Part 5: Partially Observed Markov Decision Process

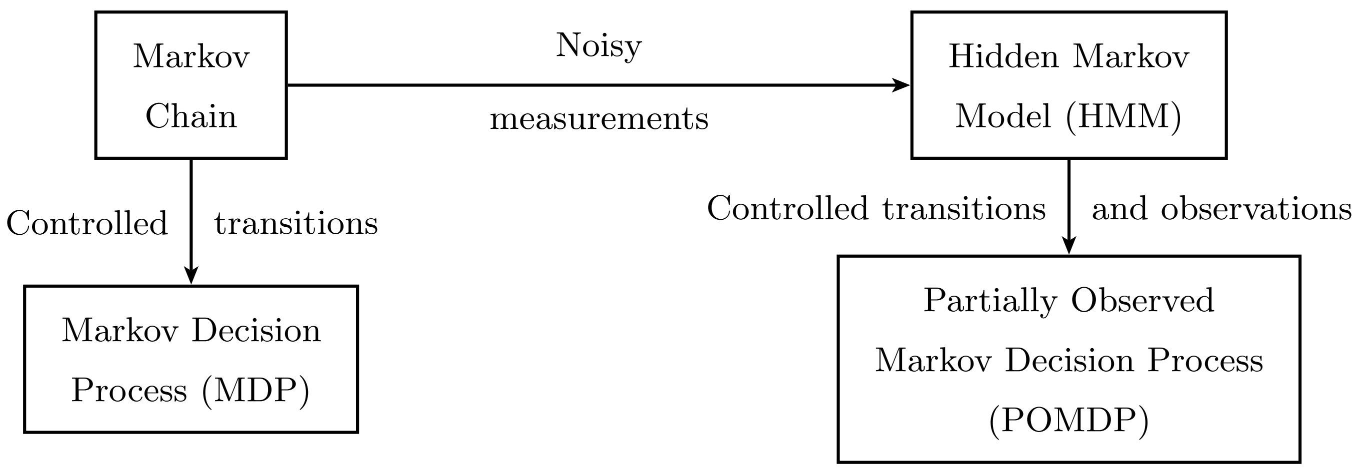
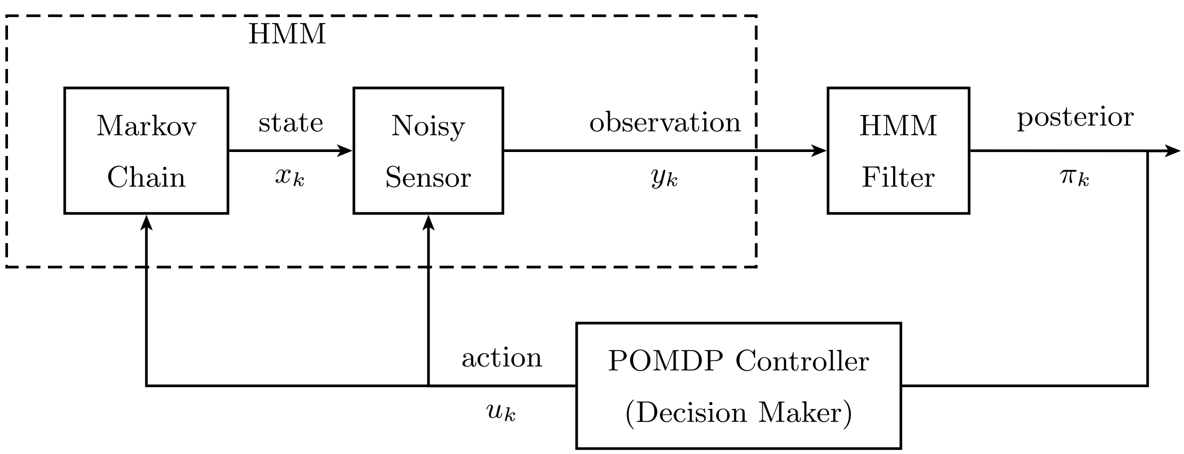


Figure 1: Terminology of HMMs, MDPs and POMDPs

Main Idea: Convert a POMDP into a fully observed MDP. The resulting fully observed problem is in terms of the belief (information) state, namely, HMM filtered density.



Finite horizon POMDP

A POMDP with finite horizon N is a 7-tuple

$$(\mathcal{X}, \mathcal{U}, \mathcal{Y}, P(u), B(u), c(u), c_N).$$

1. $\mathcal{X} = \{1, 2, \ldots, X\}$ denotes the state space and $x_k \in \mathcal{X}$ is controlled Markov chain $k = 0, 1, \ldots, N$.
2. $\mathcal{U} = \{1, 2, \ldots, U\}$ denotes the action space with $u_k \in \mathcal{U}$ denoting the action chosen at time k
3. $\mathcal{Y}$ is observation space (finite or a subset of $\mathbb{R}$) . $y_k \in \mathcal{Y}$ is observation at time $k \in \{1, 2, \ldots, N\}$.
4. For action $u \in \mathcal{U}$, $P(u)$ is transition matrix

$$P_{ij}(u) = \mathbb{P}(x_{k+1} = j | x_k = i, u_k = u), \quad i, j \in \mathcal{X}.$$

5. For $u \in \mathcal{U}$, $B(u)$ is observation distribution

$$B_{iy}(u) = \mathbb{P}(y_{k+1} = y | x_{k+1} = i, u_k = u), \quad i \in \mathcal{X}, y \in \mathcal{Y}.$$

6. For state x_k and action u_k, incurs cost $c(x_k, u_k)$.
7. A terminal cost $c_N(x_N)$ is incurred.

Aim: $\boldsymbol{\mu}^* = \operatorname{argmin}_{\boldsymbol{\mu}} J_{\boldsymbol{\mu}}(\pi_0)$ for any initial prior π_0 where

$$J_{\boldsymbol{\mu}}(\pi_0) = \mathbb{E}_{\boldsymbol{\mu}}\{\sum_{k=0}^{N-1} c(x_k, u_k) + c_N(x_N) \mid \pi_0\}.$$

$\mathbb{E}_{\boldsymbol{\mu}}$ wrt $(x_0, y_0, x_1, y_1, \ldots, x_{N-1}, y_{N-1}, x_N, y_N)$.

Controller does not observe x_k. Only observes y_k

Knows cost matrix $c(x, u)$ but not cost at time k.

Belief State Formulation

Fully observed MDP: $u_k = \mu_k^*(x_k)$.

POMDP: $u_k = \mu_k^*(\mathcal{I}_k)$, where $\mathcal{I}_k = (\pi_0, u_0, y_1, \dots, u_{k-1}, y_k)$

Define the posterior distribution of Markov chain given $\mathcal{I}_k$

$\pi_k(i) = \mathbb{P}(x_k = i | \mathcal{I}_k)$, $i \in \mathcal{X}$ where $\mathcal{I}_k = \{\pi_0, u_0, y_1, \dots, u_{k-1}, y_k\}$

X-dimensional probability vector $\pi_k = [\pi_k(1), \dots, \pi_k(X)]'$ is the *belief state* or *information state* at time k.
Computed via HMM filter $\pi_k = T(\pi_{k-1}, y_k, u_{k-1})$ where

$$T(\pi, y, u) = \frac{B_y(u) P'(u) \pi}{\sigma(\pi, y, u)}, \quad \text{where } \sigma(\pi, y, u) = \mathbf{1}_X' B_y(u) P'(u) \pi,$$

$$B_y(u) = \operatorname{diag}\big(B_{1y}(u), \cdots, B_{Xy}(u)\big).$$

Main point: optimal controller operates on belief state

$$u_k = \mu_k^*(\pi_k).$$

Belief space: $\mathit{\Pi}(X)$ is called the *belief space.*

$$\mathit{\Pi}(X) \stackrel{\text{defn}}{=} \Big\{\pi \in \mathbb{R}^X : \mathbf{1}'\pi = 1, \quad 0 \leq \pi(i) \leq 1 \text{ for all } i \in \mathcal{X}\Big\}.$$

$\mathit{\Pi}(2)$ is a one dimensional simplex (unit line segment),

$\Pi(3)$ is equilateral triangle, $\Pi(4)$ is tetrahedron.

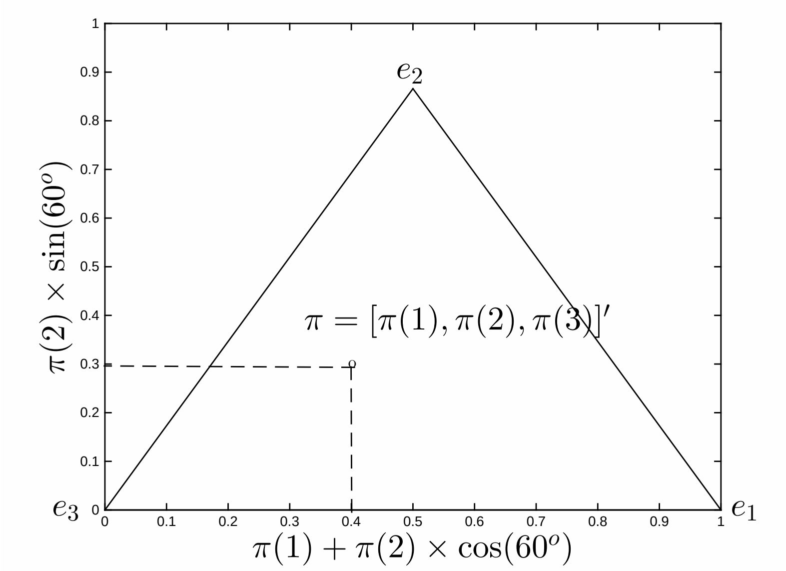


Figure 2: $\Pi(X)$ for $X = 3$.

Belief State Formulation of POMDP objective:

$$J_{\boldsymbol{\mu}}(\pi_0) = \mathbb{E}_{\boldsymbol{\mu}}\{\sum_{k=0}^{N-1} c(x_k, u_k) + c_N(x_N) \mid \pi_0\}$$

$$\overset{(a)}{=} \mathbb{E}_{\boldsymbol{\mu}}\{\sum_{k=0}^{N-1} \mathbb{E}\{c(x_k, u_k) \mid \mathcal{I}_k\} + \mathbb{E}\{c_N(x_N) \mid \mathcal{I}_N\} \mid \pi_0\}$$

$$= \mathbb{E}_{\boldsymbol{\mu}}\{\sum_{k=0}^{N-1} \sum_{i=1}^{X} c(i, u_k)\, \pi_k(i) + \sum_{i=1}^{X} c_N(i)\, \pi_N(i) \mid \pi_0\}$$

$$= \mathbb{E}_{\boldsymbol{\mu}}\{\sum_{k=0}^{N-1} c'_{u_k}\, \pi_k + c'_N\, \pi_N \mid \pi_0\}$$

(a) uses smoothing property of conditional expectations

$$c_u = \begin{bmatrix} c(1,u) & \cdots & c(X,u) \end{bmatrix}',\ c_N = \begin{bmatrix} c_N(1) & \cdots & c_N(X) \end{bmatrix}'.$$

Real Time POMDP Controller

State x_0 is simulated from initial distribution π_0.

For time $k = 0, 1, \ldots, N-1$:

- Step 1: Based on belief π_k, choose $u_k = \mu_k(\pi_k) \in \mathcal{U}$.
- Step 2: The decision maker incurs a cost $c'_{u_k}\pi_k$.
- Step 3: The state evolves with transition probability $P_{x_k x_{k+1}}(u_k)$ to the next state x_{k+1} at time $k+1$.

$$P_{ij}(u) = \mathbb{P}(x_{k+1} = j | x_k = i, u_k = u).$$

- Step 4: The decision-maker records $y_{k+1} \in \mathcal{Y}$

$$\mathbb{P}(y_{k+1} = y | x_{k+1} = i, u_k = u) = B_{iy}(u).$$

- Step 5: Update belief state $\pi_{k+1} = T(\pi_k, y_{k+1}, u_k)$ using HMM filter

Theorem 1. *For finite horizon POMDP*

1. $J_{\boldsymbol{\mu}^*}(\pi)$ *is achieved by deterministic policies*

$$\boldsymbol{\mu}^* = (\mu_0^*, \mu_1^*, \ldots, \mu_{N-1}^*), \qquad \textit{where } u_k = \mu_k^*(\pi_k).$$

2. *Optimal policy* $\boldsymbol{\mu}^* = (\mu_0, \mu_1, \ldots, \mu_{N-1})$ *satisfies Bellman's DP: Initialize* $J_N(\pi) = c'_N \pi$.

$$J_k(\pi) = \min_{u \in \mathcal{U}} \{c'_u \pi + \sum_{y \in \mathcal{Y}} J_{k+1}\left(T(\pi, y, u)\right) \sigma(\pi, y, u)\}$$

$$\mu_k^*(\pi) = \operatorname*{argmin}_{u \in \mathcal{U}} \{c'_u \pi + \sum_{y \in \mathcal{Y}} J_{k+1}\left(T(\pi, y, u)\right) \sigma(\pi, y, u)\}.$$

for $k = N-1, \ldots, 0$.

Toy Example: Machine Replacement

$\mathcal{X} = \{1 \text{ (poor) }, 2\text{(good)}\}$, $\mathcal{U} \in \{1\text{(replace) }, 2 \text{ (keep)}\}$.

$$P(1) = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad P(2) = \begin{bmatrix} 1 & 0 \\ \theta & 1-\theta \end{bmatrix}.$$

$\theta \in [0,1]$: probability that machine deteriorates.
$y_k \in \mathcal{Y} = \{1 \text{ bad product}, 2 \text{ good product}\}$:

$$B = \begin{bmatrix} p & 1-p \\ 1-q & q \end{bmatrix}.$$

Replacement cost: $c(x, u=1) = R$.
Aim: Minimize $\mathbb{E}_{\boldsymbol{\mu}}\{\sum_{k=0}^{N-1} c(x_k, u_k) | \pi_0\}$
DP equation: Initialize $J_N(\pi) = 0$ (no terminal cost) and for $k = N-1, \ldots, 0$:

$$J_k(\pi) = \min\{c_1'\pi + J_{k+1}(e_2), \;\; c_2'\pi + \sum_{y \in \{1,2\}} J_{k+1}(T(\pi, y, 2))\sigma(\pi, y, 2)\}$$

where $T(\pi, y, 2) = \dfrac{B_y P'(2)\pi}{\sigma(\pi, y, 2)}$, $\sigma(\pi, y, 2) = \mathbf{1}' B_y P'(2)\pi$, $y \in \{1,2\}$,

$$B_1 = \begin{bmatrix} p & 0 \\ 0 & 1-q \end{bmatrix}, \;\; B_2 = \begin{bmatrix} 1-p & 0 \\ 0 & q \end{bmatrix}.$$

$\mathit{\Pi}(X)$ is a one dimensional simplex $[0,1]$. So $J_k(\pi)$ can be expressed in terms of $\pi_2 \in [0,1]$, because $\pi_1 = 1 - \pi_2$.

Implement DP numerically by discretizing $\pi_2 \in [0,1]$.

