POMDP Sensor Scheduling – Structural Results

Vikram Krishnamurthy
University of British Columbia

- POMDP Formulation for Sensor Scheduling
- POMDPs with MLR Threshold Policies
- Summary and Extensions: Multi-armed Bandit POMDPs
POMDP Sensor Scheduling Problem

1. Markov Chain Dynamics $X_k \in \{e_1, \ldots, e_S\}$

$$A = [a_{ij}]_{S \times S} \text{ where } a_{ij} = P(X_k = e_j | X_{k-1} = e_i).$$

2. Sensor Choice: $u_{k+1} \in \{1, \ldots, L\}$ from stationary policy $\mu : Z_k \rightarrow u_{k+1}$.

3. Noisy Measurement: (partially observed state)

$$b_i(u_{k+1}, y_{k+1}) = P(y_{k+1} | X_{k+1} = e_i, u_{k+1}).$$

Update History: $Z_{k+1} = Z_k \cup \{y_{k+1}, u_{k+1}\}$.

4. Bayesian State Estimation (Hidden Markov filter):

$$\pi_{k+1} \triangleq \mathbb{E}\{X_{k+1}|Z_{k+1}\} = \frac{B(y_{k+1}, u_{k+1})A'\pi_k}{1'_{S}B(y_{k+1}, u_{k+1})A'\pi_k}, \quad \pi_0 \in \Pi$$

5. Scheduling Cost:

$$\alpha(u_{k+1}) \|X_k - \pi_k\|_D + c(X_k, u_{k+1})$$

**Aim:** Compute $\inf_{\mu} J_\mu = \mathbb{E} \left\{ \sum_{k=0}^{\infty} \beta^k \left[ \alpha(u_{k+1}) \|X_k - \pi_k\|^2 + c(X_k, u_{k+1}) \right] | \pi_0 \right\}$
POMDP Sensor Scheduling Problem

1. Markov Chain Dynamics \( X_k \in \{e_1, \ldots, e_S\} \)
   \[
   A = [a_{ij}]_{S \times S} \text{ where } a_{ij} = P(X_k = e_j|X_{k-1} = e_i).
   \]

2. Sensor Choice: \( u_{k+1} \in \{1, \ldots, L\} \) from stationary policy \( \mu : Z_k \rightarrow u_{k+1} \).

3. Noisy Measurement: (partially observed state)
   \[
   b_i(u_{k+1}, y_{k+1}) = P(y_{k+1}|X_{k+1} = e_i, u_{k+1}).
   \]
   Update History: \( Z_{k+1} = Z_k \cup \{y_{k+1}, u_{k+1}\} \).

4. Bayesian State Estimation (Hidden Markov filter):
   \[
   \pi_{k+1} \triangleq \mathbb{E}\{X_{k+1}|Z_{k+1}\} = \frac{B(y_{k+1}, u_{k+1})A'\pi_k}{1'SB(y_{k+1}, u_{k+1})A'\pi_k}, \quad \pi_0 \in \Pi
   \]

5. Scheduling Cost:
   \[
   \alpha(u_{k+1}) \|X_k - \pi_k\|_D + c(X_k, u_{k+1})
   \]
   \( \text{Aim: Compute } \inf_{\mu} J_\mu = \mathbb{E}\left\{ \sum_{k=0}^{\infty} \beta^k \left[ \alpha(u_{k+1}) \|X_k - \pi_k\|^2 + c(X_k, u_{k+1}) \right] | \pi_0 \right\} \)

Remarks: (i) Nonstandard POMDP.
(ii) Kalman filter scheduling not interesting.
(iii) Average cost problem.
(iv) Straightforward extension: for \( A(u) \).

How to efficiently implement scheduling policy?
1. Myopic heuristic (instantaneous cost)
2. Myopic optimal (e.g., multiarmed bandits)
3. Threshold policy.
Ex 1: ESA Radar/Passive Sensor Bank Sensor Management

Each scheduling slot comprises of $\Delta$ time points.

$u = 1$: Use full kinematic mode tracker for $\Delta$ points.

$u = 2$: Use IR imager for $N_1$ points followed by full tracker for $\Delta - N_1$ points.

