8 Feedback and control

We have defined Markov chains as autonomous stochastic systems whose state at each time t is a function of the state at time $t-1$ and a fresh independent random input. The word "autonomous" indicates that this random input cannot be manipulated or even directly observed. In this lecture, we will consider a simple modification of this model that includes an additional external input that can be manipulated, giving us some ability to steer the state in a desired direction. In particular, by introducing a feedback connection from the state to this external input, we may *control* the state of the Markov chain.

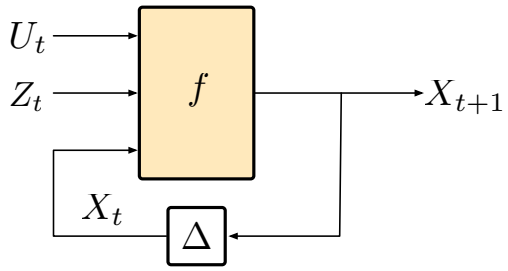

Figure 1: A controlled Markov chain.

Consider the set-up shown in Figure 1. It depicts a stochastic system whose state X_t evolves according to the model

$$X_{t+1} = f(X_t, U_t, Z_t), \tag{8.1}$$

where, as before, X_0 is the initial state, $U_0, U_1, \ldots$ is a sequence of i.i.d. random variables independent of X_0, and Z_t is an additional input at time t. This input, which takes values in some set Z, may be deterministic or stochastic, where in the latter case it can only depend on $X_0^t = (X_0, \ldots, X_t)$ and $Z_0^{t-1} = (Z_0, \ldots, Z_{t-1})$. Any stochastic system described by Eq. (8.1) is called a *controlled Markov chain*, and we refer to Z_t as the *control* or *action* at time t, and we call Z the *control space* or *action space*. To explain the term "controlled Markov chain," let us first consider the simplest case: $Z_t = z$, where z is a fixed deterministic element of the action space Z. That is, the state evolves according to the rule

$$X_{t+1} = f(X_t, U_t, z). \tag{8.2}$$

It is easy to see that this is a time-homogeneous Markov chain. Indeed, fix any t and consider the joint distribution of X_0^t (we will assume for simplicity that the state space X is discrete):

$$\begin{aligned}\mathbf{P}[X_0^t = x_0^t] &= \mathbf{P}[X_0 = x_0, f(x_0, U_0, z) = x_1, \ldots, f(x_{t-1}, U_{t-1}, z) = x_t] \\ &= \mathbf{P}[X_0 = x_0] \prod_{s=1}^{t} \mathbf{P}[f(x_{s-1}, U_{s-1}, z) = x_s],\end{aligned}$$

where we have used the fact that $X_0, U_0, \ldots, U_{t-1}$ are independent. Now, for any pair $x, y \in \mathsf{X}$ define the quantity $M_z(x,y) \triangleq \mathbf{P}[f(x, U_0, z) = y]$. Then we can write

$$\mathbf{P}[X_0^t = x_0^t] = \mathbf{P}[X_0 = x_0] \prod_{s=1}^{t} M_z(x_{s-1}, x_s),$$

which shows that $X = (X_t)_{t \in \mathbb{Z}_+}$ is indeed a time-homogeneous Markov chain with transition probabilities $M_z(x,y)$. Note, however, that the transition probabilities are a function of (or are *controlled by*) the external input z! In other words, we have a separate transition probability matrix $M_z = (M_z(x,y))_{x,y \in \mathsf{X}}$ for each available control action $z \in \mathsf{Z}$. By the same token, we can consider a time-varying deterministic sequence $z = (z_t)_{t \in \mathbb{Z}_+}$ and let

$$X_{t+1} = f(X_t, U_t, z_t). \tag{8.3}$$

In this case, X is a *time-inhomogeneous* Markov chain with $\mathbf{P}[X_{t+1} = y | X_t = x] = M_{z_t}(x,y)$:

$$\mathbf{P}[X_0^t = x_0^t] = \mathbf{P}[X_0 = x_0] \prod_{s=1}^{t} M_{z_{s-1}}(x_{s-1}, x_s). \tag{8.4}$$

That is, we can use a time-varying control signal $z = (z_t)_{t \in \mathsf{Z}}$ to choose a different transition probability matrix at every time step t.

The models described by Eqs. (8.2) and (8.3) demonstrate the utility of having an external control input: we can use it to *manipulate* the probabilities of state transitions. However, these models are *open-loop*: all the control inputs are fixed to some predetermined values ahead of time. A more flexible architecture is the one where close the loop and allow the control input at time t to depend on the entire state sequence X_0^t up to that time and on the past control inputs Z_0^{t-1}. Thus, we may consider the following model:

$$X_{t+1} = f_t(X_t, U_t, Z_t) \tag{8.5a}$$

$$Z_t = g_t(X_0^t, Z_0^{t-1}), \tag{8.5b}$$

where Eq. (8.5a) is the state update, and the sequence of functions $g = (g_t)_{t \in \mathbb{Z}_+}$ in (8.5b) is called the *feedback control strategy*. The strategy g is something that can be designed with particular goals in mind, and we will investigate this in detail shortly. The joint probability distribution of states X_0^t and actions Z_0^t at any time t depends on g, and we will indicate this expicitly by writing $\mathbf{P}^g[\cdot]$. From (8.5), we have

