Partially Observable Markov Decision Processes

Sequential decision-making with imperfect observation

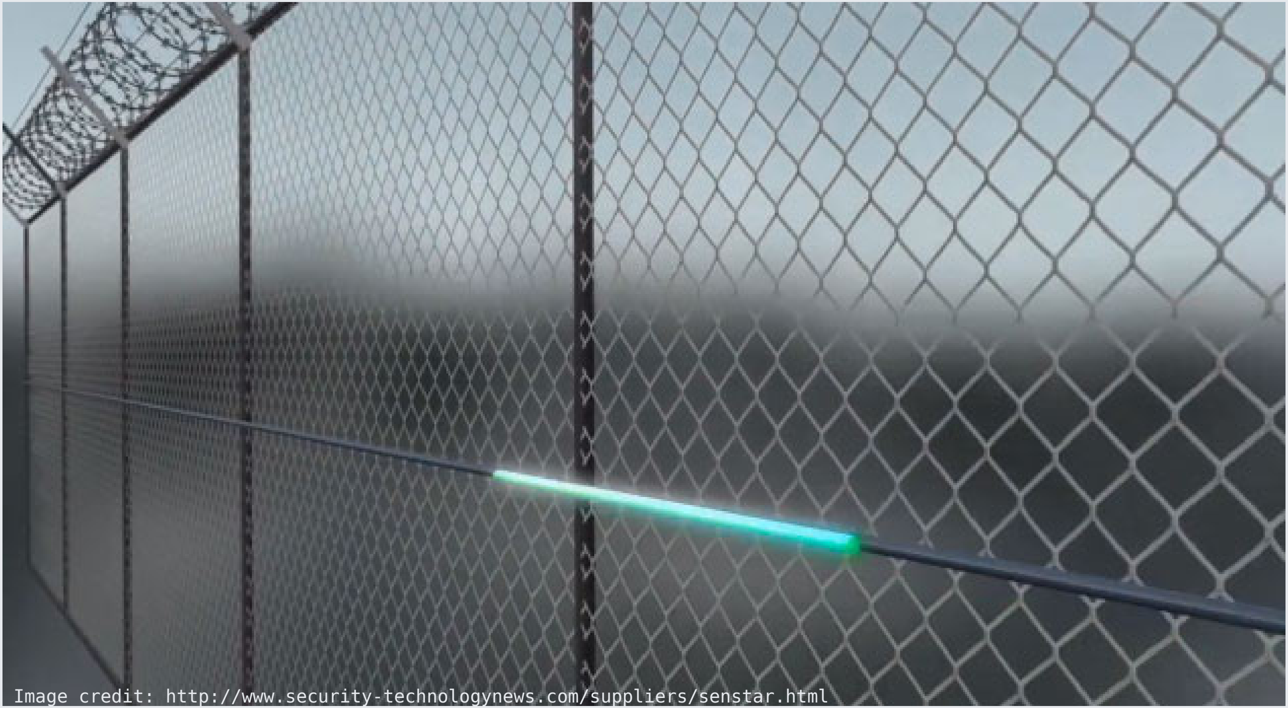
Aditya Mahajan

McGill University

Lecture Notes for ECSE 506: Stochastic Control and Decision Theory

March 15, 2014

POMDP Example: Sequential hypothesis testing



POMDP example: Sequential hypothesis testing

Description A decision maker (DM) makes a series of i.i.d. observations which may be distributed according to PDF f_0 or f_1. Let Y_t denote the decision maker's t-th observation. In this example, time denotes the number of observations that the DM has made so far.

Example:

$h_0 : Y_t \sim \mathcal{N}(0, \sigma^2)$

$h_1 : Y_t \sim \mathcal{N}(\mu, \sigma^2)$

Example:

$h_0 : Y_t \sim \text{Ber}(p)$

$h_1 : Y_t \sim \text{Ber}(q)$

The DM wants to differentiate between the two hypothesis:

$$h_0 : Y_t \sim f_0, \quad \text{and} \quad h_1 : Y_t \sim f_1.$$

Let the random variable H denote the value of the hypothesis. The a priori probability $\mathbb{P}(H = h_0) = p$.

The system continues for a finite time T. At each $t < T$, the DM has three options: stop and declare h_0, stop and declare h_1, or continue and take another measurement. At time T, the last alternative is unavailable.

Cost per obs. c

Type-I error $\ell(h_1, h_0)$

Type-II error $\ell(h_0, h_1)$

Usually: $\ell(h_0, h_0) = \ell(h_1, h_1) = 0$.