Finite Dimensional Controller for POMDP

Even though $\Pi(X)$ is continuum, Bellman's equation

$$J_k(\pi) = \min_{u \in \mathcal{U}} \{ c_u' \pi + \sum_{y \in \mathcal{Y}} J_{k+1} \left(\frac{B_y(u) P'(u) \pi}{\mathbf{1}' B_y(u) P'(u) \pi} \right) \mathbf{1}' B_y(u) P'(u) \pi \}$$

for finite horizon POMDP has a finite dimensional characterization when $\mathcal{Y}$ is finite. (Sondik)

Theorem 2. *Consider POMDP with $\mathcal{U} = \{1, 2, \ldots, U\}$ and finite observation space $\mathcal{Y} = \{1, 2, \ldots, Y\}$. Then $J_k(\pi)$ and $\mu_k^*(\pi)$ have finite dimensional characterization:*

1. *$J_k(\pi)$ is piecewise linear and concave wrt $\pi \in \Pi(X)$:*

   $$J_k(\pi) = \min_{\gamma \in \Gamma_k} \gamma' \pi.$$

   Γ_k is a finite set of X-dim vectors.
   $J_N(\pi) = c_N' \pi$ and $\Gamma_N = \{c_N\}$
2. *$\Pi(X)$ can be partitioned into at most $|\Gamma_k|$ convex polytopes. In each polytope $\mathcal{R}_l = \{\pi : J_k(\pi) = \gamma_l' \pi\}$, $\mu_k^*(\pi)$ is a constant corresponding to a single action.*

   $$\mu_k^*(\pi) = u(\operatorname*{argmin}_{\gamma_l \in \Gamma_k} \gamma_l' \pi)$$

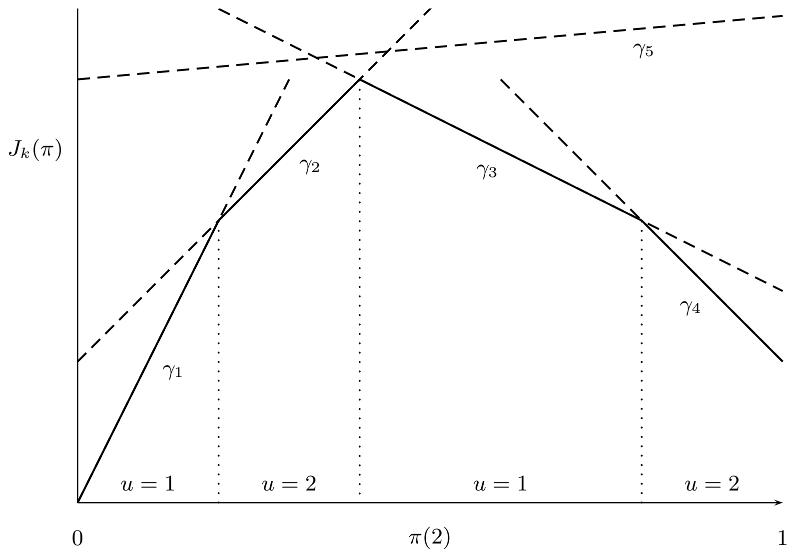


Figure 3: Example of piecewise linear concave value function $J_k(\pi)$ of 2-state POMDP. Here $J_k(\pi) = \min\{\gamma_1'\pi, \gamma_2'\pi, \gamma_3'\pi, \gamma_4'\pi\}$ is depicted by solid lines. Belief space can be partitioned into 4 regions. Each region where line segment $\gamma_l'\pi$ is active (i.e., is equal to the solid line) corresponds to a single action, $u=1$ or $u=2$. Note that γ_5 is never active.

Proof: By backward induction for $k = N, \ldots, 0$.
Clearly, $J_N(\pi) = c_N'\pi$ is linear in π.
Assume $J_{k+1}(\pi)$ is piecewise linear and concave in π: so

$$J_{k+1}(\pi) = \min_{\bar{\gamma} \in \Gamma_{k+1}} \bar{\gamma}'\pi$$

Substituting this in DP yields

$$J_k(\pi) = \min_{u\in\mathcal{U}}\{c_u'\pi + \sum_{y\in\mathcal{Y}} \min_{\bar{\gamma}\in\Gamma_{k+1}} \frac{\bar{\gamma}' B_y(u)P'(u)\pi}{\cancel{\sigma(\pi,u,y)}} \cancel{\sigma(\pi,u,y)}\}$$

$$= \min_{u\in\mathcal{U}}\left\{\sum_{y\in\mathcal{Y}} \min_{\bar{\gamma}\in\Gamma_{k+1}} \left\{\left[\frac{c_u}{Y} + P(u)B_y(u)\bar{\gamma}\right]'\pi\right\}\right\}.$$

RHS is the minimum (over u) of the sum (over y) of piecewise linear concave functions. These preserve piecewise linear concave property. So $J_k(\pi)$ is piecewise linear and concave

$$J_k(\pi) = \min_{\gamma\in\Gamma_k} \gamma'\pi,$$

$$\text{where} \quad \Gamma_k = \cup_{u\in\mathcal{U}} \oplus_{y\in\mathcal{Y}} \left\{\frac{c_u}{Y} + P(u)B_y(u)\,\bar{\gamma} \mid \bar{\gamma}\in\Gamma_{k+1}\right\}.$$

Here $A \oplus B$ consists of all pairwise additions of vectors from these two sets.

Lemma 3. *The value function of a POMDP is positive homogeneous. That is for $\alpha \geq 0$, $J_n(\alpha\,\pi) = \alpha J_n(\pi)$. As a result, Bellman's equation*

$$J_k(\pi) = \min_{u\in\mathcal{U}}\{c_u'\pi + \sum_{y\in\mathcal{Y}} J_{k+1}\left(\frac{B_y(u)P'(u)\pi}{\mathbf{1}'B_y(u)P'(u)\pi}\right)\mathbf{1}'B_y(u)P'(u)\pi\}.$$

becomes

$$J_k(\pi) = \min_{u \in \mathcal{U}} \{ c_u' \pi + \sum_{y \in \mathcal{Y}} J_{k+1}\left(B_y(u) P'(u) \pi \right) \}.$$

Exact Algorithms for Finite horizon POMDPs

Bellman's dynamic programming recursion

$$\begin{aligned} Q_k(\pi, u, y) &= \frac{c_u' \pi}{Y} + J_{k+1}\left(T(\pi, y, u)\right) \sigma(\pi, y, u) \\ Q_k(\pi, u) &= \sum_{y \in \mathcal{Y}} Q_k(\pi, u, y) \\ J_k(\pi) &= \min_u Q_k(\pi, u). \end{aligned} \tag{1}$$

Construct Γ_k that form piecewise linear value function

$$\begin{aligned} \Gamma_k(u, y) &= \left\{ \frac{c_u}{Y} + P(u) B_y(u) \gamma \mid \gamma \in \Gamma^{(k+1)} \right\} \\ \Gamma_k(u) &= \oplus_y \Gamma_k(u, y) \\ \Gamma_k &= \cup_{u \in \mathcal{U}} \Gamma_k(u). \end{aligned} \tag{2}$$

$A \oplus B$ consists of all pairwise additions of vectors.

Γ_k constructed by (2) may contain vectors that never arise in the value function $J_k(\pi) = \min_{\gamma_l \in \Gamma_k} \gamma_l' \pi$.

Incremental Pruning Algorithm.

Given Γ_{k+1} generate Γ_k as follows:

For each $u \in \mathcal{U}$

For each $y \in \mathcal{Y}$

$$\Gamma_k(u,y) \leftarrow \text{prune}\left(\left\{\frac{c_u}{Y} + P(u)B_y(u)\gamma \mid \gamma \in \Gamma^{(k+1)}\right\}\right)$$

$$\Gamma_k(u) \leftarrow \text{prune}\left(\Gamma_k(u) \oplus \Gamma_k(u,y)\right)$$

$$\Gamma_k \leftarrow \text{prune}\left(\Gamma_k \cup \Gamma_k(u)\right)$$

Linear programming dominance test can be used to identify inactive vectors:

$$\begin{aligned} &\min\ x \\ \text{subject to: } &(\gamma - \bar{\gamma})'\pi \geq x, \quad \forall \bar{\gamma} \in \Gamma - \{\gamma\} \\ &\pi(i) \geq 0, i \in \mathcal{X}, \quad \mathbf{1}'\pi = 1, \quad \text{i.e. } \pi \in \Pi(X). \end{aligned} \tag{3}$$

If LP yields solution $x \geq 0$, then γ dominates all other vectors in $\Gamma - \{\gamma\}$. Then vector γ is inactive and can be eliminated from Γ. In the worst case, it is possible that all vectors are active and none can be pruned.

Examples: Monahan's Algorithm, Witness Algorithm.

Lovejoy's Suboptimal Algorithm.

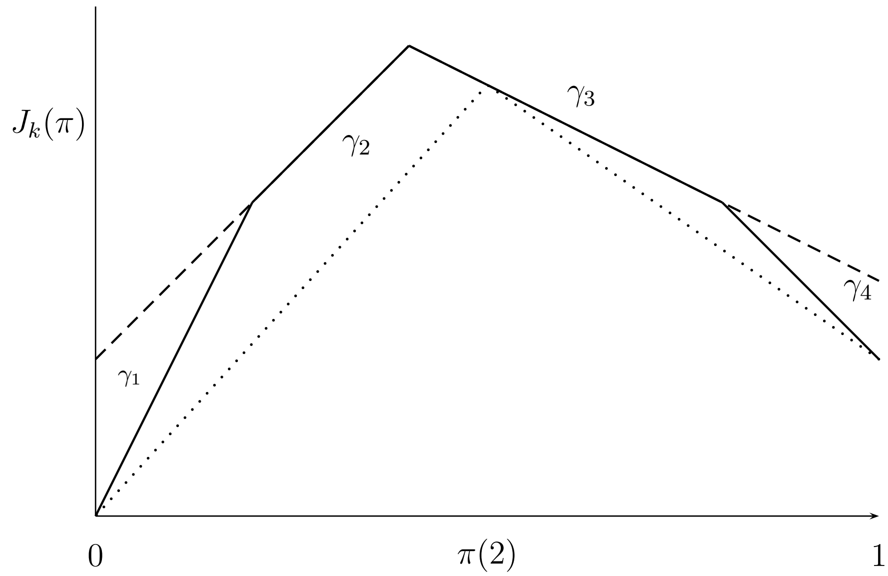


Figure 4: Interpolation (dotted lines) yields a lower bound to the value function. Omitting any piecewise linear segments leads to an upper bound (dashed lines).

Upper bound is computed as follows:

Initialize: $\bar{\Gamma}_N = \Gamma_N = \{c_N\}$.
Step 1: Given Γ_k, construct set $\bar{\Gamma}_k$ by pruning: Pick any R belief states $\pi_1, \pi_2, \ldots, \pi_R$. Set

$$\bar{\Gamma}_k = \{\arg\min_{\gamma \in \Gamma_k} \gamma' \pi_r, \quad r = 1, 2, \ldots, R\}.$$

Step 2: Given $\bar{\Gamma}_k$, compute the set of vectors Γ_{k-1} using a standard POMDP algorithm.

Step 3: $k \to k-1$ and go to Step 1.

Open Loop Feedback Control

Open Loop Feedback Control (OLFC) for POMDP with finite horizon N

For $n = 0, \dots, N-1$

1. Given belief π_n, evaluate expected cumulative cost

$$C_n(u_n, \dots, u_{N-1}) = \mathbb{E}\{\sum_{k=n}^{N-1} c(x_k, u_k) + c_N(x_N)|\mathcal{I}_n\}$$

   where $\mathcal{I}_n = (\pi_0, u_0, y_1, \dots, u_{n-1}, y_n)$

$$= c'_{u_n}\pi_n + c'_{u_{n+1}}P'_{u_n}\pi_n + c'_{u_{n+2}}P'_{u_{n+1}}P'_{u_n}\pi_n + \cdots + c'_N P'_{u_{N-1}} \cdots P'_{u_{n+2}}P'_{u_{n+1}}\pi_n$$

   for $|U|^{N-n}$ possible sequences $u_n, \dots, u_{N-1}$.
2. Evaluate $(u_n^*, \dots, u_{N-1}^*) = \operatorname{argmin} C_n(u_n, \dots, u_{N-1})$
3. Use action u_n^* to obtain observation y_{n+1}
4. Update belief $\pi_{n+1} = T(\pi_n, y_{n+1}, u_n^*)$ using HMM filter.

Since π_{n+1} depends on u_n^* and affects the choice of action u_{n+1}^* there is *feedback control* in the algorithm.

Open loop control is a special case where only first iteration $n = 0$ is performed yielding $(u_0^*, \dots, u_{N-1}^*)$

Theorem 4. *OLFC results in an expected cumulative cost that is at least as small as open loop control.*

- $\bar{J}_0(\pi_0)$: expected cost incurred with OLFC and $(\bar{\mu}_0, \ldots, \bar{\mu}_{N-1})$ is policy. Then for $n = N-1, \ldots, 0$
$$\bar{J}_n(\pi_n) = c'_{\bar{\mu}_n(\pi_n)}\pi_n + \bar{J}_{n+1}(T(\pi_n, y_{n+1}, \bar{\mu}_n(\pi_n))),$$
initialized by terminal cost $\bar{J}_N(\pi_N) = c'_N \pi_N$.
- $J_n^{\mathrm{h}}(\pi_0)$: expected cost incurred by hybrid strategy:
  - Apply OLFC from time 0 to $n-1$ and compute π_n.
  - Then apply loop control from time n to $N-1$.

  Clearly, open loop control cumulative cost is $J_0^{\mathrm{h}}(\pi_0)$.

Then need to prove $\bar{J}_0(\pi_0) \leq J_0^{\mathrm{h}}(\pi_0)$. We will prove $\bar{J}_n(\pi_0) \leq J_n^{\mathrm{h}}(\pi_0)$ by backward induction for $n = N, \ldots 0$. By definition $\bar{J}_N(\pi_0) = J_N^{\mathrm{h}}(\pi_0)$. Next assume the induction hypothesis $\bar{J}_{n+1}(\pi) \leq J_{n+1}^{\mathrm{h}}(\pi)$ for all $\pi \in \Pi(X)$.

$$\begin{aligned}
\bar{J}_n(\pi_n) &= c'_{\bar{\mu}_n(\pi_n)}\pi_n + \bar{J}_{n+1}(T(\pi_n, y_{n+1}, \bar{\mu}_n(\pi_n))) \\
&\leq c'_{\bar{\mu}_n(\pi_n)}\pi_n + J_{n+1}^{\mathrm{h}}(T(\pi_n, y_{n+1}, \bar{\mu}_n(\pi_n))) \quad \text{(induction hypothesis)} \\
&= c'_{\bar{\mu}_n(\pi_n)}\pi_n + \mathbb{E}\left\{\min_{u_{n+1},\ldots,u_{N-1}} C_{n+1}(u_{n+1}, \ldots, u_{N-1}) | \mathcal{I}_n\right\} \\
&= c'_{\bar{\mu}_n(\pi_n)}\pi_n + \mathbb{E}\left\{\min_{u_{n+1},\ldots,u_{N-1}} \mathbb{E}\{\sum_{k=n+1}^{N-1} c(x_k, u_k) + c_N(x_N) | \mathcal{I}_{n+1}\} \mid \mathcal{I}_n\right\} \\
&\leq c'_{\bar{\mu}_n(\pi_n)}\pi_n + \min_{u_{n+1},\ldots,u_{N-1}} \mathbb{E}\left\{\sum_{k=n+1}^{N-1} c(x_k, u_k) + c_N(x_N) | \mathcal{I}_n\right\}
\end{aligned}$$

POMDPs in Controlled Sensing

To incorporate uncertainty of the state estimate,

$$c(x_k, u_k) + d(x_k, \pi_k, u_k), \quad u_k \in \mathcal{U} = \{1, 2, \ldots, U\}.$$

(i) *Sensor Usage Cost*: $c(x_k, u_k)$
(ii) *Sensor Performance Loss*: $d(x_k, \pi_k, u_k)$ explicit function of π_k captures uncertainty in state estimate.

Accurate sensors: high usage but low performance loss.

Denote $\mathcal{I}_k = \{\pi_0, u_0, y_1, \ldots, u_{k-1}, y_k\}$. Then

$$C(\pi_k, u_k) = \mathbb{E}\{c(x_k, u_k) + d(x_k, \pi_k, u_k) | \mathcal{I}_k\}$$

$$= c'_{u_k} \pi_k + D(\pi_k, u_k), \text{ where } c_u = (c(u,1), \ldots, c(u,X))',$$

$$D(\pi_k, u_k) \stackrel{\text{defn}}{=} \mathbb{E}\{d(x_k, \pi_k, u_k) | \mathcal{I}_k\} = \sum_{i=1}^{X} d(e_i, \pi_k, u_k)\, \pi_k(i).$$

$$D_N(\pi) = \mathbb{E}\{d_N(x, \pi_N) | \mathcal{I}_N\} = \sum_{i=1}^{X} d_N(e_i, \pi_N) \pi_N(i),$$

$$C_N(\pi) = c'_N \pi_N + D_N(\pi_N).$$

Example. Mean Square, l_1 and l_∞ Loss:

$d(x,\pi,u) = \alpha(u)(x-\pi)'M(x-\pi)+\beta(u),\ x \in \{e_1,\dots,e_X\}, \pi \in \Pi$.

$$
\begin{aligned}
D(\pi_k,u_k) &= \mathbb{E}\{d(x_k,\pi_k,u_k)|\mathcal{I}_k\} \\
&= \alpha(u_k)\Big(\sum_{i=1}^{X} M_{ii}\pi_k(i) - \pi_k' M \pi_k\Big) + \beta(u_k)
\end{aligned}
$$

because
$\mathbb{E}\{(x_k-\pi_k)'M(x_k-\pi_k)|\mathcal{I}_k\} = \sum_{i=1}^{X}(e_i-\pi)'M(e_i-\pi)\pi(i)$.
Alternatively, if $d(x,\pi,u) = \|x-\pi\|_1$ then
$D(\pi,u) = 2(1-\pi'\pi)$ is also quadratic in the belief. Also, choosing $d(x,\pi,u) = \|x-\pi\|_\infty$ yields $D(\pi,u) = (1-\pi'\pi)$.