$$\begin{aligned}\mathbf{P}^g[X_0^t = x_0^t, Z_0^t = z_0^t] &= \mathbf{P}[X_0 = x_0, f(x_0, U_0, z_0) = x_1, \ldots, f(x_{t-1}, U_{t-1}, z_{t-1}) = x_t] \\ &\qquad \cdot 1\{z_0 = g_0(x_0), z_1 = g_1(x_0^1, z_0), \ldots, z_t = g_t(x_0^t, z_0^{t-1})\} \\ &= \mathbf{P}[X_0 = x_0] \prod_{s=1}^{t} M_{z_{s-1}}(x_{s-1}, x_s) \prod_{s=0}^{t} 1\{z_s = g_s(x_0^s, z_0^{s-1})\},\end{aligned} \tag{8.6}$$

where $1\{\cdot\}$ takes the value 1 if the statement in the curly braces is true and 0 otherwise. From Eq. (8.6), it is evident that, for a general g, the state signal $X = (X_t)_{t \in \mathbb{Z}_+}$ may not be a Markov chain

because the action z_t that determines the state transition probabilities at time t may depend on all states up to time t and all actions up to time $t-1$. However, the state signal will be a Markov chain if the control action at time t depends only on the state at time t, i.e., if $Z_t = g_t(X_t)$ for each t. Indeed, in that case we can deduce the following from Eq. (8.6):

$$\mathbf{P}^g[X_0^t = x_0^t] = \mathbf{P}[X_0 = x_0] \prod_{s=1}^{t} M_{g_{s-1}(x_{s-1})}(x_{s-1}, x_s), \tag{8.7}$$

which is indeed a time-inhomogeneous Markov chain. We refer to any control strategy of the form $Z_t = g_t(X_t)$ as a *Markov strategy*. Compared to the open-loop Markov model of (8.4), where all the control actions are fixed ahead of time, the closed-loop Markov model of (8.7) allows for the manipulation of transition probabilties via state feedback. Now we are ready to formulate the optimal control problem.

8.1 Finite-horizon optimal control problems

To motivate the discussion, consider the following concrete realization of the model in Eq. (8.1). We take $\mathsf{X} = \mathsf{U} = \mathsf{Z} = \{0, 1\}$, and consider the update rule of the form

$$f(x, u, z) \stackrel{\triangle}{=} \begin{cases} x \oplus u, & \text{if } z = 0 \\ x \oplus 1, & \text{if } z = 1. \end{cases} \tag{8.8}$$

Here, $\oplus$ denotes the Boolean XOR operation. In other words, if the control input z is set to 0, then f outputs the XOR of the current state x with the disturbance input u; if z is set to 1, then f flips the current state. We assume that the initial state X_0 is a Bern($\frac{1}{2}$) random variable, and that the disturbance inputs $U_0, U_1, \ldots$ are i.i.d. Bern(p) random variables that are also independent of X_0. Thus, the two possible transition probability matrices corresponding to the control inputs $z = 0$ and $z = 1$ are

$$M_0 = \begin{pmatrix} M_0(0,0) & M_0(0,1) \\ M_0(1,0) & M_0(1,1) \end{pmatrix} = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix} \quad \text{and} \quad M_1 = \begin{pmatrix} M_1(0,0) & M_1(0,1) \\ M_1(1,0) & M_1(1,1) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{8.9}$$

Now suppose that we want the first four states X_0, X_1, X_2, X_3 to track the pattern $0, 1, 0, 1$ by issuing an appropriate sequence of control actions Z_0, Z_1, Z_2, Z_3. We incur a unit cost every time the actual state differs from the prescribed value (i.e., when $X_0 = 1$, $X_2 = 0$, $X_3 = 1$, or $X_4 = 0$) and also every time the control action is set to 1. Otherwise, we incur no cost. Thus, if we define four state-action cost functions

$$c_0(x_0, z_0) = 1\{x_0 = 1\} + 1\{z_0 = 1\} \tag{8.10a}$$
$$c_1(x_1, z_1) = 1\{x_1 = 0\} + 1\{z_1 = 1\} \tag{8.10b}$$
$$c_2(x_2, z_2) = 1\{x_2 = 1\} + 1\{z_2 = 1\} \tag{8.10c}$$
$$c_3(x_3, z_3) = 1\{x_3 = 0\} + 1\{z_3 = 1\}. \tag{8.10d}$$

For any control strategy $g = (g_0, g_1, g_2, g_3)$, we can compute its *expected cost*

$$J(g) = J(g_0, g_1, g_2, g_3) = \mathbf{E}^g \left[\sum_{t=0}^{3} c_t(X_t, Z_t) \right],$$

where the control input at time t is given by $Z_t = g_t(X_0^t, Z_0^{t-1})$. The *optimal control problem* is to find a control strategy $g^* = (g_0^*, g_1^*, g_2^*, g_3^*)$, such that

$$J(g^*) = \min_g J(g),$$

where the minimum on the right-hand side is over all valid control strategies $g = (g_0, g_1, g_2, g_3)$, i.e., those under which the control action Z_t at time t is a function only of X_0^t and Z_0^{t-1}. We call the integer $T = 4$ the *horizon* of the problem, and we say that any control strategy $g = (g_0, \ldots, g_3)$ is a 4-step strategy.

