2 Dynamic programming and the value function

We consider the stochastic dynamical system on state space $\mathcal{X}$ and action space $\mathcal{U}$. For the sake of simplicity, we assume that $\mathcal{X}$ and $\mathcal{U}$ are both finite. The system dynamics depends on the state-action pair (x_t, u_t) at current time and a random disturbance W_t:

$$x_{t+1} = f_t(x_t, u_t, W_t), \quad t = 0, 1, \dots \tag{1}$$

We assume that the initial state X_0 and disturbance process $W_0, W_1, \dots$ are mutually independent. The problem of stochastic optimal control over a finite horizon is to minimize the following expected cost of g over horizon T:

$$J_T(g) := \mathbf{E}^g \left[\sum_{t=0}^{T-1} c_t(X_t, U_t) + c_T(X_T) \right] \tag{2}$$

During the last week's lecture, we observed that the optimal policy to the stochastic control problem $\min_g J_T(g)$ is always a Markov policy, that is, it suffices to evaluate (2) on the set of Markov policies:

$$\min_{g=(g_0,\dots,g_{T-1})} \mathbf{E}^g \left[\sum_{t=0}^{T-1} c_t(X_t, g_t(X_t)) + c_T(X_T) \right] \tag{3}$$

where the function $g_t : \mathcal{X} \to \mathcal{U}$ maps the state at time t into the choice of action at time t. Now the question is, how do we find the optimal Markov policy g^* such that $J_T(g^*) = \min_g J_T(g)$?

2.1 Dynamic programming

Dynamic programming, first proposed by R. Bellman in 1950s, is *a recursive algorithm* to find the optimal Markov policy of the optimization problem (3) subject to (1). We start by defining the *value functions*:

Definition 2.1 *The value functions $\{V_t\}_{t=0}^T$ are a sequence of functions $V_t : \mathcal{X} \to \mathbb{R}$, defined by the following recursion:*

1. *At terminal time, $V_T(x) = c_T(x)$, $\forall x$ by definition.*
2. *At time $t = T-1, T-1, \dots, 0$, $Q_t : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$, the Q-function, is defined by:*

$$Q_t(x, u) = c_t(x, u) + \mathbf{E}\big(V_{t+1}(X_{t+1}) | X_t = x, U_t = u\big)$$

3. *At time t, the value function is:*

$$V_t(x) = \min_{u \in \mathcal{U}} Q_t(x, u).$$

After we obtain the value function, the optimal policy is simply given by

$$g_t^*(x) = \arg\min_{u \in \mathcal{U}} Q_t(x, u) \tag{4}$$

The optimality of (4) is illustrated by the following theorem.

Theorem 2.1 (Comparison principle) *Fix a Markov policy* $g = (g_0, g_1, \ldots, g_{T-1})$. *Define* cost-to-go *by*

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

Then for all t, x,

$$J_t(x\,;g) \geq V_t(x) \tag{5}$$

where equality holds iff $(g_t, g_{t+1}, \ldots, g_{T-1})$ *satisfy* $g_s(x) = \arg\min_{u\in\mathcal{U}} Q_s(x,u)$, $\forall s \geq t$, $\forall x$.

Proof: (Backward induction)

- At time T,

$$\begin{aligned} J_T(x\,;g) &= \mathbf{E}^g\big[c_T(X_T)|X_T = x\big] \\ &= c_T(x) = V_T(x) \end{aligned}$$

  Therefore the inequality obviously holds.

- Suppose (5) is true for time $t+1$. Then

$$\begin{aligned} V_t(x) &= \min_{u\in\mathcal{U}} Q_t(x,u) && \text{(6a)}\\ &\leq Q_t(x, g_t(x)) && \text{(6b)}\\ &= c_t(x, g_t(x)) + \mathbf{E}\big(V_{t+1}(X_{t+1})|X_t = x, U_t = g_t(x)\big) && \text{(6c)}\\ &\leq c_t(x, g_t(x)) + \mathbf{E}\big(J_{t+1}(X_{t+1}\,;g)|X_t = x, U_t = g_t(x)\big) && \text{(6d)}\\ &= J_t(x\,;g) && \text{(6e)} \end{aligned}$$

  The inductive assumption is applied on (6d). The equality condition is a direct consequence of (6b).