Let τ be the time when the DM stops and υ be his final decision. The cost of running the system is $c\tau + \ell(\upsilon, H)$. Find the optimal stopping strategy for the DM that minimizes expected value of this cost.

POMDP example: Sequential hypothesis testing

Notation

State: $X_t = (H, S_t) \in \{h_0, h_1\} \times \{0, 1\}$

$S_t = 1$ implies that the process has stopped.

Observation: Under $H = h_0 : Y_t \sim f_0$; under $H = h_1 : Y_t \sim f_1$.

Control: For $t < T$, $U_t \in \{h_0, h_1, C\}$

For $t = T$, $U_t \in \{h_0, h_1\}$

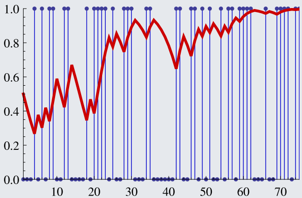
Dynamics $S_{t+1} = \mathbb{1}\{S_t = 1\} + \mathbb{1}\{S_t = 0\}\,\mathbb{1}\{U_t \in \{h_0, h_1\}\}$, where $U_t = g_t(Y_{1:t})$

Costs

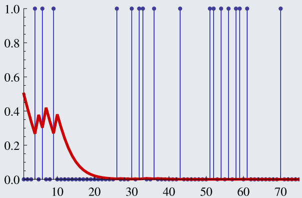
Measurement cost, $u_t \in C$: $c_t(X_t, C) = c$

Stopping cost $u_t \in \{h_0, h_1\}$: $c_t(X_t, u_t) = \ell(H, u_t)$

Illustration Observations $Y_t \sim \mathrm{Ber}(q_i)$, where $q_0 = 0.5$ and $q_1 = 0.3$.



$Y_t \sim \mathrm{Ber}(0.5)$



$Y_t \sim \mathrm{Ber}(0.3)$

— denotes $\mathbb{P}(H = h_0 | Y_{1:t})$

Sequential hypothesis testing is a POMDP

	POMDP Dynamic Model	Sequential Hypothesis Testing
System Dynamics	$X_{t+1} = f_t(X_t, U_t, W_t)$	$X_t = (H_t, S_t)$, $H_{t+1} = H_t, \quad S_{t+1} = \text{Func}(S_t, U_t)$
Observation	$Y_t = h_t(X_t, N_t)$	$Y_t = \text{Func}(H_t, N_t)$
Information Structure	$U_t = g_t(Y_{1:t}, U_{1:t-1})$	$U_t = g_t(Y_{1:t}), \quad \because \forall t' < t, U_{t'} = C,$
Objective Function	$\mathbb{E}\left[\sum_{t=1}^{T} c_t(X_t, U_t)\right]$	$\mathbb{E}\left[c\tau + \ell(H, U_\tau)\right]$

Per-step cost function

Define a per-step cost function $\rho(x_t, u_t)$ as

$$\rho((h, s), u) = \begin{cases} 0 & \text{if } s = 1 \\ c & \text{if } s = 0 \text{ and } u = C \\ \ell(h, u) & \text{if } s = 0 \text{ and } u \in \{h_0, h_1\} \end{cases}$$

Sequential hypothesis testing is a POMDP

Information state

The state X_t has two components, an unobservable H and observable S_t. Define information state (π_t, s_t) where

$$\pi_t(h) = \mathbb{P}(H = h \mid Y_{1:t}).$$

π_t is equivalent to $p_t = \pi_t(0)$, which evolves as follows:

$$p_{t+1} = \varphi(p_t, y_t) = p_t f_0(y_t) \big/ \big(p_t f_0(y_t) + (1 - p_t) f_1(y_t) \big)$$

Structure of Controller

Since we only take a decision when $S_t = 0$, there is no loss of optimality in using strategies of the form:

$$U_t = g_t(p_t)$$

Dynamic program

$$V_T(p) = \max \big\{ p\ell(h_0, h_0) + (1 - p)\ell(h_1, h_0), \\ p\ell(h_0, h_1) + (1 - p)\ell(h_1, h_1) \big\}$$

$$V_t(p) = \max \Big\{ c + \mathbb{E}[V_{t+1}(\varphi(p, Y_{t+1})) \mid p_t = p], \\ p\ell(h_0, h_0) + (1 - p)\ell(h_1, h_0). \\ p\ell(h_0, h_1) + (1 - p)\ell(h_1, h_1) \Big\}$$