This is an instance of a *finite-horizon optimal control problem*: Given a controlled Markov chain with controlled transition probabilities $M_z(x, y)$, a positive integer T, and a sequence of *state-action cost functions* $c_t : \mathsf{X} \times \mathsf{Z} \to \mathbb{R}$, for $t \in \{0, 1, \ldots, T-1\}$, we seek a T-step control strategy $g^* = (g_0^*, \ldots, g_T^*)$, such that

$$J(g^*) = J(g_0^*, \ldots, g_{T-1}^*) = \min_g J(g), \tag{8.11}$$

where the expected cost of any T-step strategy g is defined as

$$J(g) = \mathbf{E}^g \left[\sum_{t=0}^{T-1} c_t(X_t, Z_t) \right],$$

with $Z_t = g_t(X_0^t, Z_0^{t-1})$ for every t.

On the face of it, finding such a g^* is a messy affair, especially when the horizon T is large — at each time t, the control action Z_t may depend on the entire state history X_0^t as well as on all of past actions Z_0^{t-1}. The complexity of finding the best g_t quickly gets out of hand already in the binary case, when $\mathsf{X} = \mathsf{Z} = \{0, 1\}$. Indeed, at time t, there are two options for the control action for each possible realization of (X_0^t, Z_0^{t-1}), and there are $2^{t+1} \times 2^t = 2^{2t+1}$ possible realizations. Thus, the number of possibilities grows exponentially with time!

Fortunately, as we will see next, because the cost at each time t depends only on the current state X_t and the current action Z_t, and because the state update rule is Markov, there is no loss of optimality if we restrict our consideration only to Markov control strategies. That is, for any candidate control strategy g, we can always find a Markov control strategy g^+ that will perform at least as well, in the sense that $J(g^+) \leq J(g)$. This entails huge savings in complexity, both when finding the optimal strategy and when implementing. Moreover, armed with this result, we will prove that the optimal strategy can be constructed via an explicit recursive scheme called *dynamic programming*.

8.2 Blackwell's principle of irrelevant information and optimality of Markov strategies

The proof of optimality of Markov strategies is a deep consequence of the so-called *principle of irrelevant information*, which was derived in 1964 by David Blackwell.[1] While the result is very general and applies to continuous state and action spaces, it is very simple to state in the finite case:

> Let X and Y be two jointly distributed random variables taking values in finite spaces X and Y. Let Z be a finite action space. Consider a real-valued cost function $c(x,z)$ of $x \in \mathsf{X}$ and $z \in \mathsf{Z}$. Then, for any strategy $Z = g(X,Y)$ we can always find a strategy $Z = g^*(X)$ that uses only X, such that
>
> $$\mathbf{E}[c(X, g^*(X))] \leq \mathbf{E}[c(X, g(X,Y))].$$

The construction of g^* is painfully obvious: just take

$$g^*(x) = \arg\min_{z \in \mathsf{Z}} c(x,z).$$

Then $c(x, g(x,y)) \geq c(x, g^*(x))$ by definition, for any pair x, y, and we are done. (Blackwell's result applies more generally when the sets X, Y, and Z are continuous, and when the strategies are randomized, but the construction of g^* in such cases is much less straightforward.) The interpretation here is as follows: we get observe a state variable X and some additional information Y correlated with X, and then get to choose an action Z on the basis of X and Y. However, because the cost $c(X,Z)$ directly depends only on the state X and the action Z, the additional piece of information Y is irrelevant, and we can always do without it.

We will now use Blackwell's principe to prove optimality of Markov strategies for an arbitrary horizon T. We will first prove this optimality statement for $T = 1, 2, 3$ and then derive the general case. For $T = 1$, there is nothing to prove since $Z_0 = g_0(X_0)$, and the strategy is already Markov. For $T = 2$, consider an arbitrary strategy $g = (g_0, g_1)$ with $Z_0 = g_0(X_0)$ and $Z_1 = g_1(X_0, X_1, Z_0)$. Then the expected cost is

$$\begin{aligned} J(g_0, g_1) &= \mathbf{E}^g[c_0(X_0, Z_0) + c_1(X_1, Z_1)] \\ &= \mathbf{E}^g[c_0(X_0, Z_0)] + \mathbf{E}^g[c_1(X_1, Z_1)]. \end{aligned} \tag{8.12}$$

We will show that we can replace g_1 with another function g_1^* that depends only on X_1 and achieves smaller expected cost: $J(g_0, g_1^*) \leq J(g_0, g_1)$. Since the first term in (8.12) depends only on g_0, we focus on the second term

$$\mathbf{E}^g[c_1(X_1, Z_1)] = \mathbf{E}[c_1(X_1, g_1(X_0, X_1, Z_0))].$$

Applying Blackwell's principle to $X = X_1$, $Y = (X_0, Z_0)$, and $c = c_1$, we see that we can replace $g_1(X_0, X_1, Z_0)$ with $g_1^*(X_1) = \arg\min_{z \in \mathsf{Z}} c_1(X_1, z)$ to get

$$\mathbf{E}[c_1(X_1, g_1^*(X_1))] \leq \mathbf{E}[c_1(X_1, g_1(X_0, X_1, g_0(X_0)))], \tag{8.13}$$