Ex 2: Unattended Ground Sensor Network:

Each sensor: 2 webcams + bearings measurement.
Lower level: How to schedule subunits?
Higher level: How to activate sensors? (game)

Ex 3: Optimal search for moving target – stochastic shortest path problem.
Threshold structure for 2 boxes proved [Mcphee, Jordan: 1995].
Singh, Krishnamurthy, IEEE Auto control 2003: conditions for finite stopping time a.s.

Ex 4: Dynamic Spectrum Allocation.
Information State Formulation

\[ \pi_k \triangleq \mathbb{E}\{X_k|Z_k\} = P(X_k = e_1|Z_k) \ldots P(X_k = e_S|Z_k) = \frac{B(y_k,u_k)A'\pi_{k-1}}{1^T S(y_k,u_k)A'\pi_{k-1}}. \]

Simplex: \( \pi_k \in \Pi \triangleq \left\{ \pi \in \mathbb{R}^S : \sum_{i=1}^{S} \pi(i) = 1, \quad 0 \leq \pi(i) \leq 1 \text{ for all } i \in \{1 \ldots, S\} \right\} \) (1)

\[ \pi(1) + \pi(2) = 1 \]

\[ e_2 \quad \pi(1) \quad e_1 \]

\[ 0 \quad \pi(1) \quad 1 \]

\( \Pi = 1 \text{ dim simplex} \)
Information State Formulation

\[ \pi_k \triangleq \mathbb{E}\{X_k | Z_k\} = P(X_k = e_1 | Z_k) \ldots P(X_k = e_S | Z_k) = \frac{B(y_k, u_k) A^T \pi_{k-1}}{1_S B(y_k, u_k) A^T \pi_{k-1}}. \]

Simplex: \( \pi_k \in \Pi \triangleq \left\{ \pi \in \mathbb{R}^S : \sum_{i=1}^{S} \pi(i) = 1, \ 0 \leq \pi(i) \leq 1 \text{ for all } i \in \{1 \ldots, S\} \right\} \) (1)

\[ \pi(1) + \pi(2) = 1 \]

\( \Pi = 1 \text{ dim simplex} \)
Information State Formulation

\[ \pi_k \triangleq \mathbb{E}\{X_k | Z_k\} = P(X_k = e_1 | Z_k) \ldots P(X_k = e_S | Z_k) = \frac{B(y_k, u_k) A' \pi_{k-1}}{\sum_{s} B(y_k, u_k) A' \pi_{k-1}}. \]

Simplex: \( \pi_k \in \Pi \triangleq \left\{ \pi \in \mathbb{R}^S : \sum_{i=1}^{S} \pi(i) = 1, \quad 0 \leq \pi(i) \leq 1 \text{ for all } i \in \{1 \ldots, S\} \right\} \) (1)
Information State Formulation

\[ \pi_k \triangleq \mathbb{E}\{X_k | Z_k\} = P(X_k = e_1 | Z_k) \ldots P(X_k = e_S | Z_k) = \frac{B(y_k, u_k) A' \pi_{k-1}}{\mathbf{1}_S^TB(y_k, u_k)A' \pi_{k-1}}. \]

Simplex: \( \pi_k \in \Pi \triangleq \left\{ \pi \in \mathbb{R}^S : \sum_{i=1}^{S} \pi(i) = 1, \ 0 \leq \pi(i) \leq 1 \text{ for all } i \in \{1 \ldots, S\} \right\} \) (1)
Information State Formulation

$$\pi_k \triangleq \mathbb{E}\{X_k | Z_k\} = P(X_k = e_1 | Z_k) \ldots P(X_k = e_S | Z_k) = \frac{B(y_k, u_k) A' \pi_{k-1}}{\mathbf{1}_S^T B(y_k, u_k) A' \pi_{k-1}}.$$ 

Simplex: \( \pi_k \in \Pi \triangleq \left\{ \pi \in \mathbb{R}^S : \sum_{i=1}^{S} \pi(i) = 1, \quad 0 \leq \pi(i) \leq 1 \text{ for all } i \in \{1 \ldots, S\} \right\} \) (1)

$${\pi(1) + \pi(2) = 1}$$

$$\Pi = 1 \text{ dim simplex}$$

$$\Pi = 2 \text{ dim simplex}$$

$$\Pi = 3 \text{ dim simplex}$$

Mean Square Error QoS: \( \|X_k - \pi_k\|_D = \|X_k - \pi_k\|_2^2 \text{ or } \|X_k - \pi_k\|_\infty \text{ or } \|X_k - \pi_k\|_1 : \)

