Participation in the Panel MDPs: AI versus OR Workshop on Decision Making in Adversarial Domains Greenbelt, MD

Eugene A. Feinberg

eugene.feinberg@sunysb.edu

Department of Applied Mathematics and Statistics

State University of New York at Stony Brook

Markov Decision Process (MDP)

- I: state space;
- A: action space;
- $A(i)$: action set available at state i;
- $p(i, a, j)$: transition probabilities;
- $r_k(i, a)$: one-step rewards.

For a stationary policy, a selected action depends only on the current state. We also consider randomized stationary policies. General policies may be randomized and depend on the past.

Introduction

- Consider a problem with $K+1$ criteria $W_0(\pi)$,$W_1(\pi), \ldots, W_K(\pi)$, where π is a policy. A natural approach to dynamic optimization is

$$\text{maximize } W_0(\pi)$$

subject to

$$W_k(\pi) \geq C_k, \qquad k = 1, \ldots, K.$$

- For $K > 0$ this approach typically leads to the optimality of randomized policies with the number of randomizations is limited by the number of constraints.
- For unconstrained problems ($K = 0$) there exists a nonrandomized stationary policy. This policy is usually optimal for all initial states.

Performance Criteria

The most common criteria are:

- Expected total rewards over the finite horizon.
- Average rewards per unit time.
- Expected total discounted rewards.

Let $r_k(i,a)$ be the one-step reward for criterion k if an action a is used in state i.

Expected total rewards over N steps:

$$W_k(i_0,\pi,N) := \mathbb{E}_{i_0}^{\pi} \sum_{t=0}^{N-1} r_k(i_t,a_t),$$

where i_0 is the initial state and π is the policy.

Performance Criteria: Continuation

Average rewards per unit time:

$$W_k(i_0, \pi) := \liminf_{N \to \infty} \frac{W_k(i_0, \pi, N)}{N}.$$

Total discounted rewards:

$$W_k(i_0, \pi) := \mathbb{E}_{i_0}^{\pi} \sum_{t=0}^{\infty} \beta^t r_k(i_t, a_t),$$

where $\beta \in [0, 1)$ is a discount factor.

In some problems, the initial state is given by an initial distribution μ. Similar criteria can be considered for continuous-time problems.

LP Formulation

$$\text{maximize} \sum_{i \in I} \sum_{a \in A(i)} r_0(i,a) x_{i,a}$$

subject to

$$\sum_{a \in A(j)} x_{j,a} - \beta \sum_{i \in I} \sum_{a \in A(i)} p(j,a,i) x_{i,a} = \mu(j), \qquad j \in I,$$

$$\sum_{i \in I} \sum_{a \in A(i)} r_k(i,a) x_{i,a} \geq C_k, \qquad k = 1, \ldots, K,$$

$$x_{i,a} \geq 0, \qquad i \in I,\ a \in A(i).$$

Optimal policy

$$\phi(a|i) = \begin{cases} x_{i,a} / \sum\limits_{b \in A(i)} x_{i,b}, & \text{if the denominatior is positive;} \\ arbitrary, & \text{otherwise.} \end{cases}$$

Interpretation: $x_{i,a}$ are so-called occupation measures,

$$x_{i,a} = \mathbb{E}_{\mu}^{\phi} \sum_{t=0}^{\infty} \beta^t I\{i_t = i, a_t = a\}.$$

For average rewards per unit time, $x_{i,a}$ are state-action frequencies,

$$x_{i,a} = \lim_{N \to \infty} \frac{1}{N} \mathbb{E}_{\mu}^{\phi} \sum_{t=0}^{N-1} I\{i_t = i, a_t = a\}.$$

Number of Randomizations

Let $Rand(\phi)$ be the number of randomizations for a randomized stationary policy ϕ,

$$Rand(\phi) = \sum_{i \in I} \{-1 + \sum_{a \in A(i)} \mathbf{I}\{\phi(a|i) > 0\}\}.$$

Then

$$Rand(\phi) \leq K, \quad (1)$$

where K is the number of constraints.

- For finite I, (1) follows from the LP arguments (Ross 1989).
- For countable I: F $\&$ Shwartz (1996) (Borkar 1992 for average rewards per unit time).
- For uncountable I: an open problem.

