

# **Non-Myopic Multi-Aspect Sensing with Partially Observable Markov Decision Processes**

<sup>1</sup>Shihao Ji, <sup>2</sup>Ronald Parr and <sup>1</sup>Lawrence Carin

<sup>1</sup>Department of Electrical & Computer Engineering

<sup>2</sup>Department of Computer Engineering

Duke University

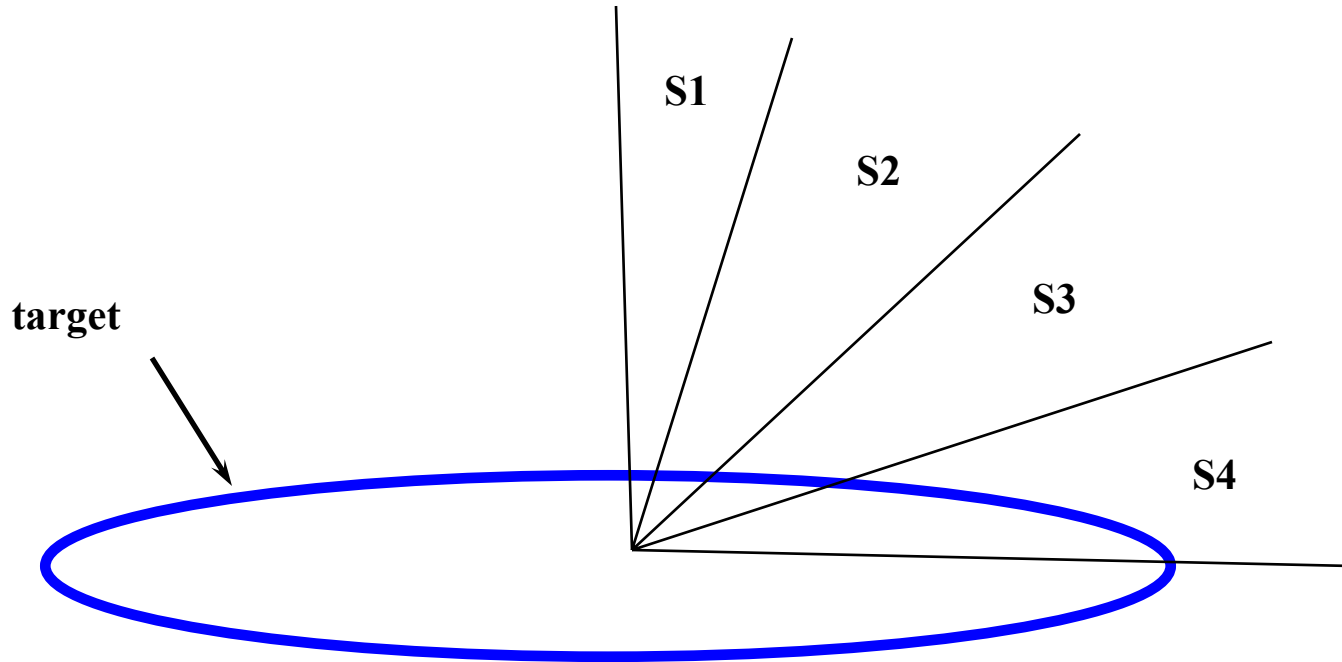
Durham, NC 27708-0291

[www.ee.duke.edu/~lcarin](http://www.ee.duke.edu/~lcarin)

# Outline

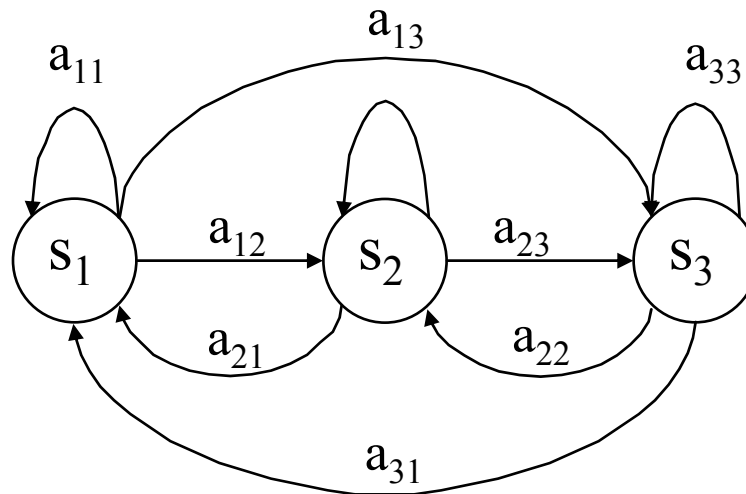
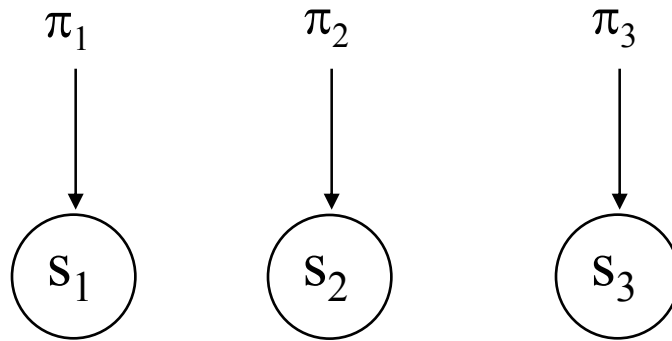
- Summary of the underlying partially-observed Markov model, with corresponding actions
- Partially observed Markov decision processes (POMDPs) and belief states, costs and Bayes risk
- Learning a POMDP policy via value iteration, with a policy defining the optimal action for a given belief state, accounting for discounted infinite horizon (non-myopic)
- Two POMDP implementation strategies for multi-target scattering data
- Myopic or greedy sensing alternative, with a stop criterion
- Example results on scattering data measured by NRL
  - Actions: Selection of optimal target-sensor orientation, fullband data
  - Actions: Selection of optimal target-sensor orientation and frequency subband