$$J_\mu = \mathbb{E} \left\{ \sum_{k=0}^{\infty} \beta^k \left[ \alpha(u_{k+1}) \|X_k - \pi_k\|_2^2 + c(X_k, u_{k+1}) \right] |\pi_0 \right\}$$

$$= \mathbb{E} \left\{ \sum_{k=0}^{\infty} \beta^k \left[ \alpha(u_{k+1}) \left(1 - \pi'_k \pi_k\right) + c'(u_{k+1} \pi_k) \right] |\pi_0 \right\}$$
DP formulation for POMDPs

Belman’s equation (value iteration) for Mean Square Error QoS:

\[ V_{k+1}(\pi) = \min_{u \in U} \left( c(u)'\pi + \alpha(u)(1 - \pi'\pi) + \beta \sum_{Y \in Y} V_k \left( \frac{B(y_{k+1}, u_{k+1})A'\pi}{1' SB(y_{k+1}, u_{k+1)}A'\pi} \right) 1'_S B(y_{k+1}, u_{k+1})A'\pi \right) \]

Entropy Error QoS: \( \|X_k - \pi_k\|_D = \sum_i \pi(i) \log \pi(i) \)

\[ \inf_{\mu} J_\mu = \mathbb{E} \left\{ \sum_{k=0}^{\infty} \beta^k \left[ \alpha(u_{k+1}) \sum_{i=1}^{S} \pi_k(i) \log \pi_k(i) + c'(u_{k+1})\pi_k \right] | \pi_0 \right\} \]

MAP probability estimator QoS: \( \|X_k - \pi_k\|_D = 1 - I(X_k = e_{i^*}), i^* = \text{argmax}_i \pi_k(i) \).

\[ \inf_{\mu} J_\mu = \mathbb{E} \left\{ \sum_{k=0}^{\infty} \beta^k \left[ \alpha(u_{k+1})(1 - \max_i \{\pi_k(i)\}) + c'(u_{k+1})\pi_k \right] | \pi_0 \right\} \]

For standard POMDP: Sondik 1970s: \( V_k(\pi) = \min_{\gamma_i, k \in \Gamma_k} \gamma_i, k, \pi \) (PWLC).

PSPACE hard but numerous algorithms: e.g., Witness algorithms [1998], etc.

For nonstandard POMDP: No closed form solution to DP recursion.

**Aim:** (i) Prove existence of threshold Scheduling Policy (under suitable conditions)
(ii) Compute parameterized threshold via simulation based stochastic approximation.
Approx Value Iteration for nonstandard POMDP

How to deal with quadratic term? Construct a piecewise linear approximation.

(Krishnamurthy, IEEE Signal Proc, 2002)
Lower bound: Interpolation.
Upper bound: Tangents.

In higher dimensions – Freudenthal triangularization

\[ e_2 = [0, 1, 0]' \]
\[ e_3 = [0, 0, 1]' \]
\[ e_1 = [1, 0, 0]' \]
Approx Value Iteration for nonstandard POMDP

How to deal with quadratic term? Construct a piecewise linear approximation.

(Krishnamurthy, IEEE Signal Proc, 2002)
Lower bound: Interpolation.
Upper bound: Tangents.

In higher dimensions – Freudenthal triangularization
Approx Value Iteration for nonstandard POMDP

How to deal with quadratic term? Construct a piecewise linear approximation.

(Krishnamurthy, IEEE Signal Proc, 2002)
Lower bound: Interpolation.
Upper bound: Tangents.

In higher dimensions – Freudenthal triangularization

Sandwich Theorem: For $R$th degree interpolation

$$ V_k(\pi) \leq \bar{V}_k(\pi) \leq \tilde{V}_k(\pi), \quad k = N, N - 1, \ldots, 0 $$

$$ \max_{\pi \in \mathcal{P}} \| V_0(\pi) - \tilde{V}_0(\pi) \| \leq \frac{K}{R^2} $$

Also Lovejoy’s [Journal OR 1991] approximation applies.
Approx Value Iteration for nonstandard POMDP

**How to deal with quadratic term?** Construct a piecewise linear approximation.

(Krishnamurthy, IEEE Signal Proc, 2002)

Lower bound: Interpolation.
Upper bound: Tangents.