[1] D.P. Blackwell, "Memoryless strategies in finite-stage dynamic programming," *Annals of Mathematical Statistics*, vol. 35, no. 2, pp. 863–865, 1964. The term "Blackwell's principle of irrelevant information" was coined by Peter Whittle.

and therefore $J(g_0, g_1^*) \le J(g_0, g_1)$, as claimed. Now we consider the case of $T = 3$. Consider a strategy $g = (g_0, g_1, g_2)$, where g_2 is Markov, i.e., $Z_2 = g_2(X_2)$ and g_1 is arbitrary, i.e., $Z_1 = g_1(X_0, X_1, Z_0)$. We will show that we can replace g_1 with a Markov strategy g_1^* and guarantee smaller expected cost: $J(g_0, g_1^*, g_2) \le J(g_0, g_1, g_2)$. The expected cost of g is

$$\begin{aligned} J(g_0, g_1, g_2) &= \mathbf{E}^g[c_0(X_0, Z_0) + c_1(X_1, Z_1) + c_2(X_2, Z_2)] \\ &= \mathbf{E}^g[c_0(X_0, Z_0)] + \mathbf{E}^g[c_1(X_1, Z_1) + c_2(X_2, Z_2)]. \end{aligned} \tag{8.14}$$

Again, the first term in (8.14) involves only g_0, so we can focus on the second and on the third terms only. The second term involves g_1, but, since $X_2 = f(X_1, U_1, Z_1)$ and Z_1 is determined by g_1, the choice of g_1 affects both the second and the third terms. Let us look at these terms more closely:

$$\begin{aligned} &\mathbf{E}^g[c_1(X_1, Z_1) + c_2(X_2, Z_2)] \\ &\quad = \sum_{x_0^2, z_0^2} \mathbf{P}^g[X_0^2 = x_0^2, Z_0^2 = z_0^2]\{c_1(x_1, z_1) + c_2(x_2, z_2)\} \\ &\quad = \sum_{x^2, z_0^2} \mathbf{P}[X_0 = x_0] M_{z_0}(x_0, x_1) M_{z_1}(x_1, x_2) \\ &\qquad\qquad 1\{z_0 = g_0(x_0)\} 1\{z_1 = g_1(x_0^1, z_0)\} 1\{z_2 = g_2(x_2)\} \cdot \{c_1(x_1, z_1) + c_2(x_2, z_2)\} \\ &\quad = \sum_{x_0^1, z_0^1} \mathbf{P}[X_0 = x_0] M_{z_0}(x_0, x_1) 1\{z_0 = g_0(x_0)\} 1\{z_1 = g_1(x_0^1, z_0)\} \\ &\qquad\qquad \cdot \sum_{x_2, z_2} M_{z_1}(x_1, x_2) 1\{z_2 = g_2(x_2)\} \cdot \{c_1(x_1, z_1) + c_2(x_2, z_2)\}. \end{aligned} \tag{8.15}$$

Now, if we take a close look at the second summation in (8.15), we see that it is a function of x_1 and z_1 only, since x_2 and z_2 are marginalized away. Moreover, it does not depend on g_1. With that in mind, if we define

$$\begin{aligned} \tilde{c}(x_1, z_1) &\triangleq \sum_{x_2, z_2} M_{z_1}(x_1, x_2) 1\{z_2 = g_2(x_2)\} \cdot \{c_1(x_1, z_1) + c_2(x_2, z_2)\} \\ &\equiv \sum_{x_2} M_{z_1}(x_1, x_2) \{c_1(x_1, z_1) + c_2(x_2, g_2(x_2))\}, \end{aligned}$$

then we can write

$$\mathbf{E}^g[c_1(X_1, Z_1) + c_2(X_2, Z_2)] = \mathbf{E}[\tilde{c}(X_1, g_1(X_0, X_1, g_0(X_0)))].$$

Applying the Blackwell principle to $X = X_1$, $Y = X_0$, and $c = \tilde{c}$, we see that we can replace $g_1(X_0, X_1, Z_0)$ with $g_1^*(X_1) = \arg\min_{z \in \mathsf{Z}} \tilde{c}(X_1, z)$, such that the new strategy $g^* = (g_0, g_1^*, g_2)$ satisfies

$$\mathbf{E}^{g^*}[c_1(X_1, Z_1) + c_2(X_2, Z_2)] \le \mathbf{E}^g[c_1(X_1, Z_1) + c_2(X_2, Z_2)].$$

This shows that $J(g_0, g_1^*, g_2) \le J(g_0, g_1, g_2)$, as claimed.

With these two results in hand, we can prove optimality of Markov strategies for finite-horizon optimal control problems. Given the horizon $T > 3$ and a sequence of one-stage state-action costs $c_0, \ldots, c_{T-1}$, consider an arbitrary control strategy $g = (g_0, \ldots, g_{T-1})$ with the expected cost

$$J(g) = J(g_0, \ldots, g_{T-1}) = \mathbf{E}^g \left[\sum_{t=0}^{T-1} c_t(X_t, Z_t) \right].$$

We will now prove that there exists a Markov control strategy $g^* = (g_0^*, \ldots, g_{T-1}^*)$, such that $J(g^*) \leq J(g)$. Since g is arbitrary, this implies that we can optimize only over Markov policies, which yields tremendous savings in implementation complexity — at each time t, we only need to feed back the most recent state X_t, and there is no need to store the entire history X_0^t, Z_0^{t-1}.