Entropy based Performance Loss:

$$
D(\pi,u) = -\alpha(u)\sum_{i=1}^{S}\pi(i)\log_2\pi(i) + \beta(u), \qquad \pi \in \Pi. \quad (4)
$$

Theorem 5. *Consider a POMDP with possibly continuous-valued observations. Assume that for each action u, the instantaneous cost $C(\pi,u)$ and terminal cost $C_N(\pi,u)$ are concave and continuous with respect to $\pi \in \Pi(X)$. Then the value function $J_k(\pi)$ is concave in π.*

Discounted Inifinite Horizon POMDP

$$J_\mu(\pi_0) = \mathbb{E}_\mu \left\{ \sum_{k=0}^{\infty} \rho^k c(x_k, u_k) \right\}, \qquad \text{where } u_k = \mu(\pi_k)$$
$$= \mathbb{E}_\mu \left\{ \sum_{k=0}^{\infty} \rho^k c'_{\mu(\pi_k)} \pi_k \right\}$$

For any finite horizon N that

$$J_k(\pi) = \min_{u \in \mathcal{U}} \{ \rho^k c'_u \pi + \sum_{y \in \mathcal{Y}} J_{k+1}\left(T(\pi, y, u)\right) \sigma(\pi, y, u) \}$$

Convenient to use forward iteration of indices. Define:

$$V_n(\pi) = \rho^{n-N} J_{N-n}(\pi), \quad 0 \leq n \leq N, \ \pi \in \Pi(X).$$

Then $V_n(\pi)$ satisfies DP equation

$$V_n(\pi) = c'_u \pi + \rho \sum_{y \in Y} V_{n-1}\left(T(\pi, y, u)\right) \sigma(\pi, y, u), \quad V_0(\pi) = 0.$$

Theorem 6. *The optimal policy $\mu^*(\pi)$ and value function $V(\pi)$ satisfy Bellman's dynamic programming equation*

$$\mu^*(\pi) = \operatorname*{argmin}_{u \in \mathcal{U}} Q(\pi, u), \quad V(\pi) = \min_{u \in \mathcal{U}} Q(\pi, u),$$
$$Q(\pi, u) = c'_u \pi + \rho \sum_{y \in Y} V\left(T(\pi, y, u)\right) \sigma(\pi, y, u).$$

Value function $V(\pi)$ is continuous and concave in π.

Existence & uniqueness of a solution to Bellman's equation for infinite horizon discounted cost POMDP.
We prove Bellman's equation is a contraction mapping.
So by the fixed point theorem a unique solution exists.

Banach Fixed Pt Thm:

- Given Banach space (X, d), $T : X \to X$ is contraction mapping if $\exists \alpha \in [0, 1)$ s.t. $x_1, x_2 \in X$ implies $d\big(T(x_1), T(x_2)\big) < \alpha\, d(x_1, x_2)$.
- Given Banach space (X, d) and contraction mapping $T : X \to X$. Then T admits a unique fixed point $x^* \in X$, i.e. $T(x^*) = x^*$. Also x^* can be computed by fixed point iteration: Start with arbitrary x_0: set $x_n = T(x_{n-1})$. Then $x_n \to x^*$.

For any bounded function ϕ on $\Pi(X)$ denote the dynamic programming operator $L : \phi \to \mathbb{R}$ as

$$L\phi(\pi) = \min_u \left\{ c_u' \pi + \rho \sum_{y \in Y} \phi\left(T\left(\pi, y, u\right)\right) \sigma\left(\pi, y, u\right) \right\}.$$

$\mathcal{B}(X)$: set of bounded real-valued functions on $\Pi(X)$.
Then for any ϕ and $\psi \in \mathcal{B}(X)$, define the sup-norm metric

$$\|\phi - \psi\|_\infty = \sup_{\pi \in \Pi(X)} |\phi(\pi) - \psi(\pi)|.$$

Then $\mathcal{B}(X)$ is a Banach space (complete metric space).

Theorem 7. *For $\rho \in [0,1)$, L is a contraction mapping:*

$$\|L\phi - L\psi\|_\infty \leq \rho\|\phi - \psi\|_\infty, \qquad \phi, \psi \in \mathcal{B}(X).$$

So Banach's fixed point theorem implies that there exists a unique solution V satisfying Bellman's equation $V = LV$.

Proof. Suppose ϕ and ψ are such that for a fixed π, $L\psi(\pi) \geq L\phi(\pi)$. Let u^* denote the minimizer for $L\phi(\pi)$,

$$u^* = \operatorname*{argmin}_u \left\{ c_u'\pi + \rho \sum_{y \in Y} \phi\left(T\left(\pi, y, u\right)\right) \sigma\left(\pi, y, u\right) \right\}.$$

Then clearly,
$L\psi(\pi) \leq c_{u^*}'\pi + \rho \sum_{y \in Y} \psi\left(T\left(\pi, y, u^*\right)\right) \sigma\left(\pi, y, u^*\right)$ since u^* is not necessarily the minimizer for $L\psi(\pi)$. So

$$\begin{aligned}
0 &\leq L\psi(\pi) - L\phi(\pi) \\
&\leq \rho \sum_{y \in Y} \left[\psi\left(T\left(\pi, y, u^*\right)\right) - \phi\left(T\left(\pi, y, u^*\right)\right)\right] \sigma\left(\pi, y, u^*\right) \\
&\leq \rho\|\psi - \phi\|_\infty \sum_{y \in Y} \sigma\left(\pi, y, u^*\right) \\
&= \rho\|\psi - \phi\|_\infty
\end{aligned}$$

A similar argument holds for the set of beliefs π for which $L\psi(\pi) \leq L\phi(\pi)$. Therefore, for all $\pi \in \Pi(X)$,

$$|L\psi(\pi) - L\phi(\pi)| \leq \rho\|\psi - \phi\|_\infty.$$

Take sup over $\pi \in \Pi(X)$ proves L is contraction

Value Iteration Algorithm for POMDP

Initialize $V_0(\pi) = 0$. For iterations $n = 1, 2, \ldots, N$,

$$V_n(\pi) = \min_{u \in \mathcal{U}} Q_n(\pi, u), \quad \mu_n^*(\pi) = \operatorname*{argmin}_{u \in \mathcal{U}} Q_n(\pi, u),$$

$$Q_n(\pi, u) = c_u' \pi + \rho \sum_{y \in Y} V_{n-1}\left(T\left(\pi, y, u\right)\right) \sigma\left(\pi, y, u\right).$$

Stationary policy μ_N^* is used at each time instant k.

How to choose iterations N in value iteration algorithm?

Theorem 8. *Consider the value iteration algorithm with discount factor ρ and N iterations. Then:*

1. $\sup_\pi |V_N(\pi) - V_{N-1}(\pi)| \leq \epsilon$ *implies that* $\sup_\pi |V_N(\pi) - V(\pi)| \leq \frac{\epsilon\rho}{1-\rho}$.
2. $|V_N(\pi) - V(\pi)| \leq \frac{\rho^{N+1}}{1-\rho} \max_{x,u} |c(x, u)|$.

$$\begin{aligned}
\|V - V_N\| &= \|LV - LV_N + LV_N - V_N\| \\
&\leq \|LV - LV_N\| + \|LV_N - V_N\| \\
&= \|LV - LV_N\| + \|LV_N - LV_{N-1}\| \\
&\leq \rho\|V - V_N\| + \rho\|V_N - V_{N-1}\| \\
&\implies \|V - V_N\| \leq \frac{\rho\|V_N - V_{N-1}\|}{1-\rho}.
\end{aligned}$$

Successive Approximation: Examples

Successive approx is used in linear algebra, establishing existence of soln of Lipschitz ODEs (Picard iteration).

Example 1: Solve linear system $Ax = b$. Assume $A_{ii} = 1$.

$$Ax = b \quad \equiv \quad x = (I - A)x + b$$

$$\text{Successive approx: } x_{k+1} = (I - A)x_k + b$$

When does it converge?
Define norm and matrix induced norm

$$\|x\|_\infty = \max_i |x_i|, \quad \|B\| = \max_i \sum_j |B_{ij}|$$

Theorem 9. *If $A_{ii} > \sum_{j \neq i} |A_{ij}|$ then S.A. converges.*

Proof

$$\|T(x) - T(y)\|_\infty = \|(A - I)(x - y)\|_\infty \leq \|A - I\|_\infty \|x - y\|_\infty$$

$$\|A - I\|_\infty = \max_i \sum_{j \neq i} |A_{ij}| = \alpha < 1.$$

Example 2: Picard iteration: existence of soln to initial valued ODE.

$$\frac{dx}{dt} = f(x,t), \qquad \text{given } x(0) = x_0$$

Suppose f is Lipschitz on $[t_0, t_1]$ if $\exists$ bounded α s.t.

$$|f(x_1,t) - f(x_2,t)| \leq \alpha\, |x_1 - x_2|$$

Then unique soln exists and can be found by S.A.

Proof: Note that ODE equiv to integral eqn

$$x(t) = x_0 + \int_{t_0}^{t} f\big(x(\tau), \tau\big) d\tau$$

Consider space $X = C[t_0, t_1]$. Define

$$T(x) = \int_{t_0}^{t} f(x(\tau), \tau)\, d\tau$$

$$\begin{aligned}\|T(x_1) - T(x_2)\| &= \| \int_{t_0}^{t} \big(f(x_1,\tau) - f(x_2,\tau)\big) dt \| \\ &\leq \int_{t_0}^{t} \alpha\, \|x_1 - x_2\| d\tau \leq \alpha(t_1 - t_0) \|x_1 - x_2\|\end{aligned}$$

Then T is contraction if $\alpha < 1/(t_1 - t_0)$.

Optimal Search for Moving Target

1. State: Target moves among cells $\{1, 2, \ldots, X\}$ as Markov chain x_k with transition matrix P. Add fictitious state $T\}$ when search is terminated before time N.

2. Action: At time k, searcher chooses cell u_k to search.

3. Observation: $y_k \in \mathcal{Y} = \{F, \bar{F}, b\}$.

$$y_k = \begin{cases} F & \text{target is found,} \\ \bar{F} & \text{target is not found,} \\ b & \text{search blocked due to insufficient resources.} \end{cases}$$

Blocking and overlook probabilities:

$$q(u) = \mathbb{P}(\text{insufficient resources to perform action } u \text{ at epoch } k),$$
$$\beta(u) = \mathbb{P}(\text{target not found}|\text{target is in the cell } u).$$

Then, observation y_k received is characterized as follows.
For all $u \in \mathcal{U}$ and $j = 1, \ldots, X$,

$$\mathbb{P}(y_k = F|x_k = j, u_k = u) = \begin{cases} (1 - q(u))(1 - \beta(u)) & \text{if cell } j \text{ searched,} \\ 0 & \text{otherwise,} \end{cases}$$

$$\mathbb{P}(y_k = \bar{F}|x_k = j, u_k = u) = \begin{cases} 1 - q(u) & \text{if cell } j \text{ not serached,} \\ \beta(u)(1 - q(u)) & \text{otherwise,} \end{cases}$$

$$\mathbb{P}(y_k = b|x_k = j, u_k = u) = q(u).$$

$$\mathbb{P}(y_k = F|x_k = T, u_k = u) = 1.$$

Obs dependent transition matrix:
$\mathbb{P}(x_{k+1} = j | x_k = i, y_k = y) = P^y_{ij}$.

$$P^F = \begin{bmatrix} 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}, \quad P^{\bar{F}} = P^b = \begin{bmatrix} P & \mathbf{0} \\ \mathbf{0}' & 1 \end{bmatrix}.$$

4. Cost: $c(x_k, u_k)$ Three examples are of interest.
1. Maximize Probability of Detection.

$$c(x_k = j, u_k = u) = -\mathbb{P}(y_k = F | x_k = j, u_k = u) \quad \text{for } j = 1, \ldots, X,$$

$c(x_k = T, u_k = u) = 0.$

2. Minimize Search Delay An instantaneous cost of 1 unit is accrued for every action taken until the target is found, i.e., until the target reaches the terminal state T:

$$\begin{aligned} c(x_k = j, u_k = u) &= 1 \quad \text{for } j = 1, \ldots, X, \\ c(x_k = T, u_k = u) &= 0. \end{aligned}$$

3. Minimize Search Cost. The instantaneous cost depends only on the action taken.

$$c(x_k = j, u_k = u) = c(u) \quad \text{for } j = 1, \ldots, X,$$

$$c(x_k = T, u_k = u) \quad = \quad 0.$$

5. Performance criterion:

$\mathcal{I}_0 = \{\pi_0\}, \quad \mathcal{I}_k = \{\pi_0, u_0, y_0, \dots, u_{k-1}, y_{k-1}\} \quad$ for $k = 1, \dots, N$.

The performance criterion considered is

$$J_{\boldsymbol{\mu}}(\pi_0) = \mathbb{E}_{\boldsymbol{\mu}} \left\{ \sum_{k=0}^{N-1} c(x_k, \mu_k(\mathcal{I}_k)) \Big| \pi_0 \right\}.$$

$$\boldsymbol{\mu}^* = \operatorname*{argmin}_{\boldsymbol{\mu} \in \mathcal{U}} J_{\boldsymbol{\mu}}(\pi_0), \qquad \forall\, \pi_0 \in \mathit{\Pi}(X).$$

In terms of the belief state $\pi_k(i) = \mathbb{P}(x_k = i | \mathcal{I}_k)$ as

$$\begin{aligned} J_{\boldsymbol{\mu}}(\pi_0) &= \mathbb{E}_{\boldsymbol{\mu}} \left\{ \sum_{k=0}^{N} c(x_k, \mu_k(\mathcal{I}_k)) \Big| \pi_0 \right\}, \\ &= \mathbb{E}_{\boldsymbol{\mu}} \left\{ \sum_{k=0}^{N} \mathbb{E}\{c(x_k, \mu_k(\mathcal{I}_k)) | \mathcal{I}_k\} \Big| \pi_0 \right\} = \sum_{k=0}^{N} \mathbb{E}\{c'_{u_k} \pi_k\} \end{aligned}$$

Belief state is updated by the HMM predictor[a]:

$$\pi_{k+1} = T(\pi_k, y_k, u_k) = \frac{P^{y_k\prime} \tilde{B}_{y_k}(u_k)\pi_k}{\sigma(\pi_k, y_k, u_k)}, \quad \sigma(\pi, y, u) = \mathbf{1}'\tilde{B}_y(u)\pi$$

$$\tilde{B}_y(u) = \text{diag}(\mathbb{P}(y_k = y|x_k = 1, u_k = u), \ldots,$$
$$\mathbb{P}\big(y_k = y|x_k = X, u_k = u), \mathbb{P}(y_k = y|x_k = T, u_k = u)\big).$$

Search problem differs from standard POMDP in 2 ways:

Timing of the events: In POMDP, $\mathbb{P}(y_{k+1}|x_{k+1}, u_k)$ In search problem, $\mathbb{P}(y_k|x_k, u_k)$.

Transition to the new state: In POMDP $\mathbb{P}(x_{k+1}|x_k, u_k)$. In search problem, $\mathbb{P}(x_{k+1}|x_k, y_k)$

Search problem can be reformulated as a POMDP with augmented state $s_{k+1} = (y_k, x_{k+1})$.