■

2.1.1 Example: Replacing equipment

Let $x_t \in \mathcal{X} = \{0, 1, \ldots, M\}$ be the age of an equipment, where M is the maximum lifetime of it. The action space is defined by $\mathcal{U} = \{0, 1\}$, and at each time instance, the operator may decide to keep using the machine ($u = 0$) or to replace it ($u = 1$). After x periods of use, the machine has the probability of breakdown $q(x)$. Hence the state transition is defined by:

$$(x, 1) \mapsto 0 \tag{7a}$$

$$(x, 0) \mapsto \begin{cases} x+1, & \text{if no breakdown} \\ 0, & \text{if device breaks down and needs to be replaced} \end{cases} \tag{7b}$$

The system (7) can be modeled as Markov decision process (1) by defining $W = \{W(x)\}_{x=0}^{M}$, where $W(x)$ for $x = 0, 1, \ldots, M$ are independent Bern$(q(x))$, and letting

$$(x, u) \mapsto x' = \begin{cases} 0, & \text{if } u = 1 \\ x + 1, & \text{if } u = 0, W(x) = 0 \\ 0, & \text{if } u = 0, W(x) = 1 \end{cases}$$

or succinctly, $f(x, u, W) = (x + 1)1_{\{u=0\}\times\{W(x)=0\}}$. We assume that $q(0) \leq q(1) \leq \ldots \leq q(M) = 1$.

If it is decided to pro-actively replace the machine, it costs α, but if the replacement is due to the breakdown, it incurs γ amount of cost. If the device does not break, then the machine requires an operating cost $h(x)$ to run. The cost function is thus formulated by:

$$c(x, 1) = \alpha \tag{8a}$$
$$c(x, 0) = q(x)\gamma + (1 - q(x))h(x) \tag{8b}$$

Remark 2.1 *Note that (8b) is in fact the expected value of the cost since the breakdown occurs randomly, which may or may not depend on the primitive random variables.*

It is natural to assume $\gamma > \alpha$ and $h(0) \leq h(1) \leq \ldots h(M - 1)$. We also set $c_T(\cdot) = 0$.

Upon applying the dynamic programming, we have $V_T(x) \equiv 0$ and for $t = T - 1, T - 2, \ldots, 0$,

$$\begin{aligned} Q_t(x, 1) &= c(x, 1) + \mathbf{E}\big(V_{t+1}(X_{t+1})|X_t = x, U_t = 1\big) \\ &= \alpha + V_{t+1}(0) \\ Q_t(x, 0) &= c(x, 0) + \mathbf{E}\big(V_{t+1}(X_{t+1})|X_t = x, U_t = 0\big) \\ &= q(x)\gamma + (1 - q(x))h(x) + q(x)V_{t+1}(0) + (1 - q(x))V_{t+1}(x + 1) \end{aligned}$$

and therefore the value function is recursively obtained by:

$$V_t(x) = \min\big\{\alpha + V_{t+1}(0),\ q(x)\gamma + (1 - q(x))h(x) + q(x)V_{t+1}(0) + (1 - q(x))V_{t+1}(x + 1)\big\} \tag{9}$$

For finite state-action space, we cannot obtain an explicit formula for the value function—which is the downside of this setting—and (9) is as far as we can go. However, we can claim a particularly interesting property of the value function for this example.

Claim: The value function and Q-function of the previous example satisfy the following:

1. For all t, $V_t(x)$ is monotone in x, that is:

$$V_t(0) \leq V_t(1) \leq \ldots \leq V_t(M) \tag{10}$$

2. For each t and $u \in \{0, 1\}$, define $\mathcal{X}_t^u$ by:

$$\begin{aligned} \mathcal{X}_t^0 &:= \{x \in \mathcal{X} : Q_t(x, 0) \leq Q_t(x, 1)\} \\ \mathcal{X}_t^1 &:= \{x \in \mathcal{X} : Q_t(x, 0) > Q_t(x, 1)\} \end{aligned}$$

then $\exists k_t \in \mathcal{X}$ such that $\mathcal{X}_t^1 = \{k_t, k_{t+1}, \ldots, M\}$

Proof: We formulate backward induction to prove the first claim.