In higher dimensions – Freudenthal triangularization

**Sandwich Theorem:** For $R$th degree interpolation

$$\bar{V}_k(\pi) \leq V_k(\pi) \leq \bar{V}_k(\pi), \quad k = N, N - 1, \ldots, 0$$

$$\max_{\pi \in \mathcal{P}} \|V_0(\pi) - \bar{V}_0(\pi)\| \leq \frac{K}{R^2}$$

Also Lovejoy’s [Journal OR 1991] approximation applies.
Approx Value Iteration for nonstandard POMDP

How to deal with quadratic term? Construct a piecewise linear approximation.

(Krishnamurthy, IEEE Signal Proc, 2002)
Lower bound: Interpolation.
Upper bound: Tangents.

In higher dimensions – Freudenthal triangularization

Sandwich Theorem: For $R$th degree interpolation

$$\bar{V}_k(\pi) \leq V_k(\pi) \leq \tilde{V}_k(\pi), \quad k = N, N - 1, \ldots, 0$$

$$\max_{\pi \in \mathcal{P}} \|V_0(\pi) - \bar{V}_0(\pi)\| \leq \frac{K}{R^2}$$

Also Lovejoy’s [Journal OR 1991] approximation applies.
**Approx Value Iteration for nonstandard POMDP**

*How to deal with quadratic term?* Construct a piecewise linear approximation.

(Krishnamurthy, IEEE Signal Proc, 2002)

Lower bound: Interpolation.
Upper bound: Tangents.

In higher dimensions – Freudenthal triangularization

**Sandwich Theorem:** For $R$th degree interpolation

\[
\bar{V}_k(\pi) \leq V_k(\pi) \leq \bar{V}_k(\pi), \quad k = N, N - 1, \ldots, 0
\]

\[
\max_{\pi \in P} \|V_0(\pi) - \bar{V}_0(\pi)\| \leq \frac{K}{R^2}
\]

Also Lovejoy’s [Journal OR 1991] approximation applies.
Approx Value Iteration for nonstandard POMDP

How to deal with quadratic term? Construct a piecewise linear approximation.

(Krishnamurthy, IEEE Signal Proc, 2002)
Lower bound: Interpolation.
Upper bound: Tangents.

In higher dimensions – Freudenthal triangularization

Sandwich Theorem: For $R$th degree interpolation

$$
\bar{V}_k(\pi) \leq V_k(\pi) \leq \tilde{V}_k(\pi), \quad k = N, N - 1, \ldots, 0
$$

$$
\max_{\pi \in \mathcal{P}} \| V_0(\pi) - \tilde{V}_0(\pi) \| \leq \frac{K}{R^2}
$$

Also Lovejoy’s [Journal OR 1991] approximation applies.
Approx Value Iteration for nonstandard POMDP

How to deal with quadratic term? Construct a piecewise linear approximation.

(Krishnamurthy, IEEE Signal Proc, 2002)
Lower bound: Interpolation.
Upper bound: Tangents.

In higher dimensions – Freudenthal triangularization

Sandwich Theorem: For $R$th degree interpolation

$$\tilde{V}_k(\pi) \leq V_k(\pi) \leq V_{k-1}(\pi), \quad k = N, N-1, \ldots, 0$$

$$\max_{\pi \in \mathcal{P}} \| V_0(\pi) - \tilde{V}_0(\pi) \| \leq \frac{K}{R^2}$$

Also Lovejoy’s [Journal OR 1991] approximation applies.
Approx Value Iteration for nonstandard POMDP

How to deal with quadratic term? Construct a piecewise linear approximation.

(Krishnamurthy, IEEE Signal Proc, 2002)
Lower bound: Interpolation.
Upper bound: Tangents.

