

Optimization of Markov Decision Processes with Model Ambiguity

Brian Denton

Department of Industrial and Operations Engineering

University of Michigan

(Work with **Lauren Steimle**, UM/GA Tech)

Sequential decision-making under uncertainty

Finance



Inventory management

Machine maintenance



Medical decision making

Prevention of cardiovascular disease (CVD) involves balancing the benefits and harms of treatment



Uncertain Future Benefits

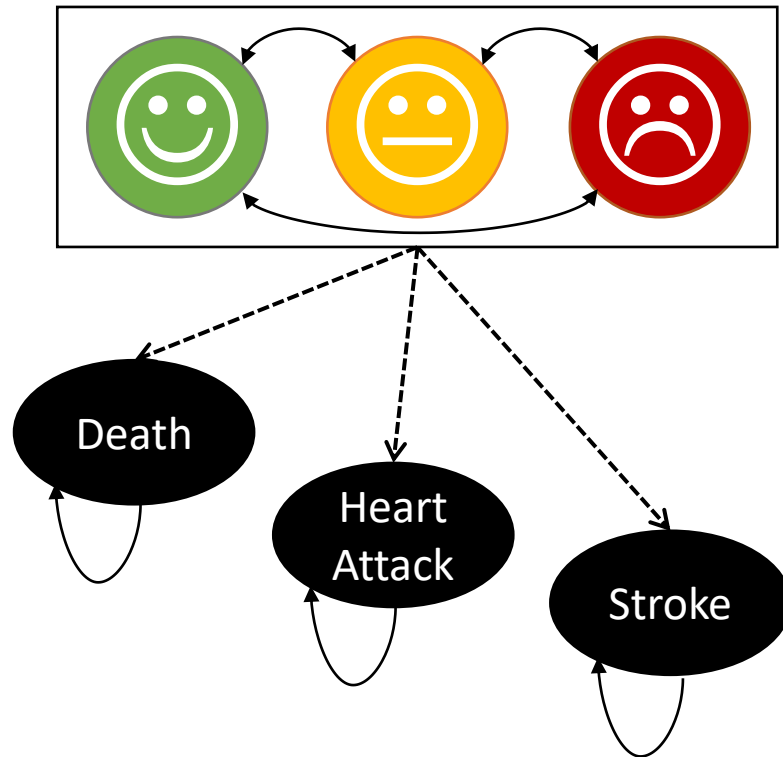
- Delay the onset of potentially deadly and debilitating heart attacks and strokes



Immediate harms

- Side effects (e.g., muscle pain, frequent urination)

