Markov Decision Processes

Sequential decision-making over time

Aditya Mahajan

McGill University

Lecture Notes for ECSE 506: Stochastic Control and Decision Theory

February 19, 2014

Theory of Markov decision processes

Sequential decision-making over time

MDP functional models

Perfect state observation

MDP probabilistic models

Stochastic orders

MDP Theory: Functional models

Functional model for stochastic dynamical systems

Notation

$X_t \in \mathcal{X}$: State of the system at time t
$Y_t \in \mathcal{Y}$: Observation of controller at time t
$U_t \in \mathcal{U}$: Control action taken by the controller at time t
$W_t \in \mathcal{W}$: Noise in system dynamics at time t
$N_t \in \mathcal{N}$: Observation noise at time t

Assumptions

- The system runs in discrete-time until horizon T.
- The primitive random variables $\{X_1, W_{1:T}, N_{1:T}\}$ are defined over a common probability space $(\Omega, \mathfrak{F}, P)$.
- The primitive variables $\{X_1, W_{1:T}, N_{1:T}\}$ are mutually independent with known probability distribution.

System dynamics

- $X_{t+1} = f_t(X_t, U_t, W_t)$
- The dynamic functions $\{f_t\}_{t=1}^{T}$ are known.

Observations

- $Y_t = h_t(X_t, N_t)$
- The observation functions $\{h_t\}_{t=1}^{T}$ are known.

The control strategy and its performance

Control design

- $U_t = g_t(Y_{1:t}, U_{1:t-1})$
- The control strategy $\mathbf{g} = \{g_t\}_{t=1}^{T}$ is to be determined.
- The controller has classical information structure (i.e., it remembers everything that has been observed and done in the past).

Cost

- Per step-cost at time $t \in \{1, \dots, T-1\}$: $c_t(X_t, U_t)$.
- Terminal cost at time T: $c_T(X_T)$.

Total expected cost

- $J(\mathbf{g}) = \mathbb{E}^{\mathbf{g}} \Big[\sum_{t=1}^{T-1} c_t(X_t, U_t) + c_T(X_T) \Big]$

Alternative formulation: Reward maximization

- In some applications, it is more natural to model per-step and terminal reward functions $r_t(X_t, U_t)$ and $r_T(X_T)$.
- In such applications, the objective is to maximize the total expected reward

$$J(\mathbf{g}) = \mathbb{E}^{\mathbf{g}} \Big[\sum_{t=1}^{T-1} r_t(X_t, U_t) + r_T(X_T) \Big]$$

The problem of optimizing over time

Objective

Given

- The spaces $(\mathcal{X}, \mathcal{Y}, \mathcal{U}, \mathcal{W}, \mathcal{N})$
- Horizon T
- Probability distribution of $\{X_1, W_{1:T}, N_{1:T}\}$
- Dynamics functions $\{f_t\}_{t=1}^{T}$
- Observation functions $\{h_t\}_{t=1}^{T}$
- Cost functions $\{c_t\}_{t=1}^{T}$

Choose

- Control strategy g to minimize the total expected cost $J(\mathbf{g})$.
  (Alternatively, to maximize the total expected reward).

Application domains

- Systems and Control
- Communication
- Power Systems
- Artificial Intelligence
- Operations Research
- Financial Engineering
- Natural Resource Management

Perfect and imperfect observations at the controller

Perfect state observation
Perfect state observation refers to the scenario when $\mathcal{Y} = \mathcal{X}$ and $h_t(X_t, N_t) = X_t$; thus, at each time the controller perfectly observes the state. Such a model is also called Markov decision process (MDP).

Imperfect state observation
Imperfect state observation refers to the general model described above (when $Y_t \neq X_t$). Such a model is also called partially observed Markov decision process (POMDP).

Solution approach
First focus on problems with perfect state observation and identify the structure of optimal controllers and a recursive algorithm, called dynamic programming decomposition, to find an optimal strategy

Then show that an appropriate state expansion converts problems with imperfect state observations to a problem with perfect state observation. Thus, it is possible to reuse the results for models with perfect state observation in models with imperfect state observation.

MDP Theory: Perfect state observation

Structure of optimal strageies

Theorem (Structural result)

A strategy $\mathbf{g} = \{g_t\}_{t=1}^{T}$ is called Markov if it only uses X_t at time t to pick U_t i.e.,

$$U_t = g_t(X_t)$$

Restricting attention to Markovian strategies is without any loss of optimality.