# Basic Construct



- **Scattering Data Can be Segmented into Angular Bins Characterized by particular physics**
- **Each such angular range termed a state (S1, S2, ... , SN)**

# Hidden Markov Models

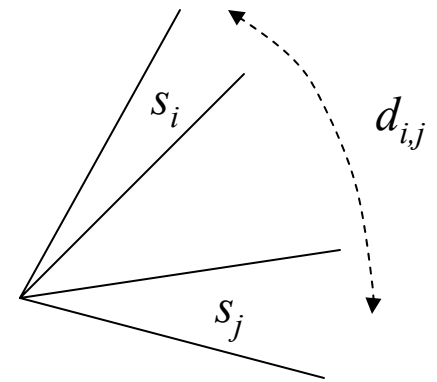


# Action-Dependent State-Transition Matrix

- Let  $d_{i,j}$  represent the angular distance between the centers of states  $i$  and  $j$ , in a prescribed direction (e.g., clockwise)
- The probability of transition from state  $i$  to state  $j$  after moving angular distance  $\Delta\phi$  is

$$p(s_j | s_i, \Delta\phi) \equiv \frac{w_j(d_{i,j} - \Delta\phi)}{\sum_{j=1}^K w_j(d_{i,j} - \Delta\phi)}$$

$$w_j(\phi) = \frac{1}{\sqrt{2\pi(\sigma_j)^2}} \exp\left[-\frac{1}{2}(\phi / \sigma_j)^2\right]$$



- The standard deviation  $\sigma_j$  is dictated by the width of state  $j$

$$\sigma_j = (\phi_j - \phi_{j-1}) / 2$$

# Outline

- Summary of the underlying partially-observed model, with corresponding actions
- Partially observed Markov decision processes (POMDPs) and belief states, costs and Bayes risk
- Learning a POMDP policy via value iteration, with a policy defining the optimal action for a given belief state, accounting for discounted infinite horizon (non-myopic)
- Two POMDP implementation strategies for multi-target scattering data
- Myopic or greedy sensing alternative, with a stop criterion
- Example results on scattering data measured by NRL
  - Actions: Selection of optimal target-sensor orientation, fullband data
  - Actions: Selection of optimal target-sensor orientation and frequency subband

# Belief State as a Sufficient Statistic

- The belief state quantifies the probability that the sensor is in state  $s$  given a sequence of  $T$  actions and corresponding observations
- The belief state at time  $T$  is a sufficient statistic for all actions and observations up to that point

$$b_T(s|o_1, \dots, o_T, a_1, \dots, a_T) = \Pr(s|o_T, a_T, b_{T-1})$$

- Very important for practical implementation: Needn't store all previous actions & observations
- Belief state computed readily, using underlying target POMDP model

$$\begin{aligned} b_T(s') &= \frac{\Pr(o_T|s', a_T, b_{T-1}) \Pr(s'|a_T, b_{T-1})}{\Pr(o_T|a_T, b_{T-1})} \\ &= \frac{\Pr(o_T|s', a_T, b_{T-1}) \sum_s \Pr(s'|a_T, b_{T-1}, s) \Pr(s|a_T, b_{T-1})}{\Pr(o_T|a_T, b_{T-1})} \\ &= \frac{p(o_T|s', a_T) \sum_s p(s'|a_T, s) b_{T-1}(s)}{\Pr(o_T|a_T, b_{T-1})} \end{aligned}$$

# Belief States and Bayes Risk

- Belief state may also be used to compute the probability that target  $n$  is being interrogated, based on  $T$  previous actions and observations

$$p(n|o_1, \dots, o_T, a_1, \dots, a_T) = p(n|b_T) = \sum_{s \in S_n} b_T(s)$$

- This fact plays a key role in subsequent policy design, which maps a belief state to a corresponding action, because the belief state may be used to compute the Bayes risk of a classification decision

$$\text{Target} = \arg \min_u \sum_{v=1}^N C_{uv} p(v|b_T) = \arg \min_u \sum_{v=1}^N C_{uv} \sum_{s \in S_v} b_T(s)$$

# Actions and Sensing Costs

- Two types of actions:
  - Sensing actions,  $a$ , that select next angle of observation and/or frequency of operation
  - Decision actions  $\hat{a}$  for which sensing is stopped and a classification decision is made
- Cost for sensing action:  $c(a)$ , independent of what target state is visited, this represents the cost of performing measurement, possibly sensor dependent
- Introduce a risk-based *terminal reward* for making a decision, this termed action  $\hat{a}$

# Classification Costs

- Upon performing classification action  $\hat{a}$  we move into a new state  $s_{ij}$ , corresponding to declaring target  $i$  when the actual target is target  $j$
- The cost associated with state  $s_{ij}$  is represented as  $C_{ij}$
- The probability of interrogating target  $T_j$  given belief state  $b(s)$ , where  $s$  are the underlying states of the targets, is

$$p(T_j|b) = \sum_{s \in S^{(j)}} b(s)$$

- The expected immediate cost of taking terminal classification action  $\hat{a}$  in belief state  $b(s)$  may therefore be represented as

$$C_1(b) = \max_{\hat{a}_i} \sum_j C_{ij} p(T_j|b) = \max_{\hat{a}_i} \sum_j \sum_{s \in S^{(j)}} C_{ij} b(s)$$

- Immediate expected value of terminating sensing and declaring target  $i$ , action  $\hat{a}_i$ , driven by Bayes' risk