Now for the proof. Let us rewrite the expected cost $J(g)$ as

$$J(g) = \mathbf{E}^g \left[\sum_{t=0}^{T-2} c_t(X_t, Z_t) \right] + \mathbf{E}^g [c_{T-1}(X_{T-1}, Z_{T-1})].$$

Here, the first term is not affected by the choice of g_{T-1}, while the second term will be affected by the entire g. If we think of the state tuple X_0^{T-2} as a "superstate" $\tilde{X}_0$ and of the action tuple Z_0^{T-2} as a "superaction" $\tilde{Z}_0$ and let $\tilde{X}_1 = X_{T-1}$, $\tilde{Z}_1 = Z_{T-1}$, then, by the previously proved result for $T = 2$, we can replace $g_{T-1}(X_0^{T-1}, Z_0^{T-2}) \equiv g_{T-1}(\tilde{X}_0, \tilde{X}_1, \tilde{Z}_0)$ with $g_{T-1}^*(\tilde{X}_1) \equiv g_{T-1}^*(X_{T-1})$, while guaranteeing that $J(g_0, \ldots, g_{T-2}, g_{T-1}^*) \leq J(g_0, \ldots, g_{T-2}, g_{T-1})$. Thus, if we assume that the last action Z_{T-1} depends only on the state X_{T-1}, we can only reduce the expected cost. With that in mind, consider an arbitrary strategy $g = (g_0, \ldots, g_{T-2}, g_{T-1})$, where g_{T-1} is a function acting on X_{T-1} only, while all others are not restricted in such a way. Now let us rewrite the expected cost $J(g)$ as

$$J(g) = \mathbf{E}^g \left[\sum_{t=0}^{T-3} c_t(X_t, Z_t) \right] + \mathbf{E}^g [c_{T-2}(X_{T-2}, Z_{T-2})] + \mathbf{E}^g [c_{T-1}(X_{T-1}, Z_{T-1})].$$

The first term is not affected by the choice of g_{T-2} and g_{T-1}, while the second and the third terms are. Thus, we can lump the state tuple X_0^{T-3} into a "superstate" $\tilde{X}_0$, the action tuple Z_0^{T-3} into a "superaction" $\tilde{Z}_0$, and take $\tilde{X}_1 = X_{T-2}$, $\tilde{Z}_1 = Z_{T-2}$, $\tilde{X}_2 = X_{T-1}$, and $\tilde{Z}_2 = Z_{T-1}$. Since $Z_{T-1} = g_{T-1}(X_{T-1})$, we see that $\tilde{Z}_2$ is a function of $\tilde{X}_2$ only, so we can apply the previously proved result for the $T = 3$ case and replace $g_{T-2}(X_0^{T-2}, Z_0^{T-3}) = g_{T-2}(\tilde{X}_0^1, \tilde{Z}_0)$ with $g_{T-2}^*(\tilde{X}_1) \equiv g_{T-2}^*(X_0^{T-3})$ with $J(g_0, \ldots, g_{T-2}^*, g_{T-1}) \leq J(g_0, \ldots, g_{T-2}, g_{T-1})$. Repeating these two operations until we reach the beginning ($T = 1$), we can replace the entire strategy g by a Markov strategy g^*, while reducing the expected cost.

Thus, we have proved a very important result: When looking for an optimal control strategy for a finite-horizon control problem of the type (8.11), we can limit the search to Markov strategies only. However, it doesn't tell us how to find such an optimal strategy. We take this up next.

8.3 Dynamic programming

Now that we have proved that there is no loss of optimality in restricting the optimization in (8.11) to Markov strategies, we will use this result to derive a general recursive procedure for actually

finding the optimal strategy. This procedure, which goes by the name of *dynamic programming*, works by breaking the multistage optimal control problem into smaller one-stage control problems. Dynamic programming was invented in the 1940's and then refined in the 1950's by Richard Bellman.[2]

We start by breaking the problem into the first $T-1$ stages and the final stage:

$$\begin{aligned}\min_g J(g) &= \min_{g_0^{T-2}} \min_{g_{T-1}} J(g_0^{T-2}, g_{T-1}) \\ &= \min_{g_0^{T-2}} \min_{g_{T-1}} \mathbf{E}\left[\sum_{t=0}^{T-1} c_t(X_t, g_t(X_t))\right] \\ &= \min_{g_0^{T-2}} \min_{g_{T-1}} \left\{\mathbf{E}\left[\sum_{t=0}^{T-2} c_t(X_t, g_t(X_t))\right] + \mathbf{E}\left[c_{T-1}(X_{T-1}, g_{T-1}(X_{T-1}))\right]\right\}. \end{aligned} \tag{8.16}$$

Now notice the following: if we fix $g_0, \ldots, g_{T-2}$ and vary g_{T-1}, then this will affect only the second expectation in (8.16). That is, we can move the minimization over g_{T-1} to that second term:

$$\begin{aligned}&\min_{g_0^{T-2}} \min_{g_{T-1}} \left\{\mathbf{E}\left[\sum_{t=0}^{T-2} c_t(X_t, g_t(X_t))\right] + \mathbf{E}\left[c_{T-1}(X_{T-1}, g_{T-1}(X_{T-1}))\right]\right\} \\ &\qquad = \min_{g_0^{T-2}} \left\{\mathbf{E}\left[\sum_{t=0}^{T-2} c_t(X_t, g_t(X_t))\right] + \min_{g_{T-1}} \mathbf{E}\left[c_{T-1}(X_{T-1}, g_{T-1}(X_{T-1}))\right]\right\}.\end{aligned}$$