Implication

Let $\mathcal{G}_{1:T}^{H}$ denote the family of all history dependent strategies and $\mathcal{G}_{1:T}^{M}$ denote the family of all Markov strategies. The above theorem asserts that

$$\min_{\mathbf{g} \in \mathcal{G}_{1:T}^{M}} J(\mathbf{g}) = \min_{\mathbf{g} \in \mathcal{G}_{1:T}^{H}} J(\mathbf{g})$$

Note that LHS $\leqslant$ RHS because $\mathcal{G}_{1:T}^{M} \subset G_{1:T}^{H}$. The above theorem is asserting equality.

This result reduces the solution space and thereby simplifies the optimization problem.

When is extra information irrelevant for optimal control?

Blackwell's principle of irrelevant information

Let $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{U}$ be standard Borel spaces and $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be random variables defined on a common probability space $(\Omega, \mathfrak{F}, P)$.

A decision maker observes (X, Y) and chooses U to minimize $\mathbb{E}[c(X, U)]$ where $c \colon \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ is a measurable function.

Then, choosing U just as a function of X is without loss of optimality.

Formally, $\exists g^* \colon \mathcal{X} \to \mathbb{R}$ such that $\forall g \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$

$$\mathbb{E}[c(X, g^*(X))] \leqslant \mathbb{E}[c(X, g(X, Y))]$$

Proof

We prove the result for the case when $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{U}$ are finite valued.

- Define $g^*(x) = \arg\min_{u \in \mathcal{U}} c(x, u)$.
- Then, $\forall x \in \mathcal{X}$ and $\forall u \in \mathcal{U}$: $c(x, g^*(x)) \leqslant c(x, u)$.
- Hence, $\forall g \colon \mathcal{X} \times \mathcal{Y} \to \mathcal{U}$ and $\forall y \in \mathcal{Y}$: $c(x, g^*(x)) \leqslant c(x, g(x, y))$.

The above point-wise inequality implies the inequality in expectation.

How to identiy irrelevant information in dynamic setups?

Two-step Lemma Let $T = 2$. For any control strategy $\mathbf{g} = (g_1, g_2)$ there exists a Markov control law $g_2^*: \mathcal{X} \to \mathcal{U}$ such that $J(g_1, g_2^*) \leqslant J(g_1, g_2)$.

Proof

- Define $J_1(g_1) = \mathbb{E}[c_1(X_1, U_1)]$ and $J_2(g_1, g_2) = \mathbb{E}[c_2(X_2, U_2)]$.
- Then $J(g_1, g_2) = J_1(g_1) + J_2(g_1, g_2)$
- $J_2(g_1, g_2) = \mathbb{E}[c_2(X_2, g_2(X_2, X_1, U_1))]$. By Blackwell's principle of irrelevant information, $\exists g_2^*: X_2 \mapsto U_2$ such that $J_2(g_1, g_2^*) \leqslant J_2(g_1, g_2)$.

Three-step Lemma Let $T = 3$. For any control strategy $\mathbf{g} = (g_1, g_2, g_3)$ such that g_3 is Markov, there exists a Markov control law $g_2^*: \mathcal{X} \to \mathcal{U}$ such that $J(g_1, g_2^*, g_3) \leqslant J(g_1, g_2, g_3)$.

Proof

- Define $J_t(g_{1:t}) = \mathbb{E}[c_t(X_t, U_t)]$. Then $J(g_{1:3}) = J_1(g_1) + J_2(g_{1:2}) + J_3(g_{1:3})$.
- Define $\tilde{c}_3(x, u; g_3) = \mathbb{E}[c_3(X_3, g_3(X_3)) \mid X_2 = x, U_2 = u]$.
- Then, $J_3(g_{1:3}) = \mathbb{E}[\mathbb{E}[c_3(X_3, g_3(X_3)) \mid X_2, U_2]] = \mathbb{E}[\tilde{c}_3(X_2, U_2; g_3)]$.
- Define $\tilde{c}_2(x, u; g_3) = c_2(x, u) + \tilde{c}_3(x, u; g_3)$.
- Then, $J_2(g_{1:2}) + J_3(g_{1:3}) = \mathbb{E}[\tilde{c}_2(X_2, g_2(X_2, X_1, U_1); g_3)]$. Use Blackwell's principle of irrelevant information, as in the two-step lemma.