Markov decision processes generalize Markov chains to include decisions

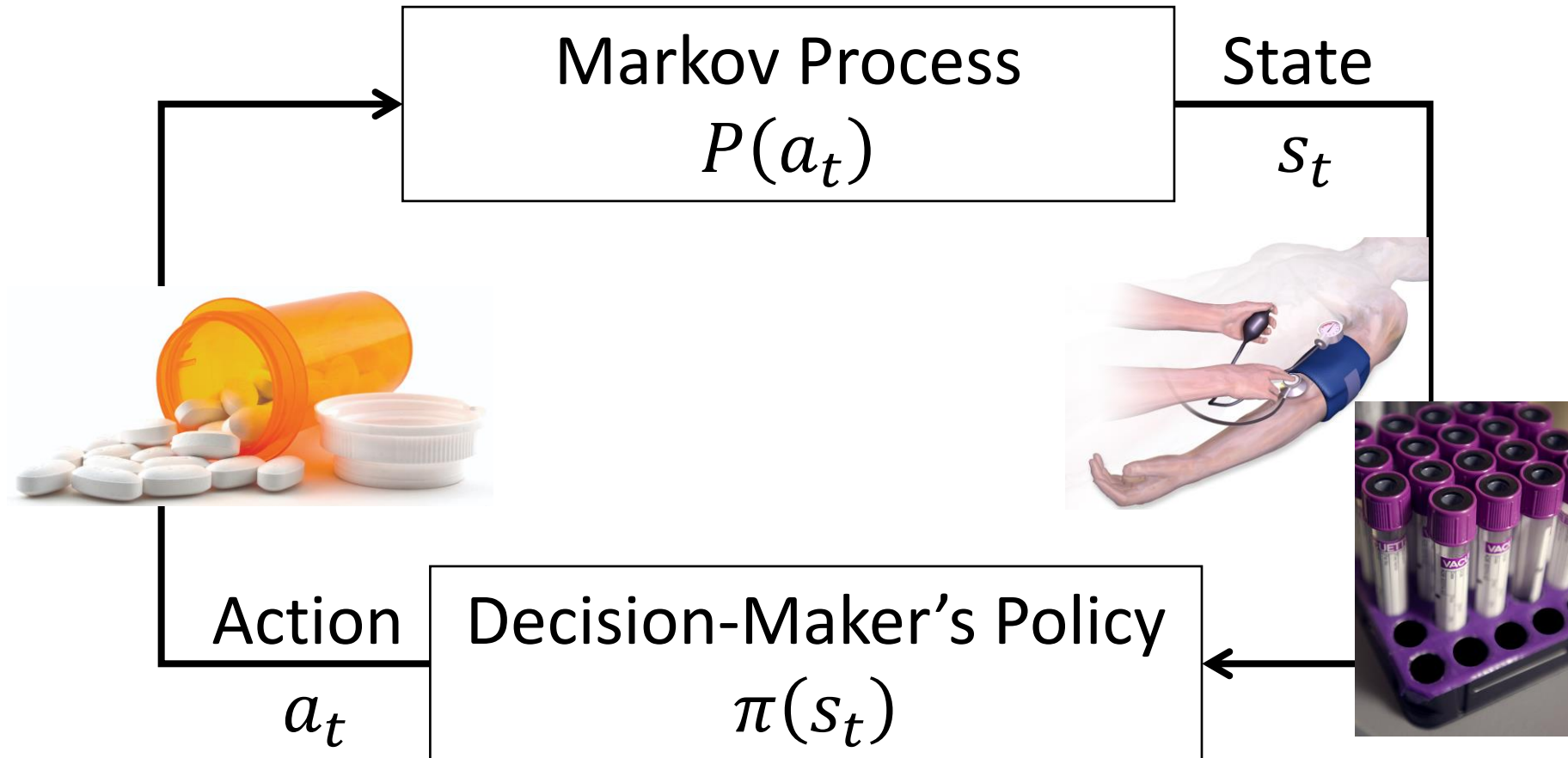


Health states

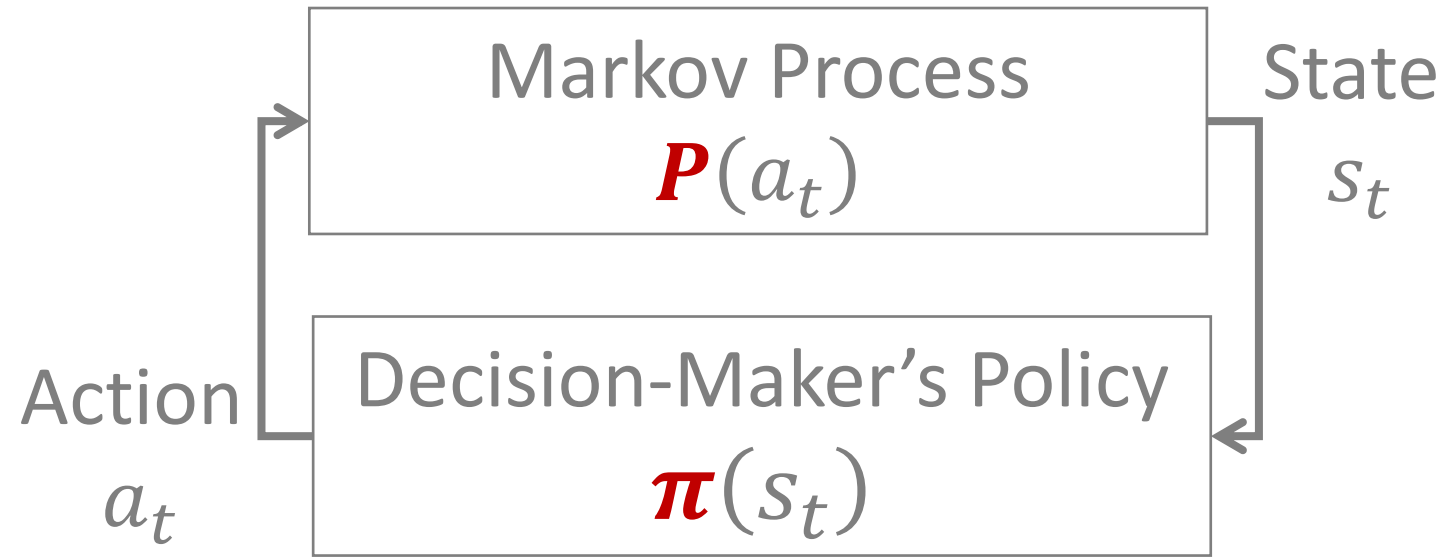
- Blood pressure levels
- Cholesterol levels
- Current medications

Steimle, L. N., & Denton, B. T. (2017). Markov decision processes for screening and treatment of chronic diseases. In *Markov Decision Processes in Practice* (pp. 189-222), Springer.

Markov decision process sequence of steps

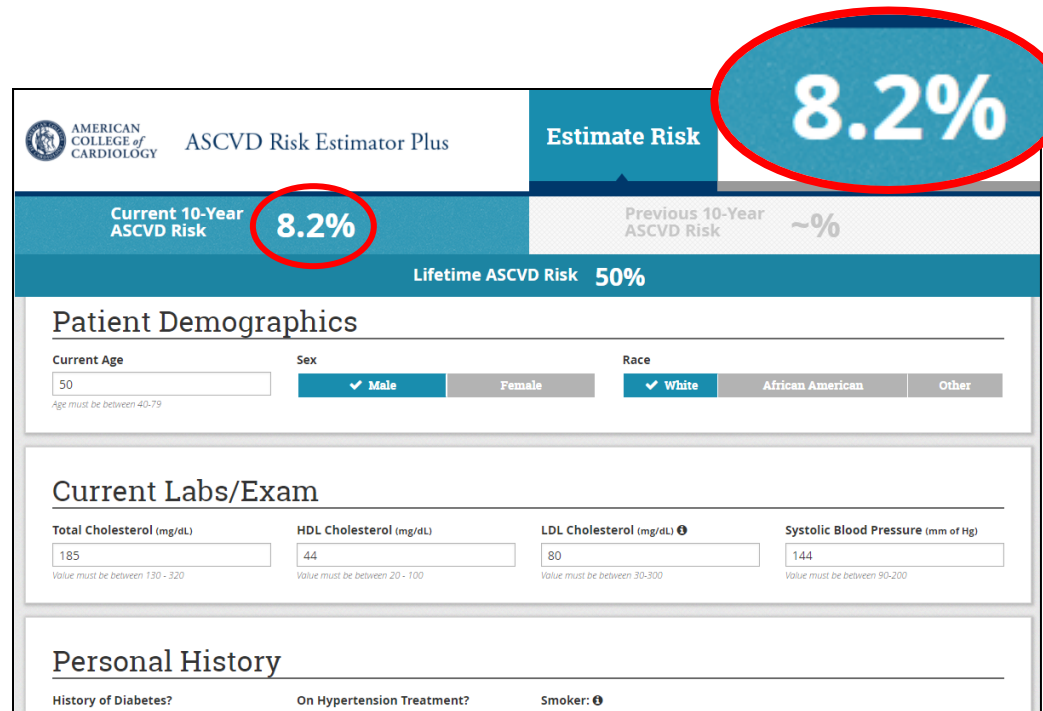


Markov decision process optimal policy



$$\max_{\pi \in \Pi} \left\{ \mathbb{E}^{\pi, P} \left[\sum_{t=1}^T r_t(s_t, a_t) + r_{T+1}(s_{T+1}) \right] \right\}$$

Clinical risk calculators are used to estimate a patient's risk



AMERICAN COLLEGE of CARDIOLOGY ASCVD Risk Estimator Plus

Estimate Risk **8.2%**

Current 10-Year ASCVD Risk **8.2%** Previous 10-Year ASCVD Risk ~%

Lifetime ASCVD Risk **50%**

Patient Demographics

Current Age: 50 Sex: Male Female Race: White African American Other

Current Labs/Exam

Total Cholesterol (mg/dL): 185 HDL Cholesterol (mg/dL): 44 LDL Cholesterol (mg/dL): 80 Systolic Blood Pressure (mm of Hg): 144

Personal History

History of Diabetes? On Hypertension Treatment? Smoker:

Inputs:

- Age
- Sex
- Race
- Cholesterol
- Blood Pressure
- History of Diabetes
- On Hypertensive Treatment
- Smoking status

Output:

Current 10-Year Risk

Well-established clinical studies give conflicting estimates about CVD risk



8.2%



17.8%

AMERICAN COLLEGE of CARDIOLOGY ASCVD Risk Estimator Plus

Estimate Risk Therapy Impact Advice

Current 10-Year ASCVD Risk **8.2%** Previous 10-Year ASCVD Risk ~%

Lifetime ASCVD Risk **50%**

Patient Demographics

Current Age: 50 Sex: Male Female Race: White African American Other

Current Labs/Exam

Total Cholesterol (mg/dL): 185	HDL Cholesterol (mg/dL): 44	LDL Cholesterol (mg/dL): 80	Systolic Blood Pressure (mm of Hg): 144
--------------------------------	-----------------------------	-----------------------------	---

Personal History

History of Diabetes? On Hypertension Treatment? Smoker:

General CVD Risk Prediction Using Framingham Heart Study

Sex: M F

Age (years): 50

Systolic Blood Pressure (mmHg): 144

Treatment for Hypertension: Yes No

Current smoker: Yes No

Diabetes: Yes No

HDL: 44

Total Cholesterol: 185

Calculate

Your Heart/Vascular Age: **67**

10 Year Risk

Your risk	17.8%
Normal	7.7%
Optimal	4.1%

1 Wilson et. al. Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation*. 1998

2 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. 2014

Research Questions

How can we improve Markov decision processes to account for model ambiguity?

How much benefit is there really?