In higher dimensions – Freudenthal triangularization

Sandwich Theorem: For $R$th degree interpolation

\[ \bar{V}_k(\pi) \leq V_k(\pi) \leq \overline{V}_k(\pi), \quad k = N, N - 1, \ldots, 0 \]
\[ \max_{\pi \in \mathcal{P}} \| V_0(\pi) - \bar{V}_0(\pi) \| \leq \frac{K}{R^2} \]

Also Lovejoy’s [Journal OR 1991] approximation applies.
Outline

• Nonstandard POMDP Formulation for Sensor Scheduling
• POMDPs with MLR Threshold Policies
• Summary and Extensions: Multi-armed Bandit POMDPs
Monotone Policies for POMDPs

Unstructured optimal policy

What structure in $A$, $B$, $C$ leads to a threshold optimal policy for a POMDP?

$$V_{k+1}(\pi) = \max_{u \in \mathbf{u}} C(\pi, u) + \beta \sum_{Y \in \mathbf{Y}} V_k \left( \frac{B(y_{k+1}, u_{k+1}) A' \pi_k}{1' S B(y_{k+1}, u_{k+1}) A' \pi_k} \right) 1' S B(y_{k+1}, u_{k+1}) A' \pi_k.$$

$$\mu^*_{k+1}(\pi) = \arg\max_{u \in \mathbf{u}} Q_{k+1}(\pi, u).$$

Threshold Policy

$$u_{k+1} = \mu^*(\pi_k) = \begin{cases} 1 & \text{if } \pi_k \leq T_1 \\ 2 & \text{if } T_1 \leq \pi_k \leq T_2, \text{ etc.} \end{cases}$$

Q1: $\pi_k$ is a probability vector – how do we order probability vectors $\pi$?
Define problem on poset $[\Pi, \leq_r]$ where $\leq_r$ is the likelihood ratio ordering.

Q2: Under what conditions is $\mu^*(\pi) = \arg\max_{u \in \mathbf{u}} Q_{k+1}(\pi, u)$ increasing wrt $\leq_r$ on $\Pi$?
Supermodularity.
Some Definitions

**Monotone Likelihood Ratio ordering**: Why? Ideal for POMDPS since conditioning on any info preserves MLR ordering.

**Defn**: Consider $\pi_1 \in \Pi, \pi_2 \in \Pi$. Then $\pi_1 \succeq_r \pi_2$, if $\pi_1(i)/\pi_2(i)$ is increasing in $i$, i.e,

$$\pi_1(i)\pi_2(j) \leq \pi_2(i)\pi_1(j), \quad i < j, i, j \in \{1, \ldots, S\}.$$ 

**Example**: $\pi_1 = \begin{bmatrix} 0.1 & 0.2 & 0.7 \end{bmatrix}$ and $\pi_2 = \begin{bmatrix} 0.3 & 0.2 & 0.5 \end{bmatrix}$, then $\pi_1 \succeq_r \pi_2$

$\pi_1 = \begin{bmatrix} 0.1 & 0.2 & 0.7 \end{bmatrix}$ and $\pi_2 = \begin{bmatrix} 0.1 & 0.3 & 0.6 \end{bmatrix}$ are not MLR orderable.

**Characterization of MLR orderable sequences** on information state poset $[\Pi, \leq_r]$:

(i) For all $\pi \in \Pi$, $e_1 \leq_r \pi \leq_r e_S$.
(ii) $\pi \leq_r \tilde{\pi} \implies \pi \leq_r \gamma \pi + (1 - \gamma)\tilde{\pi} \leq_r \tilde{\pi}, \quad 0 \leq \gamma \leq 1$.
(iii) All points on line connected to $e_S$ or $e_1$ are MLR comparable. $e_1 \leq \pi_1 \leq \pi_2 \leq \pi_3$.

**Supermodularity** [Topkis, 1970s] $Q(\pi, u)$ is supermodular if it has increasing differences:

$$Q(\pi, u) - Q(\pi, \bar{u}) \geq Q(\bar{\pi}, u) - Q(\bar{\pi}, \bar{u}), \quad \bar{u} \leq u, \bar{\pi} \leq_r \pi.$$ 

**Result**: [Topkis] If $Q : \Pi \times u \rightarrow \mathbb{R}$ is supermodular, then $u^*(\pi) = \arg\max_{\pi \in \Pi} Q(\pi, u)$ is MLR increasing on $\Pi$, i.e., $\pi \leq_r \tilde{\pi} \implies u^*(\pi) \leq u^*(\tilde{\pi})$.