Ex 1. Dynamic (Active) hypothesis testing: so far no false alarms. When target not in cell, then observation recorded is "not found". In active hypothesis testing,

- If target in cell u then observation $y \sim \phi(y)$
- If the target **not** in cell u, then observation $y \sim \bar{\phi}(y)$. (In classical search $\bar{\phi}(y)$ is dirac measure on $\bar{F}$.)

[a] Note difference between the information pattern of a standard POMDP, namely, $\mathcal{I}_k = (\pi_0, u_0, y_1, \ldots, u_{k-1}, y_k)$ and the information pattern $\mathcal{I}_k$ for the search problem. In search, $\mathcal{I}_k$ has observations until time $k-1$, requiring the HMM predictor.

Ex 2. Optimal Observer Trajectory Moving observer (sensor) measures target's position in noise. Noise depends on the relative distance between the target and the observer. How should the observer move amongst the X-cells in order to locate where the target is? One possible metric: observer moves to maximize the stochastic observability of the target.

Multiple searchers? Pursuit-Evasion game?

Search for static target

Assume: $P = I$, $\mathcal{X} = \mathcal{U} = \{1, 2, \dots, X\}$,
cost is action dependent only $c(u)$, $y \in \{F, \bar{F}\}$.
Overlook prob: $\beta(u) = \mathbb{P}(\text{target not found} \mid \text{target is in } u)$

Belief update given $y_k = \bar{F}$ is

$$\pi_{k+1} = T(\pi_k, y_k = \bar{F}, u_k) = \frac{B_{\bar{F}}(u_k)\pi_k}{\sigma(\pi_k, \bar{F}, u_k)}$$

$$= \begin{cases} \frac{\pi_k(i)}{\sigma(\pi_k, \bar{F}, u_k)}, & i \neq u_k \\ \frac{\pi_k(i)\beta(u_k)}{\sigma(\pi_k, \bar{F}, u_k)}, & i = u_k \end{cases}$$

$$\sigma(\pi, \bar{F}, u) = \mathbf{1}' B_{\bar{F}}(u)\pi = 1 - \pi(u)(1 - \beta(u))$$

If $y_k = F$, then target found and problem terminates.

Given π_0, optimal search strategy is as follows:

Theorem 10. *Given π_k, optimal to search location*

$$u_k = \mu^*(\pi_k) = \operatorname*{argmax}_{i \in \mathcal{U}} \frac{\pi_k(i)(1 - \beta(i))}{c(i)}$$

where the belief π_k is updated according to Bayes rule.

Proof uses the "interchange argument".
$J_{u_1, u_2, \boldsymbol{\mu}}$: first search cell i, then cell j, then search with $\boldsymbol{\mu}$.
$J_{u_2, u_1, \boldsymbol{\mu}}$: first search cell j, then cell i, then search with $\boldsymbol{\mu}$.

Lemma 11.

$$J_{u_1,u_2,\boldsymbol{\mu}} \leq J_{u_2,u_1,\boldsymbol{\mu}} \iff \frac{\pi(u_1)(1-\beta(u_1))}{c(u_1)} \geq \frac{\pi(u_2)(1-\beta(u_2))}{c(u_2)}$$

Proof. $J_{u_1,u_2,\boldsymbol{\mu}} = c'_{u_1}\pi + \sum_{y_1} c'_{u_2} T(\pi, y_1, u_1)\sigma(\pi, y_1, u_1)$

$$+\sum_{y_1}\sum_{y_2} J_{\boldsymbol{\mu}}(T(T(\pi, y_1, u_1), y_2, u_2))\,\sigma(\pi, y_1, u_1)\,\sigma(T(\pi, y_1, u_1), y_2, u_2)$$

$$= c(u_1) + c(u_2)\,\sigma(\pi, \bar{F}, u_1) + K$$

since $c_{u_1} = c(u_1)\mathbf{1}$, $c_{u_2} = c(u_2)\mathbf{1}$ if $y_1 = \bar{F}$ and zero if $y_1 = F$. Note for $y_1 = y_2 = \bar{F}$,

$$T(T(\pi, y_1, u_1), y_2, u_2)) = T(T(\pi, y_2, u_2), y_1, u_1))$$
$$\sigma(\pi, y_1, u_1)\,\sigma(T(\pi, y_1, u_1), y_2, u_2)$$
$$= \sigma(\pi, y_2, u_2)\,\sigma(T(\pi, y_2, u_2), y_1, u_1).$$

So $J_{u_1,u_2,\boldsymbol{\mu}} \leq J_{u_2,u_1,\boldsymbol{\mu}}$

$$\iff c(u_1) + c(u_2)\sigma(\pi, \bar{F}, u_1) \leq c(u_2) + c(u_1)\sigma(\pi, \bar{F}, u_2)$$

Substituting $\sigma(\pi, \bar{F}, u) = 1 - \pi(u)(1-\beta(u))$ implies that

$$J_{u_1,u_2,\boldsymbol{\mu}} \leq J_{u_2,u_1,\boldsymbol{\mu}} \iff \frac{\pi(u_1)(1-\beta(u_1))}{c(u_1)} \geq \frac{\pi(u_2)(1-\beta(u_2))}{c(u_2)}$$

$\square$

With Lemma 11, suppose

$$\frac{\pi(1)(1-\beta(1))}{c(1)} = \max_{u\in\mathcal{U}} \frac{\pi(u)(1-\beta(u))}{c(u)}.$$

From Lemma 11 it follows that a policy which does not immediately search cell 1 has a larger cumulative cost than the policy that does search cell 1.

POMDP Multi-armed Bandits

Consider L independent projects $l = 1,\dots,L$. Each project l has state space $\mathcal{X} = \{1,2,\dots,X\}$. Let $x_k^{(l)}$: state of project l. At each time instant k only one projects can be worked on:

- If project l is worked on at time k:
  1. reward $\rho^k\, r(x_k^{(l)})$ where $0 \leq \rho < 1$.
  2. $x_k^{(l)}$ evolves with transition probability P.
  3. State observed via $y_{k+1}^{(l)} \in \mathcal{Y} = \{1,2,\dots,Y\}$ with observation probability $B_{xy} = \mathbb{P}(y^{(l)} = y | x^{(l)} = x)$.
- The $(L-1)$ idle projects are unaffected: $x_{k+1}^{(l)} = x_k^{(l)}$, if project l is idle at time k. No obs for idle projects.

Denote $r(x^{(l)}, l)$ as $r(x^{(l)})$. Projects initialized: $x_0^{(l)} \sim \pi_0^{(l)}$.

$u_k \in \{1,\dots,L\}$: project worked on at time k. So $x_{k+1}^{(u_k)}$ is the state of the active project at time $k+1$.

$$\mathcal{I}_0 = \pi_0, \quad \mathcal{I}_k = \{\pi_0, y_1^{(u_0)}, \dots, y_k^{(u_{k-1})}, u_0, \dots, u_{k-1}\}.$$

Then the project at time k is chosen as $u_k = \mu(\mathcal{I}_k)$,

$$J_\mu(\pi) = \mathbb{E}_\mu\{\sum_{k=0}^{\infty} \rho^k r\left(x_k^{(u_k)}\right) |\pi_0 = \pi\}, \quad u_k = \mu(\mathcal{I}_k). \quad (5)$$

Aim: determine $\mu^*(\pi) = \operatorname{argmax}_\mu J_\mu(\pi)$.

At first sight intractable since equivalent state space dimension is X^L. The multi-armed bandit structure yields a remarkable simplification - can be solved by considering L individual POMDPs each of dimension X.

Belief State Formulation: $\pi_k^{(l)}(i) = \mathbb{P}(x_k^{(l)} = i|\mathcal{I}_k)$.
Then scheduling problem: Consider P parallel HMM state estimation filters.
If project l is active, $y_{k+1}^{(l)}$ is obtained and $\pi_{k+1}^{(l)}$ updated by HMM filter

$$\pi_{k+1}^{(l)} = T(\pi_k^{(l)}, y_{k+1}^{(l)}) \qquad \text{if project } l \text{ is worked on at time } k$$

where $T(\pi^{(l)}, y^{(l)}) = \dfrac{B_{y^{(l)}} P' \pi^{(l)}}{\sigma(\pi^{(l)}, y^{(l)})}, \quad \sigma(x^{(l)}, y^{(l)}) = \mathbf{1}' B_{y^{(l)}} P' \pi^{(l)}$

$$B_{y^{(l)}} = \operatorname{diag}\big(\mathbb{P}(y^{(l)}|x^{(l)} = 1), \dots, \mathbb{P}(y^{(l)}|x^{(l)} = X)\big).$$

The beliefs of the other $L-1$ projects remain unaffected,

$$\pi_{k+1}^{(q)} = \pi_k^{(q)} \qquad \text{if project } q \text{ is not worked on.}$$

Let $r = [r(x_k^{(l)} = 1), \ldots, r(x_k^{(l)} = X)]'$. Then

$$J_\mu(\pi) = \mathbb{E}\{\sum_{k=0}^{\infty} \rho^k \, r' \pi_k^{(u_k)} \mid (\pi_0^{(1)}, \ldots, \pi_0^{(L)}) = \pi\}, \; u_k = \mu(\pi_k^{(1)}, \ldots, \pi_k^{(L)})$$

Compute $\mu^*(\pi) = \arg\max_\mu J_\mu(\pi)$.

Gittins Index Rule. $\bar{M} \stackrel{\text{defn}}{=} \max_i r(i)/(1-\rho)$, $M \in [0, \bar{M}]$.

Optimal policy has an *indexable rule*:

Theorem 12 (Gittins index)**.** *For each project l there is a function $\gamma(\pi_k^{(l)})$* Gittins index, *s.t. optimal policy is:*

$$\mu^*(\pi_k^{(1)}, \pi_k^{(2)}, \ldots, \pi_k^{(L)}) = \underset{l \in \{1,\ldots,L\}}{\operatorname{argmax}} \left\{ \gamma(\pi_k^{(l)}) \right\} \qquad (6)$$

The Gittins index of project l with belief $\pi^{(l)}$ is

$$\gamma(\pi^{(l)}) = \min\{M : V(\pi^{(l)}, M) = M\}$$

where $V(\pi^{(l)}, M)$ satisfies Bellman's equation

$$V(\pi^{(l)}, M) = \max\left\{ r'\pi^{(l)} + \rho \sum_{y=1}^{Y} V\left(T(\pi^{(l)}, y), M\right) \sigma(\pi^{(l)}, y), \; M \right\}$$

Part 6. Structural Results for POMDPs

$$\mu^*(\pi) = \operatorname*{argmin}_{u \in \mathcal{U}} Q(\pi, u), \quad V(\pi) = \min_{u \in \mathcal{U}} Q(\pi, u),$$

$$Q(\pi, u) = \; c_u' \pi + \rho \sum_{y \in Y} V\left(T\left(\pi, y, u\right)\right) \sigma\left(\pi, y, u\right).$$

We want to prove $\mu^*(\pi) \uparrow \pi$.

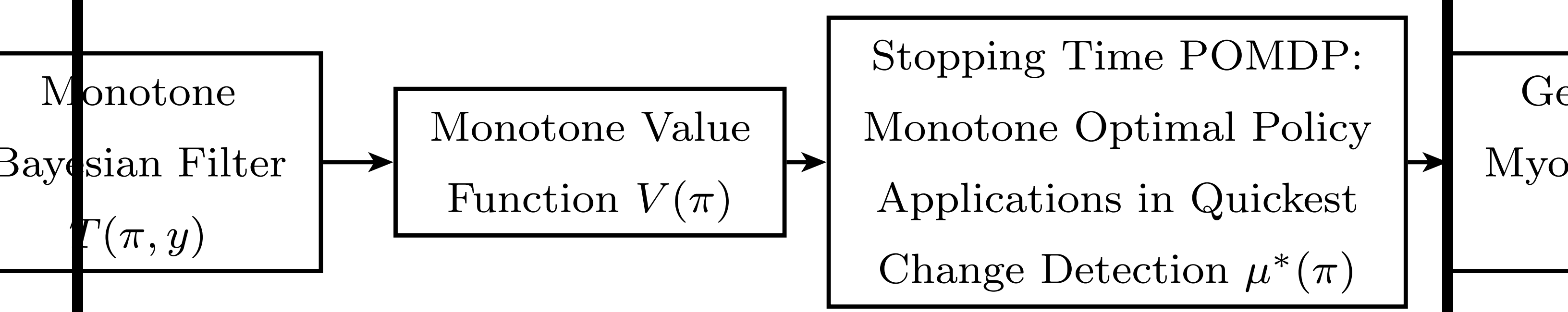


Figure 5: Organization of POMDP structural results

1. How can beliefs π be ordered in unit simplex $\Pi(X)$?
2. Under what conditions does the HMM filter $T(\pi, y, u)$ increase with belief π, observation y and action u?

1. Stopping Time POMDP - convexity of stopping region

Stopping time POMDP: $\mathcal{U} = \{1 \text{ (stop)}, 2 \text{ (continue)}\}$.

- $u = 2$: $x_k \in \{1, 2, \ldots, X\}$ has transition matrix P; $B_{xy} = \mathbb{P}(y_k = y | x_k = x)$; cost $c(x, u = 2)$. Thus for $u = 2$, $\pi_k = T(\pi_{k-1}, y_k)$.
- $u = 1$: terminal cost of $c(x, u = 1)$ and terminates.

$u_k = \mu(\pi_k) \in \{1 \text{ (stop)}, 2 \text{ (continue) }\}$, $\tau = \{\inf k : u_k = 1\}$.

$$J_\mu(\pi_0) = \mathbb{E}_\mu \left\{ \sum_{k=0}^{\tau-1} \rho^k c(x_k, 2) + \rho^\tau c(x_\tau, 1) \right\}$$

$$= \mathbb{E}_\mu \left\{ \sum_{k=0}^{\tau-1} \rho^k c_2' \pi_k + \rho^\tau c_1' \pi_\tau \right\}, \; c_u = [c(1, u), \ldots, c(X, u)]'$$

Aim: Optimal policy $\mu^* : \Pi(X) \to \mathcal{U} = \arg\inf_\mu J_\mu(\pi_0)$.
μ^* is the solution of Bellman's equation

$$\mu^*(\pi) = \operatorname*{argmin}_{u \in \mathcal{U}} Q(\pi, u), \quad V(\pi) = \min_{u \in \mathcal{U}} Q(\pi, u),$$

$$Q(\pi, 1) = c_1' \pi, \quad Q(\pi, 2) = \; c_2' \pi + \rho \sum_{y \in Y} V\left(T\left(\pi, y\right)\right) \sigma\left(\pi, y\right).$$

Stopping time POMDP = infinite horizon POMDP. Add fictitious stopping state e_{X+1} with $c(e_{X+1}, u) = 0, \forall u \in \mathcal{U}$. When $u_k = 1$, $\pi_{k+1} = e_{X+1}$ and remains indefinitely.
$J_\mu(\pi) = \mathbb{E}_\mu\{\sum_{k=0}^{\tau-1} \rho^k c_2' \pi_k + \rho^\tau c_1' \pi + \sum_{k=\tau+1}^{\infty} \rho^k c(e_{X+1}, u_k\}$.