# POMDP Formulation Summary

Actions	States	Cost
<p><b>Sensing Action:</b></p> <ul style="list-style-type: none"> <li>• Move platform angle <math>\Delta\phi</math></li> <li>• Perform measurement with one of <math>M</math> sensors</li> </ul>	<p><math>S = \{s_k^{(n)}, \forall k, n\}</math></p> <p>Target states <math>k</math> across all targets <math>n = \{1, 2, \dots, N\}</math></p>	<p><math>c(m)</math>, <math>m</math> representing one of the <math>M</math> possible sensors (independent of target state visited)</p>
<p><b>Classification Action:</b></p> <ul style="list-style-type: none"> <li>• Stop sensing, declare object under test to be one member from set <math>\{1, 2, \dots, N\}</math></li> </ul>	<p><math>s_{uv}</math>, corresponding to declaring target <math>u</math> when in reality target <math>v</math> is being sensed; both <math>u</math> and <math>v</math> members of the set <math>\{1, 2, \dots, N\}</math></p>	<p><math>C_{uv}</math>, for classification state <math>s_{uv}</math>            In terms of target states <math>s</math> in <math>S</math>, <math>c(s, a=u) = C_{uv}</math> for all <math>s</math> associated with target <math>v</math></p>

# POMDP Summary

- Algorithm has two types of states: underlying states of the targets  $s$  plus terminal states  $s_{ij}$  after making a classification decision
- Optimal policy learns what sensing actions  $a$  to take given belief state  $b(s)$ , as well as a policy as to when to make a decision (stop sensing), as a function of the belief state
- May include different costs for different sensor modalities, while also accounting via Bayes risk for costs of different misclassifications  $C_{ij}$
- Optimal policy determined via point-based algorithm that preserves the local slope of value function

# Outline

- Summary of the underlying partially-observed model, with corresponding actions
- Partially observed Markov decision processes (POMDPs) and belief states, costs and Bayes risk
- Learning a POMDP policy via value iteration, with a policy defining the optimal action for a given belief state, accounting for discounted infinite horizon (non-myopic)
- Two POMDP implementation strategies for multi-target scattering data
- Myopic or greedy sensing alternative, with a stop criterion
- Example results on scattering data measured by NRL
  - Actions: Selection of optimal target-sensor orientation, fullband data
  - Actions: Selection of optimal target-sensor orientation and frequency subband

# Implementation Issues

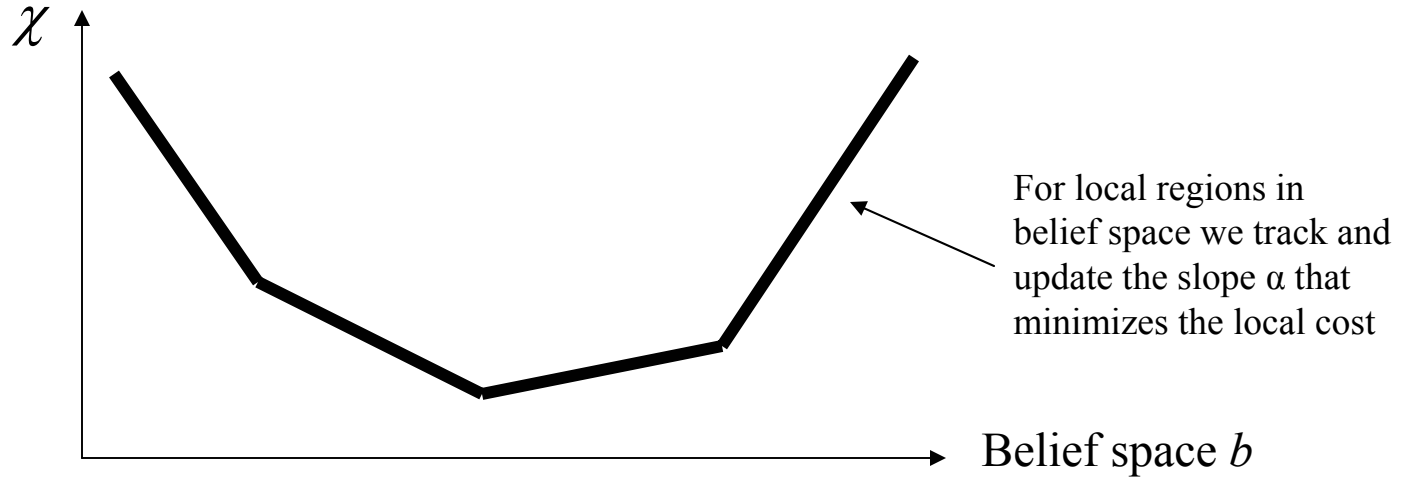
- The cost of taking action  $a$  when in belief state  $b$ ,  $t$  steps from the horizon is

$$\chi_t(b) = \min_a \left[ \underbrace{C(b, a)}_{\text{Immediate Expected Cost}} + \underbrace{\gamma \sum_{b' \in B} p(b'|b, a) \chi_{t-1}(b')}_{\text{Discounted Expected Future Costs}} \right]$$

- Becomes a dynamic-programming problem for learning the optimal policy, which maps belief states to actions (discounted infinite-horizon problem)
- Value-iteration dynamic programming stabilizes when a fixed action is defined for each belief state, defining the optimal discounted infinite-horizon policy

# Implementation Issues - 2

- The cost function is linear in the belief state, which implies that the cost function is a piecewise linear concave problem in the belief-space simplex



$$\chi_t(b) = \min_{\alpha \in C_t} \sum_{s \in S} \alpha(s) b(s)$$

$$\chi_t(b) = \min_{a \in A} \left[ C(b, a) + \gamma \sum_{o \in O} \min_{\alpha \in C_{t-1}} \sum_{s \in S} \sum_{s' \in S} p(s'|s, a) p(o|s', a) \alpha(s') b(s) \right]$$