Constrained MDPs and problems in adversarial domains

- Constrained MDP is a nice model for problems in adversarial domains because of optimality of randomized policies. It is natural for problems in adversarial domains to keep the randomization index as large as possible.
- This leads to new problem formulations.

F's current research directions relevant to MDPs

- Continuous time MDPs;
- Non-atomic discrete-time MDPs;
- Applications to inventory control, discrete optimization, queueing control, ...

Continuous time MDPs

- The time is continuous and an action a selected in state i defines a vector of transition intensities $q(i,a,j) \geq 0$, $i \neq j$.
- Let $q(i,a) = \sum_{j \neq i} q(i,a,j)$.
- If $q(i,a) = 0$ then i is an absorbing state.
- Otherwise, the system spends on average $q^{-1}(i,a)$ units of time in state i and then moves to $j \neq i$ with the probability $p(i,a,j) = q(i,a,j)/q(i,a)$.
- There are reward rates $r_k(i,a)$ when the system stays in state i and instant rewards $R_k(i,a,j)$ when the system jumps from state i to state j.
- Major motivation: control of queues and queueing networks.

Continuous time MDPs: Switching policies

- Let x be the LP solution and
$A_x(i) = \{a \in A(i) : \; x_{i,a} > 0\} = \{a(i,1), \ldots, a(i, n(i,x))\}$.
- Let $S_0(i,x) = 0$. For $\ell = 1, \ldots, n(i,x)$, we set

$$s_\ell(i,x) = -(\alpha + q(i, a(i,\ell)))^{-1} \ln(1 - x_{i,a(i,\ell)} / \sum_{j=\ell}^{n(i,x)} x_{i,a(i,j)}),$$

and $$S_\ell(i,x) = S_{\ell-1}(i,x) + s_\ell(i,x).$$

- Optimal switching stationary policy ψ:

$$\psi(i,t) = \begin{cases} a(i,\ell), \text{if } A_x(i) \neq \emptyset \text{ and } S_{\ell-1}(i,x) \leq t < S_\ell(i,x); \\ \text{arbitrary action } a, \quad \text{if } A_x(i) = \emptyset. \end{cases}$$

Optimality of switching policies

Change of intensity between jumps is equivalent to randomized decisions at jump epochs.

- Consider two independent Poisson arrival processes 1 or 2 with positive intensities λ_1 and λ_2.
- At each epoch $t \in [0, \infty[$, an observer can watch either process 1 or 2. The process stops when the observer sees an arrival.

- A policy π is a measurable function $\pi : [0, \infty[\to \{1, 2\}$.
    - Let p_i^π, $i = 1, 2$ be the probability that the first observed arrival belongs to process i.
    - Let ξ be the time when an observer sees an arrival for the first time. In other words, $p_i = P\{\pi(\xi) = i\}$.

Optimality of switching policies

Let ξ_i be the time that process i has been watched before the first detected arrival, $\xi = \xi_1 + \xi_2$,

