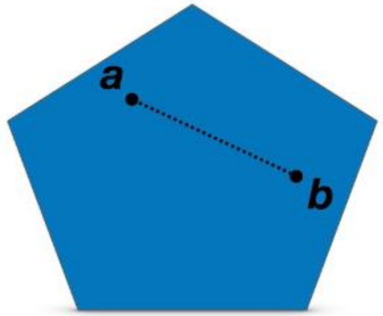
- Used in most results of the rest of the paper
- The uncertainty sets need to be convex compact.

Assumption 2.4 *The sets $\mathcal{W}^1, \ldots, \mathcal{W}^r$ are convex compact. Moreover, for any $i \in [r]$, for any cost vector $\mathbf{c} \in \mathbb{R}^S$, we can find an ϵ-optimal solution of*

$$\min_{\boldsymbol{w}_i \in \mathcal{W}^i} \mathbf{c}^\top \boldsymbol{w}_i$$

in $O\left(comp\left(\mathcal{W}^i\right) \log\left(\epsilon^{-1}\right)\right)$ arithmetic operations, where $comp(\mathcal{W}^i) \in \mathbb{R}$ depends on the structure of the uncertainty set $\mathcal{W}^i$.

- Compact: closed and bounded
- Convex: Every line segment is contained in the set



Paper 1: Evaluating a policy

Adversary MDP

- Adversary MDP: ***r*** States, ***S*** actions, ***W*** is the policy

$$Prob\left(i \underset{action\ s}{\rightarrow} j\right) = \sum_{a \in \mathbb{A}} \pi_{sa} u_{sa}^{j}, \quad Reward\,(i,\ action\ s) = \sum_{a \in \mathbb{A}} \pi_{sa} r_{sa}.$$

$$\boldsymbol{T}_{\pi} = \left(\textstyle\sum_{a \in \mathbb{A}} \pi_{sa} u_{sa}^{i}\right)_{(s,i) \in \mathbb{S} \times [r]} \in \mathbb{R}_{+}^{S \times r}.$$

- Bellman equation of this adversary gives us value function **β** for the adversary:

$$v_s^{\pi} = \sum_{a \in \mathbb{A}} \pi_{sa} \left(r_{sa} + \lambda \boldsymbol{P}_{sa}^{\top} \boldsymbol{v}^{\pi}\right), \forall\, s \in \mathbb{S}, \quad \longrightarrow \quad \beta_i = \boldsymbol{w}_i^{\top}(\boldsymbol{r}_{\pi} + \lambda \cdot \boldsymbol{T}_{\pi} \boldsymbol{\beta}), \forall\, i \in [r].$$

- R-rect is necessary to allow the adversary to independently optimize each vector.

$$\mathbb{P} = \left\{ \left(\sum_{i=1}^{r} u_{sa}^{i} w_{i,s'} \right)_{sas'} \,\middle|\, \boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_r) \in \mathcal{W} \subseteq \mathbb{R}^{S \times r} \right\}$$

Theorem of Duality

- Lemma 4.1: If $\mathbb{P}$ is r-rectangular and the sets $\mathcal{W}^1, \ldots, \mathcal{W}^r$ are convex compact, there exists a stationary, deterministic optimal policy.
- Duality Theorem:

Theorem 4.2 *Under Assumption 2.4, let* $(\pi^*, \boldsymbol{W}^*)$ *be a solution to the robust MDP problem* (1.2) *with* r*-rectangular uncertainty set, with* π^* *deterministic. Then*

$$\boldsymbol{W}^* \in \arg\min_{\boldsymbol{W}\in\mathcal{W}} R(\pi^*, \boldsymbol{W}) \text{ and } \pi^* \in \arg\max_{\pi\in\Pi} R(\pi, \boldsymbol{W}^*). \tag{4.6}$$

Moreover, the following strong min-max duality holds.

$$\max_{\pi\in\Pi} \min_{\boldsymbol{W}\in\mathcal{W}} R(\pi, \boldsymbol{W}) = \min_{\boldsymbol{W}\in\mathcal{W}} \max_{\pi\in\Pi} R(\pi, \boldsymbol{W}). \tag{4.7}$$

- Result: $(\pi^*, \boldsymbol{W}^*)$ is an equilibrium in the two-player game.
- Does not hold for s-rectangular uncertainty sets.