- At time T, $V_T(x) = 0$ for all x, thus the claim obviously holds.
- Suppose (10) is true for $t+1$. If $x \in \mathcal{X}_t^1$, $V_t(x) = Q_t(x,1) = \alpha + V_{t+1}(0)$ is constant over x. For $x \in \mathcal{X}_t^0$, we will show that

$$\{x > 0, x \in \mathcal{X}_t^0\} \Rightarrow Q_t(x-1,0) \leq Q_t(x,0)$$

Then conclude the claim via $V_t(x-1) \leq Q_t(x-1,0) \leq Q_t(x,0) = V_t(x)$. Consider

$$\begin{aligned}
Q_t(x,0) - Q_t(x-1,0) &= (1-q(x))h(x) + q(x)\gamma + (1-q(x))V_{t+1}(x+1) + q(x)V_{t+1}(0) \\
&\quad - (1-q(x-1))h(x-1) - q(x-1)\gamma \\
&\quad - (1-q(x-1))V_{t+1}(x) - q(x)V_{t+1}(0) \\
&= (q(x)-q(x-1))(\gamma + V_{t+1}(0)) + (1-q(x))\big(h(x) + V_{t+1}(x+1)\big) \\
&\quad - (1-q(x-1))\big(h(x-1) + V_{t+1}(x)\big) \\
&\geq (q(x)-q(x-1))(\gamma + V_{t+1}(0)) + (1-q(x))\big(h(x) + V_{t+1}(x+1)\big) \\
&\quad - (1-q(x-1))\big(h(x) + V_{t+1}(x+1)\big) \\
&= (q(x)-q(x-1))\big(\gamma + V_{t+1}(0) - h(x) - V_{t+1}(x+1)\big)
\end{aligned}$$

The inequality step is due to the inductive assumption $V_{t+1}(x) \leq V_{t+1}(x+1)$ and the assumption of the problem $h(x-1) \leq h(x)$. Now it suffices to show that if $x \in \mathcal{X}_t^0$ then $\big(\gamma + V_{t+1}(0) - h(x) - V_{t+1}(x+1)\big) \geq 0$. If $x \in \mathcal{X}_t^0$, we have

$$\begin{aligned}
0 &\leq Q_t(x,1) - Q_t(x,0) \\
&= \alpha + V_t(0) - q(x)\gamma - (1-q(x))h(x) - q(x)V_{t+1}(0) - (1-q(x))V_{t+1}(x+1) \\
&\leq (1-q(x))\big(\gamma + V_{t+1}(0) - h(x) - V_{t+1}(x+1)\big)
\end{aligned}$$

■

If they can be established, qualitative properties of the value function, such as monotonicity or convexity, are quite useful for analyzing the system and allow us to apply other methodologies. In the next section, the monotonicity of the value function will be further investigated in a more systematic way.

2.2 Criteria for Monotonicity of Value Functions

The following assumptions are adopted:

1. The system is time invariant
   (a) $X_{t+1} = f(X_t, U_t, W_t)$ where $f_t(\cdot) = f(\cdot)\ \forall\ t$
   (b) The disturbances $W_0, W_1, ...$ are independent and identically distributed (i.i.d.)

An interesting consequence for any Markov policy g and any t is as follows: consider the conditional probability distribution

$$\mathbf{P}^g\left[X_{t+1} = x_{t+1}\middle|X_0^t = x_0^t, U_0^t = u_0^t\right] \tag{11}$$

of the next state given the history of all states and actions. We claim the following:

$$\mathbf{P}^g\left[X_{t+1} = x_{t+1}\middle|X_0^t = x_0^t, U_0^t = u_0^t\right] = \mathbf{P}\left[X_{t+1} = x_{t+1}|X_t = x, U_t = u\right] \tag{12}$$

and so does not depend on g or t. If true, this yet is another manifestation of policy independence of conditional expectations. Specifically, consider a function $h:\ \mathcal{X} \to \mathbb{R}$. Then, by time invariance and by (12),

$$\mathbf{E}^g\left[h(X_{t+1})\middle|X_0^t = x_0^t, U_0^t = u_0^t\right] = \sum_{x' \in \mathcal{X}}\left\{\mathbf{P}\left(X_1 = x'\middle|X_0 = x, U - 0 = u\right)h(x')\right\} \tag{13}$$

which implies that, at each time t, (X_t, U_t) is *sufficient information* for forecasting the future of the controlled process.