$$\xi_i = \int_0^{\xi} I\{\pi(t) = i\} dt.$$

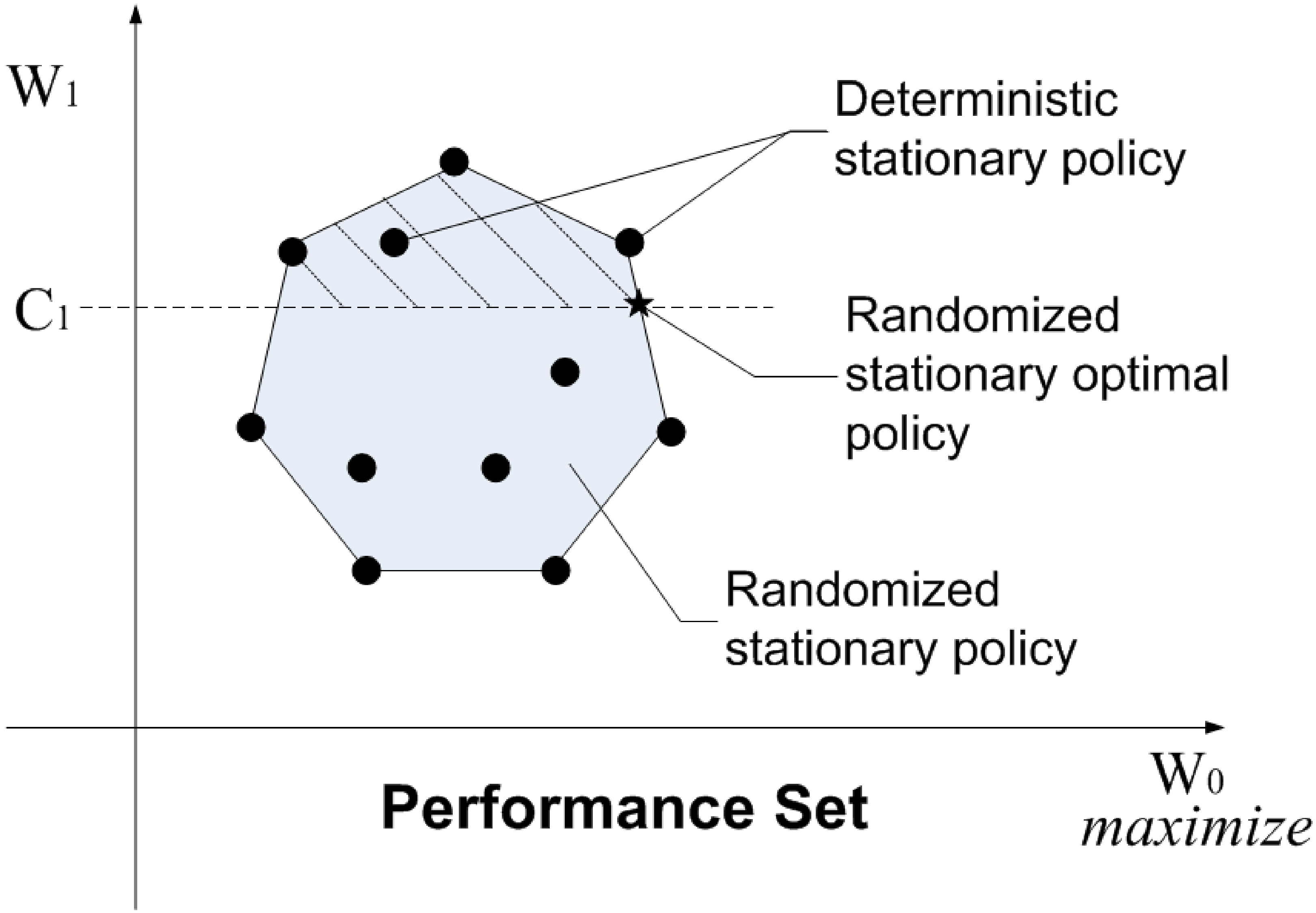
- **Lemma 1** $p_i = \lambda_i \mathbb{E}\, \xi_i$.
- **Remark 1** p_i *and* $\mathbb{E}\, \xi_i$ *depend on the policy* π
- **Remark 2** *If* $\pi(t) = 1$ *for all* t*, we get* $\mathbb{E}\, \xi_1 = \frac{1}{\lambda_1}$ *- the mean of an exponential random variable.*

Selecting intensities λ_1 and λ_2 randomly with the probabilities p_1 and p_2 respectively yields the same average characteristics as selecting intensity λ_1 during time $T = -\lambda_i^{-1} \ln(1 - p_1)$ and then switching to λ_2.