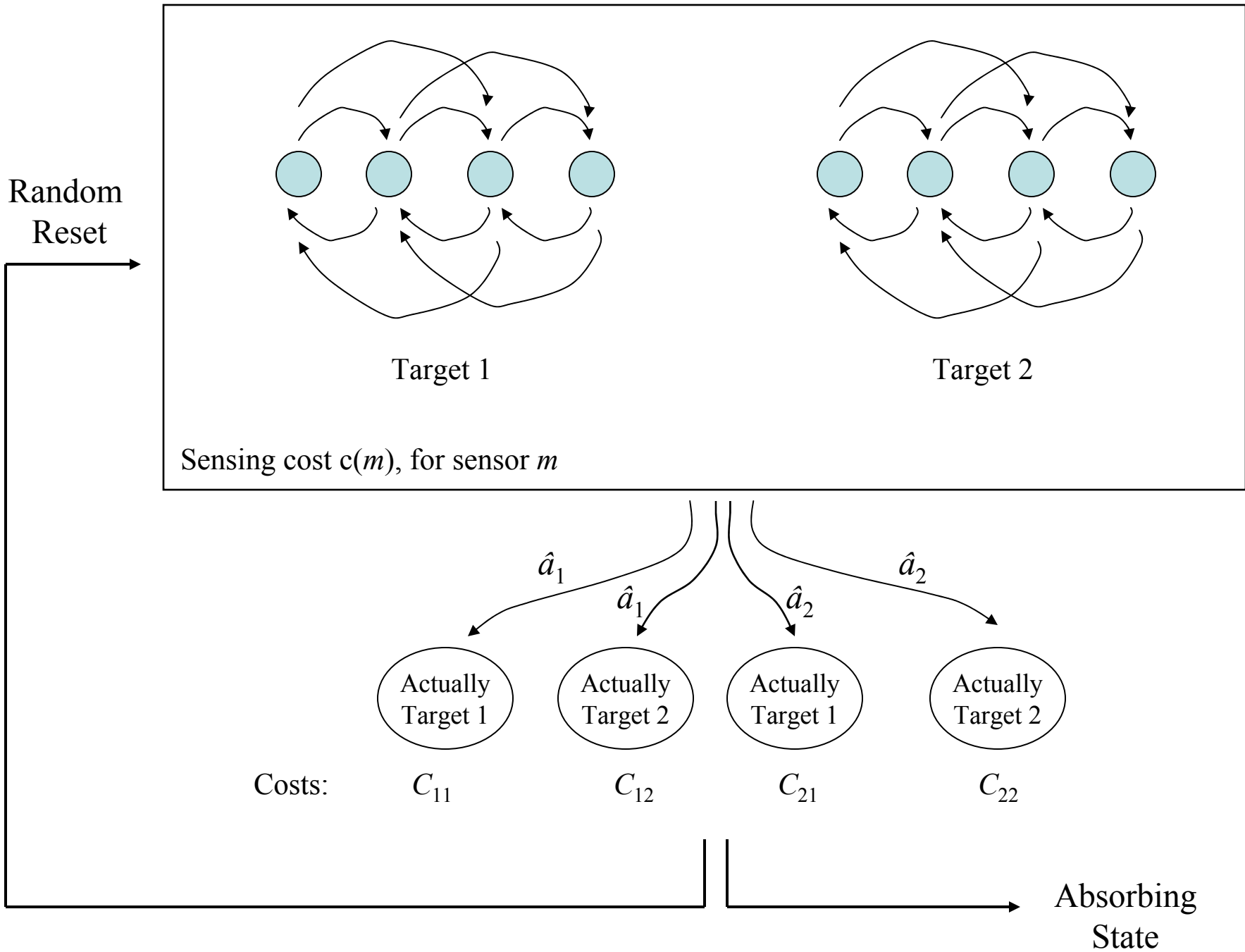
- Value iteration becomes problem of learning the belief-state local slopes  $\alpha(b)$ , for each of which there is an optimal action (policy) – Policy learned by tracking slopes approximately

# Outline

- Summary of the underlying partially-observed model, with corresponding actions
- Partially observed Markov decision processes (POMDPs) and belief states, costs and Bayes risk
- Learning a POMDP policy via value iteration, with a policy defining the optimal action for a given belief state, accounting for discounted infinite horizon (non-myopic)
- Two POMDP implementation strategies for multi-target scattering data
- Myopic or greedy sensing alternative, with a stop criterion
- Example results on scattering data measured by NRL
  - Actions: Selection of optimal target-sensor orientation, fullband data
  - Actions: Selection of optimal target-sensor orientation and frequency subband

# Two Distinct POMDP Formulations

- Infinite horizon with reset upon each classification
  - Appropriate when we wish to perform a sequence of many classifications
  - Multi-target sensing within a “budget”
  - Policy has the opportunity to “opt out” of difficult sensing cases (target ambiguity)
- Algorithm transitions into an “absorbing state” after classification
  - Finite-horizon policy, with horizon dictated by difficulty of initial belief state
  - Does not have opportunity to “opt out” of difficult classification cases



# Outline

- Summary of the underlying partially-observed model, with corresponding actions
- Partially observed Markov decision processes (POMDPs) and belief states, costs and Bayes risk
- Learning a POMDP policy via value iteration, with a policy defining the optimal action for a given belief state, accounting for discounted infinite horizon (non-myopic)
- Two POMDP implementation strategies for multi-target scattering data
- Myopic or greedy sensing alternative, with a stop criterion
- Example results on scattering data measured by NRL
  - Actions: Selection of optimal target-sensor orientation, fullband data
  - Actions: Selection of optimal target-sensor orientation and frequency subband

# Myopic Sensing with a Stop Criterion

- Given belief state  $b_T$ , the *expected* risk after taking a sensing action  $a_{T+1}$  may be expressed as

$$R_E(b_T, a_{T+1}) = \sum_{o \in O} \min_u \left[ \sum_{v=1}^N C_{uv} \sum_{s' \in S_v} \sum_{s \in S} p(o|s', a_{T+1}) p(s'|a_{T+1}, s) b_T(s) \right]$$

- We compute the difference between the cost of action  $a_{T+1}$  and the corresponding expected reduction in risk