Robust Maximum Principle

- Maximum principle: the optimal policy attains the highest value regardless of starting state.
- Robust equivalent:

Proposition 6.1 *Under Assumption 2.4, let $\mathbb{P}$ be an r-rectangular uncertainty set.*

1. *Let π be a policy and $\boldsymbol{W}^1 \in \arg\min_{\boldsymbol{W} \in \mathcal{W}} R(\pi, \boldsymbol{W})$. Then*

$$v_s^{\pi, \boldsymbol{W}^1} \leq v_s^{\pi, \boldsymbol{W}^0}, \ \forall\ \boldsymbol{W}^0 \in \mathcal{W}, \forall\ s \in \mathbb{S}.$$

2. *Let $(\pi^*, \boldsymbol{W}^*) \in \arg\ \max_{\pi \in \Pi}\ \min_{\boldsymbol{W} \in \mathcal{W}}\ R(\pi, \boldsymbol{W})$. Then*

$$\forall\ \pi \in \Pi, \forall\ \boldsymbol{W}^1 \in \arg\min_{\boldsymbol{W} \in \mathcal{W}} R(\pi, \boldsymbol{W}), v_s^{\pi, \boldsymbol{W}^1} \leq v_s^{\pi^*, \boldsymbol{W}^*}, \forall\ s \in \mathbb{S}.$$

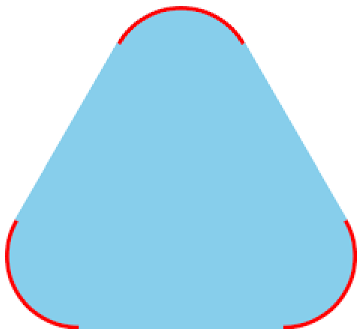
- Follows from the Duality Theorem

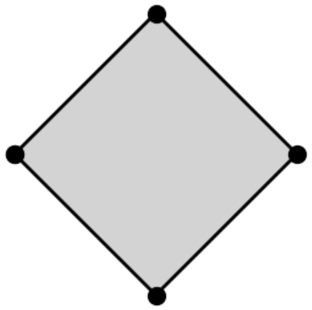
Blackwell Optimality

- Recall: A policy **π** is **Blackwell optimal** if it is optimal for all **λ** close enough to 1
- Proven for r-rect sets with finitely many extreme points.

Proposition 6.2 *Let* $\mathbb{P}$ *be an* r*-rectangular uncertainty set. Assume that the sets* $\mathcal{W}^1, \ldots, \mathcal{W}^r$ *have finitely many extreme points. Then there exists a stationary deterministic policy* π^**, a factor matrix* $\boldsymbol{W}^*$*, and a discount factor* $\lambda_0 \in (0,1)$*, such that for all* $\lambda \in (\lambda_0, 1)$*, the pair* $(\pi^*, \boldsymbol{W}^*)$ *remains an optimal solution to the robust MDP problem* (1.2).

- Proof idea: both **W** and **Π** have finitely many extreme points.
- Extreme point: A point which does not lie in any open line segment joining two points in the set.





Proof of Blackwell Optimality

Lemma G.1 *Let $(u_n)_{n\geq 0}$ be a sequence with values in a finite set $\mathcal{E}$. Then there exists an element $e \in \mathcal{E}$ that is attained infinitely often by $(u_n)_{n>0}$.*

- So, we can construct:
  $\lambda_n \to 1, \text{ and } (\pi^*, \boldsymbol{W}^*) \in \arg\max_{\pi\in\Pi} \min_{\boldsymbol{W}\in\mathcal{W}} R(\pi, \boldsymbol{W}, \lambda_n), \forall\, n \geq 0.$
- By contradiction, suppose we can also construct:
  $\gamma_n \to 1, \text{ and } (\pi^*, \boldsymbol{W}^*) \notin \arg\max_{\pi\in\Pi} \min_{\boldsymbol{W}\in\mathcal{W}} R(\pi, \boldsymbol{W}, \gamma_n), \forall\, n \geq 0.$
- Take $(\tilde{\pi}, \tilde{\boldsymbol{W}})$ to be optimal for all γ_n
- Since $(\pi^*, \boldsymbol{W}^*)$ not optimal, we can find: $v_{\gamma_n,x_1}^{\pi^*,\boldsymbol{W}^{**}} < v_{\gamma_n,x_1}^{\tilde{\pi},\tilde{\boldsymbol{W}}}$

  $v_{\gamma_n,x_1}^{\pi^*,\boldsymbol{W}^{**}} < v_{\gamma_n,x_1}^{\tilde{\pi},\tilde{\tilde{\boldsymbol{W}}}}, v_{\lambda_n,x_1}^{\tilde{\pi},\tilde{\tilde{\boldsymbol{W}}}} \leq v_{\lambda_n,x_1}^{\pi^*,\boldsymbol{W}^{**}}.$

- Continuous, rational function that takes on the value 0 an infinite number of times:

  $f : (0,1) \to \mathbb{R}, f(t) = v_{t,x_1}^{\tilde{\pi},\tilde{\tilde{\boldsymbol{W}}}} - v_{t,x_1}^{\pi^*,\boldsymbol{W}^{**}}.$

Proposition 6.3

- This proof works for any **λ**!
- We conclude the following proposition:

Proposition 6.3 *Let $\mathbb{P}$ be an r-rectangular uncertainty set. Assume that the sets $\mathcal{W}^1, ..., \mathcal{W}^r$ have finitely many extreme points. Then there exists an integer $p \in \mathbb{N}$, there exists some scalars $\lambda_0 = 0 < \lambda_1 < ... < \lambda_p = 1$ such that for all $j \in \{0, ..., p-1\}$, the same pair of stationary deterministic policy and factor matrix $(\pi_j, \boldsymbol{W}_j)$ is an optimal solution to the robust MDP problem* (1.2) *for all $\lambda \in (\lambda_j, \lambda_{j+1})$.*

Paper 1

The rest of the paper

- Algorithm to compute the optimal policy:

$$\boldsymbol{v}^0 = \mathbf{0}, v_s^{k+1} = \max_{a \in \mathbb{A}} \left\{ r_{sa} + \lambda \cdot \sum_{i=1}^{r} u_{sa}^i \min_{\boldsymbol{w}_i \in \mathcal{W}^i} \boldsymbol{w}_i^\top \boldsymbol{v}^k \right\}, \forall\, s \in \mathbb{S}, \forall\, k \geq 0.$$

- Numerical Experiments
- Conclusion: R-rectangularity outperforms s-rectangularity

Paper 2: Summary

- Average optimality for RMDPs
- **Blackwell optimality for RMDPs**
- Algorithms to compute the optimal average reward
- Numerical experiments

- **Does not talk about r-rect.**

Uncertainty set $\mathcal{U}$	Discount optimality	Average optimality	Blackwell optimality
Singleton (MDPs)	stationary, deterministic	stationary, deterministic	stationary, deterministic
sa-rectangular, compact, convex	stationary, deterministic	**stationary, deterministic**	• **may not exist** • $\exists\, \pi$ **stationary deterministic,** π ϵ**-Blackwell optimal,** $\forall \epsilon > 0$ • π **also average optimal**
sa-rectangular, compact, convex, definable	stationary, deterministic	**stationary, deterministic**	• **stationary, deterministic** • π **also average optimal**
s-rectangular, compact convex	stationary, randomized	**history-dependent, randomized**	**may not exist**

Paper 2: Introduction

Motivation

- Less assumptions made than in previous work
- Again, uncertainty set is assumed to be convex, compact.
- Adversary MDP makes the problem similar to stochastic games
- However, the adversary is restricted to stationary policies

Paper 2: Blackwell optimality

Introduction

- Reminder: A policy **π** is **Blackwell optimal** if it is optimal for all discount factors close enough to 1

$$\inf_{\boldsymbol{P}\in\mathcal{U}} R_\gamma(\pi,\boldsymbol{P}) \geqslant \sup_{\pi'\in\Pi_\mathrm{H}} \inf_{\boldsymbol{P}\in\mathcal{U}} R_\gamma(\pi',\boldsymbol{P}), \quad \forall\,\gamma\in(\gamma_0,1).$$

- ϵ-Blackwell optimality:

$$\min_{\boldsymbol{P}\in\mathcal{U}} (1-\gamma)R_\gamma(\pi,\boldsymbol{P}) \geqslant \sup_{\pi\in\Pi_\mathrm{S}} \min_{\boldsymbol{P}\in\mathcal{U}} (1-\gamma)R_\gamma(\pi',\boldsymbol{P}) - \epsilon, \quad \forall\,\gamma\in(\gamma_\epsilon,1).$$

- Normalised and a difference of ϵ is allowed

General (s,a)-rect uncertainty sets

Theorem 4.4 *There exists an sa-rectangular robust MDP instance, with a compact convex uncertainty set $\mathcal{U}$, and with no Blackwell optimal policy:*

$$\forall\, \pi \in \Pi_{\mathsf{H}}, \forall\, \gamma \in (0,1), \exists\, \gamma' \in (\gamma, 1), \min_{\boldsymbol{P}\in\mathcal{U}} R_{\gamma'}(\pi, \boldsymbol{P}) < \sup_{\pi' \in \Pi_{\mathsf{S}}} \min_{\boldsymbol{P}\in\mathcal{U}} R_{\gamma'}(\pi', \boldsymbol{P}).$$

- Based on two distinct sets $\mathcal{U}_{s_0 a_1}$ $\mathcal{U}_{s_0 a_2}$, of which the boundaries intersect infinitely often.