Non-Atomic MDPs

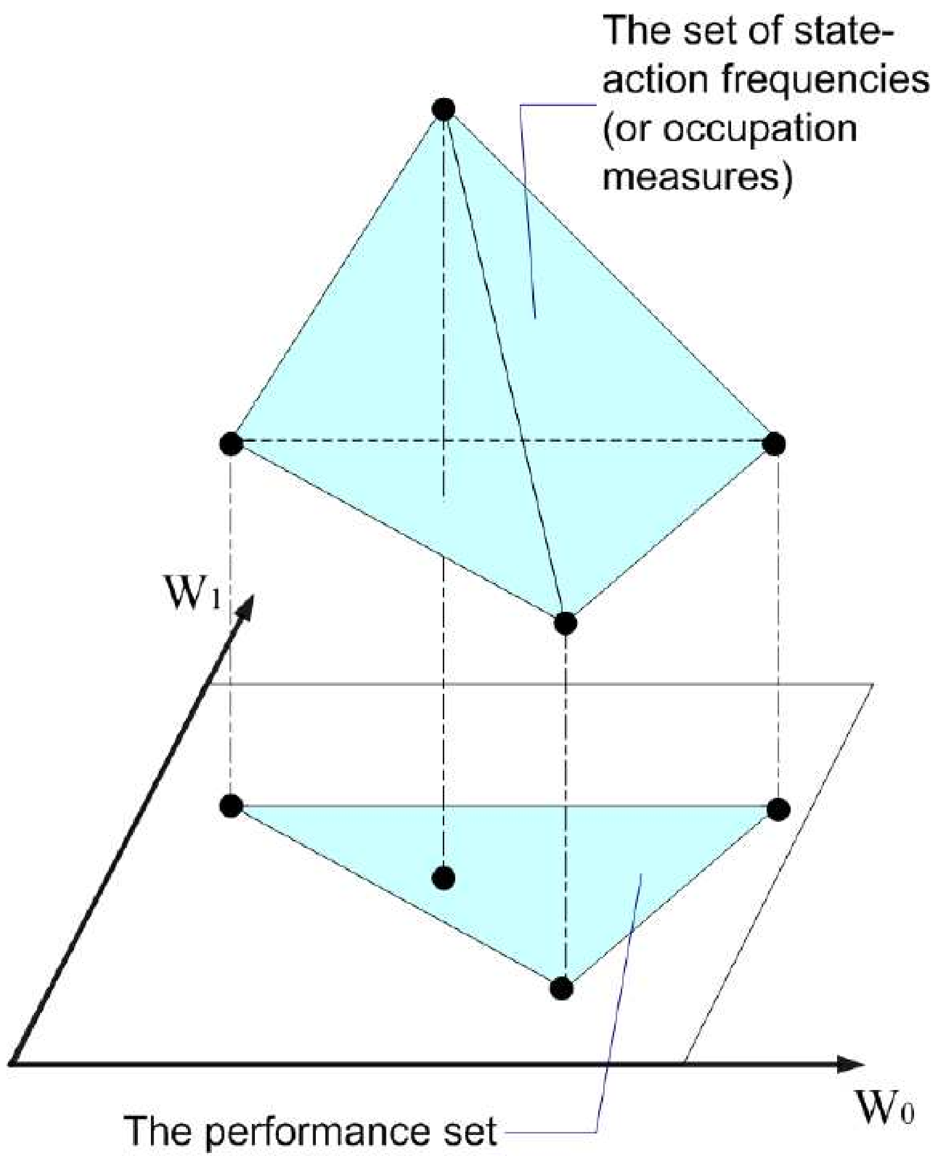
- If the state space is uncountable, we denote it by X instead of I. In this case, we use the notation $p_{x,a}(Y)$ instead of $p(i,a,j)$. Let the initial distribution μ and transition probabilities $p_{x,a}$ be non-atomic; i.e. $\mu(x) = 0$ and $p_{x,a}(y) = 0$ for all states x, y and for all actions a.
- Then for any policy π there exists a deterministic policy ϕ such that $W_k(\mu, \phi) = W_k(\mu, \pi), \quad k = 0, 1, \ldots, K.$
- F and Piunovskiy (2002, 2004).
- Examples of applications:
  - Statistical decision theory (Dvoretzky, Wald, and Wolfowitz 1951, Blackwell 1951).
  - Inventory control.
  - Portfolio management.