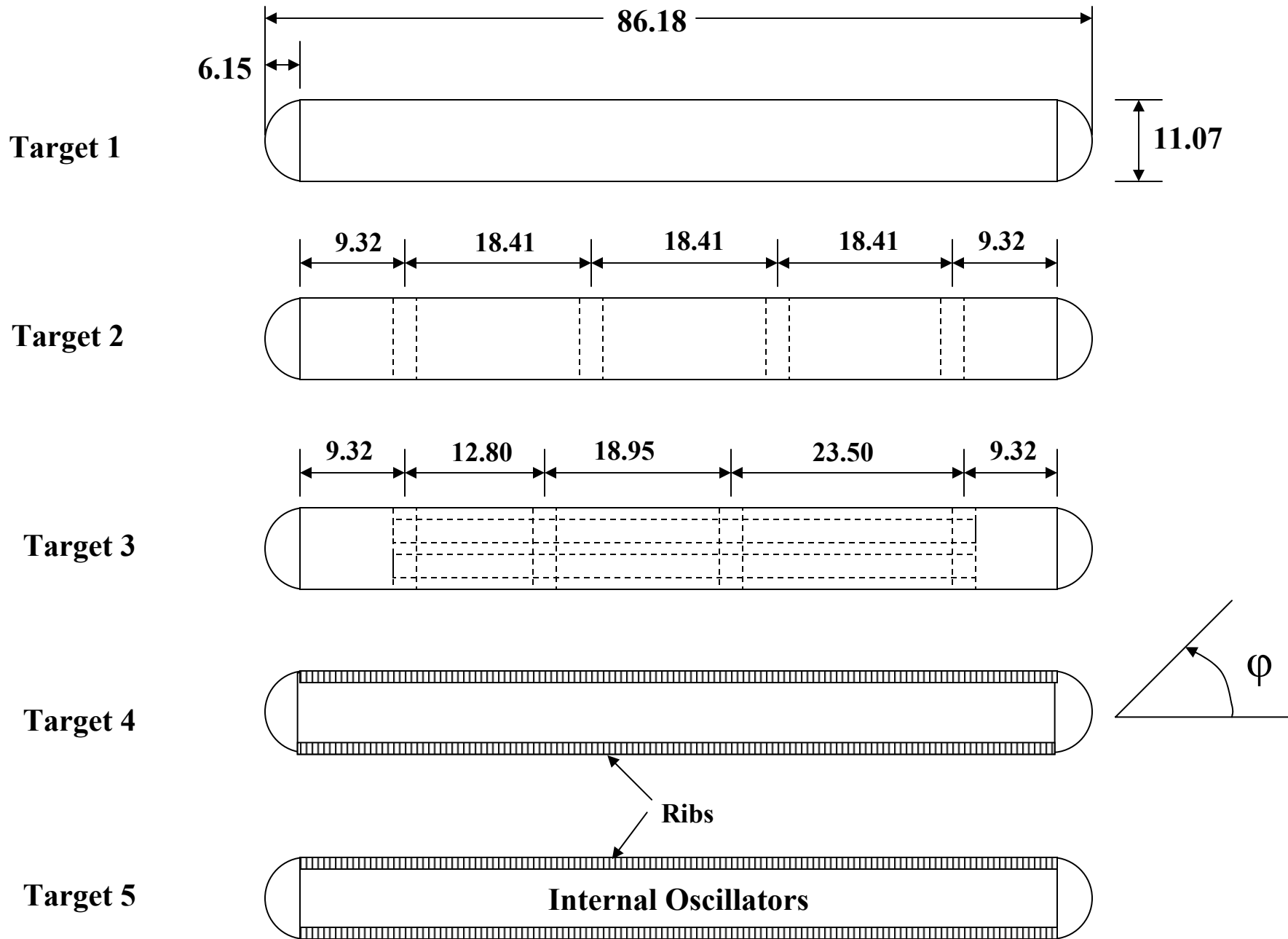
$$\hat{C}(b_T, a_{T+1}) = c(a_{T+1}) - [R(b_T) - R_E(b_T, a_{T+1})]$$

$$R(b_T) = \min_u \left[ \sum_{v=1}^N C_{uv} \sum_{s \in S_v} b_T(s) \right]$$

- We terminate sensing when this difference becomes positive (costs exceed expected reduction in risk)

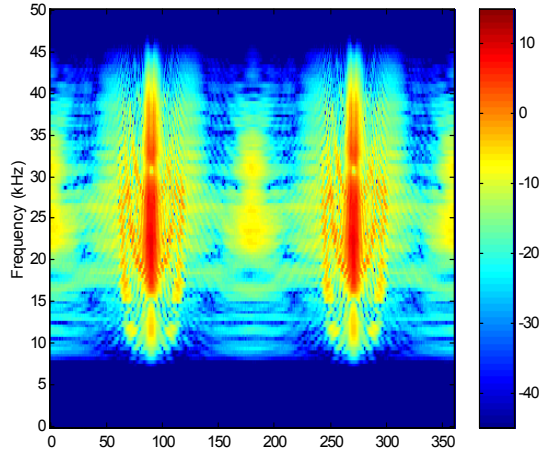
# Outline

- Summary of the underlying partially-observed model, with corresponding actions
- Partially observed Markov decision processes (POMDPs) and belief states, costs and Bayes risk
- Learning a POMDP policy via value iteration, with a policy defining the optimal action for a given belief state, accounting for discounted infinite horizon (non-myopic)
- Two POMDP implementation strategies for multi-target scattering data
- Myopic or greedy sensing alternative, with a stop criterion
- Example results on scattering data measured by NRL
  - Actions: Selection of optimal target-sensor orientation, fullband data
  - Actions: Selection of optimal target-sensor orientation and frequency subband