2.2.1 Proof of Claim

Reminder, if random variables $Y,\ Z$ are independent, then

$$\mathbf{P}\left(Y \in B|Z \in A\right) = \mathbf{P}\left(Y \in B\right) \tag{14}$$

Proof: We start with the expression

$$\begin{aligned}\mathbf{P}^g\left(X_{t+1} = x_{t+1}\middle|X_0^t = x_0^t, U_0^t = u_0^t\right) &= \mathbf{P}^g\left(f\left(x_t, u_t, W_t\right) = x_{t+1}\middle|X_0^t = x_0^t, U_0^t = u_0^t\right)\\ &= \mathbf{P}^g\left(f\left(x_t, u_t, W_t\right) = x_{t+1}\middle|X_t = x_t, U_t = u\right).\end{aligned} \tag{15}$$

Because our disturbances are i.i.d., we can replace W_t with W_0 and get

$$\mathbf{P}^g\left(f\left(x_t, u_t, W_t\right) = x_{t+1}\middle|X_t = x_t, U_t = u\right) = \mathbf{P}\left(f(x_t, u_t, W_0 = x_{t+1})\right) \tag{16}$$

■

2.3 Markov Chains

Let us fix a state $x \in \mathcal{X}$ and an action $u \in \mathcal{U}$. We define

$$P_u(x, \cdot) \in \mathcal{P}(\mathcal{X}) \tag{17}$$

where $\mathcal{P}$ is the set of all probability distributions and $P_u(x, x') = \mathbf{P}\left(X_{t+1} = x' \middle| X_t = x, U_t = u\right)$.

For clarity, we define $\mathcal{X} := \{1, ..., n\}$, $\mathcal{U} := \{1, ..., m\}$, $P_u \in \mathbb{R}^{n \times n}$, and each element of the following matrix is described as $P_u(i, j) = \mathbf{P}(X_{t+1} = j | X_t = i, U_t = u)$ where the corresponding matrix is

$$P_u = \begin{pmatrix} P_u(1,1) & P_u(1,2) & ... & P_u(1,n) \\ \vdots & \vdots & \vdots & \vdots \\ P_u(n,1) & P_u(n,2) & ... & P_u(n,n) \end{pmatrix} \tag{18}$$

2.3.1 Stochastic Dominance

If X and Y are real-valued random variables, then X stochastically dominates Y ($X \overset{s}{\geq} Y$) if

$$F_X(x) \leq F_Y(x) \ \forall \ x \tag{19}$$

where $F_X(x)$ and is the cumulative distribution function (CDF)

$$F_X(x) = \mathbf{P}(X \leq x) \tag{20}$$

Note that stochastic dominance directly implies, for example, both that $\mathbf{E}[X] \geq \mathbf{E}[Y]$ and $\text{med}(X) \geq \text{med}(Y)$ where $\text{med}(\cdot)$ is the median (50th percentile).

Here, we are using the state space $\mathcal{X} := \{1, 2, ..., n\}$ and X is an $\mathcal{X}$-valued random variable. Recall also that $\mathbf{P}(X \leq x) = \sum_{x' \in \mathcal{X}} P\left(X = x'\right)$. Also note that X Y are random variables on $\mathcal{X}$, where $X \overset{s}{\geq} Y$ if and only if

$$\mathbf{P}(X \leq x) \leq \mathbf{P}(Y \leq x) \ \forall \ x. \tag{21}$$

Let $f : \mathcal{X} \to \mathbb{R}$ be a monotone nondecreasing function $f(1) \leq f(2) \leq ... \leq f(n)$.

Theorem 2.2 *$X \overset{s}{\geq} Y$ if and only if $\mathbf{E}\left[f(X)\right] \geq \mathbf{E}\left[f(Y)\right]$ for all nondecreasing $f : \mathcal{X} \to \mathbb{R}$.*

Proof: First we assume that $X \overset{s}{\geq} Y$ and fix $f : \mathcal{X} \to \mathbb{R}$ where f is a nondecreasing function. We start by taking the expectation of $f(X)$ as

$$\begin{aligned} \mathbf{E}\left[f(X)\right] &= \sum_{i=1}^{n} \left\{f(i) p_X(i)\right\} \\ &= \sum_{i=1}^{n} \left\{f(i)\left[F_X(i) - F_X(i-1)\right]\right\} \qquad (22) \\ &= \sum_{i=1}^{n} \left\{f(i) F_X(i)\right\} + \sum_{i=1}^{n} \left\{\left(f(i-1) - f(i)\right) F_X(i-1)\right\} - \sum_{i=1}^{n} \left\{f(i-1) F_X(i-1)\right\} \end{aligned}$$

where $p_X(i)$ is the probability mass function. We expand this equation as follows

$$\begin{aligned}\mathbf{E}\left[f(X)\right] &= \sum_{i=1}^{n}\Big\{F_X(i-1)\left(f(i-1)-f(i)\right)\Big\}+\sum_{i=1}^{n}\left\{f(i)F_X(i)\right\}-\sum_{i=1}^{n}\left\{f(i-1)FX(i-1)\right\}\\ &= \sum_{i=1}^{n}\left\{F_X(i-1)\underbrace{\left(f(i-1)-f(i)\right)}_{\geq 0}\right\}+f(n)\underbrace{F_X(n)}_{1}\\ &\geq \sum_{i=1}^{n}\Big\{F_Y(i-1)\left(f(i-1)-f(i)\right)\Big\}+f(n)F_Y(n)\\ &= \mathbf{E}\left[f(Y)\right]\end{aligned} \tag{23}$$

■

Note that if $f(X)=1_{\{X\leq i\}}$ then $\mathbf{E}\left[f(X)\right]=F_X(i)$, which proves the reverse implication.

