1 Markov decision processes

Control is optimization over time. Without diving too deep into definitions for now, we have a system that has a *state* which evolves in time. The evolution of the state can be guided (or *controlled*) by means of *actions* in such a way that the future evolution of the state is affected by the current state, current action, and a random *disturbance*. The goal is to select actions to optimize some performance criterion that depends on the trajectory of the states and the actions. The presence of random disturbances is the salient feature of stochastic control. Thus, the problem of stochastic control is that of optimization over time while managing risk. Let us look at a few examples to see what is involved, and then abstract the common ideas into a general formulation.

Example 1.1 (Allocation of resources) We have a manufacturing process that converts x units of some resource into $F(x)$ units, where $x \geq 0$ is the available amount of the resource. At time $t = 0$, we start with some initial quantity x_0. At each time $t = 0, 1, 2, \ldots$, we decide how much of the available resource x_t we can consume, and then direct the remaining resource into production. Thus, denoting by $u_t \leq x_t$ the amount of resource consumed at time t, we have the following equation that relates x_t and u_t to x_{t+1}:

$$x_{t+1} = F(x_t) - u_t. \tag{1}$$

Here, the available resource x_t is the *state* at time t, the amount consumed u_t is the *action* at time t, and Eq. (1) describes the *dynamics*, i.e., the process of consumption and production over time. Now we add the element of risk: Suppose that the production process is subject to some random influences, and thus the output is given by a stochastic transformation $x \mapsto F(x, W)$, where W is some random variable with a fixed probability distribution ν. Let $W_0, W_1, W_2, \ldots$ be a sequence of independent and identically distributed (i.i.d.) random variables with common marginal distribution ν. Then, instead of the deterministic rule (1), we end up with stochastic dynamics of the form

$$x_{t+1} = F(x_t, W_t) - u_t. \tag{2}$$

Observe that the state x_{t+1} at time t is a deterministic function of x_t, the state at time t, u_t, the action at time t, and the random disturbance W_t.

Example 1.2 (Portfolio selection) This example involves decision-making by a 'small investor,' meaning an investor who cannot influence market prices. At $t = 0$, the investor has some initial wealth x_0. There are two available assets: a risk-free asset (such as a bond) with a fixed interest rate $r \in (0, 1)$ and a risky asset (such as a stock) with a random return. Let x_t denote the wealth available at time t. The investor consumes the fraction $q_t \in [0, 1]$ of x_t and allocates the remaining amount $v_t = (1 - q_t)x_t$ among the two assets. Let $p_t \in [0, 1]$ denote the fraction of v_t invested in the risky asset and let W_t denote the random rate of return of the risky asset at time t. Then the wealth available at time $t + 1$ is given by

$$x_{t+1} = [(1 - p_t)(1 + r) + p_t W_t](1 - q_t)x_t \tag{3}$$

Collectively, the pair (p_t, q_t) is the *action* u_t taken by the investor at time t. Once again, we see from Eq. (3) that x_{t+1} is a deterministic function of x_t, u_t, and W_t.

Example 1.3 (Regulation) Let x_t denote the operating point of some device or process at time t. The objective is to maintain the operating point near 0. Letting u_t denote the correction at time t, we have the following dynamics:

$$x_{t+1} = x_t + W_t - u_t, \tag{4}$$

where W_t is a random disturbance. Once again, x_{t+1} is a deterministic function of x_t, u_t, and W_t.