The remainder of this presentation



Multi-model Markov decision processes

Branch-and-bound algorithms



Alternative ambiguity-aware formulations

Multi-Model MDPs have two layers of uncertainty

Optimal control of a **stochastic system...**

- Markov decision processes

...under model ambiguity

- Robust Markov decision processes

Robust optimization approach to ambiguity in Markov decision processes can be modeled as a two-player zero-sum game

- **Decision-maker** selects an action to maximize expected rewards
- **Adversary** selects transition probabilities to minimize DM's expected rewards

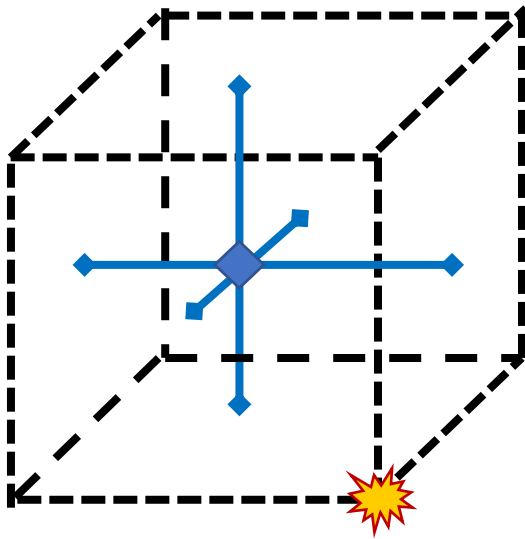
$$\max_{a \in \mathcal{A}} \min_{p_t(s,a) \in P_t(s,a)} \left\{ r_t(s, a) + \sum_{s' \in \mathcal{S}} p_t(s'|s, a) v_{t+1}(s) \right\}$$

(s,a)-rectangularity property gives a tractable model based on the assumption the adversary can select each row independently

Nilim, A. and El Ghaoui, L. "Robust control of Markov decision processes with uncertain transition matrices." *Operations Research* 53.5 (2005): 780-798.

Iyengar, G. "Robust dynamic programming." *Mathematics of Operations Research* 30.2 (2005): 257-280.

The interval model is computationally attractive, but has its drawbacks



Leads to overly-protective policies

- Optimizing for cases where all parameters take on worst-case values simultaneously

Transition matrices might lose known structure

- Ambiguity is realized independently across states, actions, and/or decision epochs

Relaxing (s,a) -rectangularity makes the max-min problem NP-hard*

*Wiesemann, Wolfram, Daniel Kuhn, and Berç Rustem. "Robust Markov decision processes." *Mathematics of Operations Research* 38.1 (2013): 153-183.

Multi-model Markov Decision Process notation

Generalizes a standard Markov decision process

- State space, $\mathcal{S} \equiv \{1, \dots, S\}$
- Decision epochs, $\mathcal{T} \equiv \{1, \dots, T\}$
- Action space, $\mathcal{A} \equiv \{1, \dots, A\}$
- Rewards, $R \in \mathbb{R}^{\mathcal{S} \times \mathcal{A} \times \mathcal{T}}$

Finite set of models, $\mathcal{M} = \{1, \dots, |\mathcal{M}|\}$

- Model m : An MDP $(\mathcal{S}, \mathcal{A}, \mathcal{T}, R, P^m)$
- Transition probabilities P^m are model-specific
- Model weights: $\lambda_1, \lambda_2, \dots, \lambda_{|\mathcal{M}|}$

The **weighted value problem** seeks a single policy that performs well in expectation

Performance of policy π in model m :

$$v^m(\pi) = \mathbb{E}^{\pi, P^m} \left[\sum_{t=1}^T r_t(s_t, a_t) + r_{T+1}(s_{T+1}) \right]$$

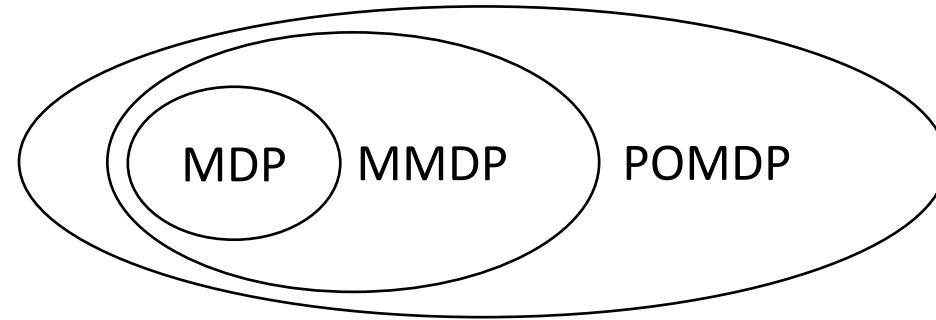
Weighted value of policy π :

$$W(\pi) = \sum_{m \in \mathcal{M}} \lambda_m v^m(\pi)$$

Weighted value problem:

$$W^* = \max_{\pi \in \Pi} W(\pi)$$

The weighted value problem is hard



The MMDP is a special case of a partially-observable MDP.

Proposition: The optimal policy may be history-dependent.

Proof by contradiction

Proposition: In general, the Weighted Value Problem is PSPACE-hard.

Reduction from *Quantified Satisfiability*

Special case of an MMDP with deterministic Markov policies

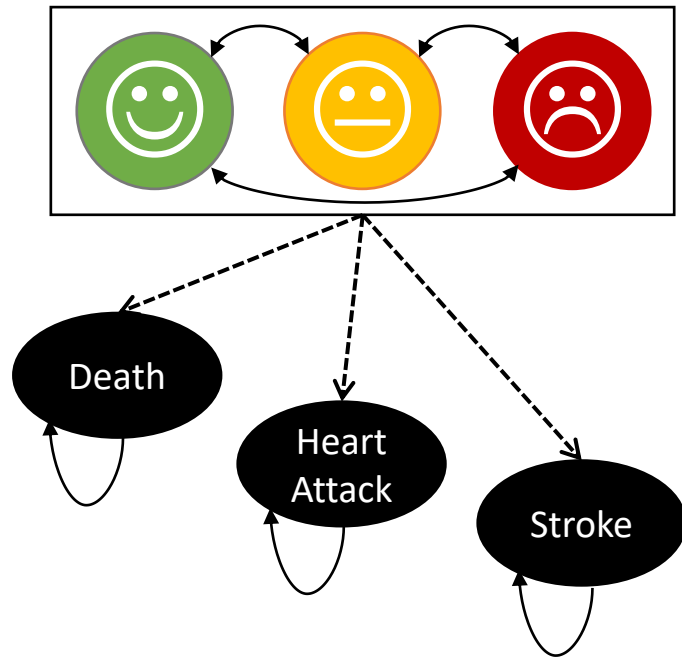
Proposition: There exists a deterministic policy that is optimal when restricting to Markov policies

Proposition: The Weighted Value Problem restricted to Markov deterministic policies is NP-hard

Reduction from 3-CNF-SAT

Initially, we focused on finding near-optimal Markov deterministic policies, $\pi \in \Pi^{\text{MD}}$, using a polynomial time approximation.

Example: approximation algorithm for cardiovascular disease prevention MMDP



Multi-model Markov decision process

- 4,096 states
- 64 actions
- 40 decision epochs
- 2 models

Case study data

- Longitudinal data from Mayo Clinic
- Framingham, ACC risk calculators
- Disutilities from medical literature

We compared our approximation algorithm policy to policies that ignore model ambiguity