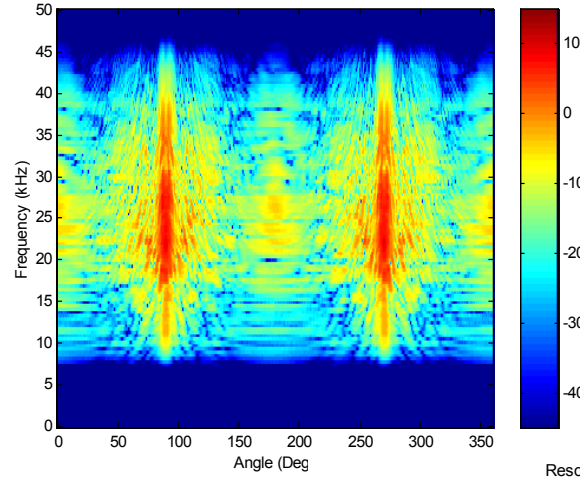


# Multi-Aspect Data

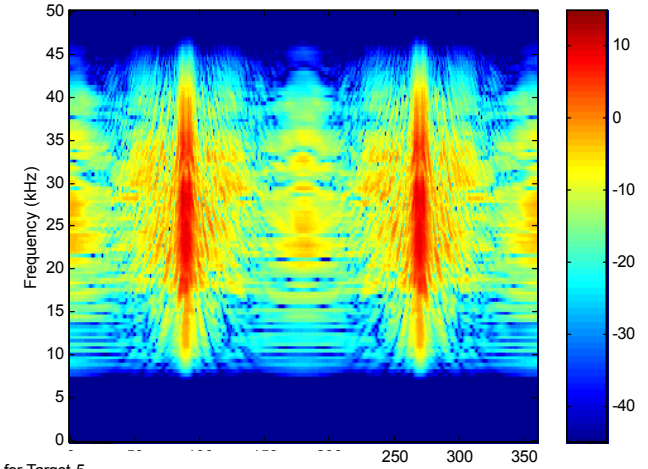
Response for Target 1



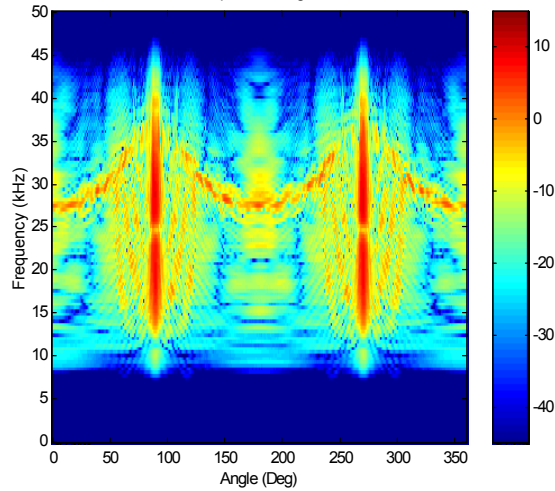
Response for Target 2



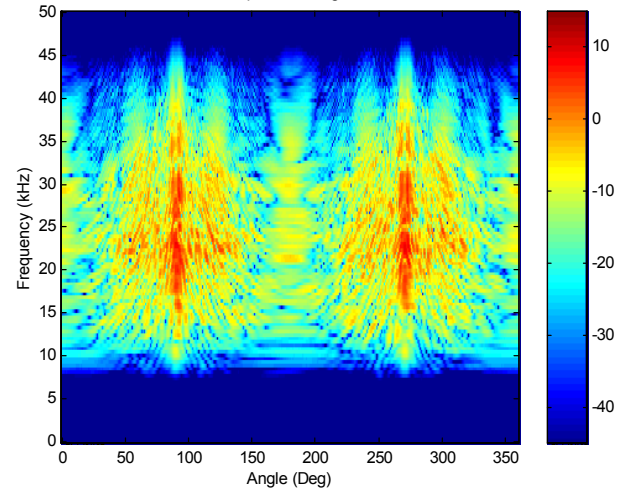
Response for Target 3



Response for Target 4



