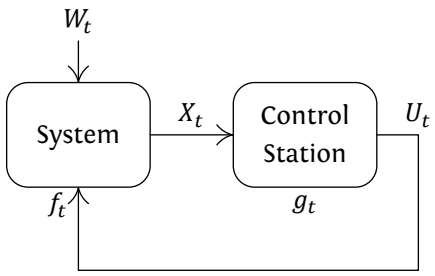
Structural results for MDP: A direct proof

Aditya Mahajan

Markov decision process (MDP) models the simplest stochastic control architecture shown in the figure on the right. The dynamic behavior of the MDP is modeled by an equation of the form

$$X_{t+1} = f_t(X_t, U_t, W_t)$$



where $X_t \in \mathbb{X}_t$ is the state, $U_t \in \mathbb{U}_t$ is the control input, and $W_t \in \mathbb{W}_t$ is noise. A control station observes the state and chooses the control input U_t. This control station can be extremely sophisticated. So, in principle, it can analyze all the past observations and all its past actions to choose the current control input. This behavior of the control station can be modeled by an equation of the form

$$U_t = g_t(X^t, U^{t-1}).$$

(X^t means the sequence $X_1, \dots, X_t$. A similar interpretation holds for U^{t-1}). The function g_t is called the *control law* at time t.

The purpose of the control is to maintain the state of the system close to a desired value. This objective is captured by a cost function of the form $c_t(X_t, U_t)$. The system operates for a time horizon T. During this time it incurs a total cost

$$\sum_{t=1}^{T} c_t(X_t, U_t)$$

The initial state X_1 of the system is random and is chosen by nature according to a known distribution. The noise process $\{W_t, t = 1, \dots, T\}$ is an independent process that is also independent of the initial state X_1.

Suppose we have to design such a control station. We are told the probability distribution of the initial state and the noise. We are also told the system update functions $f_1, \dots, f_T$ and the cost functions $c_1, \dots, c_T$. We are asked to choose a *control strategy* $g := (g_1, \dots, g_T)$ to minimize the expected total cost

$$\mathbb{E}^g \left\{ \sum_{t=1}^{T} c_t(X_t, U_t) \right\}$$

How should we proceed?

At first glance, the problem looks intimidating. It appears that we have to design a very sophisticated controller; one that can analyzes all past data to choose a control input. However, this is not the case. A remarkable result is that even the optimal control station can discard all past data and choose the control input based only on the current state of the system. Formally, we have the following:

Structural Result *For the system model described above, without loss of optimality the control input can be chosen according to*

$$U_t = g_t(X_t), \quad t = 1, \dots, T.$$

Such a control strategy is called a Markov strategy.

The above result claims that the cost incurred by the best Markovian strategy is the same as the cost incurred by a strategy that analyzes the past data in the most sophisticated manner. This appears to be a tall claim, so lets see how we can prove it. The main idea of the proof is an elementary inequality.

An Elementary Inequality *Let $\mathbb{X}$, $\mathbb{Y}$, $\mathbb{Z}$ be finite spaces and $f : \mathbb{X} \times \mathbb{Z} \mapsto \mathbb{R}$. Then, there exists a function $\hat{g} : \mathbb{X} \mapsto \mathbb{Z}$ such that for any function $g : \mathbb{X} \times \mathbb{Y} \mapsto \mathbb{Z}$,*

$$f(x, \hat{g}(x)) \leq f(x, g(x, y)), \quad \forall x \in \mathbb{X}, y \in \mathbb{Y}.$$

Proof The result is trivially true by choosing

$$\hat{g}(x) = \arg\min_{z \in \mathbb{Z}} f(x, z).$$ □

The result can be extended to continuous spaces by invoking appropriate measurability arguments. However, for the ease of exposition, I avoid the *measurability tax*, and only assume the situation when all variables are discrete valued.

The proof of the structural result is almost an immediate consequence of the above elementary inequality. To make the argument trasparent, we prooceed step-by-step.

The Two-Step Lemma *Consider the system described above that operates for two steps ($T = 2$). Without any loss of optimality the control input can be chosen according to*

$$U_2 = g_2(X_2).$$

The Three-Step Lemma *Consider the system described above that operates for three steps ($T = 3$). Assume that the control law at $t = 3$ is Markovian, i.e.,*

$$U_3 = g_3(X_3).$$

Then, without loss of optimality, the control input at time $t = 2$ can be chosen according to

$$U_2 = g_2(X_2).$$

We will prove these lemmas later. First, let us show how these lemmas lead to a proof of the structural result.

Proof of the structural result The main idea is that any system can be thought of as a two- or three-step system by agreegating time. Suppose the system operates for T steps. It can be thought of as a two step system where $t = 1, ..., T-1$ correspond to step 1 and $t = T$ corresponds to step 2. By using the two-step lemma, without loss of optimality we can choose the controller at time T to be Markovian, i.e.,

$$U_T = g_T(X_T).$$

Thus, the structural result is true for $t = T$. Moreover, the structural results are true for $t = 1$ vaccuously. So, we now only need to prove the result for $t = 2, ..., T-1$. We do this by proceeding backwards in time.