Now, if we consider

$$g_{T-1}^*(X_{T-1}) = \arg\min_{z \in Z} c(X_{T-1}, z),$$

[2]The origin of the name "dynamic programming" is very curious. Here is how Bellman recalls it in his 1984 autobiography *Eye of the Hurricane*:

> "I spent the Fall quarter (of 1950) at RAND. My first task was to find a name for multistage decision processes. An interesting question is, Where did the name, dynamic programming, come from? The 1950s were not good years for mathematical research. We had a very interesting gentleman in Washington named Wilson. He was Secretary of Defense, and he actually had a pathological fear and hatred of the word research. I'm not using the term lightly; I'm using it precisely. His face would suffuse, he would turn red, and he would get violent if people used the term research in his presence. You can imagine how he felt, then, about the term mathematical. The RAND Corporation was employed by the Air Force, and the Air Force had Wilson as its boss, essentially. Hence, I felt I had to do something to shield Wilson and the Air Force from the fact that I was really doing mathematics inside the RAND Corporation. What title, what name, could I choose? In the first place I was interested in planning, in decision making, in thinking. But planning, is not a good word for various reasons. I decided therefore to use the word "programming". I wanted to get across the idea that this was dynamic, this was multistage, this was time-varying. I thought, let's kill two birds with one stone. Let's take a word that has an absolutely precise meaning, namely dynamic, in the classical physical sense. It also has a very interesting property as an adjective, and that it's impossible to use the word dynamic in a pejorative sense. Try thinking of some combination that will possibly give it a pejorative meaning. It's impossible. Thus, I thought dynamic programming was a good name. It was something not even a Congressman could object to. So I used it as an umbrella for my activities."

then, for any g_{T-1},

$$\mathbf{E}\left[c_{T-1}(X_{T-1}, g_{T-1}(X_{T-1}))\right] \geq \mathbf{E}\left[\min_{z\in Z} c_{T-1}(X_{t-1}, z)\right] = \mathbf{E}[c_{T-1}(X_{T-1}, g^*_{T-1}(X_{T-1}))].$$

Thus, regardless of how we got to the last stage T, the optimal action at that stage is given by $g^*_{T-1}(X_{T-1})$, and therefore

$$\begin{aligned}\min_g J(g) &= \min_{g_0^{T-2}} J(g_0^{T-2}, g^*_{T-1}) \\ &= \min_{g_0^{T-2}} \mathbf{E}\left\{\sum_{t=0}^{T-2} c_t(X_t, g_t(X_t)) + \min_{z\in Z} c_{T-1}(X_{T-1}, z)\right\}.\end{aligned}$$

Now we need to optimize g_{T-2}. Observe, however, that we cannot simply take

$$g^*_{T-2}(X_{T-2}) = \arg\min_{z\in Z} c_{T-2}(X_{T-2}, z),$$

because the action Z_{T-2} will affect the future state via the update rule $X_{T-1} = f(X_{T-2}, U_{T-2}, Z_{T-2})$. Thus, we need to choose g^*_{T-2} to balance the current expected cost $\mathbf{E}[c_{T-2}(X_{T-2}, Z_{T-2})]$ and the future expected cost $\mathbf{E}[c_{T-1}(X_{T-1}, Z_{T-1})]$. To that end, let us split off the $t = T-2$ term and write

$$\begin{aligned}&\min_{g_0^{T-2}} J(g_0^{T-2}, g^*_{T-1}) \\ &= \min_{g_0^{T-2}} \mathbf{E}\left\{\sum_{t=0}^{T-2} c_t(X_t, g_t(X_t)) + \min_{z\in Z} c_{T-1}(X_{T-1}, z)\right\} \\ &= \min_{g_0^{T-2}} \mathbf{E}\left\{\sum_{t=0}^{T-3} c_t(X_t, g_t(X_t)) + c_{T-2}(X_{T-2}, g_{T-2}(X_{T-2})) + \min_{z\in Z} c_{T-1}(X_{T-1}, z)\right\} \\ &= \min_{g_0^{T-2}} \mathbf{E}\left\{\sum_{t=0}^{T-3} c_t(X_t, g_t(X_t)) + c_{T-2}(X_{T-2}, g_{T-2}(X_{T-2})) + \min_{z\in Z} c_{T-1}(f(X_{T-2}, U_{T-2}, g_{T-2}(X_{T-2})), z)\right\} \\ &= \min_{g_0^{T-3}} \left\{\mathbf{E}\left[\sum_{t=0}^{T-3} c_t(X_t, g_t(X_t))\right] + \min_{g_{T-2}} \mathbf{E}\left[c_{T-2}(X_{T-2}, g_{T-2}(X_{T-2})) + \min_{z\in Z} c_{T-1}(f(X_{T-2}, U_{T-2}, g_{T-2}(X_{T-2})), z)\right]\right\}.\end{aligned}$$