Finite State and Action MDP with Discounted Costs

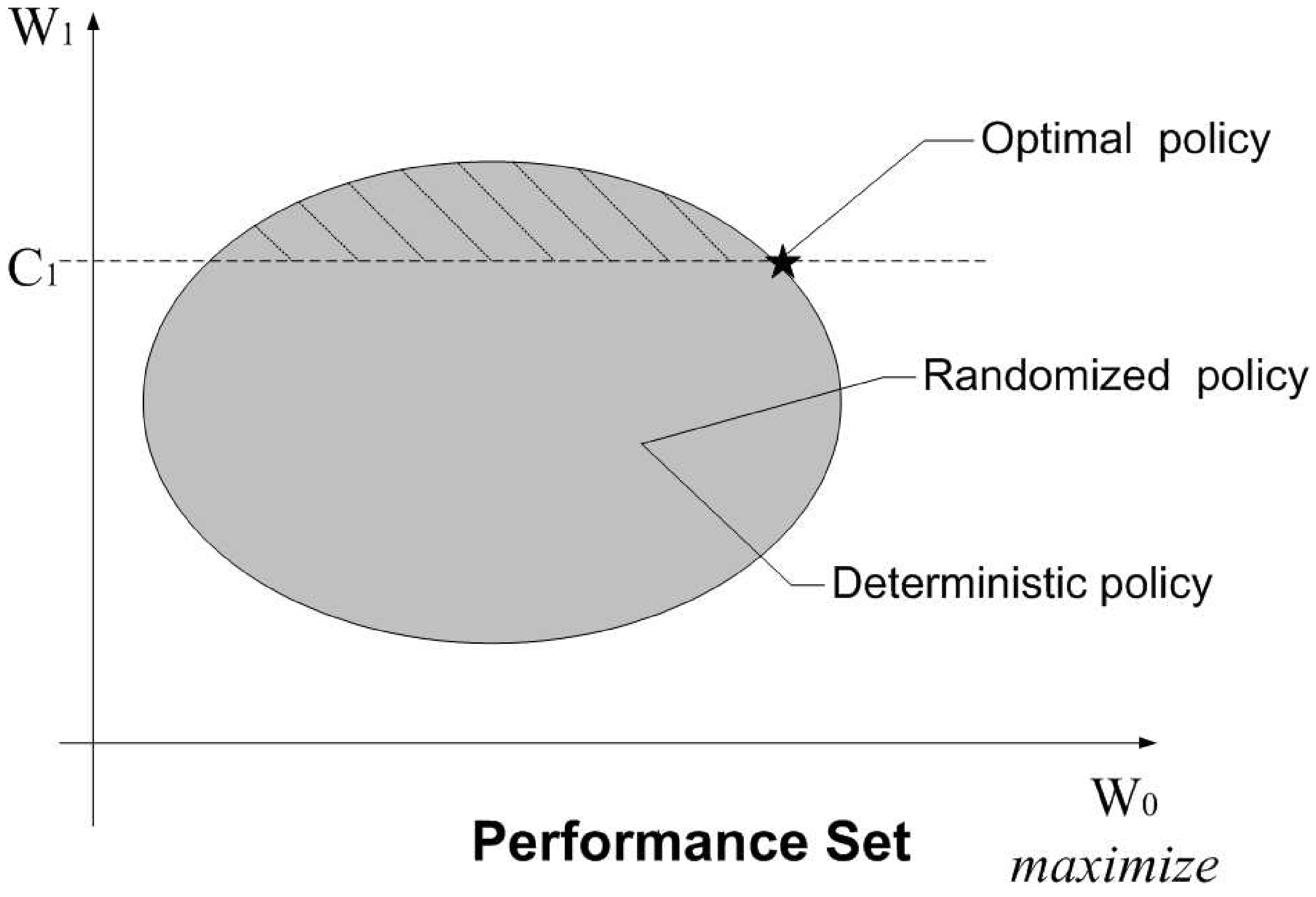


Performance Set

Explanation



Nonatomic MDP with Discounted Costs



Deterministic Statistical Decisions

- X: Borel state space;
- A: Borel action space;
- μ_n, $n = 1, \ldots, N$: non-atomic initial probabilities on X;
- $\rho(\mu_n, x, a)$: costs.

$$W(\mu, \pi) = \int_X \int_A \rho(\mu_n, x, a)\pi(da|x)\mu_n(dx).$$

- Dvoretzky, Wald, and Wolfowitz (1951): If A is finite, for any π there exists a deterministic decision rule ϕ such that