The system can be thought of a three step system where $t = 1, ..., T-2$ correspond to step 1, $t = T-1$ corresponds to step 2, and $t = T$ corresponds to step 3. Since the controller at time T is Markovian, the assumption of the three-step lemma is satisfied. Thus, by using that lemma, without loss of optimality, we can choose the controller at time $T-1$ to be Markovian, i.e.,

$$U_{T-1} = g_{T-1}(X_{T-1}).$$

Next, we again think of the system as a three step system but relabel time differently. $t = 1, ..., T-3$ correspond to step 1, $t = T-2$ corresponds to step 2, and $t = T-1, T$ corresponds to step 3. Since the

controllers at time T and $T-1$ are Markovian, the assumption of the three-step lemma is satisfied. Thus, by using that lemma, without loss of optimality, we can choose the controller at time $T-2$ to be Markovian, i.e.,

$$U_{T-2} = g_{T-2}(X_{T-2}).$$

Proceeding this way, we continue to think of the system as a three step system by different relabeling of time. Once we have shown that the controllers at time $t = s+1, s+2, \ldots, T$ are Markovian, we relabel time as follows. $t = 1, \ldots, s-1$ corresponds to step 1, $t = s$ corresponds to step 2, and $t = s+1, s+2, \ldots, T$ corresponds to step 3. Since the controllers at time $s+1, \ldots, T$ are Markovian, the assumption of the three-step lemma is satisfied. Thus, by using that lemma, without loss of optimality, we can choose the controller at time s to be Markovian, i.e.,

$$U_s = g_s(X_s).$$

Proceeding until $s = 2$ completes the proof. □

Now lets complete the proofs of the two lemmas.

Proof of the two-step lemma Fix g_1 and look at optimizing g_2. The total cost is

$$c_1(X_1, U_1) + c_2(X_2, U_2).$$

The choice of g_2 does not influence the first term. So, for a fixed g_1, the total cost is the same as minimizing the expected value of the second term. In the elementary inequality, take $f(\cdot,\cdot) = c(\cdot,\cdot)$ and $\mathbb{X} = \mathbb{X}_2$, $\mathbb{Y} = \mathbb{X}_1 \times \mathbb{U}_1$ and $\mathbb{Z} = \mathbb{U}_2$. Then, there exists a function $\hat{g}_2 : \mathbb{X}_2 \mapsto \mathbb{U}_2$ such that for any (control law) $g_2 : \mathbb{X}_1 \times \mathbb{X}_2 \times \mathbb{U}_1 \mapsto \mathbb{U}_2$

$$c_2(x_2, \hat{g}_2(x_2)) \leq c_2(x_2, g_2(x_1, x_2, u_1)), \quad x_2 \in \mathbb{X}_2, (x_1, u_1) \in \mathbb{X}_1 \times \mathbb{U}_1.$$

Consequently,

$$\mathbb{E}\{c_2(X_2, \hat{g}_2(X_2))\} \leq \mathbb{E}\{c_2(X_2, g_2(X_1, X_2, U_1))\}.$$

This implies that we can pick control input at time 2 according to

$$U_2 = \hat{g}_2(X_2)$$

without any loss. □

Proof of the three-step lemma Fix g_1 and g_3 and look at optimizing g_2. The total cost is

$$c_1(X_1, U_1) + c_2(X_2, U_2) + c_3(X_3, U_3).$$

The choice of g_2 does not affect the first term. So, for a fixed g_1 and g_3, minimizing the total cost is the same as minimizing the expected value of the last two term. Let us look at the expected value of the last term carefully. By the law of iterated expectations, we have

$$\mathbb{E}\{c_3(X_3, U_3)\} = \mathbb{E}\big\{\mathbb{E}\{c_3(X_3, U_3) \,\big|\, X_2, U_2\}\big\}.$$

Now,

$$\begin{aligned}
\mathbb{E}\{c_3(X_3, U_3) \,|\, X_2 = x_2, U_2 = u_2\} &= \mathbb{E}\{c_3(X_3, g_3(X_3)) \,|\, X_2 = x_2, U_2 = u_2)\} \\
&= \sum_{x_3 \in \mathbb{X}} c_3(x_3, g_3(x_3)) \,\text{Prob}\,(X_3 = x_3 \,|\, X_2 = x_2, U_2 = u_2) \\
&= \sum_{x_3 \in \mathbb{X}} c_3(x_3, g_3(x_3)) \,\text{Prob}\,(w_2 \in \mathbb{W}_2 : f_3(x_2, u_2, w_2) = x_3) \\
&=: h_2(x_2, u_2).
\end{aligned}$$

Thus, the total expected cost affected by the choice of g_2 can be written as

$$\begin{aligned}\mathbb{E}\{c_2(X_2, U_2) + c_3(X_3, U_3)\} &= \mathbb{E}\{c_2(X_2, U_2) + h_2(X_2, U_2)\} \\ &=: \mathbb{E}\{\tilde{c}_2(X_2, U_2)\} \qquad (\star)\end{aligned}$$

This cost has the same form as the cost to be minimized in the proof of the two-step lemma. Thus, by a similar argument, we can pick the control input at time 2 according to

$$U_2 = \hat{g}_2(X_2)$$

without any loss. □

Discussion

The above argument is a modification of the proof given in [1]. Contrast this with the standard proof of the structural result like [2, Comparison principle, pg 74]. The proof presented here does not require us to find a dynamic programming decomposition of the problem. The assumption about future control laws is needed in the three-step lemma only to establish something similar to ($\star$), specifically,

$$\begin{aligned}&\mathbb{E}\Big\{\text{“dependent” cost} \,\Big|\, \text{all observations, current control}\Big\} \\ &\quad = \mathbb{E}\Big\{\text{“dependent” cost} \,\Big|\, \text{current “state”, current control}\Big\}.\end{aligned}$$

Then, we can choose

$$\text{current control} = g(\text{current “state”}).$$

This argument can be extended to decentralized systems [3].

Reference

[1] H. S. Witsenhausen, “On the structure of real-time source coders,” *Bell System Technical Journal*, vol. 58, no. 6, pp. 1437-1451, July-August 1979.

[2] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation Identification and Adaptive Control*, Prentice Hall, 1986.

[3] A. Mahajan and S. Tatikonda, “Sequential team form and its simplification using graphical models,” in *Proceedings of the 47th Allerton conference on communication, control and computation*, 2009.