2.3.2 Controlled Markov Chain and Stochastic Dominance

Let us define $\mathcal{X}:=\{1,...,n\}$ where $X,\ Y$ are random variables in $\mathcal{X}$. Also we assume that $X\overset{\text{s}}{\geq}Y$ which is equivalent to $\mathbf{P}\left(X\leq x\right)\leq\mathbf{P}\left(Y\leq x\right)\ \forall\ x$.

Theorem 2.3 *Stochastic dominance of X of Y ($X\overset{\text{s}}{\geq}Y$) holds if and only if $\mathbf{E}\left[f(X)\right]\geq\mathbf{E}\left[f(Y)\right]$ for all nondecreasing $f:\mathcal{X}\to\mathbb{R}$ ($f(1)\leq f(2)\leq....\leq f(n)$)*

So now consider the following Controlled Markov Chain with the above state space as the action space $\mathcal{U}=\{1,..,m\}$ where

$$\begin{aligned}P_u &= \left\{P_u(i,j):\ i,j\in\mathcal{X}\right\},\\ P_u(i,j) &= \mathbf{P}\left(X_{t+1}=j|X_t=i,U_t=u\right),\\ &P(u(i,\cdot)\in\mathcal{P}\left(X\right).\end{aligned}$$

This controlled markov chain is *stochastically monotone* if $\forall\ u\in\mathcal{U}$

$$i<j:\ X_{u,j}\overset{\text{s}}{\geq}X_{u,i}$$

where

$$\begin{aligned}X_{u,i}&\sim P_u(i,\cdot)\\ P_u(i,\cdot)&=\big(P_u(i,1),P_u(i,1),...,P_u(i,n)\big)\\ P_u(j,\cdot)&=\big(P_u(j,1),P_u(j,1),...,P_u(j,n)\big)\end{aligned}$$

Theorem 2.4 *The Markov chain $(P_u)_{u\in\{1,\ldots,m\}}$ is stochastically monotone if and only if*

$$\mathbf{E}\left[f(X_{t+1})\middle|X_t = i, U_t = u\right] \geq \mathbf{E}\left[f(X_{t+1})\middle|X_t = j, U_t = u\right] \tag{24}$$

for all u.

We can also state

Theorem 2.5 *Let $(P_u)_{u\in\{1,..,m\}}$ be given. Suppose the following holds*

1. *(P_u) is stochastically monotone*
2. *For each $u \in \mathcal{U}\ \forall\ t$, $c_t(x,u)$ is nondecreasing in x*
3. *$c(1) \leq c(2) \leq .... \leq c(n)$*

Then, for each t, $V_t(1) \leq V_t(2) \leq ... \leq f_t(n)$

Proof: [backward induction] Let $t = T:\ V_T = c_T$ be monotone by hypothesis. Now we focus on the case where we go from $t+1 \to t$. We fix $u \in \mathcal{U}$ and use the following Q-function

$$Q_t(x,u) = \underbrace{c_t(x,u)}_{\text{monotone by hypothesis}} + \underbrace{\mathbf{E}\left[V_{t+1}(X_{t+1})\middle|X_t = x, U_t = u\right]}_{\text{monotone by induction hypothesis}}. \tag{25}$$

Thus, with x increasing, and $x > x'$

$$V_t(x) = Q_t(x,u^*) \geq Q_t(x',u^*) \geq V_t(x'), \tag{26}$$

where u^* is optimal for x.

■

In closing, note that not all of the assumptions in the preceding theorem hold in the equipment use/replace problem we considered last class. One such condition is stochastic dominance. Take for example the cost,

$$c_t(x,0) = (1-q(x))h(x) + q(x)\gamma. \tag{27}$$

It is easy to see that $x < x'$ does not imply $c_t(x,0) \leq c_t(x',0)$, and yet the value functions are monotone in this case anyway.