Quality-Adjusted Life Years Gained
Over No Treatment, per 1000 Men

Optimal Decisions for FHS Model

MMDP Decisions

Optimal Decisions for ACC Model

In some cases, ignoring ambiguity has relatively minor implications

Quality-Adjusted Life Years Gained
Over No Treatment, per 1000 Men

Optimal Decisions for FHS Model

1,881

Framingham Heart Study Model

In some cases, ignoring ambiguity has relatively minor implications

Quality-Adjusted Life Years Gained
Over No Treatment, per 1000 Men

Optimal Decisions for FHS Model

1,881

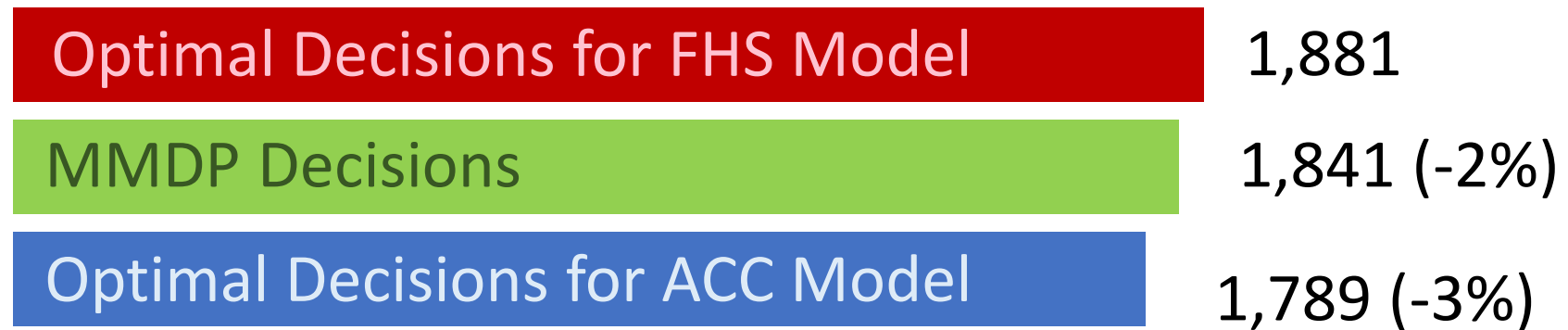
Optimal Decisions for ACC Model

1,789 (-3%)

Framingham Heart Study Model

In some cases, ignoring ambiguity has relatively minor implications

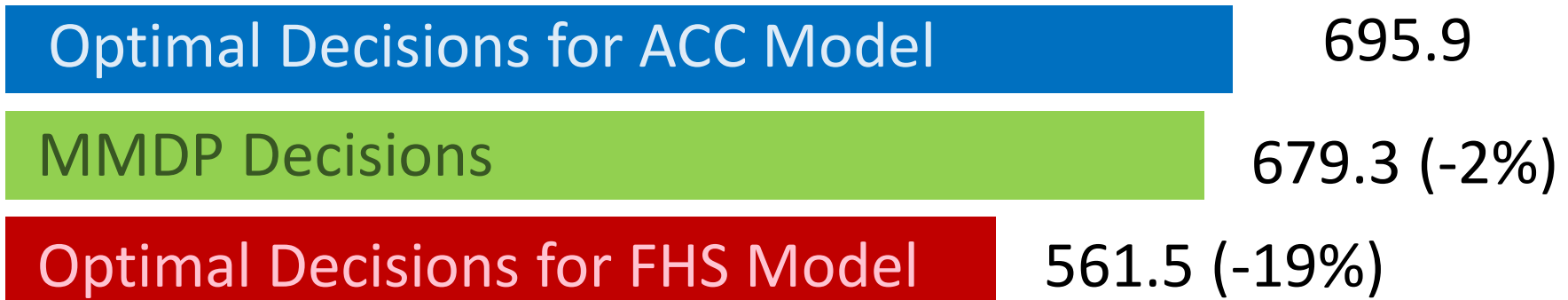
Quality-Adjusted Life Years Gained
Over No Treatment, per 1000 Men



Framingham Heart Study Model

But in other cases, ignoring ambiguity can have major implications

Quality-Adjusted Life Years Gained
Over No Treatment, per 1000 Men



American College of Cardiology Model

Observations

- MMDPs are difficult to solve computationally, but a polynomial-time approximation algorithm can provide near-optimal solutions in many instances
- Based on a CVD case study, it can be important to address ambiguity when there are multiple plausible models

Steimle, Lauren N., David L. Kaufman, and Brian T. Denton. "Multi-model Markov decision processes." *IIE Transactions* 53, no. 10 (2021): 1124-1139.

The remainder of this presentation



Multi-model Markov decision processes

Branch-and-bound algorithms

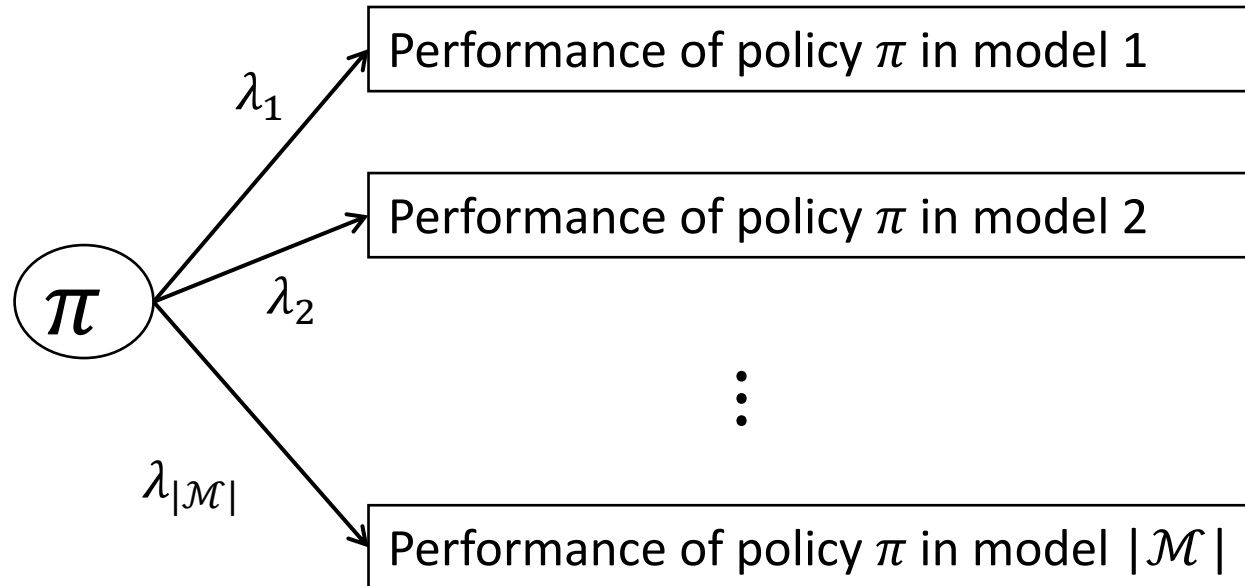


Alternative ambiguity-aware formulations

Approaches to solve the weighted value problem

- Mixed-integer programming (MIP)
- Branch-and-cut on a 2-stage stochastic integer program formulation
- Custom branch-and-bound that exploits MMDP structure

The connection between MMDP and two-stage stochastic program

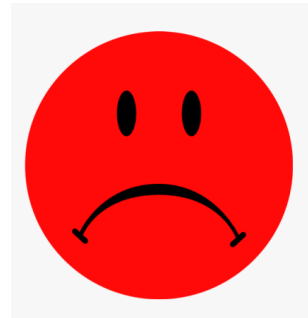


Stochastic program	MMDP
Scenarios	Model of MDP
Binary first-stage decision variables	Policy
Continuous second-stage decision variables	MDP model value functions

The MMDP is largely decomposable but Big-Ms in logic-based constraints cause trouble

Big-M's in logic-based constraints cause difficulty for standard stochastic programming methods

- Weak linear programming relaxation for the MIP
- Weak optimality cuts in Benders Decomposition



MMDPs are decomposable

- Evaluation of a fixed policy is easily done by solving $|\mathcal{M}|$ independent MDPs