- However: we can find a policy that is ϵ-Blackwell optimal for every $\epsilon > 0$:

Theorem 4.5 *Let $\mathcal{U}$ be an sa-rectangular compact uncertainty set. Then there exists a stationary deterministic policy that is ϵ-Blackwell optimal for all $\epsilon > 0$, i.e., $\exists\, \pi \in \Pi_{\mathsf{SD}}, \forall\, \epsilon > 0, \exists\, \gamma_\epsilon \in (0,1)$ such that*

$$\min_{\boldsymbol{P}\in\mathcal{U}} (1-\gamma) R_\gamma(\pi, \boldsymbol{P}) \geqslant \sup_{\pi' \in \Pi_{\mathsf{S}}} \min_{\boldsymbol{P}\in\mathcal{U}} (1-\gamma) R_\gamma(\pi', \boldsymbol{P}) - \epsilon, \forall\, \gamma \in (\gamma_\epsilon, 1).$$

Definability

- A subset of $\mathbb{R}^n$ is definable if it is of the form:

  $$\{\boldsymbol{x} \in \mathbb{R}^n \,|\exists k \in \mathbb{N}, \exists \boldsymbol{y} \in \mathbb{R}^k,\ P\left(x_1, \ldots, x_n, y_1, \ldots, y_k, \exp(x_1), \ldots, \exp(x_n), \ldots, \exp(y_1), \ldots, \exp(y_k)\right) = 0\}$$

- A function $f : \Omega \to \mathbb{R}^m$, $\Omega \subset \mathbb{R}^n$ is definable if its graph is definable:

  $$\{(\boldsymbol{x}, \boldsymbol{y}) \in \Omega \times \mathbb{R}^m \mid \boldsymbol{y} = f(\boldsymbol{x})\}$$

- Intuitively, a set is definable if it is constructed based on polynomials, the exponential function, and canonical projections (elimination of variables).

- Simple example:

  $$x \geq 0 \longrightarrow P(x, y) = x - y^2$$

Definability

Lemma 4.15 1. *The only definable subsets of* $\mathbb{R}$ *are the* finite *union of open intervals and singletons.*

2. *If* $A, B \subset \mathbb{R}^n$ *are definable sets, then* $A \cup B$, $A \cap B$ *and* $\mathbb{R}^n \backslash A$ *are definable sets.*

3. *Let* f, g *be definable functions. Then* $f \circ g$, $-f$, $f + g$, $f \times g$ *are definable.*

4. *For* A, B *two definable sets and* $g : A \times B \to \mathbb{R}$ *a definable function, then the functions* $\boldsymbol{x} \mapsto \inf_{\boldsymbol{y} \in B} g(\boldsymbol{x}, \boldsymbol{y})$ *and* $\boldsymbol{x} \mapsto \sup_{\boldsymbol{y} \in B} g(\boldsymbol{x}, \boldsymbol{y})$ *(defined over* A*) are definable functions.*

5. *If* $A, B \subseteq \mathbb{R}$ *and* $g : A \to \mathbb{R}$ *are definable then* $g^{-1}(B)$ *is a definable set.*

An example

$$\{\boldsymbol{x} \in \mathbb{R}^n \mid \exists k \in \mathbb{N}, \exists \boldsymbol{y} \in \mathbb{R}^k,\ P\,(x_1, \ldots, x_n, y_1, \ldots, y_k, \exp(x_1), \ldots, \exp(x_n), \ldots, \exp(y_1), \ldots, \exp(y_k)) = 0\}$$

Example 4.16 (ℓ_p-norms are definable.) *Consider an ℓ_p-norm for $p \in \mathbb{N}$. Then its graph is $\{(\boldsymbol{x}, y) \in \mathbb{R}^{\mathcal{S}} \times \mathbb{R} \mid \sum_{s \in \mathcal{S}} |x_s|^p - y^p = 0, y \geqslant 0\}$, which is a definable set. Therefore, ℓ_p-norms are definable.*

- They can have infinitely many extreme points but are definable!

Definable (s,a)-rect uncertainty sets

Theorem 4.17 (Theorem 2.1, Coste [2000]) *Let $f:(a,b)\to\mathbb{R}$ be a definable function. Then there exists a finite subdivision of the interval (a,b) as $a=a_1<a_2<\cdots<a_k=b$ such that on each (a_i,a_{i+1}) for $i=1,\ldots,k-1$, f is continuous and either constant or strictly monotone.*

Proposition 4.19 *Assume that $\mathcal{U}$ is sa-rectangular and definable. Then for any policy $\pi\in\Pi_{\mathsf{S}}$, the function $\gamma\mapsto \boldsymbol{v}_{\gamma}^{\pi,\mathcal{U}}$ is a definable function.*

Theorem 4.25 *Consider an sa-rectangular robust MDP with a definable compact uncertainty set $\mathcal{U}$. Then there exists a stationary deterministic Blackwell optimal policy:*

$$\exists\,\pi\in\Pi_{\mathsf{SD}},\exists\,\gamma_0\in(0,1),\forall\,\gamma\in(\gamma_0,1),\min_{\boldsymbol{P}\in\mathcal{U}}R_\gamma(\pi,\boldsymbol{P})\geqslant\sup_{\pi'\in\Pi_{\mathsf{S}}}\min_{\boldsymbol{P}\in\mathcal{U}}R_\gamma(\pi',\boldsymbol{P}).$$

- Proof based on $v_{\gamma,s}^{\pi,\mathcal{U}}-v_{\gamma,s}^{\pi',\mathcal{U}}$