Qualitative properties of the value function

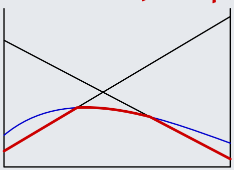
Definition

$$W_T(p) = \infty$$

$$W_t(p) = c + \mathbb{E}[V_{t+1}(\varphi(p, Y_t) \mid p_t = p]$$

Theorem $V_t(p)$ and $W_t(p)$ are ▶ $\forall t$, concave in p ▶ $\forall p$, increasing in t

Proof of concavity in p



Minimum of two linear and one concave function

Proceed by backward induction.

- Basis: $V_T(p)$ is minimum of two linear functions, and hence concave. $W_T(p)$ is a constant, and hence concave.
- Induction hypothesis: $V_{t+1}(p)$ and $W_{t+1}(p)$ are concave in p.
- Induction step: Properties of convex functions: (i) if $f(x)$ is concave in x, then $tf(x/t)$, the perspective of f, is concave in (x, t) for $t > 0$. (ii) sum of concave functions is concave. Hence,

$$W_t(p) = c + \int_y [pf_0(y) + (1-p)f_1(y)] V_{t+1}\left(\frac{pf_0(y)}{pf_0(y) + (1-p)f_1(y)}\right) dy$$

is concave in p. Thus, $V_t(p)$ is a minimum of three functions, two linear in p and one concave in p. Hence, $V_t(p)$ is also concave in p.

Qualitative properties of the value function

Definition $L_i(p) = p\ell(h_i, h_0) + (1-p)\ell(h_i, h_1), \quad i \in \{1, 2\}$

Proof of increasing in t Proceed by backward induction.

- Basis: By construction, $W_{T-1}(p) \leqslant W_T(p)$. Moreover,

$$V_{T-1}(p) = \min\{W_{T-1}(p), L_0(p), L_1(p)\}$$
$$\leqslant \min\{L_0(p), L_1(p)\} = V_T(p)$$

- Induction hypothesis: $V_{t+1}(p) \leqslant V_{t+2}(p)$ and $W_{t+1}(p) \leqslant W_{t+2}(p)$.
- Induction step:

$$W_t(p) = c + \mathbb{E}[V_{t+1}(\varphi(p, Y_t)) \mid p_t = p]$$
$$\leqslant c + \mathbb{E}[V_{t+2}(\varphi(p, Y_{t+1})) \mid p_{t+1} = p] = W_{t+1}(p)$$

and

$$V_t(p) = \min\{W_t(p), L_0(p), L_1(p)\}$$
$$\leqslant \min\{W_{t+1}(p), L_0(p), L_1(p)\} = V_{t+1}(p)$$

Alternate proof The set of strategies increases with horizon. Hence $V_t(p) \leqslant V_{t+1}(p)$. Monotonicity of expectation implies $W_t(p) \leqslant W_{t+1}(p)$.

Qualitative properties of optimal control law

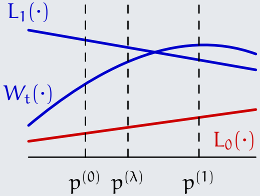
Definition Stopping set $S_t(h) = \{p \in [0,1] : g_t(p) = h\}, \quad h \in \{h_0, h_1\}$.

Theorem For all t and $h \in \{h_0, h_1\}$, the set $S_t(h)$ is convex.

Proof To show that $S_t(h_0)$ is convex, it suffices to show that:

For any $p^{(0)}, p^{(1)} \in S_t(h_0)$ and $\lambda \in [0,1]$,

the information state $p^{(\lambda)} = (1-\lambda)p^{(0)} + \lambda p^{(1)}$ is in $S_t(h_0)$.



- Since $p^{(i)} \in S_t(h_0)$, $i = 0, 1$:

$$L_0(p^{(i)}) \leqslant \min\{L_1(p^{(i)}), W_t(p^{(i)})\}, \quad i = 0, 1.$$

- Since $L_i(p)$ is linear in p, $i = 0, 1$:

$$(1-\lambda)L_i(p^{(0)}) + \lambda L_i(p^{(1)}) \leqslant L_i(p^{(\lambda)}), \quad i = 0, 1.$$

- Since $W_t(p)$ is concave in p:

$$(1-\lambda)W_t(p^{(0)}) + \lambda W_t(p^{(1)}) \leqslant W_t(p^{(\lambda)})$$

Combining the above three, we have

$$L_0(p^{(\lambda)}) \leqslant \min\{L_1(p^{(\lambda)}), W_t(p^{(\lambda)})\}$$

Hence, $p^{(\lambda)} \in S_t(h_0)$. Consequently, $S_t(h_0)$ is convex.