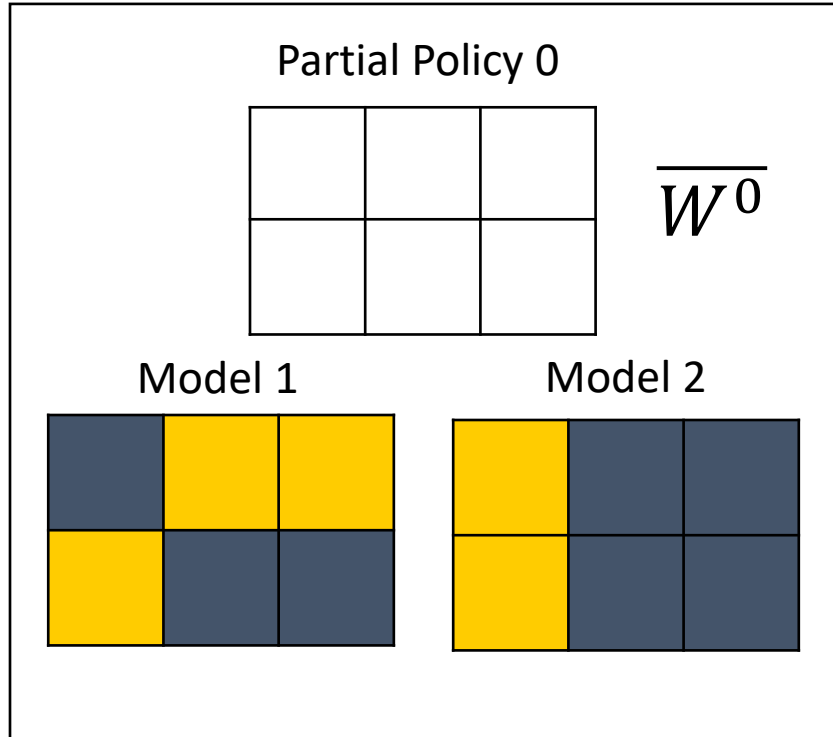
Branch-and-bound searches for policies that match across all models

Root Node: Relax requirement that policy must be same in each model

Goal: Find an *implementable policy* (policy is the same in all models) that maximizes weighted value



Branch & Bound begins by solving each model independently

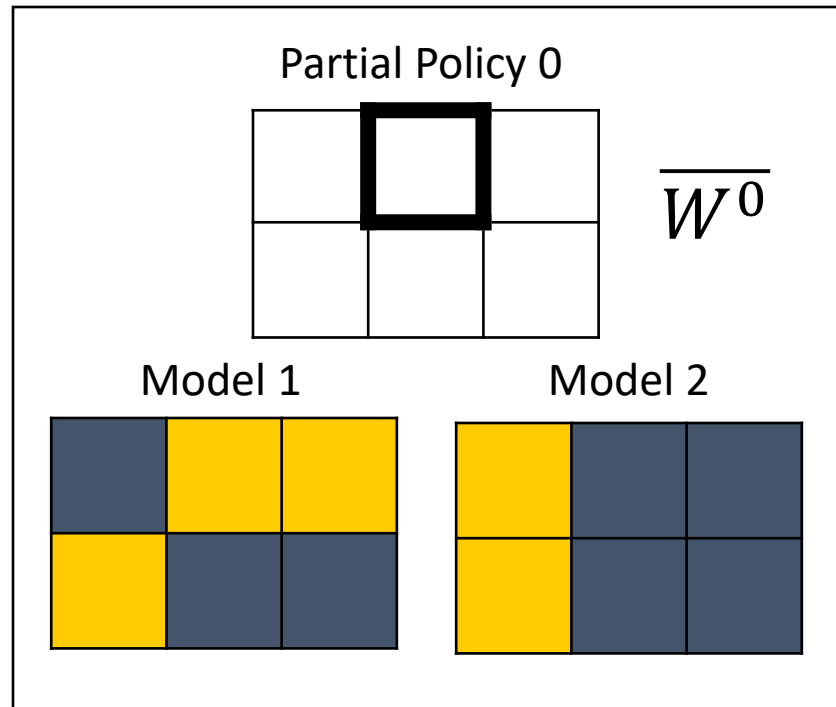


No actions have been fixed at the **root node**

Each model solved independently via backward induction

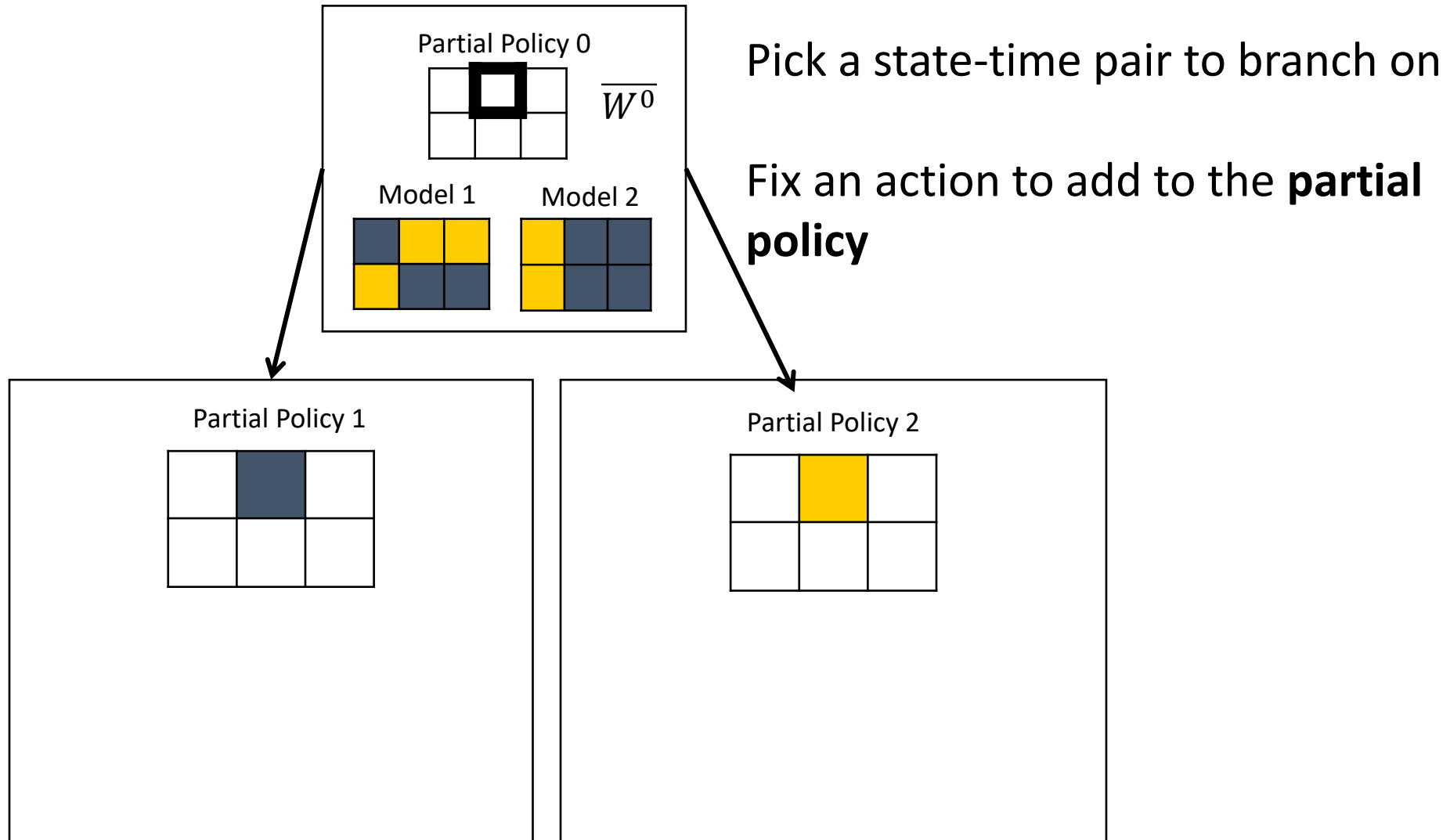
Gives an upper bound \overline{W}^0

Branch & Bound proceeds by fixing a part of the policy that must match in all models