Proof of Blackwell Optimality

Lemma G.1 *Let $(u_n)_{n\geq 0}$ be a sequence with values in a finite set $\mathcal{E}$. Then there exists an element $e \in \mathcal{E}$ that is attained infinitely often by $(u_n)_{n\geq 0}$.*

- So, we can construct:
  $\lambda_n \to 1, \text{ and } (\pi^*, \boldsymbol{W}^*) \in \arg\max_{\pi\in\Pi} \min_{\boldsymbol{W}\in\mathcal{W}} R(\pi, \boldsymbol{W}, \lambda_n), \forall\, n \geq 0.$
- By contradiction, suppose we can also construct:
  $\gamma_n \to 1, \text{ and } (\pi^*, \boldsymbol{W}^*) \notin \arg\max_{\pi\in\Pi} \min_{\boldsymbol{W}\in\mathcal{W}} R(\pi, \boldsymbol{W}, \gamma_n), \forall\, n \geq 0.$
- Take $(\tilde{\pi}, \tilde{\boldsymbol{W}})$ to be optimal for all γ_n
- Since $(\pi^*, \boldsymbol{W}^*)$ not optimal, we can find: $v_{\gamma_n, x_1}^{\pi^*, \boldsymbol{W}^{**}} < v_{\gamma_n, x_1}^{\tilde{\pi}, \tilde{\boldsymbol{W}}}$

  $v_{\gamma_n, x_1}^{\pi^*, \boldsymbol{W}^{**}} < v_{\gamma_n, x_1}^{\tilde{\pi}, \tilde{\tilde{\boldsymbol{W}}}}, v_{\lambda_n, x_1}^{\tilde{\pi}, \tilde{\tilde{\boldsymbol{W}}}} \leq v_{\lambda_n, x_1}^{\pi^*, \boldsymbol{W}^{**}}.$
- Continuous, rational function that takes on the value 0 an infinite number of times:

  $f : (0,1) \to \mathbb{R}, f(t) = v_{t, x_1}^{\tilde{\pi}, \tilde{\tilde{\boldsymbol{W}}}} - v_{t, x_1}^{\pi^*, \boldsymbol{W}^{**}}.$

Paper 2

The rest of the paper: average optimality results

- Blackwell optimal policies also average optimal
- Algorithms to compute the optimal gain, but not the actual policies
- No polynomial-time algorithm known for (s,a)-rect RMDPs.
- Experiments on the newly defined algorithms.

Uncertainty set $\mathcal{U}$	Discount optimality	Average optimality	Blackwell optimality
Singleton (MDPs)	stationary, deterministic	stationary, deterministic	stationary, deterministic
sa-rectangular, compact, convex	stationary, deterministic	**stationary, deterministic**	• **may not exist** • $\exists \pi$ **stationary deterministic,** π ϵ**-Blackwell optimal,** $\forall \epsilon > 0$ • π **also average optimal**
sa-rectangular, compact, convex, definable	stationary, deterministic	**stationary, deterministic**	• **stationary, deterministic** • π **also average optimal**
s-rectangular, compact convex	stationary, randomized	**history-dependent, randomized**	**may not exist**

Questions?

Paper 2: Average optimality

Introduction

$$R_{\mathsf{avg}}(\pi, \boldsymbol{P}) = \mathbb{E}_{\pi,\boldsymbol{P}}\left[\limsup_{T\to+\infty}\frac{1}{T+1}\sum_{t=0}^{T} r_{s_t a_t s_{t+1}} \mid s_0 \sim \boldsymbol{p}_0\right].$$

- To solve: $\sup_{\pi\in\Pi_{\mathsf{H}}} \inf_{\boldsymbol{P}\in\mathcal{U}} R_{\mathsf{avg}}(\pi, \boldsymbol{P})$
- Inf is used instead of min because it might not exist

Proposition 3.2 *There exists a robust MDP instance with an sa-rectangular compact convex uncertainty set, for which* $\inf_{\boldsymbol{P}\in\mathcal{U}} R_{\mathsf{avg}}(\pi, \boldsymbol{P})$ *is not attained for any* $\pi \in \Pi_{\mathsf{S}}$.