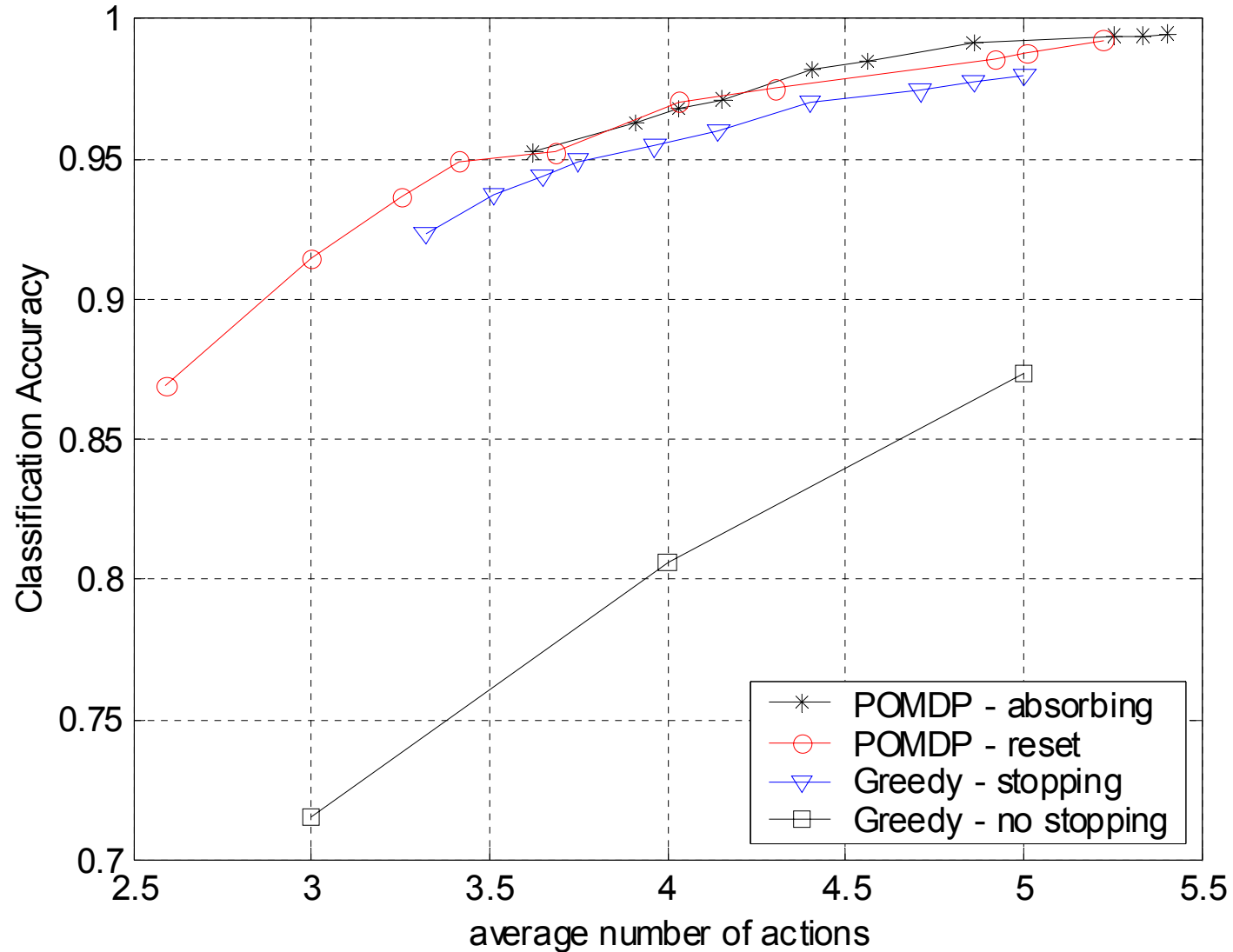
Response for Target 5



# Full-Band Data

## Classification Accuracy vs. Average Number of Actions

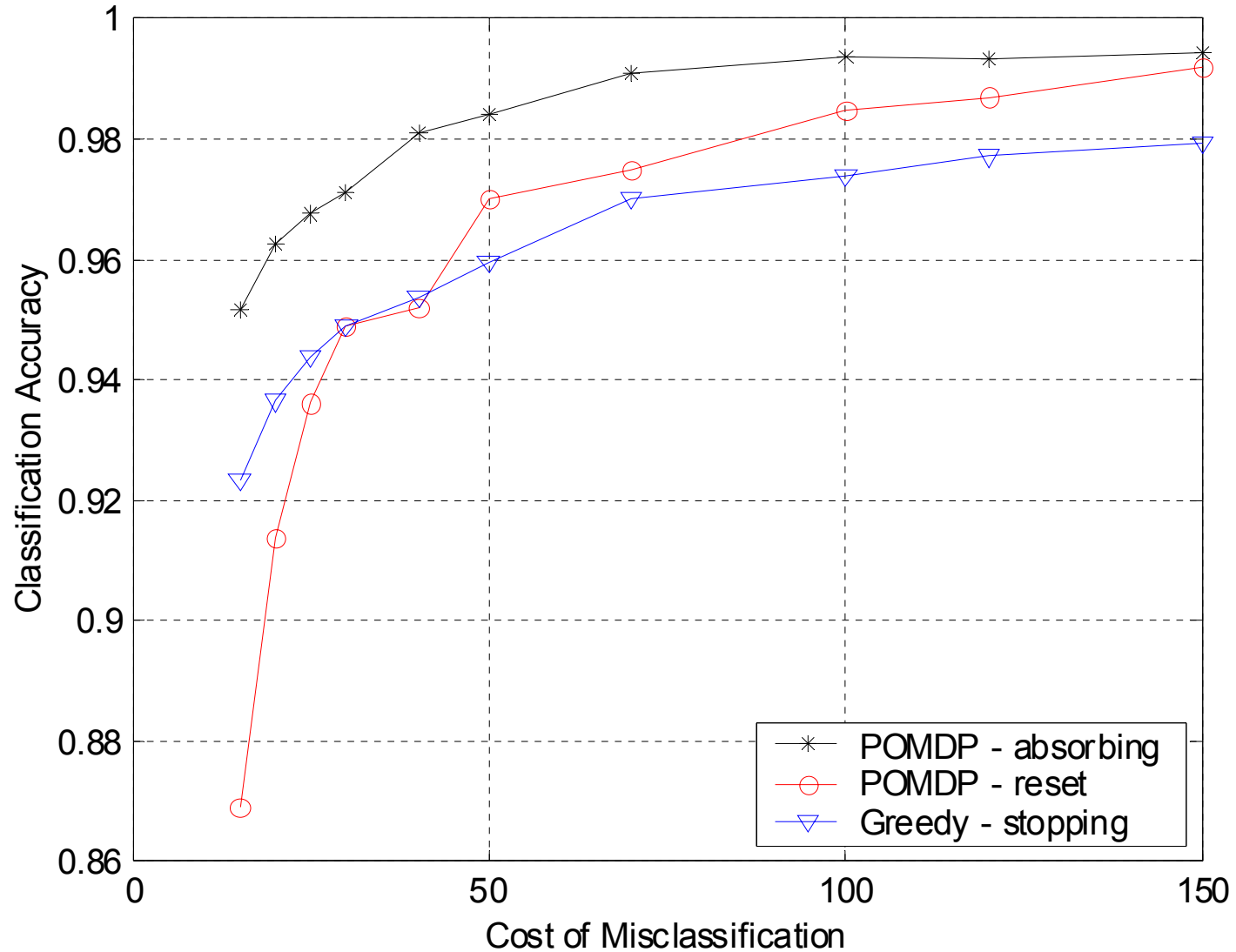
$C_s=1$ ,  $C_{uu}=-10$  and  $C_{uv}=C_c$  with  $C_c$  variable from 15 to 150