Convexity of Stopping Region. Define stopping set

$$\mathcal{R}_1 = \{\pi : \mu^*(\pi) = 1 \text{ (stop) } \}$$

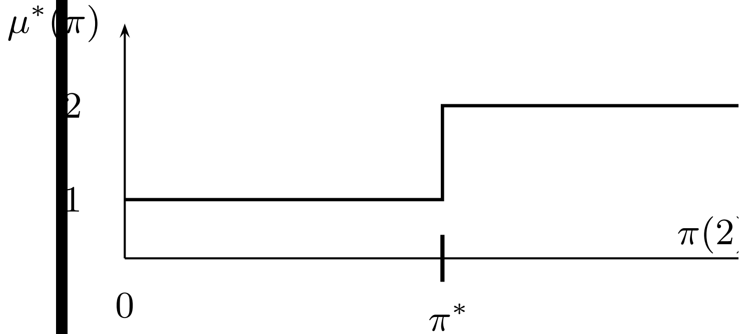
Theorem 13 (Lovejoy 1987)**.** *Consider the stopping-time POMDP with linear cost. Then $\mathcal{R}_1$ is convex.*

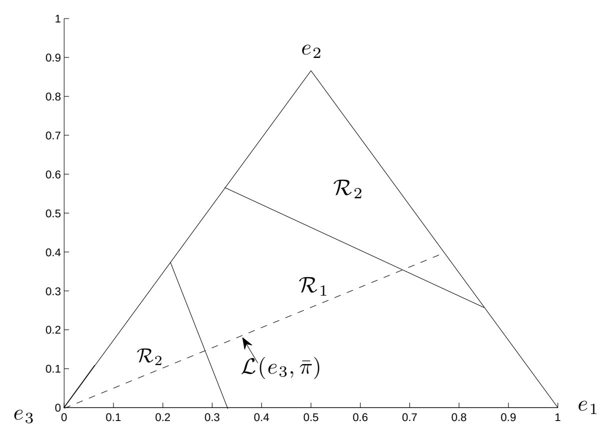
Proof Pick any two belief states $\pi_1, \pi_2 \in \mathcal{R}_1$. Need to show for any $\lambda \in [0,1]$, $\lambda\pi_1 + (1-\lambda)\pi_2 \in \mathcal{R}_1$.
Since $V(\pi)$ is concave,

$$\begin{aligned}
&V(\lambda\pi_1 + (1-\lambda)\pi_2) \geq \lambda V(\pi_1) + (1-\lambda)V(\pi_2) \\
&= \lambda Q(\pi_1, 1) + (1-\lambda)Q(\pi_2, 1) \text{ (since } \pi_1, \pi_2 \in \mathcal{R}_1) \\
&= Q(\lambda\pi_1 + (1-\lambda)\pi_2, 1) \text{ (since } Q(\pi, 1) \text{ is linear in } \pi) \\
&\geq V(\lambda\pi_1 + (1-\lambda)\pi_2) \text{ (since } V(\pi) \text{ is value function)}
\end{aligned}$$

So inequalities above are equalities: $\lambda\pi_1 + (1-\lambda)\pi_2 \in \mathcal{R}_1$.

Theorem says nothing about the "continue" region $\mathcal{R}_2$.





Example 1. Quickest Change Detection

Process x jump changes at geometric distributed time τ^0. Observations y_k, $\{k \leq \tau^0\} \sim B_{1y}$ and $\{y_k, k > \tau^0\} \sim B_{2y}$.

Kolmogorov–Shiryayev criterion Detect change time τ^0 to min false alarm & delay

$$J_\mu(\pi) = d\,\mathbb{E}_\mu\{(\tau - \tau^0)^+\} + \mathbb{P}_\mu(\tau < \tau^0), \quad \pi_0 = \pi.$$

POMDP model: $\tau^0 \sim$ two state Markov chain $\mathcal{X} = \{1, 2\}$,

$$P = \begin{bmatrix} 1 & 0 \\ 1 - P_{22} & P_{22} \end{bmatrix}, \ \pi_0 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \tau^0 = \inf\{k : x_k = 1\}.$$

τ^0 is geometrically distributed with mean $1/(1 - P_{22})$.
Cost vectors $c_1 = [0,\ 1]'$, $c_2 = [d,\ 0]'$, $\rho = 1$.
$\mathcal{U} = \{1 \text{ (stop)}, 2 \text{ (continue)}\}$, obs prob B_{xy},

Corollary 14. *The optimal policy μ^* for quickest detection has a* threshold *structure: $\exists \pi^* \in [0, 1]$ such that*

$$u_k = \mu^*(\pi_k) = \begin{cases} 2 \ \textit{(continue)} & \textit{if } \pi_k(2) \in [\pi^*, 1] \\ 1 \ \textit{(stop)} & \textit{if } \pi_k(2) \in [0, \pi^*). \end{cases}$$

Proof. Since $X = 2$, $\Pi(X) = [0, 1]$, and $\pi(2) \in [0, 1]$. Theorem 13 implies $\mathcal{R}_1 = [a^*, \pi^*)$ for $0 \leq a < \pi^* \leq 1$.

Bellman's equation applied at $\pi = e_1$ implies

$$\mu^*(e_1) = \operatorname*{argmin}_u \{\underbrace{c(1, u = 1)}_{0},\quad d(1 - \pi(2)) + V(e_1)\} = 1.$$

So e_1 or equivalently $\pi(2) = 0 \in \mathcal{R}_1$. So $\mathcal{R}_1 = [0, \pi^*)$. $\square$

Example 2. Instruction Problem

Student is instructed and examined repeatedly until stopping time τ when instruction is stopped.
$\mathcal{U} = \{1(\text{stop}), 2(\text{instruct})\}$, $\mathcal{X} = \{1(\text{learnt}), 2(\text{not learnt})\}$
$\mathcal{Y} = \{1 \text{ (correct answer)}, 2 \text{ (wrong answer)}\}$.
x_k: status of the student at time k.
y_k: outcome of an exam at each time k.
If $u = 2$ (instruct) is chosen, then instruction cost $= 1$,

$$P = \begin{bmatrix} 1 & 0 \\ 1-p & p \end{bmatrix},\ B = \begin{bmatrix} 1 & 0 \\ 1-q & q \end{bmatrix},\ c_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

$1 - p$: probability student learns when instructed;
$1 - q$: prob student guesses correct when not learnt.
If instruction terminated then stopping cost vector $c_1 = [0,\ f]'$ where f: cost instruction was stopped but student has not yet learnt.

By Theorem 13 the stopping region is an interval. Since state 1 is absorbing, optimal policy is threshold.

2. Monotone Likelihood Ratio (MLR) Stochastic Order

Recall belief space is $X-1$ dimensional unit simplex

$$\Pi(X) = \Big\{\pi \in \mathbb{R}^X : \mathbf{1}'\pi = 1, \quad 0 \leq \pi(i) \leq 1,\ i \in \mathcal{X} = \{1,2,\ldots,X\}$$

Definition 15 (MLR Dominance). $\pi_1 \geq_r \pi_2$ *if*

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

So $\pi_1 \geq_r \pi_2$ if likelihood ratio $\pi_1(i)/\pi_2(i) \uparrow i$

Definition 16. $\phi : \Pi(X) \to \mathbb{R}$ *is MLR increasing if* $\pi_1 \geq_r \pi_2$ *implies* $\phi(\pi_1) \geq \phi(\pi_2)$.

Definition 17 (First order stochastic dominance). $\pi_1 \geq_s \pi_2$ *if* $\sum_{i=j}^{X} \pi_1(i) \geq \sum_{i=j}^{X} \pi_2(i)$ *for* $j = 1,\ldots,X$.

Theorem 18. π_1 *and* π_2, $\pi_1 \geq_r \pi_2$ *implies* $\pi_1 \geq_s \pi_2$.

Proof. $\pi_1 \geq_r \pi_2$ implies $\pi_1(x)/\pi_2(x)$ is increasing in x. Denote the corresponding cdfs as F_1, F_2. Define $t = \{\sup x : \pi_1(x) \leq \pi_2(x)\}$. Then $\pi_1 \geq_r \pi_2$ implies that for $x \leq t$, $\pi_1(x) \leq \pi_2(x)$ and for $x \geq t$, $\pi_1(x) \geq \pi_2(x)$. So for $x \leq t$, $F_1(x) \leq F_2(x)$. Also for $x > t$, $\pi_1(x) \geq \pi_2(x)$ implies $1 - \int_x^\infty \pi_1(x)dx \leq 1 - \int_x^\infty \pi_2(x)dx$ or equivalently, $F_1(x) \leq F_2(x)$. Therefore $\pi_1 \geq_s \pi_2$. $\square$

Remarks on MLR Dominance.

(i) For $X = 2$, MLR *complete* order & equiv to first order.

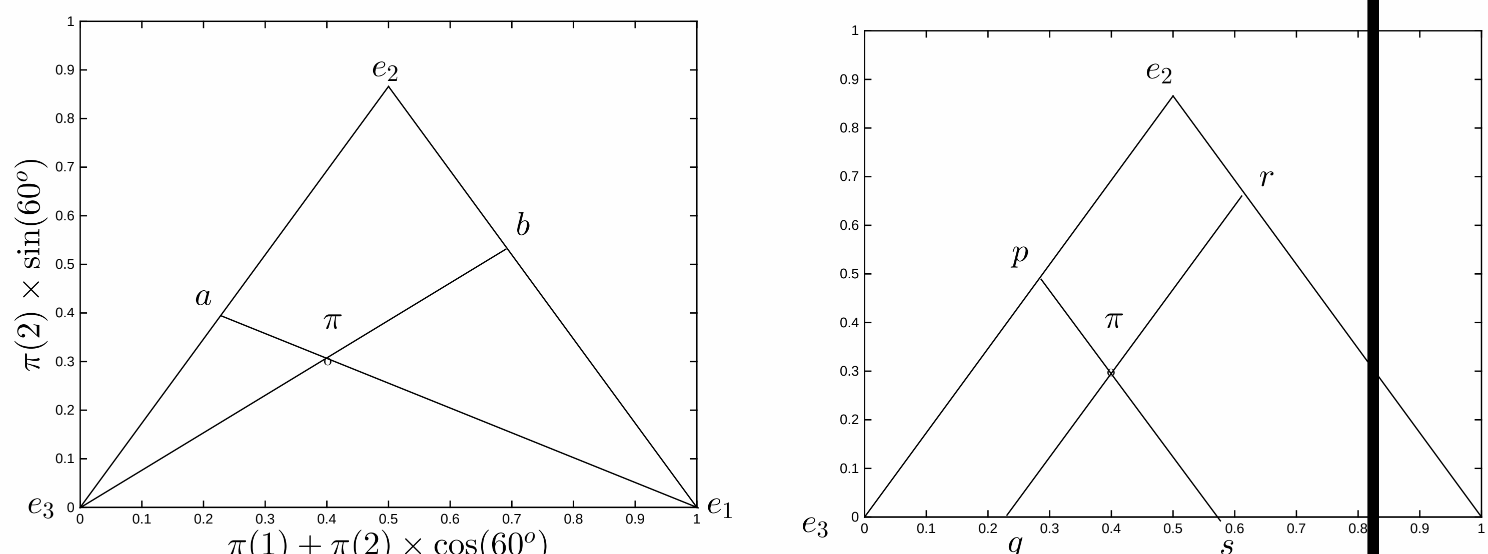
$$X = 2, \quad \pi_1 \geq_r \pi_2 \iff \pi_1 \geq_s \pi_2 \iff \pi_1(2) \geq \pi_2(2).$$

(ii) For state space dimension $X \geq 3$, MLR and first order dominance are *partial orders* on poset $[\mathit{\Pi}(X), \geq_r]$.

(iii) Examples: $[0.2, 0.3, 0.5]' \geq_r [0.4, 0.5, 0.1]'$

$[0.3, 0.2, 0.5]'$ & $[0.4.0.5.0.1]'$ not MLR comparable.

(iv) Geometric Interpretation for $X = 3$.



(a) MLR dominance (b) First order dominance

MLR is preserved by Bayes rule

Theorem 19. *Given observation likelihoods* $B_y = \text{diag}(B_{1y}, \ldots, B_{Xy})$, $B_{xy} = p(y|x)$, *then*

$$\pi_1 \geq_r \pi_2 \iff \frac{B_y \pi_1}{\mathbf{1}' B_y \pi_1} \geq_r \frac{B_y \pi_2}{\mathbf{1}' B_y \pi_2}$$

providing $\mathbf{1}' B_y \pi_1$ *and* $\mathbf{1}' B_y \pi_2$ *are non-zero.*[a]

Proof. RHS is

$$\cancel{B_{iy} B_{i+1,y}} \pi_1(i) \pi_2(i+1) \leq \cancel{B_{iy} B_{i+1,y}} \pi_1(i+1) \pi_2(i)$$

which is equivalent to $\pi_1 \geq_r \pi_2$. □

First order stochastic dominance is not preserved under conditional expectations and so is not useful for POMDPs.

[a] A notationally elegant way of saying this is: Given two random variables X and Y, then $X \leq_r Y$ iff $X|X \in A \leq_r Y|Y \in A$ for all events A providing $P(X \in A) > 0$ and $P(Y \in A) > 0$. Requiring $\mathbf{1}' B_y \pi > 0$ avoids pathological cases such as $\pi = [1, 0]'$ and $B_y = \text{diag}(0, 1)$, i.e., prior says state 1 with certainty, while observation says state 2 with certainty.

Examples of MLR Dominance

1. *First order stochastic dominance is not closed under Bayes rule*: $\pi_1 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})'$, $\pi_2 = (0, \frac{2}{3}, \frac{1}{3})'$. $\pi_1 \leq_s \pi_2$. Suppose $P = I$ and $\mathbb{P}(y|x=1) = 0$, $\mathbb{P}(y|x=2) = 0.5$, $\mathbb{P}(y|x=3) = 0.5$. Then the filtered updates are $T(\pi_1, y, u) = (0, \frac{1}{2}, \frac{1}{2})'$ and $T(\pi_2, y, u) = (0, \frac{2}{3}, \frac{1}{3})'$. Thus $T(\pi_1, y, u) \geq_s T(\pi_2, y.u)$

2. Examples of pmfs that satisfy MLR dominance:

$$\text{Poisson:} \quad \frac{\lambda_1^k}{k!}\exp(\lambda_1) \leq_r \frac{\lambda_2^k}{k!}\exp(\lambda_2), \quad \lambda_1 \leq \lambda_2$$

$$\text{Binomial:} \quad \binom{n_1}{k} p_1^k (1-p_1)^{n_1-k} \leq_r \binom{n_2}{k} p_1^k (1-p_2)^{n_2-k},$$

$$n_1 \leq n_2, p_1 \leq p_2$$

$$\text{Geometric:} \quad (1-p_1)\, p_1^k \leq_r (1-p_2)\, p_2^k, \quad p_1 \leq p_2.$$

3. MLR order for pdfs: $p \geq_r q$ if $p(x)/q(x) \uparrow x$. If the pdfs are differentiable, this is equivalent to saying $\frac{d}{dx}\frac{p(x)}{q(x)} \geq 0$. Examples include:

$$\text{Normal:} \quad \boldsymbol{N}(x; \mu_1, \sigma^2) \leq_r \boldsymbol{N}(x; \mu_2, \sigma^2), \quad \mu_1 \leq \mu_2$$

$$\text{Exponential:} \quad \lambda_1 \exp(\lambda_1(x - a_1)) \leq_r \lambda_2 \exp(\lambda_2(x - a_2)),$$

$$a_1 \leq a_2, \lambda_1 \geq \lambda_2.$$

Uniform pdfs $U[a,b] = I(x \in [a,b])/(b-a)$ are not MLR comparable with respect to a or b.

Total Positivity and Copositivity

Why? to show that $T(\pi, y, u) \uparrow \pi, y, u$.

Definition 20 (Totally Positive of Order 2 (TP2))**.** *Stochastic matrix M TP2 if all second order minors ≥ 0:*

$$\begin{vmatrix} M_{i_1 j_1} & M_{i_1 j_2} \\ M_{i_2 j_1} & M_{i_2 j_2} \end{vmatrix} \geq 0 \text{ for } i_2 \geq i_1, j_2 \geq j_1. \tag{7}$$

Equivalently, if $M_{i,:} \geq_r M_{j,:}$ for every $i > j$.

Definition 21 (Copositive Ordering of Transition Matrices)**.** $P(u) \preceq P(u+1)$ *if sequence of $X \times X$ matrices $\Gamma^{j,u}$, $j = 1 \ldots, X-1$ are copositive, i.e.,*

$$\pi' \Gamma^{j,u} \pi \geq 0, \quad \forall \pi \in \Pi(X), \quad \text{for each } j, \text{ where}$$
$$\gamma^{j,u}_{mn} = P_{m,j}(u) P_{n,j+1}(u+1) - P_{m,j+1}(u) P_{n,j}(u+1).$$

(F1) $B(u)$ with elements $B_{xy}(u)$ is TP2 for each $u \in \mathcal{U}$.

(F2) $P(u)$ is TP2 for each action $u \in \mathcal{U}$.

(F3) $P(u) \preceq P(u+1)$ (copositivity condition).