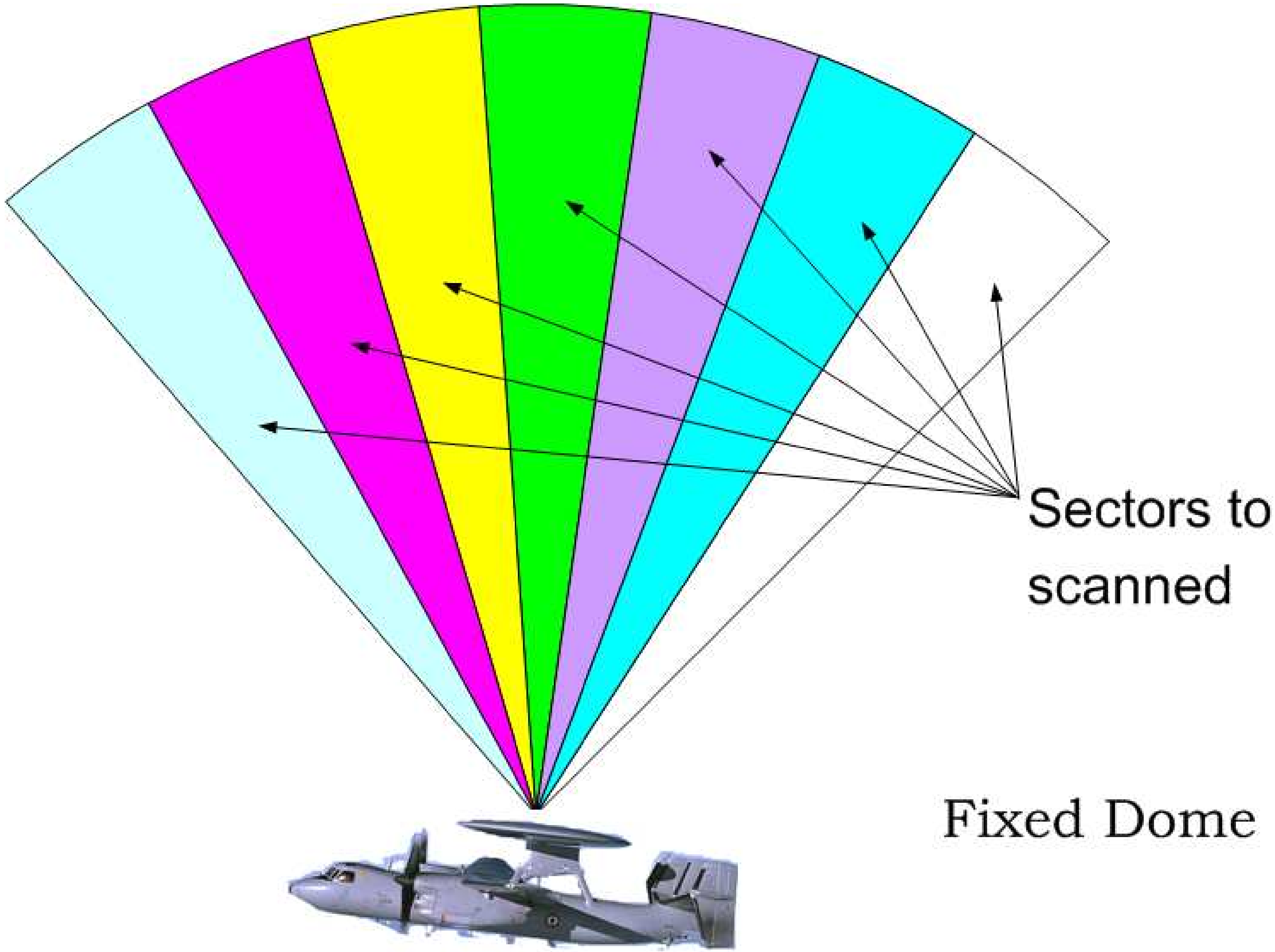
$$(W(\mu_1, \phi), \ldots, W(\mu_N, \phi)) = (W(\mu_1, \pi), \ldots, W(\mu_N, \pi)).$$

- F $\&$ Piunovskiy (2004): A could be arbitrary.