# Full-Band Data

## Classification Accuracy vs. Cost of Misclassification

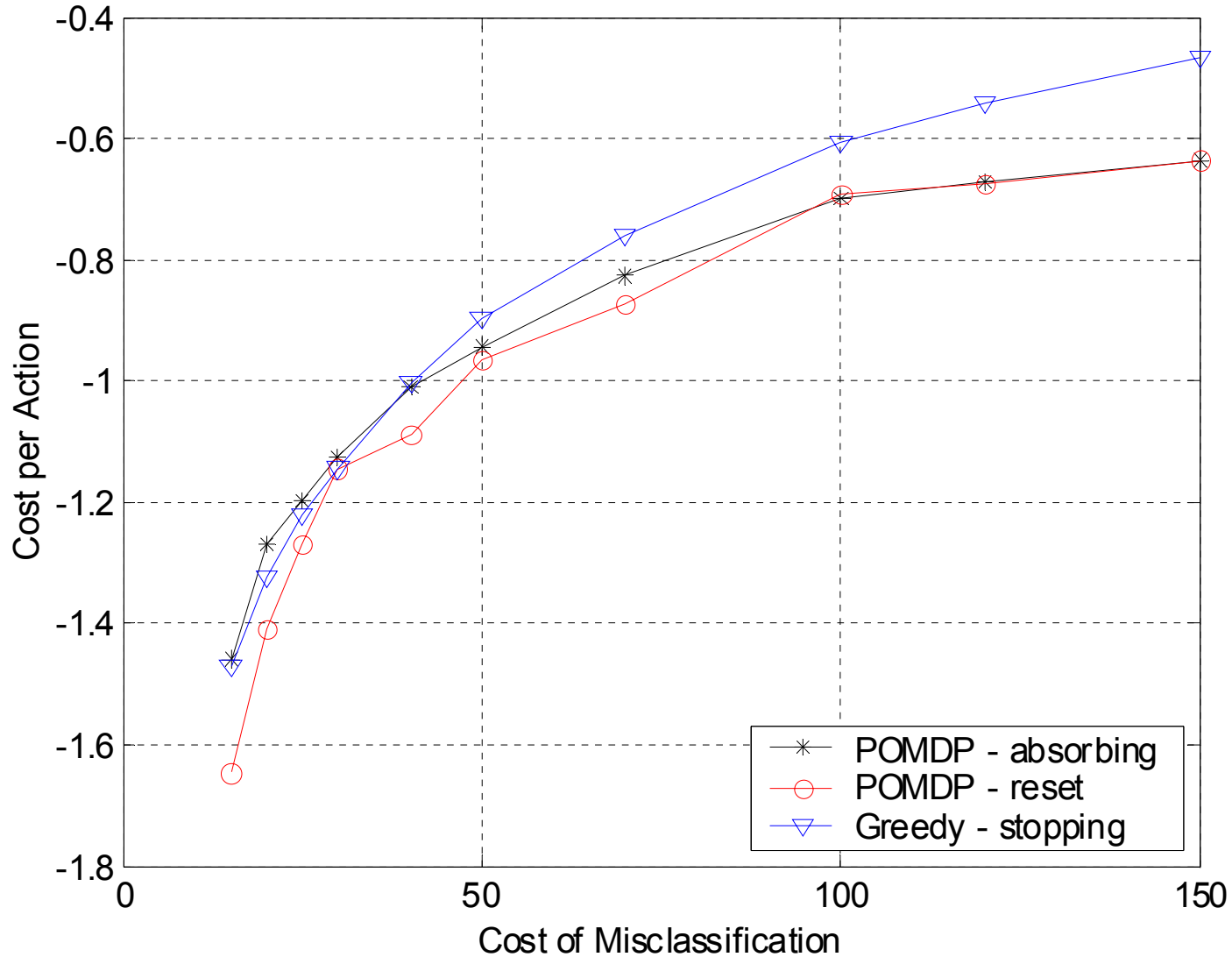
$C_s=1$ ,  $C_{uu}=-10$  and  $C_{uv}=C_c$  with  $C_c$  variable from 15 to 150



# Full-Band Data

## Cost Per Action vs. Cost of Misclassification

$C_s=1$ ,  $C_{uu}=-10$  and  $C_{uv}=C_c$  with  $C_c$  variable from 15 to 150



# Outline

- Summary of the underlying partially-observed model, with corresponding actions
- Partially observed Markov decision processes (POMDPs) and belief states, costs and Bayes risk
- Learning a POMDP policy via value iteration, with a policy defining the optimal action for a given belief state, accounting for discounted infinite horizon (non-myopic)
- Two POMDP implementation strategies for multi-target scattering data
- Myopic or greedy sensing alternative, with a stop criterion
- Example results on scattering data measured by NRL
  - Actions: Selection of optimal target-sensor orientation, fullband data
  - Actions: Selection of optimal target-sensor orientation and frequency subband

# Actions: Select Subband *and* Angle

$$C_s=1, C_{uu}=-10 \text{ and } C_{uv}=C_c \text{ with } C_c=40$$

	Fixed angular sampling, 5°	Angle Selection
Fixed subband: LL	86.11%	91.94% -- 91.33% (2.34)
Fixed subband: HL	72.67%	74.28% -- 86.28% (5.08)
Fixed subband: LH	73.72%	80.72% -- 91.22% (4.94)
Fixed subband: HH	77.72%	87.50% -- 92.67% (3.89)
Fixed band: Fullband	76.50%	84.67% -- 94.61% (4.08)
Subband Selection	90.72% -- 94.72% (3.1406)	93.50% -- 97.17% (2.51)

Black: HMM with fixed angular sampling of 5°, five actions

Blue: Myopic POMDP with a fixed number of five actions

Red: Non-myopic POMDP with reset (average number of actions)

# Summary and Future Work

- Have developed POMDP formulation for general sensing problems, with a policy designed to define the optimal action for a given belief state, accounting for discounted infinite horizon (non-myopic)
- Algorithm operates in real time, and optimally integrates the sensing and signal processing tasks (perfect match for UUVs, for example)
- Key point: The POMDP formulation assumes access to models for the targets, to learn the optimal policy; may not be realistic in many settings
- Reinforcement learning (RL) is a generalization of POMDPs, wherein the sensing actions are not performed simply to *exploit* an underlying model optimally, but the actions also address *exploration* to learn more about a given environment/target that it may not have seen previously
- RL POMDPs optimally execute actions, in a non-myopic setting, to address the exploitation-exploration tradeoff; now extending the research toward RL POMDPs