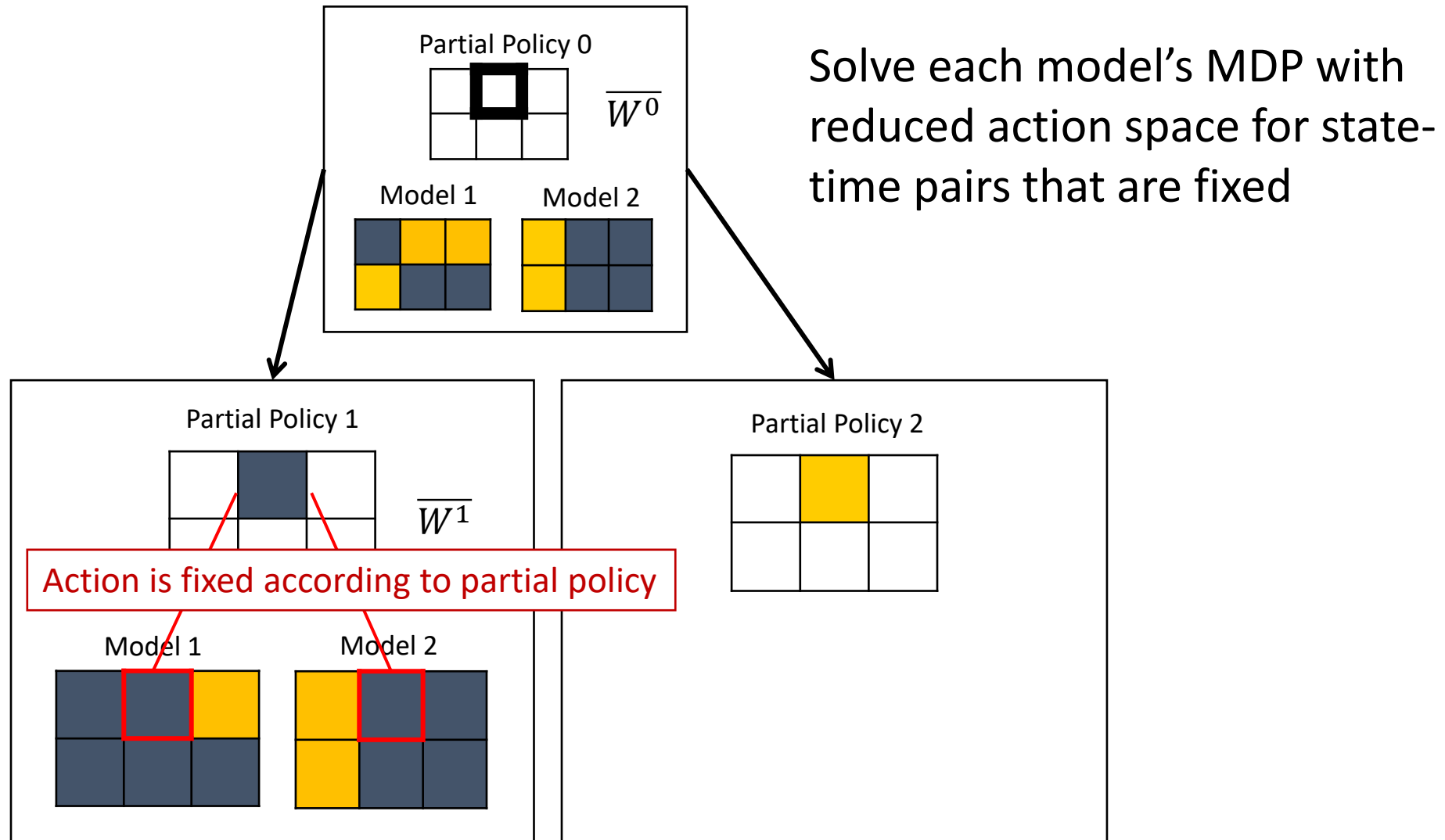


Pick a state-time pair to branch on

Branch & Bound proceeds by fixing a part of the policy that must match in all models



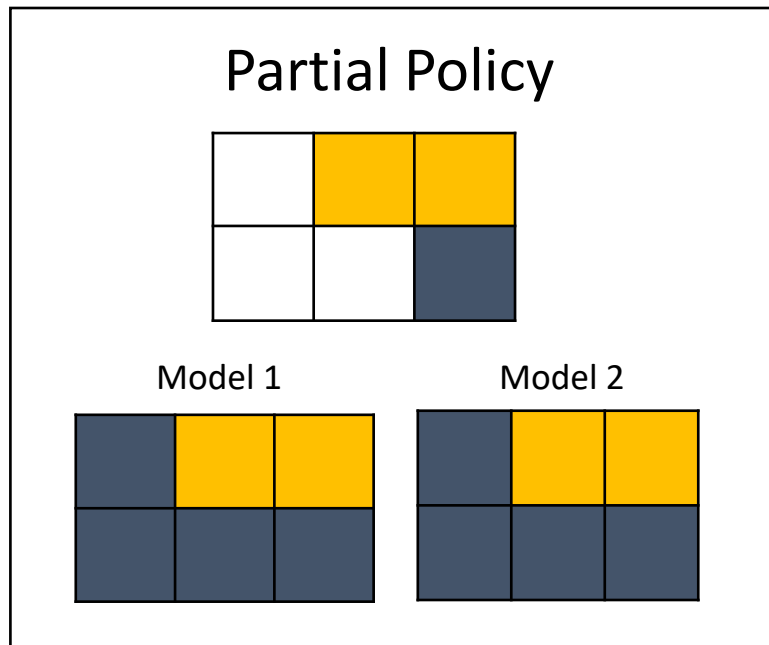
Branch & Bound solves a relaxation using backward induction to obtain upper bound



Pruning eliminates the need to explore all possible policies

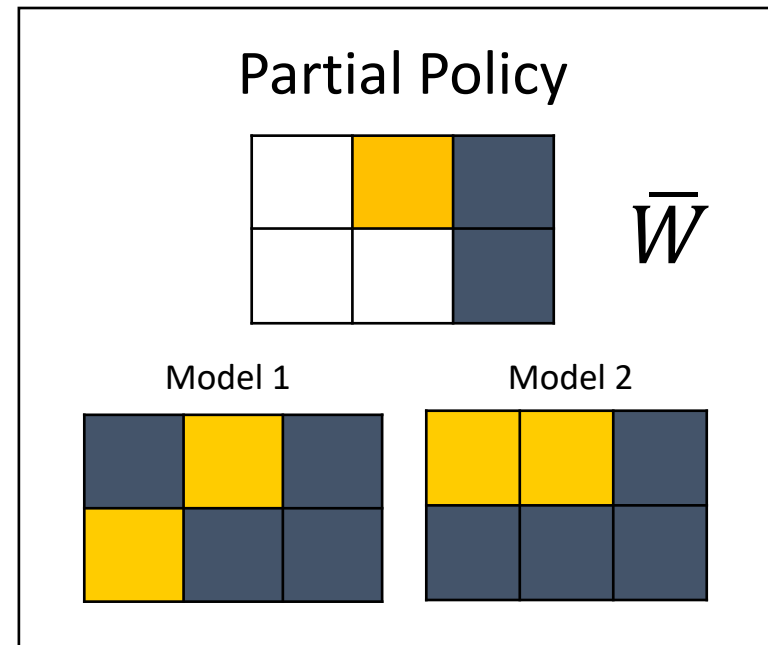
Prune by optimality

Solving the relaxation gives
an *implementable policy*



Prune by bound

The incumbent is better than
any possible completion of
the partial policy



We compared 3 exact methods on 240 instances of MMDPs

Solution Method	Implementation	% solved in 5 minutes?	Optimality Gap (avg.)
MIP Extensive Form	Gurobi		
MIP Branch-and-cut	Gurobi with Callbacks		
Branch-and-Bound	Custom code in C++		

[1] Steimle, L. N., Ahluwalia, V., Kamdar, C., and Denton B.T. (2018) "Decomposition methods for solving Multi-model Markov decision processes." *IIE Transactions*, 2022.