This looks formidable, but if we define

$$V_{T-1}(x) \triangleq \min_{z\in Z} c_{T-1}(x, z), \tag{8.17}$$

then, using the fact that $X_{T-1} = f(X_{T-2}, U_{T-2}, Z_{T-2})$, we can write

$$\begin{aligned}&\min_{g_0^{T-2}} J(g_0^{T-2}, g^*_{T-1}) \\ &= \min_{g_0^{T-3}} \left\{\mathbf{E}\left[\sum_{t=0}^{T-3} c_t(X_t, g_t(X_t))\right] + \min_{g_{T-2}} \mathbf{E}\left[c_{T-2}(X_{T-2}, g_{T-2}(X_{T-2})) + V_{T-1}(f(X_{T-2}, U_{T-2}, g_{T-2}(X_{T-2})))\right]\right\}.\end{aligned}$$

Let us define the function

$$V_{T-2}(x) \triangleq \min_{z\in\mathsf{Z}} \Big\{ c_{T-2}(x,z) + \mathbf{E}[V_{T-1}(f(x,U_{T-2},z))] \Big\}, \tag{8.18}$$

where the expectation is only with respect to the random disturbance U_{T-2}. Then, using the fact that X_{T-2} and U_{T-2} are independent, we can write

$$\begin{aligned}
&\min_{g_{T-2}} \mathbf{E}\Big[c_{T-2}(X_{T-2}, g_{T-2}(X_{T-2})) + V_{T-1}(f(X_{T-2}, U_{T-2}, g_{T-2}(X_{T-2}))) \Big] \\
&\quad = \min_{g_{T-2}} \sum_x \mathbf{P}^g[X_{T-2}=x] \Big\{ c_{T-2}(x, g_{T-2}(x)) + \mathbf{E}\Big[V_{T-1}(f(x, U_{T-2}, g_{T-2}(x))) \Big] \Big\} \\
&\quad = \sum_x \mathbf{P}^g[X_{T-2}=x] \min_{z\in\mathsf{Z}} \Big\{ c_{T-2}(x,z) + \mathbf{E}[V_{T-1}(f(x,U_{T-2},z))] \Big\} \\
&\quad = \sum_x \mathbf{P}^g[X_{T-2}=x] V_{T-2}(x) \\
&\quad = \mathbf{E}[V_{T-2}(X_{T-2})].
\end{aligned} \tag{8.19}$$

This derivation requires some explanation. The first equality is a consequence of the fact that X_{T-2} and U_{T-2} are independent; thus, the expected value of any function $h(X_{T-2}, U_{T-2})$ can be evaluated by first computing the expectation with respect to U_{T-2} only for every fixed value of X_{T-2} and then doing the expectation with respect to X_{T-2}. That is (assuming, for simplicity, that U_{T-2} is also discrete), we can write

$$\begin{aligned}
\mathbf{E}[h(X_{T-2}, U_{T-2})] &= \sum_{x,u} \mathbf{P}[X_{T-2}=x, U_{T-2}=u] h(x,u) \\
&= \sum_x \mathbf{P}[X_{T-2}=x] \sum_u \mathbf{P}[U_{T-2}=x] h(x,u) \\
&= \sum_x \mathbf{P}[X_{T-2}=x] \mathbf{E}[h(x, U_{T-2})],
\end{aligned}$$

and the same reasoning applies if U_{T-2} is continuous-valued and has a pdf. Note that $\mathbf{E}[h(x, U_{T-2})]$ is a function of x. The next line is a consequence of the identity

$$\min_g \mathbf{E}[H(X, g(X))] = \mathbf{E}[\min_{z\in\mathsf{Z}} H(X,z)],$$

where the minimization on the left-hand side is over *functions* $g : \mathsf{X} \to \mathsf{Z}$, while in the right-hand side we pull the minimization inside the expectation, but now choose the best z for every x. The remaining lines follow from the definition of V_{T-2} in (8.18). Now, (8.19) tells us that the optimal strategy at time $T-2$ is given by

$$g^*_{T-2}(x_{T-2}) = \arg\min_{z\in\mathsf{Z}} \{ c_{T-2}(x_{T-2}, z) + \mathbf{E}[V_{T-1}(f(x_{T-2}, U_{T-2}, z))] \}.$$

Continuing in this manner, we see that, if we start with V_{T-1} defined in Eq. (8.17) and recursively define the functions

$$V_{T-k}(x) \triangleq \min_{z \in \mathsf{Z}} \Big\{ c_{T-k}(x,z) + \mathbf{E}[V_{T-k+1}(f(x, U_{T-k+1}, z))] \Big\}, \tag{8.20}$$

for $k = 2, \ldots, T$, then the optimal action at time $T-k$ is given by

$$g^*_{T-k}(x) = \arg\min_{z \in \mathsf{Z}} \Big\{ c_{T-k}(x,z) + \mathbf{E}[V_{T-k+1}(f(x, U_{T-k}, z))] \Big\}, \tag{8.21}$$

and that the minimum expected cost is given by

$$\min_g J(g) = \mathbf{E}[V_0(X_0)].$$

We have thus shown how to compute the optimal Markov strategy via dynamic programming. Note the following intuitive interpretation of Eqs. (8.20) and (8.21): if the state at time $T-k$ is equal to x, then the best action at that time is the one that minimizes the sum of the current cost $c_{T-k}(x,z)$ and the expected *future* cost (or the *cost-to-go*) $\mathbf{E}[V_{T-k+1}(f(x, U_{T-k}, z))]$ starting from $X_{T-k} = x$. Note also that the cost-to-go functions $V_0, V_1, V_2, \ldots, V_{T-1}$ are computed via a backward pass: we first determine V_{T-1}, then use it to compute V_{T-2}, and all the day down to V_0. Once the cost-to-go functions are available, we compute the optimal strategy via a forward pass.