(s,a)-rectangular uncertainty sets

- The optimal policy is stationary and deterministic:

$$\sup_{\pi \in \Pi_{\mathsf{H}}} \inf_{\boldsymbol{P} \in \mathcal{U}} R_{\mathsf{avg}}(\pi, \boldsymbol{P}) = \max_{\pi \in \Pi_{\mathsf{SD}}} \inf_{\boldsymbol{P} \in \mathcal{U}} R_{\mathsf{avg}}(\pi, \boldsymbol{P}).$$

- Strong duality also holds.

$$\sup_{\pi \in \Pi_{\mathsf{H}}} \inf_{\boldsymbol{P} \in \mathcal{U}} R_{\mathsf{avg}}(\pi, \boldsymbol{P}) = \inf_{\boldsymbol{P} \in \mathcal{U}} \sup_{\pi \in \Pi_{\mathsf{H}}} R_{\mathsf{avg}}(\pi, \boldsymbol{P})$$

$$\max_{\pi \in \Pi_{\mathsf{SD}}} \inf_{\boldsymbol{P} \in \mathcal{U}} R_{\mathsf{avg}}(\pi, \boldsymbol{P}) = \inf_{\boldsymbol{P} \in \mathcal{U}} \max_{\pi \in \Pi_{\mathsf{SD}}} R_{\mathsf{avg}}(\pi, \boldsymbol{P})$$

- From these results, it also follows that: $\sup_{\pi \in \Pi_{\mathsf{H}}} \inf_{\boldsymbol{P} \in \mathcal{U}_{\mathsf{H}}} R_{\mathsf{avg}}(\pi, \boldsymbol{P}) = \sup_{\pi \in \Pi_{\mathsf{H}}} \inf_{\boldsymbol{P} \in \mathcal{U}} R_{\mathsf{avg}}(\pi, \boldsymbol{P}).$
- So we can justifiably restrict the Adversary to stationary policies

Paper 2: Average optimality

S-rectangular uncertainty sets

- History-dependent policies may be optimal
- Not true for discounted s-rect, or avg (s,a)-rect!

Paper 1: R-rectangularity

Proposition 2.5

- Under Assumption 2.4, if r-rect, there exists an optimal stationary policy
- Proven using the duality result of later in the paper. $\max_{\pi \in \Pi_S} \min_{\boldsymbol{P} \in \mathbb{P}} R(\pi, \boldsymbol{P}) = \min_{\boldsymbol{P} \in \mathbb{P}} \max_{\pi \in \Pi_S} R(\pi, \boldsymbol{P}).$

$$
\begin{aligned}
\max_{\pi \in \Pi_S} \min_{\boldsymbol{P} \in \mathbb{P}} R(\pi, \boldsymbol{P}) &\leq \max_{\pi \in \Pi} \min_{\boldsymbol{P} \in \mathbb{P}} R(\pi, \boldsymbol{P}) \\
&\leq \min_{\boldsymbol{P} \in \mathbb{P}} \max_{\pi \in \Pi} R(\pi, \boldsymbol{P}) \\
&= \min_{\boldsymbol{P} \in \mathbb{P}} \max_{\pi \in \Pi_S} R(\pi, \boldsymbol{P}) \\
&= \max_{\pi \in \Pi_S} \min_{\boldsymbol{P} \in \mathbb{P}} R(\pi, \boldsymbol{P})
\end{aligned}
$$

Paper 1: Min-Max Duality

Lemma's 4.1 and 4.3

Lemma 4.1 *Let $\mathbb{P}$ be an r-rectangular uncertainty set. Under Assumption 2.4, there exists a stationary and deterministic policy solution of the policy improvement problem.*

Lemma 4.3 *Let $\pi \in \Pi$ and $\boldsymbol{W} \in \mathcal{W}$. Let $\boldsymbol{v}$ be the value function of the decision maker and $\boldsymbol{\beta}$ be the value function of the adversary. Then $\boldsymbol{W}^\top \boldsymbol{v} = \boldsymbol{\beta}$.*

- This is the (unique) solution to the Bellman equation of the adversary
  $\beta_i = \boldsymbol{w}_i^\top (\boldsymbol{r}_\pi + \lambda \cdot \boldsymbol{T}_\pi \boldsymbol{\beta}), \forall\, i \in [r].$
- So it suffices to show that

$$\left(\boldsymbol{W}^\top \boldsymbol{v}\right)_i = \boldsymbol{w}_i^\top \left(\boldsymbol{r}_\pi + \lambda \boldsymbol{T}_\pi \boldsymbol{W}^\top \boldsymbol{v}\right), \forall\, i \in [r].$$

Proof of 4.3

- To proof: $\left(\boldsymbol{W}^\top \boldsymbol{v}\right)_i = \boldsymbol{w}_i^\top \left(\boldsymbol{r}_\pi + \lambda \boldsymbol{T}_\pi \boldsymbol{W}^\top \boldsymbol{v}\right), \forall\, i \in [r].$
- Bellman equation for the decision maker:

$$v_s = \sum_{a \in \mathbb{A}} \pi_{sa} (r_{sa} + \lambda \cdot \sum_{i=1}^{r} u_{sa}^i \boldsymbol{w}_i^\top \boldsymbol{v}), \forall\, s \in \mathbb{S}.$$

$$\boldsymbol{T}_\pi = \left(\sum_{a \in \mathbb{A}} \pi_{sa} u_{sa}^i\right)_{(s,i) \in \mathbb{S} \times [r]} \in \mathbb{R}_+^{S \times r}.$$

$$\boldsymbol{v} = \boldsymbol{r}_\pi + \lambda \boldsymbol{T}_\pi \boldsymbol{W}^\top \boldsymbol{v}.$$

$$\boldsymbol{w}_i^\top \boldsymbol{v} = \boldsymbol{w}_i^\top (\boldsymbol{r}_\pi + \lambda \cdot \boldsymbol{T}_\pi \boldsymbol{W}^\top \boldsymbol{v}), \forall\, i \in [r].$$

$$\boldsymbol{W}^\top \boldsymbol{v} = (\boldsymbol{w}_i^\top \boldsymbol{v})_{i \in [r]}.$$