POMDP Example: If $u \in \{1, 2\}$, $u^*(\pi)$ is threshold along MLR increasing trajectory!
**Sufficient Conditions for POMDP Threshold Policy**

**Aim:** To give sufficient conditions for $Q(\pi, u)$ to be supermodular.

Two sensors: $u \in \{1, 2\}$.

$$Q(\pi, u) = C(\pi, u) + \beta \sum_{Y \in Y} V_k \left( \frac{B(y_{k+1}, u_{k+1}) A' \pi_k}{1'_S B(y_{k+1}, u_{k+1}) A' \pi_k} \right) 1'_S B(y_{k+1}, u_{k+1}) A' \pi_k$$

$$\mu^*_{k+1}(\pi) = \arg\max_{u \in u} Q_{k+1}(\pi, u).$$

**Supermodularity Theorem:** If $Q(\pi, u)$ is supermodular, i.e., $Q(\pi, 2) - Q(\pi, 1)$ is MLR increasing in $\pi$, then the optimal policy is threshold:

$$u_{k+1} = \mu^*(\pi_k) = \begin{cases} 
1 & \text{if } \pi_k \leq_T T \\
2 & \text{if } T \leq_T \pi_k
\end{cases}, \quad T \in \Pi$$

How to find sufficient conditions on $A, B, C$ for $Q(\pi, u)$ to be supermodular?

Suppose $Q(\pi, 1)$ is indpt of $\pi$. Suffices to show $Q(\pi, 2)$ is MLR increasing.

**Under what conditions is $Q(\pi, 2)$ MLR increasing?**

**Sufficient condition:** The following are MLR increasing in $\pi$:

$C(\pi, 2) - C(\pi, 1)$, $V_k(\pi)$, $1'_S B(y_{k+1}, u_{k+1}) A' \pi_k$ and $\frac{B(y_{k+1}, u_{k+1}) A' \pi_k}{1'_S B(y_{k+1}, u_{k+1}) A' \pi_k}$.

$V_k(\pi)$ MLR increasing if $C(\pi, u)$ MLR decreasing.
**Flavor of Proof:** $C(\pi, u)$ is MLR decreasing in $\pi$.

Since $\pi \geq_r \tilde{\pi}$ implies $\pi \geq_s \tilde{\pi}$ it suffices to prove that $\pi \geq_s \tilde{\pi} \implies C(\pi, u) < C(\tilde{\pi}, u)$.

Parameterize

\[ \tilde{\pi} = \pi^{\epsilon_1, \epsilon_2, \ldots, \epsilon_{S-1}} \triangleq \pi + \epsilon_1(e_1 - e_2) + \epsilon_2(e_2 - e_3) + \cdots + \epsilon_{S-1}(e_{S-1} - e_S). \]

$\pi^{\epsilon_1, \epsilon_2, \ldots, \epsilon_{S-1}}$ is stochastically decreasing in $\epsilon_i, i = 1, \ldots, S-1$. So suffices to show $C'(\pi^{\epsilon_1, \epsilon_2, \ldots, \epsilon_{S-1}}, u)$ is increasing in $\epsilon_i$, i.e.,

\[ \frac{d}{d\epsilon_i} C(\pi^{\epsilon_1, \epsilon_2, \ldots, \epsilon_{S-1}}, u) \geq 0, \quad i = 1, \ldots, S - 1. \]

More general result: Stick to MLR ordering.

\[ \tilde{\pi} = \pi^\epsilon \triangleq (1 - \epsilon)\pi + \epsilon e_S, \quad 0 \leq \epsilon \leq 1 \]
Main Theorem on MLR Threshold Policy for Nonstandard POMDP:

(A1) Sensor usage cost satisfies $c(e_i, u) - c(e_{i+1}, u) \geq 2\alpha(u)$

(A2) $A$ is TP2 if all its second order minors are non-negative

(A3) Symbol probabilities satisfy $b_i(u) \geq_r b_j(u)$ for $i \geq j$

(S): Sensor usage costs satisfy $c(i, 1) - c(i, 2) \geq c(i + 1, 1) - c(i + 1, 2) + 2[\alpha(1) - \alpha(2)]$.

(I): For action $u = 1$, $\pi_{k+1}$ is indpt of $\pi_k$: This is quite restrictive! (e.g., machine replacement problem or cued target acquisition in MFR Radar scheduling).

Then the optimal policy $\mu^*(\pi)$ is MLR increasing in $\pi$.