Theorem 22. *Consider HMM filter $T(\pi, y, u)$*

$$T(\pi, y, u) = \frac{B_y(u) P'(u) \pi}{\sigma(\pi, y, u)}, \ \sigma(\pi, y, u) = \mathbf{1}' B_y(u) P'(u) \pi, \text{ where}$$

$B_y(u) = \operatorname{diag}(B_{1y}(u), \ldots, B_{Xy}(u))$.

1. (a) *For* $\pi_1 \geq_r \pi_2$, *the HMM predictor satisfies* $P'(u)\pi_1 \geq_r P'(u)\pi_2$ *iff (F2) holds.*
   (b) *So* $\pi_1 \geq_r \pi_2 \implies T(\pi_1, y, u) \geq_r T(\pi_2, y, u)$ *for any* y *iff (F2) holds*
2. *Under (F1), (F2),* $\pi_1 \geq_r \pi_2 \implies \sigma(\pi_1, u) \geq_s \sigma(\pi_2, u)$
3. $T(\pi_1, y, u) \uparrow y$ *iff (F1) holds.*
4. *Consider two HMMs* $(P(u), B)$ *and* $(P(u+1), B)$.
   (a) *(F3)* $\iff P'(u+1)\pi \geq_r P'(u)\pi$.
   (b) *(F3)* $\implies T(\pi, y, u+1) \geq_r T(\pi, y, u)$, $y \in \mathcal{Y}$.

Proof (1): If (F2), then $\pi_1 \geq_r \pi_2$ implies $P'\pi_1 \geq_r P'\pi_2$:
$P'\pi_1 \geq_r P'\pi_2 \equiv \begin{bmatrix} \pi_2' P \\ \pi_1' P \end{bmatrix}$ is TP2. But $\begin{bmatrix} \pi_2' P \\ \pi_1' P \end{bmatrix} = \begin{bmatrix} \pi_2' \\ \pi_1' \end{bmatrix} P$.

Also since $\pi_1 \geq_r \pi_2$, the matrix $\begin{bmatrix} \pi_2' \\ \pi_1' \end{bmatrix}$ is TP2. By (F2), P is TP2. But product of TP2 matrices is TP2.

(2). MLR implies first order dominance.
So by (F1), $\sum_{y \geq \bar{y}} B_{x,y}(u) \uparrow x$.
By (F2), $(P_{i,1}, \dots P_{i,X}) \leq_s (P_{j,1}, \dots, P_{j,X})$ for $i \leq j$.
So $\sum_j P_{ij}(u) \sum_{y \geq \bar{y}} B_{j,y}(u) \uparrow i$.
Therefore $\pi_1 \geq_r \pi_2 \implies \sigma(\pi_1, u) \geq_s \sigma(\pi_2, u)$.

(3). Let $P'(u)\pi_1 = \bar{\pi}$. $T(\pi_1, y, u) \geq_r T(\pi_1, \bar{y}, u)$ equiv to
$(B_{i,y} B_{i+1,\bar{y}} - B_{i+1,y} B_{i,\bar{y}}) \cancel{\bar{\pi}(i)\bar{\pi}(i+1)} \leq 0, \quad y > \bar{y}$.
Equivalent to B being TP2, namely (F1).

Examples

$$P(1) = \begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.2 & 0.5 & 0.3 \\ 0.1 & 0.3 & 0.6 \end{bmatrix}, P(2) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \pi_1 = \begin{bmatrix} 0.2 \\ 0.2 \\ 0.6 \end{bmatrix}, \pi_2 = \begin{bmatrix} 0.3 \\ 0.2 \\ 0.5 \end{bmatrix}.$$

Then $P(1)$ is TP2, $P(2)$ is not TP2. Also, $\pi_1 \geq_r \pi_2$: the ratio of elements $[2/3, 1, 6/5] \uparrow$.

Ex (i): $P'(1)\pi_1 \geq_r P'(1)\pi_2$: ratio $[0.8148, 1, 1.1282]' \uparrow$
Ex (ii): Suppose $B(1) = P(1)$ so (F1), (F2) . Then
$\sigma(\pi_1, 1) = [0.2440, 0.3680, 0.3880]'$,
$\sigma(\pi_2, 1) = [0.2690, 0.3680, 0.3630]'$.. $\sigma(\pi_1, 1) \geq_s \sigma(\pi_2, 1)$.

Ex: Reduced Complexity HMM filter

$$\underbrace{\pi_{k+1} = T(\pi_k, y_{k+1}; P)}_{\text{optimal}}, \quad \underbrace{\underline{\pi}_{k+1} = T(\underline{\pi}_k, y_{k+1}; \underline{P})}_{\text{lower bound}}$$

HMM filter with P has $O(X^2)$ multiplications.
Aim: Construct sparse $\underline{P}$ with rank r s.t. $\underline{P} \preceq P$. So $O(Xr)$ mults (since $\underline{P}'\pi = \sum_{i=1}^{r} \lambda_i v_i v_i' \underline{\pi}(i)$) and $\underline{\pi}_k \leq_r \pi_k$ for all k.
State levels $g = (1, 2, \ldots, X)'$, then

$$\hat{x}_k = \mathbb{E}\{x_k | y_{0:k}; P\} = g'\pi_k, \quad \underline{x}_k \stackrel{\text{defn}}{=} \mathbb{E}\{x_k | y_{0:k}; \underline{P}\} = g'\underline{\pi}_k.$$

$$\hat{x}^{\text{MAP}} \stackrel{\text{defn}}{=} \operatorname*{argmax}_i \pi_k(i), \quad \underline{x}_k^{\text{MAP}} \stackrel{\text{defn}}{=} \operatorname*{argmax}_i \underline{\pi}_k(i).$$

Then $\underline{x}_k \leq \hat{x}_k$ and $\underline{x}_k^{\text{MAP}} \leq \hat{x}_k^{\text{MAP}}$ for all k.

Theorem 23 (Stochastic Dominance Sample-Path Bounds). *Consider $T(\pi, y; P)$ and $T(\pi, y; \underline{P})$*

1. *For any P, there exist $\underline{P}$ st $\underline{P} \preceq P$*
2. *Suppose $\underline{P} \preceq P$. Then $T(\pi, y; \underline{P}) \leq_r T(\pi, y; P)$.*
3. *Suppose P is TP2. Assume $T(\pi, y; P)$, $T(\pi, y; \underline{P})$ initialized with π_0. Then*

$$\underline{\pi}_k \leq_r \pi_k \leq_r \bar{\pi}_k, \quad \textit{for all time } k = 1, 2, \ldots$$

As a consequence (a) $\underline{x}_k \leq \hat{x}_k$. (b) $\underline{x}_k^{MAP} \leq \hat{x}_k^{MAP}$.

Proof. 1. Choose $\underline{P} = \Big[e_1, \ldots, e_1\Big]'$, $\bar{P} = \Big[e_X, \ldots, e_X\Big]'$. Then $\underline{P} \succeq P \succeq \bar{P}$.

2. Statement 2 is from Theorem 22.

3. Suppose $\underline{\pi}_k \leq_r \pi_k$. Statement 2 implies $T(\underline{\pi}_k, y_{k+1}; \underline{P}) \leq_r T(\underline{\pi}_k, y_{k+1}; P)$. Since P is TP2, $\underline{\pi}_k \leq_r \pi_k \implies T(\underline{\pi}_k, y_{k+1}; P) \leq_r T(\pi_k, y_{k+1}; P)$. Combining two ineqs $T(\underline{\pi}_k, y_{k+1}; \underline{P}) \leq_r T(\pi_k, y_{k+1}; P)$, or equivalently $\underline{\pi}_{k+1} \leq_r \pi_{k+1}$. Finally, MLR dominance implies first order dominance. So 3(a)

3(b): RTP $\underline{\pi} \leq_r \pi$ implies $\arg\max_i \underline{\pi}(i) \leq \arg\max_i \pi(i)$. Shown by contradiction: Let $i^* = \operatorname{argmax}_i \pi(i)$, $j^* = \operatorname{argmax}_j \underline{\pi}(j)$. Suppose $i^* \leq j^*$. Then $\pi \geq_r \underline{\pi} \implies \pi(i^*) \leq \frac{\underline{\pi}(i^*)}{\underline{\pi}(j^*)} \pi(j^*)$. Since $\frac{\underline{\pi}(i^*)}{\underline{\pi}(j^*)} \leq 1$, we have $\pi(i^*) \leq \pi(j^*)$ which is a contradiction.

Computing $\underline{P}$ for tightest bound is convex optimization problem.

$$\text{Minimize rank of } X \times X \text{ matrix } \underline{P} \tag{8}$$

subject to the constraints $\mathbf{Cons}(\Pi(X), \underline{P}, m)$ for $m = 1, 2, \ldots, X-1$, where for $\epsilon > 0$,

$$\mathbf{Cons}(\Pi(X), \underline{P}, m) \equiv \begin{cases} \Gamma^{(m)} \text{ is copositive on } \Pi(X) & \text{(9a)} \\ \|P'\pi - \underline{P}'\pi\|_1 \leq \epsilon, \pi \in \Pi(X) & \text{(9b)} \\ \underline{P} \geq 0, \quad \underline{P}\mathbf{1} = \mathbf{1}. & \text{(9c)} \end{cases}$$

Objective (8) is replaced with the reweighted nuclear norm (sum of the singular values of a matrix) which is convex.

Structural result 2: Monotone Value Function for POMDP

Why? Essential step for establishing monotone policy

Setup: Consider infinite horizon discounted cost POMDP:

$$J_\mu(\pi_0) = \mathbb{E}_\mu \left\{ \sum_{k=1}^{\infty} \rho^{k-1} C(\pi_k, \mu(\pi_k)) \right\}.$$

$\pi_k = T(\pi_{k-1}, y_k, u_k)$

$$\mu^*(\pi) = \underset{u \in \mathcal{U}}{\operatorname{argmin}}\, Q(\pi, u), \quad V(\pi) = \min_{u \in \mathcal{U}} Q(\pi, u),$$
$$Q(\pi, u) = C(\pi, u) + \rho \sum_{y \in Y} V\left(T\left(\pi, y, u\right)\right) \sigma\left(\pi, y, u\right).$$

Assumptions

(C) $\pi_1 \geq_s \pi_2$ implies $C(\pi_1, u) \leq C(\pi_2, u)$.
Linear costs $C(\pi, u) = c_u'\pi$: $c(x, u) \downarrow x$ for each u.
(F1) $B(u)$ is TP2 for each action $u \in \{1, 2, \ldots, U\}$.
(F2) $P(u)$ is TP2 for each action u.

Recall (F1) $\implies T(\pi, y, u) \uparrow y$, (F2) $\implies T(\pi, y, u) \uparrow \pi$.

Theorem 24. *If (C1), (F1), (F2) hold, then $V(\pi) \downarrow \pi$ wrt MLR*

Proof. By math induction on value iteration algorithm:
Initialize $V_0(\pi) = 0$ and $V_n(\pi) = \min_{u \in \mathcal{U}} Q_n(\pi, u)$,

$$Q_n(\pi, u) = C(\pi, u) + \rho \sum_{y \in Y} V_{n-1}\left(T\left(\pi, y, u\right)\right) \sigma\left(\pi, y, u\right).$$

Assume $V_{n-1}(\pi) \downarrow \pi$ by induction hypothesis.
Under (F1), $T(\pi, y, u) \uparrow y$. So $V_{n-1}\left(T\left(\pi, y, u\right)\right) \downarrow y$.
Under (F1), (F2) $\sigma(\pi, u) \uparrow \pi \geq_s$. So $\pi \geq_r \bar{\pi} \implies$

$$\sum_y V_{n-1}\left(T\left(\pi, y, u\right)\right) \sigma(\pi, y, u) \leq \sum_y V_{n-1}\left(T\left(\pi, y, u\right)\right) \sigma(\bar{\pi}, y, u)$$

Next, (F2) implies $T(\pi, y, u) \uparrow \pi$. Since $V_{n-1}(\pi) \downarrow \pi$,

$$\pi \geq_r \bar{\pi} \implies V_{n-1}\left(T\left(\pi, y, u\right)\right) \leq V_{n-1}\left(T\left(\bar{\pi}, y, u\right)\right)$$

$$\implies \sum_y V_{n-1}\left(T\left(\pi, y, u\right)\right) \sigma(\bar{\pi}, y, u) \leq \sum_y V_{n-1}\left(T\left(\bar{\pi}, y, u\right)\right) \sigma(\bar{\pi}, y, u)$$

Therefore $\pi \geq_r \bar{\pi} \implies$

$$\sum_y V_{n-1}\left(T\left(\pi, y, u\right)\right) \sigma(\pi, y, u) \leq \sum_y V_{n-1}\left(T\left(\bar{\pi}, y, u\right)\right) \sigma(\bar{\pi}, y, u).$$

Finally, under (C1), $C(\pi, u) \downarrow \pi$ So $\pi \geq_r \bar{\pi} \implies$

$$C(\pi, u) + \sum_y V_{n-1}\left(T\left(\pi, y, u\right)\right) \sigma(\pi, y, u) \\ \leq C(\bar{\pi}, u) + \sum_y V_{n-1}\left(T\left(\bar{\pi}, y, u\right)\right) \sigma(\bar{\pi}, y, u)$$

i.e., $Q_n(\pi, u) \leq Q_n(\bar{\pi}, u)$. So $Q_n(\pi, u) \downarrow \pi$. So $V_n(\pi) \downarrow \pi$.

Example: 2 state POMDP

Consider discounted cost POMDP $(\mathcal{X}, \mathcal{U}, \mathcal{Y}, P(u), B(u), c(u), \rho)$ where $\mathcal{X} = \{1, 2\}$, $\mathcal{Y}$ can be continuous or discrete, and $\rho \in [0, 1)$.

(C) $c(x, u)$ is decreasing in $x \in \{1, 2\}$ for each $u \in \mathcal{U}$.

(F1) B is totally positive of order 2 (TP2).

(F2) $P(u)$ is totally positive of order 2 (TP2).

(F3) $P_{12}(u+1) - P_{12}(u) \leq P_{22}(u+1) - P_{22}(u)$ (tail-sum supermodularity).

(S) The costs are submodular: $c(1, u+1) - c(1, u) \geq c(2, u+1) - c(2, u)$.

Theorem 25. *Under (C), (F1), (F2), (F3), (S), optimal policy $\mu^*(\pi) \uparrow \pi$ Thus $\mu^*(\pi(2))$ has the following finite dimensional characterization: There exist $U + 1$ thresholds $0 = \pi_0^* \leq \pi_1^* \leq \cdots \leq \pi_U^* \leq 1$ such that*

$$\mu^*(\pi) = \sum_{u \in \mathcal{U}} u \, I\left(\pi(2) \in (\pi_{u-1}^*, \pi_u^*]\right).$$

Proof exploits $V(\pi) \downarrow \pi$ and concave to show $Q(\pi, u)$ is submodular.

$$Q(\pi, u) - Q(\pi, \bar{u}) - Q(\bar{\pi}, u) + Q(\bar{\pi}, \bar{u}) \leq 0, \quad u > \bar{u}, \ \pi \geq_r \bar{\pi}.$$

Recall for $X = 2$, $\geq_s = \geq_r$ = completely ordered (so submod defn wrt total order).

Structural Result 3: Monotone Policy for Stopping time POMDP

$$\mu^*(\pi) = \underset{u \in \mathcal{U}}{\operatorname{argmin}}\, Q(\pi, u), \quad V(\pi) = \min_{u \in \mathcal{U}} Q(\pi, u),$$

$$Q(\pi, 1) = c_1'\pi, \quad Q(\pi, 2) = c_2'\pi + \rho \sum_{y \in Y} V\left(T\left(\pi, y\right)\right) \sigma\left(\pi, y\right).$$

Aim: Sufficient conditions so that $\mu^*(\pi) \uparrow \pi$. But MLR is partial order, how to interpret on simplex?