Optimal control law has a threshold property

Assumption (A1) $\ell(h_0, h_0) \leqslant c \leqslant \ell(h_0, h_1)$ and $\ell(h_1, h_1) \leqslant c \leqslant \ell(h_1, h_0)$.

Theorem Under (A1): $0 \in S_t(h_1)$ and $1 \in S_t(h_0)$

Proof $L_0(0) = \ell(h_0, h_1)$, $L_1(0) = \ell(h_1, h_1)$, and $W_t(0) \geqslant c$. Thus

$$L_1(0) \leqslant \min\{L_0(0), W_t(0)\} \implies 0 \in S_t(1).$$

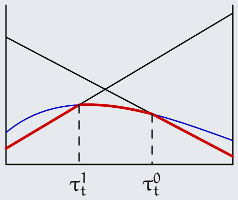
Definition

$$\tau_t^1 = \max\{p \in [0,1] : g_t(p) = h_1\}$$

$$\tau_t^0 = \min\{p \in [0,1] : g_t(p) = h_0\}$$

Threshold property Under (A1), the optimal control law has the following form

$$g_t(p) = \begin{cases} h_1, & \text{if } p \leqslant \tau_t^1 \\ C, & \text{if } \tau_t^1 < p < \tau_t^0 \\ h_0, & \text{if } \tau_t^0 \leqslant p \end{cases}$$



Optimality of sequential likelihood ratio test

Likelihood ratio $\pi_t(0)/\pi_t(1) = p_t/(1-p_t) = \lambda_t$

Likelihood ratio test Under (A1), the optimal control law has the following form

$$g_t(\lambda) = \begin{cases} h_1, & \text{if } \lambda \leqslant \tau_t^1/(1-\tau_t^1) \\ C, & \text{if } \tau_t^1/(1-\tau_t^1) < \lambda < \tau_t^0/(1-\tau_t^0) \\ h_0, & \text{if } \tau_t^0/(1-\tau_t^0) \leqslant \lambda \end{cases}$$

Proof of optimality For $a, b \in [0, 1]$,

$$a \leqslant b \iff \frac{a}{1-a} \leqslant \frac{b}{1-b}.$$

Decision thresholds are monotone in time

Theorem For all t, $\tau^1_t \leqslant \tau^1_{t+1}$ and $\tau^0_t \geqslant \tau^0_{t+1}$

Proof Since $W_t(p)$ is monotone increasing in t: $W_t(\tau^1_t) \leqslant W_{t+1}(\tau^1_t)$. Hence,

$$L_1(\tau^1_t) \leqslant \min\{L_0(\tau^1_t), W_t(\tau^1_t)\} \leqslant \min\{L_0(\tau^1_t), W_{t+1}(\tau^1_t)\}$$

Therefore, $\tau^1_t \in S_{t+1}(h_1)$ which implies $\tau^1_t \leqslant \tau^1_{t+1}$.
By a similar argument, $\tau^0_t \in S_{t+1}(h_0)$ which implies $\tau^0_t \geqslant \tau^0_{t+1}$.

Infinite horizon setup

Model Assume $T \to \infty$ so that the continuation alternative is always available.

Theorem An optimal stopping rule exists, is time-invariant (stationary), and is given by the solution to the following fixed point equation

$$V(p) = \min\{L_0(p), L_1(p), W(p)\}$$

where $W(p) = c + \int_y [p f_0(y) + (1-p) f_1(y)] V(\varphi(p,y)) dy$.

Proof Follows from standard results on non-negative dynamic programming.

Corollary The thresholds τ^1 and τ^0 are time-invariant.

Sequential hypothesis testing: Further Reading

1. For more details on this problem, including an approximate method to determine the thresholds, read: Abraham Wald, "Sequential tests of statistical hypothesis", Annals of Mathematical Statistics, pp. 117-186, 1945.
   http://projecteuclid.org/euclid.aoms/1177731118

2. The model described in these notes was first considered by: Arrow, Blackwell. and Girshick, "Bayes and Minimax Solutions of Sequential Decision Problems", Econometrica, pp. 213-244, Jul.-Oct., 1949.
   http://www.jstor.org/stable/1905525