Backward induction proof of the structural result

To be written

Dyanamic programming decomposition to find optimal Markov strategy

Definition of value functions

Define value functions $\{V_t\}_{t=1}^{T}$, $V_t \colon \mathcal{X} \to \mathbb{R}$ recursively as follows:

$$V_T(x) = c_T(x), \qquad x \in \mathcal{X}$$

and for $t = T-1, T-2, \dots, 1$:

$$V_t(x) = \min_{u \in \mathcal{U}(x)} \underbrace{\mathbb{E}[c(X_t, U_t) + V_{t+1}(X_{t+1}) \mid X_t = x, U_t = u]}_{= Q_t(x,u)}, \quad x \in \mathcal{X}$$

Verification step

A Markov strategy $\{g_t^*\}_{t=1}^{T}$ is optimal iff

$$g_t^*(x) \in \arg\min_{u \in \mathcal{U}(x)} Q_t(x, u), \qquad \forall x \in \mathcal{X} \text{ and } \forall t \in \{1, \dots, T\}$$

The comparison principle to prove dynamic programming

The cost-to-go functions

For any strategy g, define the cost-to-go function at time t as

$$J_t(x; \mathbf{g}) = \mathbb{E}^{\mathbf{g}} \left[\sum_{s=t}^{T-1} c_s(X_s, U_s) + c_T(X_T) \,\middle|\, X_t = x \right]$$

Note that

$$J(\mathbf{g}) = \mathbb{E}[J_1(X_1; \mathbf{g})]$$

The comparison principle

For any Markov strategy $\mathbf{g}$

$$J_t(x; \mathbf{g}) \geqslant V_t(x)$$

with equality at t iff the future strategy $g_{t:T}$ satisfies the verification step.

An immediate consequence of the comparison principle is that the strategy obtained using the dynamic programming decomposition is optimal.

Proof of comparison principle

Proof
- Basis: $J_T(x) = V_T(x)$. Thus, the comparison principle is true.
- Induction hypothesis: Comparison principle is true for $t+1$.
- Induction step:

$$J_t(x;\mathbf{g}) = \mathbb{E}^{\mathbf{g}}\left[\sum_{s=t}^{T} c_s(X_s, U_s) \,\middle|\, X_t = x\right]$$

$$= \mathbb{E}^{\mathbf{g}}\left[c_t(x, g_t(x) + \mathbb{E}^{\mathbf{g}}\left[\sum_{s=t+1}^{T} c_s(X_s, U_s) \,\middle|\, X_{t+1}\right] \,\middle|\, X_t = x\right]$$

$$= \mathbb{E}^{\mathbf{g}}\left[c_t(x, g_t(x) + J_{t+1}(X_{t+1};\mathbf{g}) \,\middle|\, X_t = x\right]$$

By the induction hypothesis

$$\geqslant \mathbb{E}^{\mathbf{g}}\left[c_t(x, g_t(x) + V_{t+1}(X_{t+1}) \,\middle|\, X_t = x, U_t = g_t(x)\right]$$

$$\geqslant V_t(x)$$

with equality iff

- first inequality: $g_{t+1:T}$ satisfies verification step (induction hypothesis)
- second inequality: $g_t \in \arg\min_{u \in \mathcal{U}(x)} Q_t(x, u)$.

MDP Theory: Probabilistic models

To be written

Stochastic orders

Stochastic dominance

Notation

- $\mathcal{X} = \{1, \dots, n\}$ and $\mathcal{Y} = \{1, \dots, m\}$ are finite spaces.
- $\Delta(\mathcal{X})$ is the space of probability measures (PMFs) over $\mathcal{X}$.

Definition (Stochastic dominance)

For any $\pi, \mu \in \Delta(\mathcal{X})$, π stochastically dominates μ (denoted by $\pi \geqslant_s \mu$) if

$$\sum_{i \geqslant k} \pi_i \geqslant \sum_{i \geqslant k} \mu_i, \quad \forall k.$$

Equivalently, if $X_1 \sim \pi$ and $X_2 \sim \mu$, then $\pi \geqslant_s \mu$ iff

$$\mathbb{P}(X_1 \geqslant x) \geqslant \mathbb{P}(X_2 \geqslant x), \quad \forall x \in \mathcal{X}.$$

Example

$$\begin{bmatrix} 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix} \geqslant_s \begin{bmatrix} \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{2} \end{bmatrix} \geqslant_s \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$$

Stochastic dominance preserves monotonicity

Lemma (Stochastic dominance and monotonicity) Let $\{v_i\}_{i=1}^{n}$ be an increasing sequence and $\pi \geqslant_s \mu$. Then,

$$\sum_{i=1}^{n} \pi_i v_i \geqslant \sum_{i=1}^{n} \mu_i v_i$$

Equivalently, if $X_1 \sim \pi$, $X_2 \sim \mu$, and $f : \mathcal{X} \to \mathbb{R}$ is an increasing function, then $\pi \geqslant_s \mu$ implies

$$\mathbb{E}[f(X_1)] \geqslant \mathbb{E}[f(X_2)]$$

Proof To be written. See Hardy, Polya, and Littlewood.