[2] Gurobi Optimization, LLC (2018) "Gurobi Optimizer Reference Manual", <http://www.gurobi.com>

Custom branch-and-bound approach is the fastest of the solution methods

Solution Method	Implementation	% solved in 5 minutes?	Optimality Gap (avg.)
MIP Extensive Form	Gurobi	0%	12.2%
MIP Branch-and-cut	Gurobi with Callbacks	0%	13.1%
Branch-and-Bound	Custom code in C++	97.9%	1.11%

Observations

- A custom branch-and-bound approach outperforms MIP-based solution methods
- MMDPs tend to be harder to solve when there is more variance in the models' parameters
- In low variance cases, the *mean value problem* provides an optimal or near-optimal solution

Steimle, Lauren N., Vinayak S. Ahluwalia, Charmee Kamdar, and Brian T. Denton. "Decomposition methods for solving Markov decision processes with multiple models of the parameters." *IIE Transactions* 53, no. 12 (2021): 1295-1310.

The remainder of this presentation



Multi-model Markov decision processes

Branch-and-bound algorithms



Alternative ambiguity-aware formulations

So far, we have considered a risk-neutral decision-maker

Weighted value problem
maximizes expectation of
model performance

$$W^* = \max_{\pi \in \Pi} \sum_{m \in \mathcal{M}} \lambda_m v^m(\pi)$$

What if the decision-maker is not risk-neutral?

Branch-and-bound algorithm is easily modified to solve other ambiguity-aware formulations

Max-min

$$\max_{\pi \in \Pi^{MD}} \min_{m \in \mathcal{M}} v^m(\pi)$$

Min-max-regret¹

$$\min_{\pi \in \Pi^{MD}} \max_{m \in \mathcal{M}} \left\{ \max_{\bar{\pi} \in \Pi} v^m(\bar{\pi}) - v^m(\pi) \right\}$$

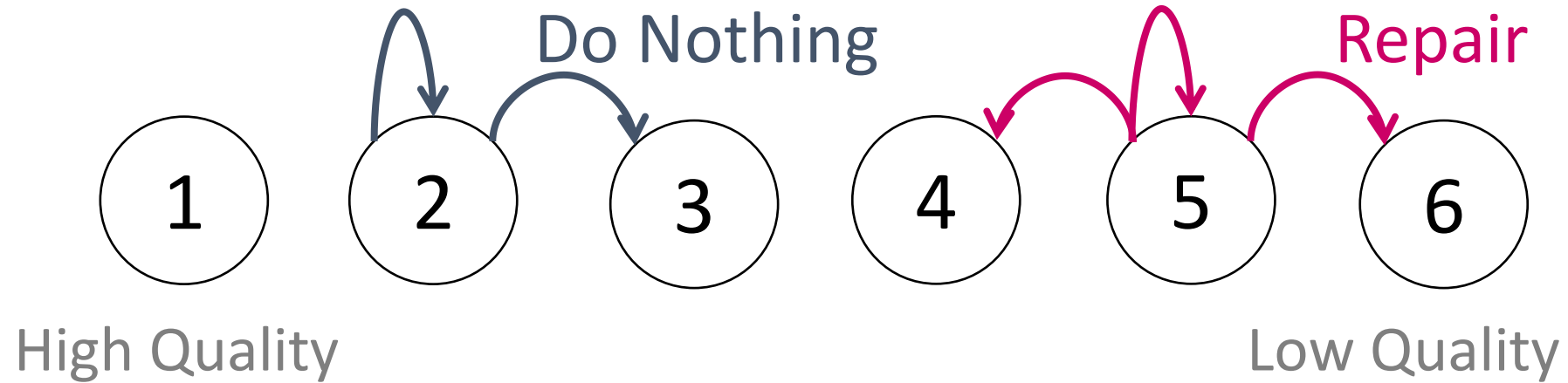
Percentile
optimization²

$$\begin{aligned} & \max_{z \in \mathbb{R}, \pi \in \Pi^{MD}} z \\ & \text{s. t.} \quad \mathbb{P}(v^m(\pi) \geq z) \geq 1 - \epsilon \end{aligned}$$

[1] Ahmed A, Varakantham P, Lowalekar M, Adulyasak Y, Jaillet P (2017) Sampling Based Approaches for Minimizing Regret in Uncertain Markov Decision Processes (MDPs). *Journal of Artificial Intelligence Research* 59:229–264

[2] Merakli, M. and Kucukyavuz, S. (2019) “Risk-Averse Markov Decision Processes under Parameter Uncertainty with an Application to Slow-Onset Disaster Relief.” *Optimization Online*.