Multiscale Threshold Policy

$u_{k+1} = \mu^*(\pi_k) = \begin{cases} 1 & \text{if } \pi_k \leq_r \tau(\bar{\pi}_1) \\ 2 & \text{if } \tau(\bar{\pi}_1) \leq_r \pi_k \leq_r e_S \end{cases}$

How to estimate MLR Threshold?

Approach 1: $\tau^* = \bigcup_{\bar{\pi} \in \mathcal{H}} \tau(\bar{\pi})$ over grid in $\mathcal{H}$ (dense in $\Pi$).

Approach 2: Estimate best parameterized MLR threshold $\tau^*(\theta)$. We will estimate best linear MLR threshold using stochastic approximation.

Constraint: All points in $\Pi$ are MLR comparable wrt threshold.
Best Linear MLR Threshold Policy

\[ \mu_\theta(\pi) = \begin{cases} 
1 & \theta'\pi \leq 0 \\
2 & \theta'\pi > 0 
\end{cases} \]

**Theorem:** Sufficient condition for a linear threshold \( \theta \) to be MLR comparable with simplex \( \Pi \) is \( \theta(1) \leq \theta(2) \leq \cdots \leq \theta(S) \).

**Threshold based Policy Gradient Algorithm**

Assume POMDP parameters satisfy assumptions (A1), (A2), (A3), (S), (I), so that the optimal policy \( \mu^*(\pi) \) is MLR threshold.

For iterations \( n = 1, 2, \ldots \)

- Evaluate sample cost \( J_n(\mu_{\hat{\theta}}) = \sum_{k=1}^{N} \beta^k C(\pi_k, \mu_\theta(\pi_k)) \) using policy \( \mu_{\hat{\theta}} \) via simulation.

- Update threshold (we used Simultaneous Perturbation Stochastic Approximation)

\[ \hat{\theta}_{n+1} = \hat{\theta}_n - \epsilon_{n+1} \nabla_{\theta} J_n[\hat{\theta}_n] \]

W.p.1 Convergence Proof: Markovian noise (e.g. Kushner and Yin).

**Remark:** We have presented a simplified story. Can obtain simpler conditions for threshold policy by restricting MLR ordering over lines (rather than entire simplex).
Target Dynamics (random walk) 

\[ A = \begin{pmatrix} 
q & 1 - q & \cdots & \cdots & 0 \\
1 - \frac{q}{2} & q & 1 - \frac{q}{2} & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
0 & \cdots & 0 & 1 - q & q 
\end{pmatrix} \]

\[ p_i = P(y_k = e_i|X_k = e_i), \quad b_i(y_t, u = l) = \begin{cases} 
0 & \text{if } |y_t - i| > 1 \\
1 - p_i/2 & \text{if } |y_t - i| = 1 \\
p_i & \text{if } y_t = i 
\end{cases} \]

Cost: \( c(\text{distance}, u) \) decreasing in distance: resources not wasted on low threat target. 
Differential threat: \( c(\text{distance, less accurate}) - c(\text{distance, more accurate}) \) increases when target is nearer. 

Then: Threshold Scheduling Policy.
Summary

1. Optimal scheduling for a POMDP is threshold wrt MLR ordering.

   \[ u_{k+1} = \mu^*(\pi_k) = \begin{cases} 
   1 & \text{if } \pi_k \leq T_1 \\
   2 & \text{if } T_1 \leq \pi_k \leq T_2, \text{ etc.} 
   \end{cases} \]

   Generalizes Lovejoy’s 1987 paper.

2. **Major limitation**: Require reset of information state for action 2.


4. Can use supermodularity to bound a “myopic” policy.

5. We have presented a simplified story. Can obtain simpler conditions for threshold policy by restricting MLR ordering over lines (rather than entire simplex).

6. Conditions for threshold policy: Can one extend to multivariate stochastic orderings?

7. Another scalable methodology: POMDP Multiarmed Bandits – for 100s of actions but few states.
POMDP Multiarmed Bandits
POMDP Multiarmed Bandits

Which target should the radar look at?
POMDP Multiarmed Bandits

Which target should the radar look at?

Ion Channel Activation on Chip

- How to dynamically schedule which ion channels to activate using laser?
- Laser: inert caged-ligands (e.g. caged glutamate) become active ligands.

Tradeoff: Desensitization and mortality vs Information Extraction.

Partially observed **Multi-armed bandits**. Playing golf with multiple balls.
References

Sensor Scheduling, MDPs, POMDPs