Theorem 4.2: Proof

$$\max_{\pi \in \Pi} \min_{\boldsymbol{W} \in \mathcal{W}} R(\pi, \boldsymbol{W}) \leq \min_{\boldsymbol{W} \in \mathcal{W}} \max_{\pi \in \Pi} R(\pi, \boldsymbol{W}).$$

Our goal is to show $\pi^* \in \arg\max_{\pi \in \Pi} R(\pi, \boldsymbol{W}^*)$.

- Equivalent to:

$$v_s^* = \max_{a \in \mathbb{A}} \left\{ r_{sa} + \lambda \cdot \sum_{i=1}^{r} u_{sa}^i \boldsymbol{w}_i^{*\top} \boldsymbol{v}^* \right\}, \forall\, s \in \mathbb{S}.$$

- Proof:

$$\begin{aligned} v_s^* &= r_{sa^*(s)} + \lambda \cdot \sum_{i=1}^{r} u_{sa^*(s)}^i \boldsymbol{w}_i^{*\top} \boldsymbol{v}^* \\ &= r_{sa^*(s)} + \lambda \cdot \sum_{i=1}^{r} u_{sa^*(s)}^i \beta_i^* \\ &= \max_{a \in \mathbb{A}} \left\{ r_{sa} + \lambda \cdot \sum_{i=1}^{r} u_{sa}^i \beta_i^* \right\} \\ &= \max_{a \in \mathbb{A}} \left\{ r_{sa} + \lambda \cdot \sum_{i=1}^{r} u_{sa}^i \boldsymbol{w}_i^{*\top} \boldsymbol{v}^* \right\}, \end{aligned}$$

Theorem 5.1

$$v_s^* = \max_{\pi_s \in \Delta} \left\{ r_{\pi_s,s} + \lambda \cdot (\boldsymbol{T}_\pi \boldsymbol{W}^{*\top} \boldsymbol{v}^*)_s \right\}, \forall\, s \in \mathbb{S},$$

$$\boldsymbol{v} = \boldsymbol{r}_\pi + \lambda \boldsymbol{T}_\pi \boldsymbol{W}^\top \boldsymbol{v}.$$

$$\beta_i^* = \min_{\boldsymbol{w}_i \in \mathcal{W}^i} \left\{ \boldsymbol{w}_i^\top (\boldsymbol{r}_{\pi^*} + \lambda \cdot \boldsymbol{T}_{\pi^*} \boldsymbol{\beta}^*) \right\}, \forall\, i \in [r].$$

- By substitution of: $\boldsymbol{W}^{*\top} \boldsymbol{v}^* = \boldsymbol{\beta}^*,$

$$v_s^* = \max_{\pi_s \in \Delta} \{ r_{\pi_s,s} + \lambda \cdot (\boldsymbol{T}_\pi (\min_{\boldsymbol{w}_i \in \mathcal{W}^i} \{ \boldsymbol{w}_i^\top \boldsymbol{v}^* \})_{i \in [r]})_s \}, \forall\, s \in \mathbb{S}.$$

$$v_s^* = \max_{a \in \mathbb{A}} \left\{ r_{sa} + \lambda \cdot \sum_{i=1}^{r} u_{sa}^i \min_{\boldsymbol{w}_i \in \mathcal{W}^i} \boldsymbol{w}_i^\top \boldsymbol{v}^* \right\}, \forall\, s \in \mathbb{S}.$$

- So, these two are equivalent:

$$F(\boldsymbol{v})_s = \max_{\pi_s \in \Delta} \{ r_{\pi_s,s} + \lambda \cdot (\boldsymbol{T}_\pi (\min_{\boldsymbol{w}_i \in \mathcal{W}^i} \{ \boldsymbol{w}_i^\top \boldsymbol{v} \})_{i \in [r]})_s \}, \forall\, s \in \mathbb{S}, \forall\, \boldsymbol{v} \in \mathbb{R}_+^S.$$