Note, by the way, that we can express the cost-to-go functions in terms of the controlled transition matrices M_z as follows: for each $k \in \{1, 2, \ldots, T-1\}$, we have

$$V_{T-k}(x) \triangleq \min_{z \in \mathsf{Z}} \Big\{ c_{T-k}(x,z) + \sum_{y \in \mathsf{X}} M_z(x,y) V_{T-k+1}(y) \Big\} \tag{8.22}$$

(exercise: prove this!).

As an illustration, let us use dynamic programming to construct the optimal strategy for the example of Section 8.1. This is an optimal control problem with horizon $T = 4$, where the controlled transition matrices are given by (8.9) and the state-action costs are given by (8.10). We start by computing $V_3(x)$ and g^*_3:

$$\begin{aligned} V_3(x) &= \min_{z \in \{0,1\}} c_3(x,z) \\ &= \min_{z \in \{0,1\}} \Big(1\{x=0\} + 1\{z=1\} \Big) \\ &= \min \Big(1\{x=0\}, 1\{x=0\} + 1 \Big) \\ &= 1\{x=0\}, \end{aligned}$$

and so the optimal action at $t = 3$ is $g^*_3(x) = 0$, regardless of the realized state. Now we compute

$V_2(x)$ and g_2^*. Using (8.22), we can write

$$\begin{aligned} V_2(x) &= \min_{z\in\{0,1\}} \Big\{ c_2(x,z) + M_z(x,0)V_3(0) + M_z(x,1)V_3(1) \Big\} \\ &= \min_{z\in\{0,1\}} \Big\{ 1\{x=1\} + 1\{z=1\} + M_z(x,0) \Big\} \\ &= \min\Big(1\{x=1\} + M_0(x,0), 1\{x=1\} + 1 + M_1(x,0) \Big). \end{aligned}$$

In particular,

$$\begin{aligned} V_2(0) = \min\Big(1-p, 1\Big) = 1-p \qquad &\Longrightarrow \qquad g_2^*(0) = 0 \\ V_2(1) = \min\Big(1+p, 3\Big) = 1+p \qquad &\Longrightarrow \qquad g_2^*(1) = 0, \end{aligned}$$

so the optimal action at time $t=2$ is $g_2^*(x)=0$, regardless of the state. We proceed to V_1 and g_1^*:

$$\begin{aligned} V_1(x) &= \min_{z\in\{0,1\}} \Big\{ c_1(x,z) + M_z(x,0)V_2(0) + M_z(x,1)V_2(1) \Big\} \\ &= \min_{z\in\{0,1\}} \Big\{ 1\{x=0\} + 1\{z=1\} + (1-p)M_z(x,0) + (1+p)M_z(x,1) \Big\} \\ &= \min\Big(1\{x=0\} + (1-p)M_0(x,0) + (1+p)M_0(x,1), 1\{x=0\} + 1 + (1-p)M_1(x,0) + (1+p)M_1(x,1) \Big). \end{aligned}$$

In particular,

$$\begin{aligned} V_1(0) = \min\Big(2p^2-p+2, p+3\Big) = 2p^2-p+2 \qquad &\Longrightarrow \qquad g_1^*(0) = 0 \\ V_1(1) = \min\Big(1-p, 2-p\Big) = 1-p \qquad &\Longrightarrow \qquad g_1^*(1) = 0 \end{aligned}$$

And finally, we compute V_0 and g_0^*:

$$\begin{aligned} V_0(x) &= \min_{z\in\{0,1\}} \Big\{ c_0(x,z) + M_z(x,0)V_1(0) + M_z(x,1)V_1(1) \Big\} \\ &= \min_{z\in\{0,1\}} \Big\{ 1\{x=1\} + 1\{z=1\} + (2p^2-p+2)M_z(x,0) + (1-p)M_z(x,1) \Big\}, \end{aligned}$$

so that

$$\begin{aligned} V_0(0) &= \min\Big((2p^2+2)(1-p), 2-p\Big) = (2p^2+2)(1-p) && \Longrightarrow g_0^*(0) = 0 \\ V_0(1) &= \min\Big(1+(2p^2-p+2)p + (1-p)^2, 4-p+2p^2\Big) = 1+(2p^2-p+2)p+(1-p)^2 && \Longrightarrow g_0^*(1) = 0. \end{aligned}$$

Thus, the open-loop strategy $g_0^*(x) = g_1^*(x) = g_2^*(x) = g_3^*(x) = 0$ is optimal, giving the expected cost

$$\mathbf{E}[V_0(X_0)] = \frac{(2p^2+2)(1-p)}{2} + \frac{1+(2p^2-p+2)p+(1-p)^2}{2} = p^2-p+2$$

(recall that $X_0 \sim \text{Bern}(\frac{1}{2})$).