Stochastic monotone Markov chains

Definition Let $\{X_t\}_{t=1}^{\infty}$ be a time-homogeneous Markov chain with transition matrix P. The Markov chain is stochastically monotone if

$$P_i \geqslant_s P_j, \quad \forall i > j$$

where P_i denotes the row-i of P.

Implication If $\{X_t\}_{t=1}^{\infty}$ is stochastically monotone and $f : \mathcal{X} \to \mathbb{R}$ is an increasing function, then

$$\mathbb{E}[f(X_{t+1}) \mid X_t = x_1] \geqslant \mathbb{E}[f(X_{t+1}) \mid X_t = x_2], \quad \forall x_1 > x_2.$$

Monotone likelihood ratio (MLR) ordering

Definition (Monotone likelihood ratio ordering)

For any $\pi, \mu \in \Delta(\mathcal{X})$, π dominates μ in monotone likelihood ratio (denoted by $\pi \geqslant_r \mu$) if

$$\pi_i \mu_j \geqslant \mu_i \pi_j, \quad \forall i > j; \qquad \text{if } \mu_i, \mu_j > 0, \text{ then } \frac{\pi_i}{\mu_i} \geqslant \frac{\pi_j}{\mu_j}$$

Examples

- $\begin{bmatrix} \frac{1}{8} & \frac{1}{8} & \frac{1}{4} & \frac{1}{2} \end{bmatrix} \geqslant_r \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$.
- $\begin{bmatrix} 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix} \not\geqslant_r \begin{bmatrix} \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{2} \end{bmatrix}$.

Monotone likelihood ratio implies stochastic dominance

Proposition For any $\pi, \mu \in \Delta(\mathcal{X})$,

$$\pi \geqslant_{r} \mu \implies \pi \geqslant_{s} \mu$$

Proof Exercise

Total positivity of order 2 (TP_2) and preserving MLR

Definition (Totally positive of order 2)

Recall for any matrix $\mathbf{A}$ and any index sets I and J

- $\mathbf{A}_{I,J}$ denotes the submatrix corresponding to the row set I and the column set J;
- The (I, J) minor of $\mathbf{A}$ is $\det \mathbf{A}_{I,J}$.

A $n \times m$ matrix is totally positive of order 2 (TP_2) if all its 2×2 submatrices have non-negative determinant.

$$\begin{bmatrix} 4 & 3 & 2 & 1 \\ 5 & 4 & 3 & 2 \\ 6 & 5 & 4 & 3 \\ 7 & 6 & 5 & 4 \end{bmatrix} \text{ is } TP_2.$$

Proposition

If Q is a row stochastic matrix that is TP_2, and $\pi, \mu \in \Delta(\mathcal{X})$ then

- $Q_i \geqslant_r Q_j$ for $i > j$. Consequently, $Q_i \geqslant_s Q_j$.
- $\pi \geqslant_r \mu \implies \pi Q \geqslant_r \mu Q$

Proof

- Let $i > j$ and $k > \ell$. Since Q is TP_2, the minor consisting of rows i, j and columns k, ℓ, $i > j$ is non-negative. Thus,

$$\begin{vmatrix} Q_{j\ell} & Q_{jk} \\ Q_{i\ell} & Q_{ik} \end{vmatrix} \geqslant 0 \implies Q_{ik} Q_{j\ell} \geqslant Q_{jk} Q_{i\ell} \implies Q_i \geqslant_r Q_j$$

- See Proposition on next page.

TP_2 ordering of functions and matrices

Definition (TP$_2$ ordering) A function $f \geqslant_{tp} g$ if $\forall x_1, x_2, y_1, y_2$

$$f(x_1 \vee x_2, y_1 \vee y_2) g(x_1 \wedge x_2, y_1 \wedge y_2) \geqslant f(x_1, y_1) g(x_2, y_2),$$

Note that $a \vee b = max(a, b)$ and $a \wedge b = min(a, b)$.

- This definition extends to matrices in a natural manner.
- A matrix Q is TP$_2$ if $Q \geqslant_{tp} Q$.

Proposition If P_1 and P_2 are row stochastic matrices such that $P_1 \geqslant_{tp} P_2$, then

$$\pi \geqslant_r \mu \implies \pi P_1 \geqslant_r \mu P_2$$

In particular,

$$\pi P_1 \geqslant_r \pi P_2, \quad \forall \pi$$

Proof See Theorem 2.4 of Samuel Karlin and Yosef Rinott, "Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions," Journal of Multivariate Analysis, vol 10, no 4 Pages 467-498, Dec 1980. http://dx.doi.org/10.1016/0047-259X(80)90065-2

MDP Theory: Functional models

MDP Theory: Perfect state observation

MDP Theory: Probabilistic models

To be written

Stochastic orders