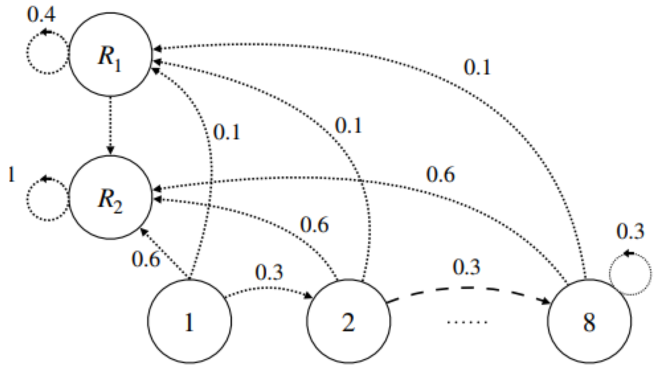
$$\boldsymbol{v}^0 = \boldsymbol{0},\, v_s^{k+1} = \max_{a \in \mathbb{A}} \left\{ r_{sa} + \lambda \cdot \sum_{i=1}^{r} u_{sa}^i \min_{\boldsymbol{w}_i \in \mathcal{W}^i} \boldsymbol{w}_i^\top \boldsymbol{v}^k \right\}, \forall\, s \in \mathbb{S}, \forall\, k \geq 0.$$

Paper 1: Numerical Experiments
Testing

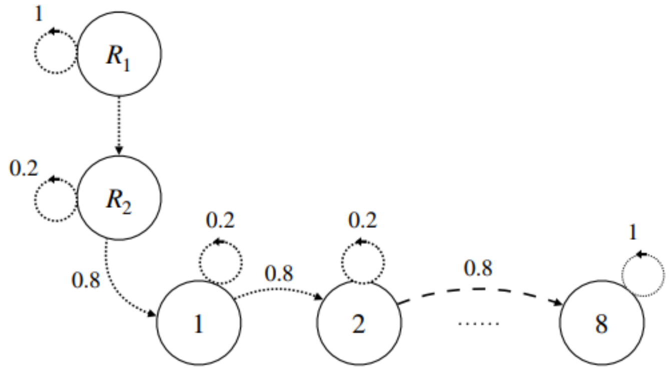
- Test 1: Machine replacement problem
- Conclusion: R-rect performs better than s-rect.

Budget of deviation τ	0.05	0.07	0.09
Worst-case of π^{nom} for $\mathbb{P}^{(\text{r})}$	94.40	92.21	90.04
Worst-case of π^{nom} for $\mathbb{P}^{(\text{s})}$	91.74	88.56	85.46

Budget of deviation τ	0.05	0.07	0.09
Nominal reward of $\pi^{\text{rob,r}}$	100.00	100.00	100.00
Worst-case of $\pi^{\text{rob,r}}$ for $\mathbb{P}^{(\text{r})}$	94.40	92.21	90.04
Nominal reward of $\pi^{\text{rob,s}}$	99.28	98.53	97.81
Worst-case of $\pi^{\text{rob,s}}$ for $\mathbb{P}^{(\text{s})}$	91.90	89.09	86.62



(a) Transition probabilities for action *repair*.



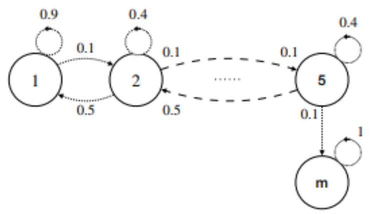
(b) Transition probabilities for action *wait*.

Paper 1: Numerical Experiments

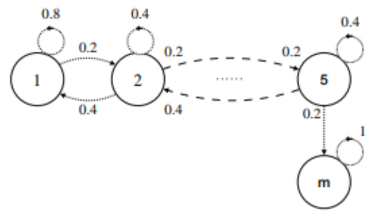
Testing

- Test 2: Inspired by healthcare
- Conclusion: Worst-case performance is actually worse, but average isn't!

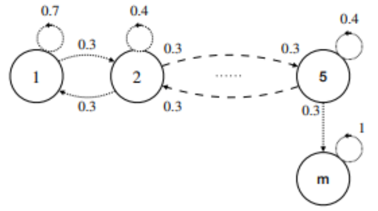
Budget of deviation τ	τ =0.05	τ =0.07	τ =0.09
Nominal reward of π^{nom}	100.00	100.00	100.00
Worst-case of π^{nom} for $\mathbb{P}^{(\mathsf{r})}$	50.26	41.74	35.63
Worst-case of π^{nom} for $\mathbb{P}^{(\mathsf{s})}$	45.75	37.37	31.51
Nominal reward of $\pi^{\mathsf{rob,r}}$	100.00	92.92	92.92
Worst-case of $\pi^{\mathsf{rob,r}}$ for $\mathbb{P}^{(\mathsf{r})}$	50.26	42.29	36.56
Nominal reward of $\pi^{\mathsf{rob,s}}$	91.48	91.35	89.56
Worst-case of $\pi^{\mathsf{rob,s}}$ for $\mathbb{P}^{(\mathsf{s})}$	52.09	44.39	38.69



(a) Transition probabilities for action *high dosage*.



(b) Transition probabilities for action *medium dosage*.



(c) Transition probabilities for action *low dosage*.