Other OR/AI approaches

- Neuro-Dynamic Programming
- Perturbation analysis
- We need more applications to test and compare different approaches
- A possible military-relevant application to test various approaches is the Generalized Pinwheel Problem

Motivation for the Generalized Pinwheel Problem



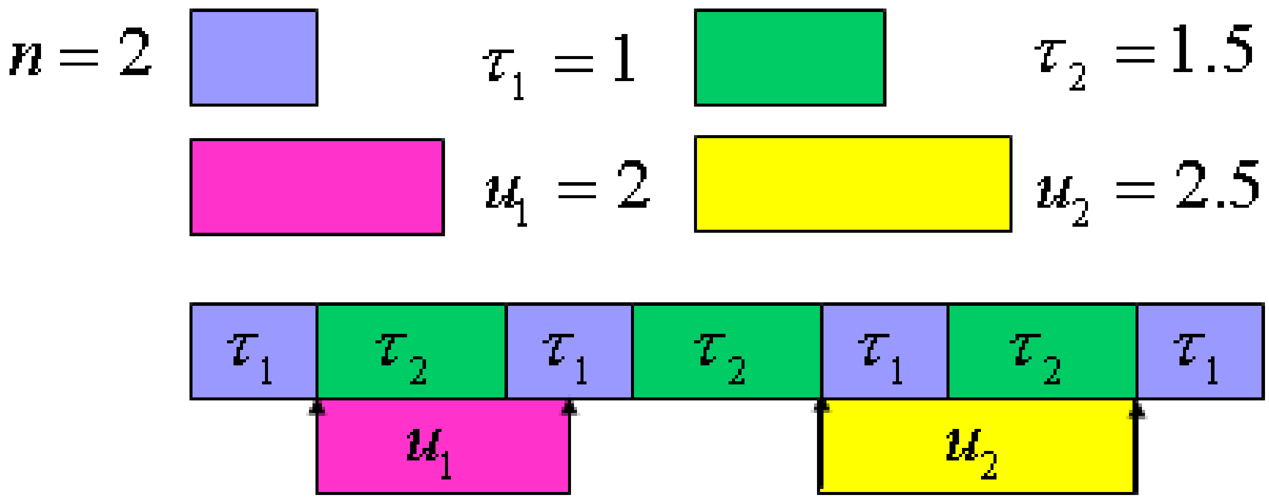
Generalized Pinwheel Problem

The radar sensor management problem can be formulated in general terms.
Consider the following infinite-horizon non-preemptive scheduling problem:

- There are N jobs. Each job i, $i = 1, \ldots, N$ is characterized by two parameters:
  - τ_i, the duration of job i,
  - u_i, the maximum amount of time between instances when job i is completed and started again.

Generalized Pinwheel Problem

- A schedule is feasible if each time job i is performed, it will be started again no more than u_i seconds after it is completed.



- Our goal is to find a feasible schedule or conclude that it does not exist.

Generalized Pinwheel Problem

This problem is NP-hard but a good heuristic is a so-called Frequency-Based Algorithm (F & Curry, 2005):

- Consider a relaxation,
- Formulate an MDP for this relaxation,
- Find optimal frequencies by solving this LP,
- Find a time-sharing policy (sequence),
- Try to cut a feasible piece of this sequence.

Time-Sharing Policies

Let I and A be finite and let stationary policies define recurrent Markov chains. Let $x_{i,a}$ be the vector of optimal state-action frequencies and ϕ is the corresponding randomized stationary optimal policy. For any finite trajectory $x_0, a_0, \ldots, x_{n-1}$, define

$$\mathbb{N}_n(i,a) = \sum_{t=0}^{n-1} \mathbf{I}\{x_t = i, a_t = a\},$$

$$\mathbb{N}_n(i) = \sum_{t=0}^{n-1} \mathbf{I}\{x_t = i\},$$

$$\delta(i,n) = argmax_{a \in A(i)}\{x_{i,a} - \frac{\mathbb{N}_n(i,a)+1}{\mathbb{N}_n(i)+1}\}.$$

Then policies ϕ and δ yield the same performances.