We want to show:

$$\pi_1, \pi_2 \in \mathcal{L}(e_i, \bar{\pi}), \quad \pi_1 \geq_r \pi_2 \implies \mu^*(\pi_1) \geq \mu^*(\pi_2), \quad i \in \{1, X\}.$$

Here $\mathcal{L}(e_i, \bar{\pi})$ denotes any line segment in $\Pi(X)$ which starts at e_1 and ends at any belief $\bar{\pi}$ in the subsimplex $\{e_2, , \dots, e_X\}$; or any line segment which starts at e_X and ends at any belief $\bar{\pi}$ in the sub simplex $\{e_1, \dots, e_{X-1}\}$. Instead of $\mu^*(\pi) \uparrow$ on $\Pi(X)$, we prove $\mu^*(\pi) \uparrow$ on special line segments $\mathcal{L}(e_i, \bar{\pi})$. On such lines MLR is total order.

1. $\mu^*(\pi)$ characterized by switching curve Γ
2. The optimal linear approx to Γ that preserves submodularity estimated via simulation based stochastic approximation

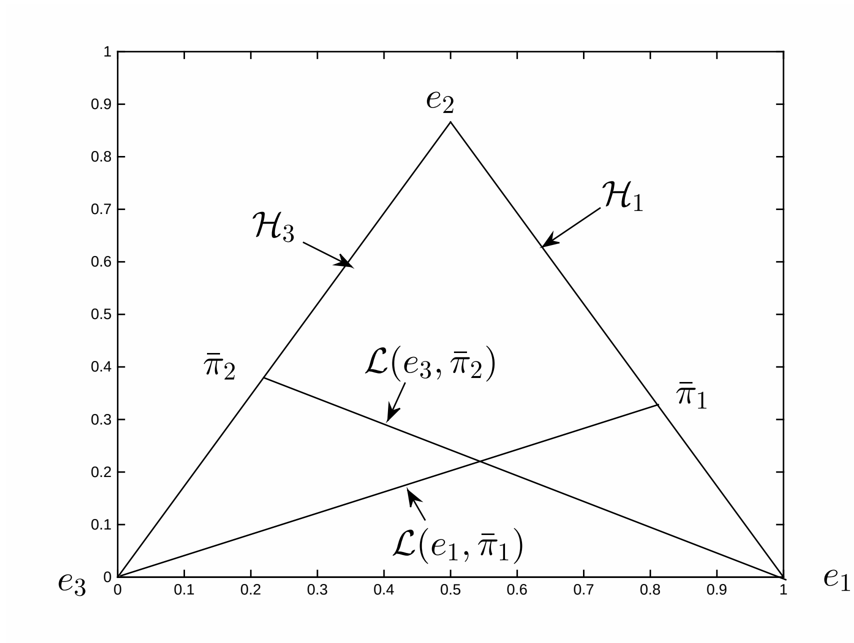


Figure 6: Examples of sub-simplices $\mathcal{H}_1$ and $\mathcal{H}_3$ and points $\bar{\pi}_1 \in \mathcal{H}_1$, $\bar{\pi}_2 \in \mathcal{H}_3$. Also shown are lines $\mathcal{L}(e_1, \bar{\pi}_1)$ and $\mathcal{L}(e_3, \bar{\pi}_2)$

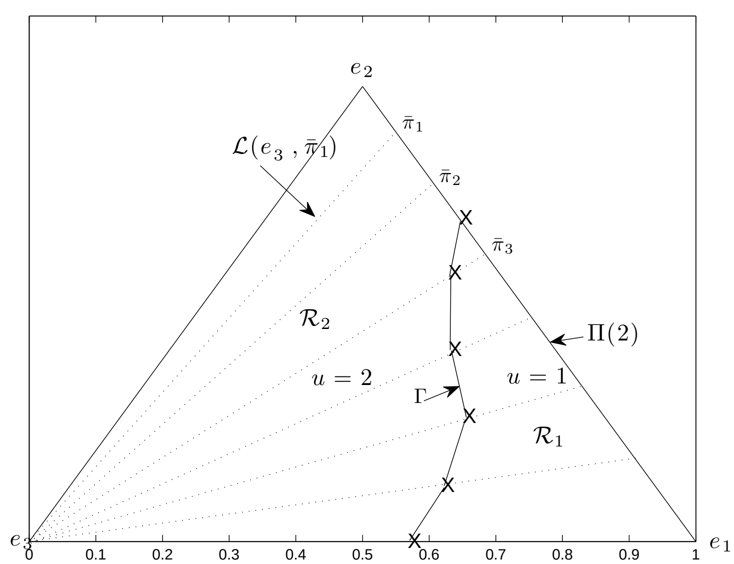


Figure 7: Both $\mathcal{R}_1$ and $\mathcal{R}_2$ are connected sets. Γ intersects each line $\mathcal{L}(e_X, \bar{\pi})$ only once.

MLR Order on Lines

Define the sub-simplices, $\mathcal{H}_1$ and $\mathcal{H}_X$:

$$\mathcal{H}_1 = \{\pi \in \Pi(X) : \pi(1) = 0\},\ \mathcal{H}_X = \{\pi \in \Pi(X) : \pi(X) = 0\} \tag{10}$$

Let $\bar{\pi} \in \mathcal{H}_1$ or $\mathcal{H}_X$. For each such $\bar{\pi} \in \mathcal{H}_i$, $i \in \{1, X\}$, construct line segment $\mathcal{L}(e_i, \bar{\pi})$ that connects $\bar{\pi}$ to e_i.

$$\mathcal{L}(e_i, \bar{\pi}) = \{\pi \in \Pi(X) : \pi = (1-\epsilon)\bar{\pi} + \epsilon e_i,\ 0 \leq \epsilon \leq 1\}, \bar{\pi} \in \mathcal{H}_i.$$

Definition 26 (MLR ordering $\geq_{L_i}$ on lines)**.** *$\pi_1 \geq_{L_i} \pi_2$, if $\pi_1, \pi_2 \in \mathcal{L}(e_i, \bar{\pi})$ for some $\bar{\pi} \in \mathcal{H}_i$, and $\pi_1 \geq_r \pi_2$.*

$[\mathcal{L}(e_1, \bar{\pi}), \geq_{L_X}]$ and $[\mathcal{L}(e_X, \bar{\pi}), \geq_{L_1}]$ are chains, i.e., totally ordered sets. All elements $\pi_1, \pi_2 \in \mathcal{L}(e_X, \bar{\pi})$ are comparable, i.e., either $\pi_1 \geq_{L_X} \pi_2$ or $\pi_2 \geq_{L_X} \pi_1$ (and similarly for $\mathcal{L}(e_1, \bar{\pi})$). The supremum of $[\mathcal{L}(e_1, \bar{\pi}), \geq_{L_X}]$ is $\bar{\pi}$ and infimum is e_1.

Definition 27 (Submodular function)**.** *Suppose $i = 1$ or X. Then $f : \mathcal{L}(e_i, \bar{\pi}) \times \mathcal{U} \to \mathbb{R}$ is submodular if $f(\pi, u) - f(\pi, \bar{u}) \leq f(\tilde{\pi}, u) - f(\tilde{\pi}, \bar{u})$, for $\bar{u} \leq u$, $\pi \geq_{L_i} \tilde{\pi}$.*

Theorem 28 (Topkis Theorem)**.** *Suppose $i = 1$ or X. If $f : \mathcal{L}(e_i, \bar{\pi}) \times \mathcal{U} \to \mathbb{R}$ is submodular, then there $\mu^*(\pi) = \operatorname{argmin}_{u \in \mathcal{U}} f(\pi, u) \uparrow$ on $[\mathcal{L}(e_i, \bar{\pi}), \geq_{L_i}]$, i.e., $\pi^0 \geq_{L_i} \pi \implies \mu^*(\pi) \leq \mu^*(\pi^0)$.*

Threshold Switching Curve

(C) $\pi_1 \geq_s \pi_2$ implies $C(\pi_1, u) \leq C(\pi_2, u)$ for each u.

(F1) B is TP2.

(F2) P is TP2

(S) $C(\pi, u)$ is submod on $[\mathcal{L}(e_X, \bar{\pi}), \geq_{L_X}]$, $[\mathcal{L}(e_1, \bar{\pi}), \geq_{L_1}]$.
For linear costs: $c(x, 2) - c(x, 1) \geq c(X, 2) - c(X, 1)$ and $c(1, 2) - c(1, 1) \geq c(x, 2) - c(x, 1)$.

Theorem 29 (Switching Curve Optimal Policy)**.** *For a stopping time POMDP under (C), (F1), (F2), (S):*

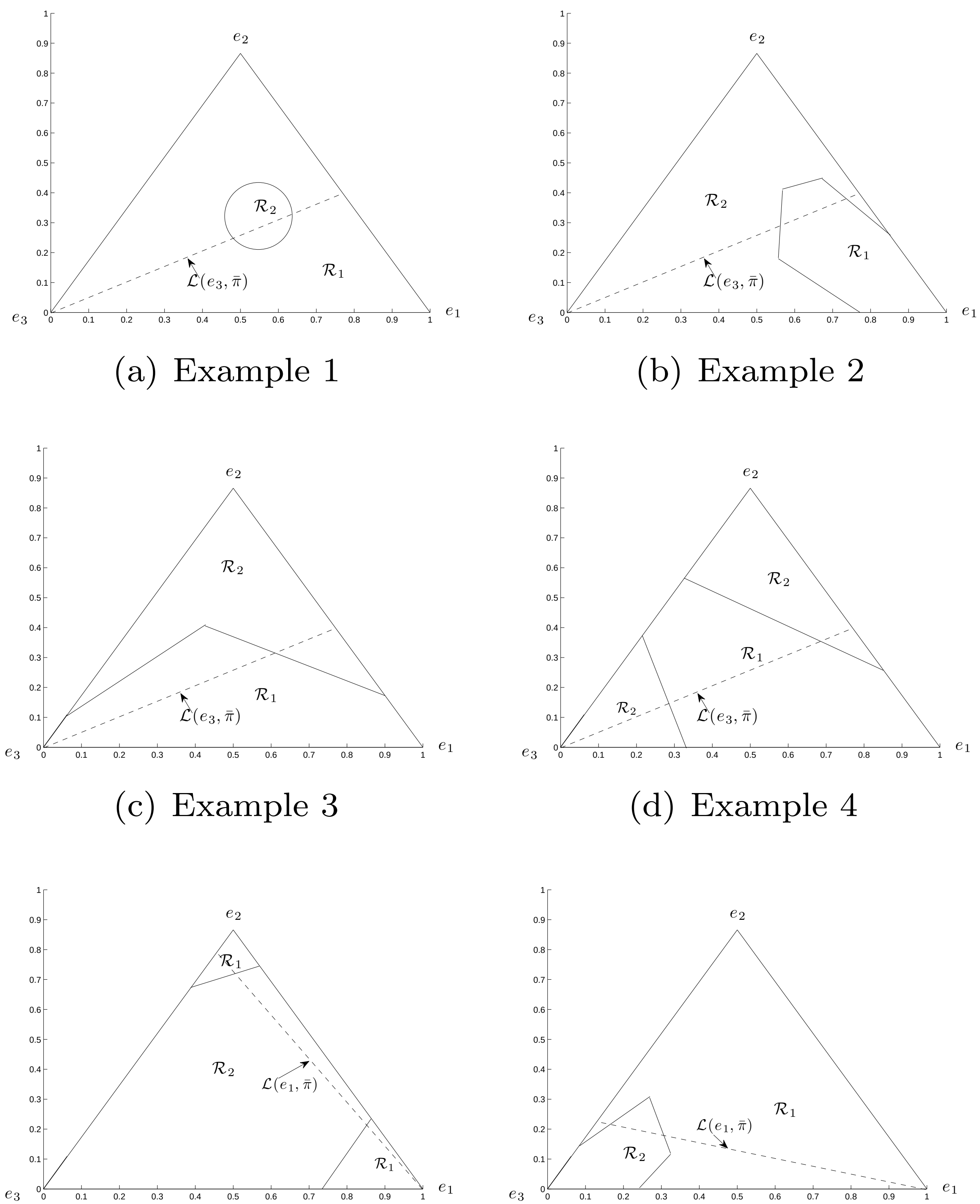
1. *There exists $\mu^*(\pi)$ that is $\geq_{L_X}$ increasing on lines $\mathcal{L}(e_X, \bar{\pi})$ and $\geq_{L_1}$ increasing on lines $\mathcal{L}(e_1, \bar{\pi})$.*
2. *Hence there exists a threshold switching curve Γ that partitions $\Pi(X)$ into two individually connected[a] regions $\mathcal{R}_1$, $\mathcal{R}_2$, such that the optimal policy is*

$$\mu^*(\pi) = \begin{cases} \textit{continue} = 2 & \textit{if } \pi \in \mathcal{R}_2 \\ \textit{stop} = 1 & \textit{if } \pi \in \mathcal{R}_1 \end{cases} \tag{11}$$

   Γ intersects each line $\mathcal{L}(e_X, \bar{\pi})$, $\mathcal{L}(e_1, \bar{\pi})$ at most once.
3. *There exists $i^* \in \{0, \ldots, X\}$, such that $e_1, \ldots, e_{i^*} \in \mathcal{R}_1$ and $e_{i^*+1}, \ldots, e_X \in \mathcal{R}_2$.*
4. *For the case $X = 2$, there exists a unique threshold point $\pi^*(2)$.*

---

[a] A set is connected if it cannot be expressed as the union of disjoint nonempty closed sets



(a) Example 1 (b) Example 2

(c) Example 3 (d) Example 4

(e) Example 5 (f) Example 6

Figure 8: Examples that violate the monotone property of Theorem 29 on $\Pi(3)$.

Example. Explicit stopping set for one-step-ahead property

$$\mathcal{S}^o \overset{\text{defn}}{=} \{\pi : c_1'\pi \leq c_2'\pi + \rho c_1' P' \pi\}.$$

$c_2'\pi + \rho c_1' P'\pi$: cost of proceeding one step ahead and then stopping, while $c_1'\pi$ is cost of stopping immediately.

Theorem 30. *Suppose* (c_1, c_2, P, B, ρ) *of stopping time POMDP satisfy one-step-ahead property:*

$$\pi \in \mathcal{S}^o \implies T(\pi, y) \in \mathcal{S}^o, \quad \forall y \in \mathcal{Y}. \tag{12}$$

Then optimal stopping set $\mathcal{R}_1 = \mathcal{S}^o$*.*

Bellman's equation: $\mathcal{S}^o \subseteq \mathcal{R}_1$. If (12) then $\mathcal{S}^o = \mathcal{R}_1$.

Proof 1. $\pi \in \mathcal{S}^o \implies V(\pi) = c_1'\pi$. So optimal to stop.
2: $\pi \notin \mathcal{S}^o \implies V(\pi) < c_1'\pi$. So optimal not to stop since if $\pi \notin \mathcal{S}^o$ then proceeding one step ahead and stopping (cost $c_2'\pi + \rho c_1' P'\pi$) is cheaper than stopping (cost $c_1'\pi$).

Under (C1), (F1), (F2), (S) (12): finite characterization.
(F1), (F2) $\implies T(\pi, y) \uparrow \pi, y$. With (C) (S) $\mu^*(\pi) \uparrow \pi$.
For finite observation $\mathcal{Y} = \{1, 2, \ldots, Y\}$, (12) equiv to:

$$T(\pi_i^*, Y) \in \mathcal{S}^o,\ i = 2, \ldots, X, \pi_i^* = \{\pi \in \mathcal{S}^0 : \pi(j) = 0, j \neq \{1, i\}\}$$

π_i^* are $X - 1$ corner points where hyperplane $c_1'\pi = c_2'\pi + \rho c_1' P'\pi$ intersect the faces of $\Pi(X)$.

Optimal Linear Decision Threshold for Stopping time POMDP

How to estimate switching curve Γ?

Since $\Pi(X) \subset \mathbb{R}^{X-1}$, linear hyperplane on $\Pi(X)$ has $X-1$ coefficients. Define linear threshold policy

$$\mu_\theta(\pi) = \begin{cases} \text{stop} = 1 & \text{if } \begin{bmatrix} 0 & 1 & \theta' \end{bmatrix}' \begin{bmatrix} \pi \\ -1 \end{bmatrix} < 0 \\ \text{continue} = 2 & \text{otherwise} \end{cases} \qquad \pi \in \Pi(X).$$

$\theta = (\theta(1), \ldots, \theta(X-1))'$ parameter vector of linear policy.