Machine maintenance: Optimal timing of machine repairs



Operating costs depend on state of machine

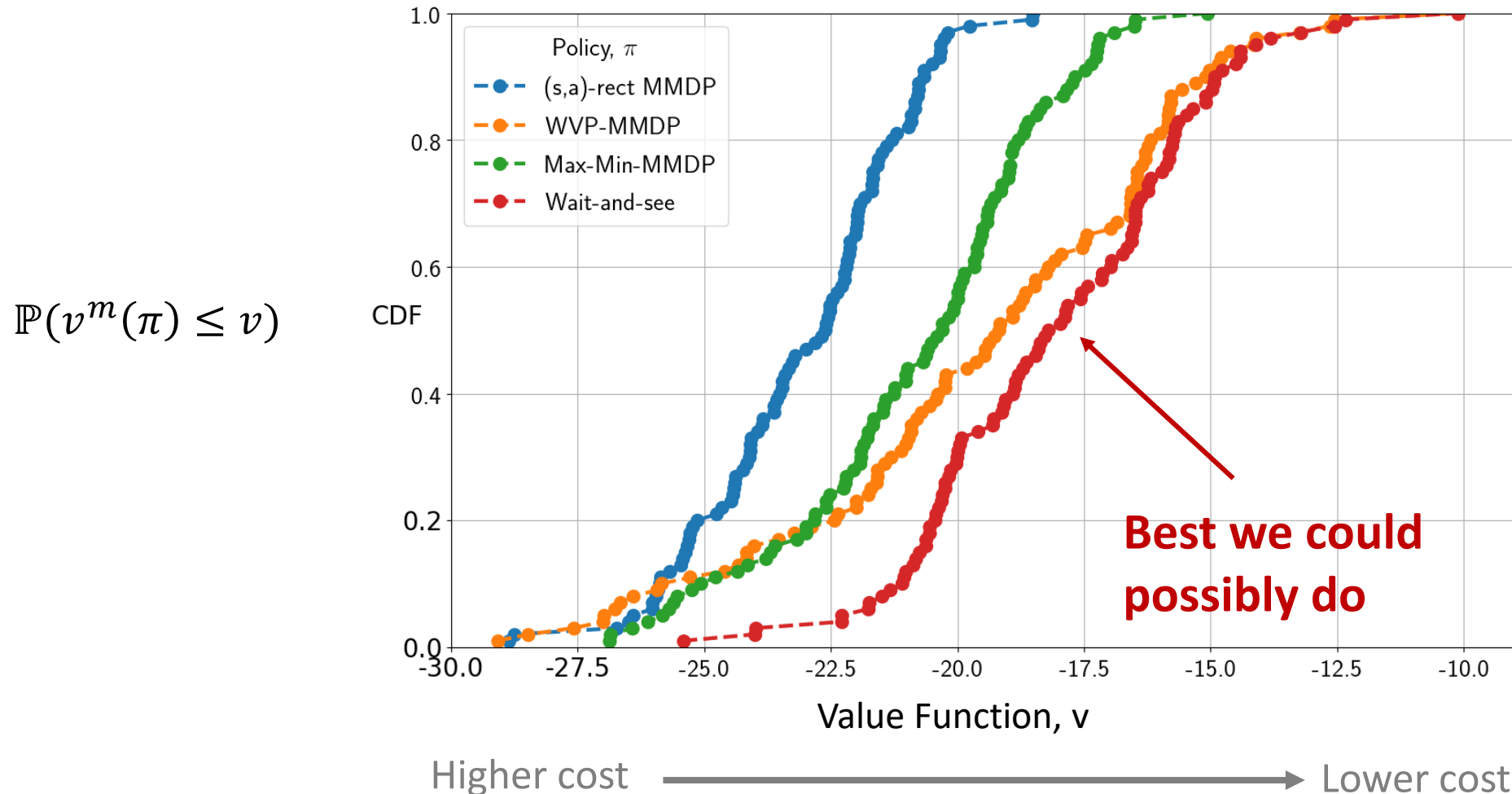


Options:

- Do Nothing at no cost
- Minor repair at low cost
- Major repair at high cost

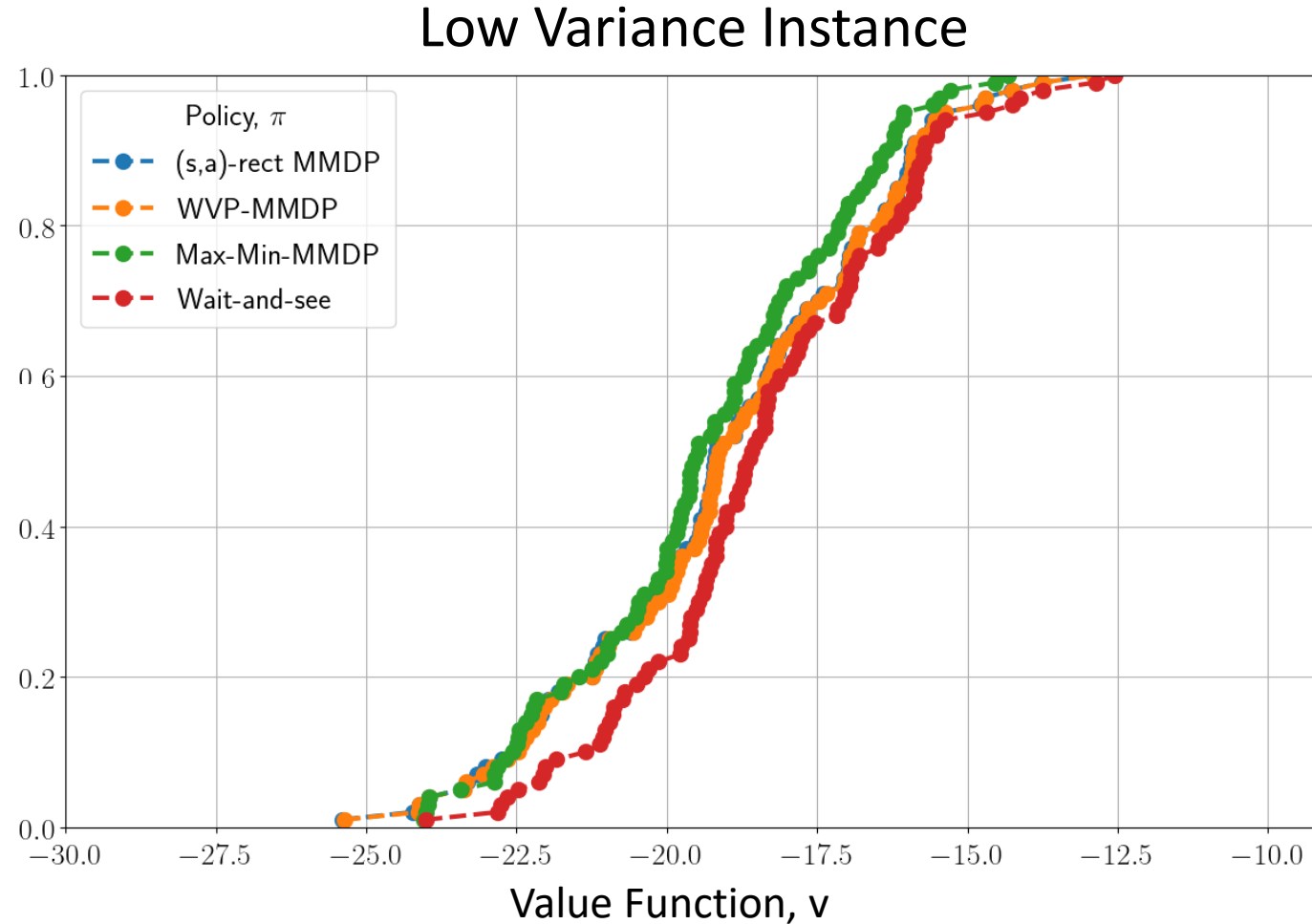
The distribution of the value function across models varies depending on the criteria selected

High Variance Instance



As **variance in models decreases**, the form of protection against ambiguity matters less

$$\mathbb{P}(v^m(\pi) \leq v)$$



Take-away messages

- Use caution before employing the interval model or assuming (s,a) -rectangularity!
- MMDPs can generate superior performance in terms of expected rewards, regret, and other performance measures.
- Branch-and-bound can be customized to leverage MMDP structure and solve practical instances.
- MMDPs are most useful when there is significant variation among models.

Related work

- Extension to partially observable Markov decision processes with multiple plausible latent Markov models
Li, W., Denton, B.T., “Multi-model Partially Observable Markov Decision Processes, working paper, 2023
- Optimization methods for “black-box” disease simulators
Zhang, Z., Denton, B.T., Morgan, T., “Optimization of Active Surveillance Strategies for Heterogeneous Patients with Prostate Cancer Journal,” *Production and Operations Management* (in press), 2022.
- Extension of algorithms to infinite horizon models
Ahluwalia, V., **Steimle, L.**, Denton, B.T., “Policy-based branch-and-bound for infinite-horizon Multi-model Markov decision processes.” *Computers and Operations Research*, 126, p. 10510, 2020.

Acknowledgments

Michigan Engineering

Lauren, Steimle, Ph.D.

Vinayak Ahluwalia

Charmee Kamdar

Mayo Clinic

Nilay Shah, Ph.D.

UM-Dearborn School of Business

David Kaufman, Ph.D.

U.S. Department of Veterans Affairs

Rodney Hayward, MD

Jeremy Sussman, MD

This material is based upon work supported by the National Science Foundation under Grant Number CMMI- 1462060 (Denton). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