Why? $\bar{\theta}'\pi \geq \gamma \implies u = 2$ and $\bar{\theta}'\pi < \gamma \implies u = 1$, $\bar{\theta} \in \mathbb{R}^n$, $\gamma \in \mathbb{R}_+$. If $\min_i \bar{\theta}(i) < 0$, $(\bar{\theta}' - \min_i \bar{\theta}(i)\mathbf{1}')\pi \geq \gamma - \min_i \bar{\theta}(i)$ implies $u = 2$. This yields above after dividing by $\gamma - \min_i \bar{\theta}(i)$.

Give necessary and sufficient conditions on θ for $\mu_\theta(\pi)$ MLR increasing on lines. Then optimizing over θ yields "optimal" linear approximation to Γ. (Assume $e_1 \in \mathcal{R}_1$).

Theorem 31 (Optimal Linear Threshold Policy)**.** *For $\pi \in \Pi(X)$, linear threshold policy $\mu_\theta(\pi)$ is*
(i) MLR increasing on lines $\mathcal{L}(e_X, \bar{\pi})$ iff $\theta(X-2) \geq 1$ and $\theta(i) \leq \theta(X-2)$ for $i < X-2$.
(ii) MLR increasing on lines $\mathcal{L}(e_1, \bar{\pi})$ iff $\theta(i) \geq 0$, for $i < X-2$. □

Proof. Given any $\pi_1, \pi_2 \in \mathcal{L}(e_X, \bar{\pi})$ with $\pi_2 \geq_{L_X} \pi_1$, we need to prove: $\mu_\theta(\pi_1) \leq \mu_\theta(\pi_2)$ iff $\theta(X-2) \geq 1$, $\theta(i) \leq \theta(X-2)$ for $i < X-2$.
Clearly $\mu_\theta(\pi_1) \leq \mu_\theta(\pi_2)$ is equivalent to
$\begin{bmatrix} 0 & 1 & \theta' \end{bmatrix}' \begin{bmatrix} \pi_1 \\ -1 \end{bmatrix} \leq \begin{bmatrix} 0 & 1 & \theta' \end{bmatrix}' \begin{bmatrix} \pi_2 \\ -1 \end{bmatrix}$, i.e.,

$$\begin{bmatrix} 0 & 1 & \theta(1) & \cdots & \theta(X-2) \end{bmatrix} (\pi_1 - \pi_2) \leq 0.$$

Now $\pi_2 \geq_{L_X} \pi_1$ implies that $\pi_1 = \epsilon_1 e_X + (1-\epsilon_1)\bar{\pi}$, $\pi_2 = \epsilon_2 e_X + (1-\epsilon_2)\bar{\pi}$ and $\epsilon_1 \leq \epsilon_2$. Substituting these into the above expression, we need to prove

$$(\epsilon_1 - \epsilon_2)\Big(\theta(X-2) - \begin{bmatrix} 0 & 1 & \theta(1) & \cdots & \theta(X-2) \end{bmatrix}' \bar{\pi}\Big) \leq 0, \forall \bar{\pi} \in \mathcal{H}_X$$

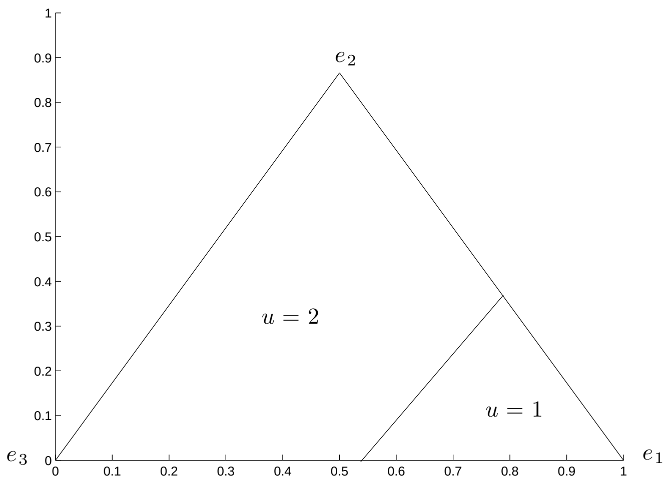
iff $\theta(X-2) \geq 1$, $\theta(i) \leq \theta(X-2)$, $i < X-2$.

A similar proof shows that on lines $\mathcal{L}(e_1, \bar{\pi})$ the linear threshold policy satisfies $\mu_\theta(\pi_1) \leq \mu_\theta(\pi_2)$ iff $\theta(i) \geq 0$ for $i < X-2$. □

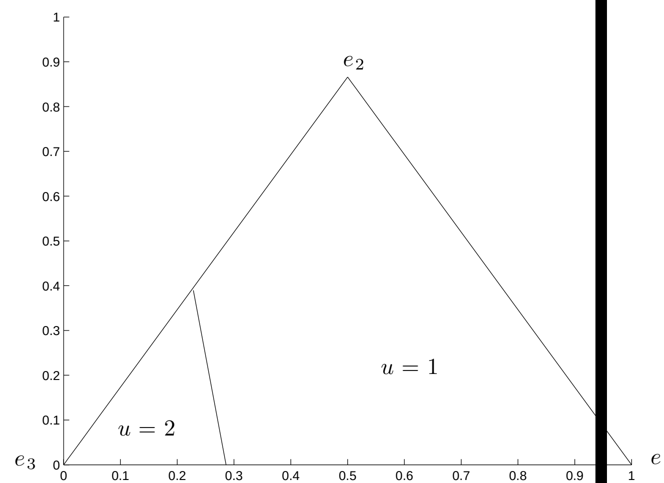
Optimal linear threshold approximation to Γ is:

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^X} J_{\mu_\theta}(\pi),$$
$$\text{st } 0 \leq \theta(i) \leq \theta(X-2),\ \theta(X-2) \geq 1 \text{ and } \theta(X-1) > 0$$

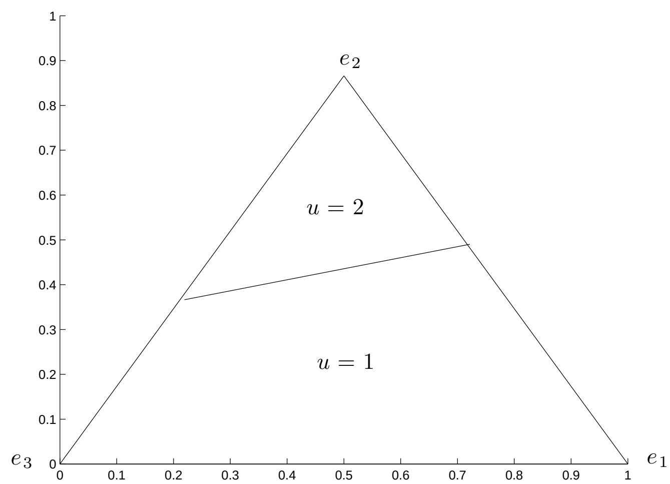
Intuition: Consider $\mathcal{X} = \{1, 2, 3\}$. Then $\theta(1) \geq 1$ implies that the linear threshold has slope of 60^o or larger.



(a) Case 1



(b) Case 2



(c) Case 3 (invalid)

Figure 9: Examples of valid MLR increasing linear threshold policies for a stopping time POMDP on belief space $\Pi(X)$ for $X = 3$ (Case 1 and Case 2). Case 3 is invalid.

Computing optimal linear threshold policy:

Compute $\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}\{J_n(\mu_\theta)\}$

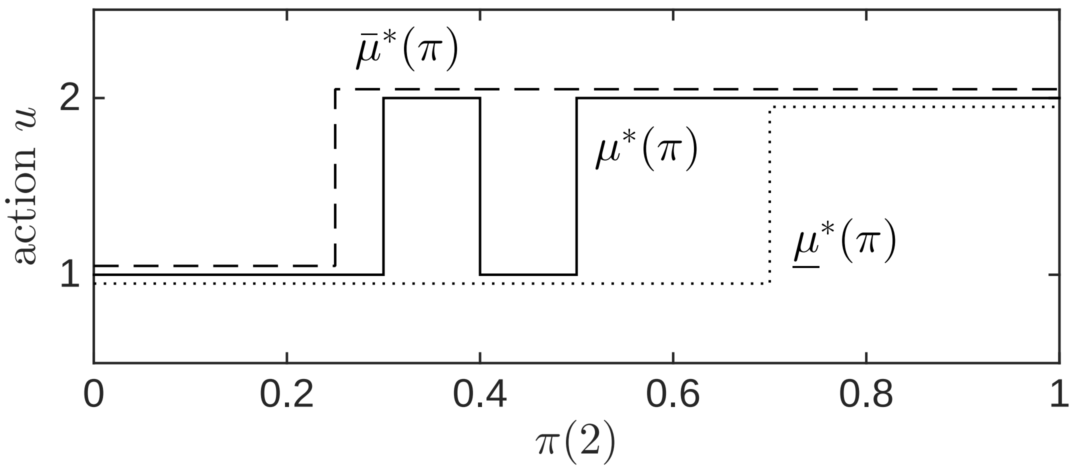
subject to $0 \leq \theta(i) \leq \theta(X-2), \theta(X-2) \geq 1$ and $\theta(X-1) > 0$.

sample path cumulative cost $J_n(\mu_\theta)$ evaluated as

$$J_n(\mu_\theta) = \sum_{k=0}^{\infty} \rho^k C(\pi_k, u_k), \quad \text{where } u_k = \mu_\theta(\pi_k)$$

with prior π_0 sampled uniformly from $\Pi(X)$. Convenient way of sampling uniformly from $\Pi(X)$ is to use the Dirichlet distribution (i.e., $\pi_0(i) = x_i / \sum_i x_i$, where $x_i \sim$ unit exponential distribution).

Myopic Bounds to Optimal Policy in Controlled Sensing



Blackwell dominance to construct lower myopic bounds to optimal policy for POMDP.

Controlled sensing: Based on π_{k-1}, choose sensing mode

$$u_k \in \{1 \text{ (low resolution sensor) }, 2 \text{ (high resolution sensor)}\}$$

$$B(u) = \big(B_{iy^{(u)}}(u), i \in \{1, 2, \ldots, X\}, y^{(u)} \in \mathcal{Y}^{(u)}\big)$$

$$\text{where } B_{iy^{(u)}}(u) = \mathbb{P}(y^{(u)} | x = e_i, u).$$

Definition. Mode 2 *Blackwell dominates* mode 1,

$$B(2) \succeq_{\mathcal{B}} B(1) \quad \text{if} \quad B(1) = B(2)\, R$$

R is stochastic matrix. So $B^{(2)}$ more accurate than $B^{(1)}$.

Aim: $\mu^*(\pi) = \operatorname{argmin}_\mu J_\mu(\pi) = \mathbb{E}_\mu\{\sum_{k=0}^{\infty} \rho^k C(\pi_k, u_k)\}$.
Main result: Define $\Pi^s = \{\pi : C(\pi, 2) < C(\pi, 1)\}$

$$\text{and myopic policy } \underline{\mu}(\pi) = \begin{cases} 2 & \pi \in \Pi^s \\ 1 & \text{otherwise} \end{cases}$$

Theorem 32. *Assume $C(\pi, u)$ concave. $B(2) \succeq_{\mathcal{B}} B(1)$. Then $\mu^*(\pi) \geq \underline{\mu}(\pi)$ and for $\pi \in \Pi^s$, $\mu^*(\pi) = \underline{\mu}(\pi)$.*

Example 1. Optimal Filter vs Predictor Scheduling:
$u = 2$ HMM filter vs $u = 1$ HMM predictor.
$B(1) = \frac{1}{Y}\mathbf{1}_{X\times Y}$. Clearly $B(1) = B(2)B(1)$ meaning that filter ($u = 2$) Blackwell dominates the predictor ($u = 1$)

Example 2. Ultrametric Matrices An $X \times X$ square matrix B is a symmetric stochastic ultrametric matrix if

1. $B_{ij} \geq \min\{B_{ik}, B_{kj}\}$ for all $i, j, k \in \{1, 2, \ldots, X\}$.
2. $B_{ii} > \max\{B_{ik}\}, k \in \{1, 2, \ldots, X\} - \{i\}$ (diagonally dominant).

If B is symmetric stochastic ultrametric matrix, then $B^{1/U}$ is stochastic matrix for any positive integer U. Then clearly $B^{1/U} \succeq_{\mathcal{B}} B^{2/(U)} \succeq_{\mathcal{B}} \cdots \succeq_{\mathcal{B}} B^{(U-1)/U} \succeq_{\mathcal{B}} B$.

Proof: We know that $C(\pi,u) \implies V(\pi)$ is concave. Next

$$T(\pi,y^{(1)},1) = \sum_{y^{(2)}\in\mathcal{Y}^{(2)}} T(\pi,y^{(2)},2)\frac{\sigma(\pi,y^{(2)},2)}{\sigma(\pi,y^{(1)},1)}P(y^{(1)}|y^{(2)})$$

$$\sigma(\pi,y^{(1)},1) = \sum_{y^{(2)}\in\mathcal{Y}^{(2)}} \sigma(\pi,y^{(2)},2)P(y^{(1)}|y^{(2)}).$$

In more detail: the j-th element is

$$T_j(\pi,y^{(1)},1) = \frac{\sum_{y^{(2)}}\sum_i \pi(i)P_{ij}\,p(y^{(2)}|j)\,p(y^{(1)}|y^{(2)})}{\sum_m\sum_{y^{(2)}}\sum_i \pi(i)P_{im}\,p(y^{(2)}|m)\,p(y^{(1)}|y^{(2)})}$$

$$= \frac{\sum_{y^{(2)}}\sum_i \pi(i)P_{ij}p(y^{(2)}|j)\frac{\sum_l\sum_m \pi(l)P_{lm}p(y^{(2)}|m)}{\sum_l\sum_m \pi(l)P_{lm}\,p(y^{(2)}|m)}\,p(y^{(1)}|y^{(2)})}{\sum_m\sum_{y^{(2)}}\sum_i \pi(i)P_{im}p(y^{(2)}|m)\,p(y^{(1)}|y^{(2)})}$$

$$= \frac{\sum_{y^{(2)}} T_j(\pi,y^{(2)},2)\,\sigma(\pi,y^{(2)},2)p(y^{(1)}|y^{(2)})}{\sum_{y^{(2)}}\sigma(\pi,y^{(2)},2)\,p(y^{(1)}|y^{(2)})}$$

Note $\frac{\sigma(\pi,y^{(2)},2)}{\sigma(\pi,y^{(1)},1)}P(y^{(1)}|y^{(2)})$ is probability measure w.r.t. $y^{(2)}$. Since $V(\cdot)$ is concave, Jensen's inequality implies

$$V(T(\pi,y^{(1)},1)) = V\left(\sum_{y^{(2)}\in\mathcal{Y}^{(2)}} T(\pi,y^{(2)},2)\frac{\sigma(\pi,y^{(2)},2)}{\sigma(\pi,y^{(1)},1)}P(y^{(1)}|y^{(2)})\right)$$

$$\geq \sum_{y^{(2)}\in\mathcal{Y}^{(2)}} V(T(\pi,y^{(2)},2))\frac{\sigma(\pi,y^{(2)},2)}{\sigma(\pi,y^{(1)},1)}P(y^{(1)}|y^{(2)})$$

$$\implies \sum_{y^{(1)}} V(T(\pi,y^{(1)},1))\sigma(\pi,y^{(1)},1) \geq \sum_{y^{(2)}} V(T(\pi,y^{(2)},2)\sigma(\pi,y^{(2)},2).$$

Therefore for $\pi \in \Pi^s$,

$$C(\pi,2) + \rho \sum_{y^{(2)}} V(T(\pi, y^{(2)}, 2)\sigma(\pi, y^{(2)}, 2)$$
$$\leq C(\pi,1) + \rho \sum_{y^{(1)}} V(T(\pi, y^{(1)}), 1)\sigma(\pi, y^{(1)}, 1).$$

So for $\pi \in \Pi^s$, $\mu^*(\pi) = \arg\min_{u \in \mathcal{U}} Q(\pi, u) = 2$.
So $\underline{\mu}(\pi) = \mu^*(\pi) = 2$ for $\pi \in \Pi^s$ and $\bar{\mu}(\pi) = 1$ otherwise, implying that $\bar{\mu}(\pi)$ is a lower bound for $\mu^*(\pi)$.