Each of the above three examples describes the following situation: We have a system that starts in some initial state x_0. At each time $t = 0, 1, 2, \ldots$, we take an action u_t (where the set of available actions may be constrained by the current state x_t). The next state x_{t+1} is a (possibly time-dependent) deterministic function of the current state-action pair (x_t, u_t) and a random disturbance W_t:

$$x_{t+1} = f_t(x_t, u_t, W_t). \tag{5}$$

In a typical stochastic control problem, at each time t we incur a *cost* $c_t(x_t, u_t)$ that depends on the system state and the action taken at time t, and the objective is to choose the actions to minimize expected cost. For instance, in Example 1.3 we could have a time-independent cost $c(x, u)$ of the form

$$c(x, u) := x^2 + ru^2, \tag{6}$$

where $r > 0$ is a fixed constant. The structure of the cost in Eq. (6) reflects the goal of keeping the operating point near 0, while at the same time keeping the control signal u small. We will make the following assumption throughout:

Assumption 1.1 *The system has a random initial state X_0 and is subject to a disturbance process $W_0, W_1, \ldots$, where the random variables $X_0, W_0, W_1, \ldots$ are mutually independent.*

At each time t, an action is taken on the basis of all information available at that time. For now, we will consider the case of *complete information*, where the information available at time t consists of the states from time 0 up to and including time t and all the actions taken at times $0, \ldots, t-1$. We assume that the system state takes values in some set $\mathcal{X}$ (the *space*) and that the available actions take values in some space $\mathcal{U}$ (the *action space* or the *control space*). For the time being, we will assume that both $\mathcal{X}$ and $\mathcal{U}$ are *finite sets*. This will allow us to introduce the key concepts without undue fuss. Later on, we will relax this assumption. We can now introduce the following key definition:

Definition 1.1 (Policy) *A* policy *(or* control strategy *or* control law*) is a sequence $g = \{g_t\}_{t \geq 0}$, where $g_t : \mathcal{X}^t \times \mathcal{U}^{t-1} \to \mathcal{U}$ is a function that maps the history of states $x_0^t := (x_0, \ldots, x_{t-1})$ and actions $u_0^{t-1} := (u_0, \ldots, u_{t-1})$ to u_t.*

The crucial aspect of Definition 1.1 is *causality*: the action at time t cannot depend on the future. Once we fix a policy g, then all the states and actions become well-defined random variables:

$$\begin{aligned} U_0 &= g_0(X_0) && \text{(7a)}\\ X_1 &= f_0(X_0, U_0, W_0) && \text{(7b)}\\ U_1 &= g_1(X_0^1, U_0) && \text{(7c)}\\ X_2 &= f_1(X_1, U_1, W_1) && \text{(7d)}\\ &\dots && \text{(7e)}\\ U_t &= g_t(X_0^t, U_0^{t-1}) && \text{(7f)}\\ X_{t+1} &= f_t(X_t, U_t, W_t) && \text{(7g)}\\ &\dots && \text{(7h)} \end{aligned}$$

In fact, if we unravel the recursion in (7), we see that, once the policy g is fixed, both the state X_t and the action U_t are *deterministic functions* of X_0 and W_0^{t-1}. For this reason, we refer to the initial state X_0 and the disturbances $W_0, W_1, \dots$ as *primitive random variables*. Moreover, each policy g defines, for each t, a joint probability distribution $\mathbf{P}_t^g[\cdot]$ of X_0^t and U_0^t, and these distributions are *consistent* in the sense that, for any t,

$$\mathbf{P}_t^g[X_0^t = x_0^t, U_0^t = u_0^t] = \sum_{x\in\mathcal{X}} \sum_{u\in\mathcal{U}} \mathbf{P}_{t+1}^g[X_0^t = x_0^t, X_{t+1} = x, U_0^t = u_0^t, U_{t+1} = u]. \tag{8}$$

From now on, we will use $\mathbf{P}^g[\cdot]$ and $\mathbf{E}^g[\cdot]$ to denote probability and expectation for a given g.

Now we can formulate the problem of *stochastic optimal control over a finite horizon*: Given an integer $T \geq 1$ and a policy g, define the *expected cost* of g over horizon T:

$$J_T(g) := \mathbf{E}^g\left[\sum_{t=0}^{T-1} c_t(X_t, U_t)\right] = \mathbf{E}^g\left[\sum_{t=0}^{T-1} c_t(X_t, g_t(X_t))\right]. \tag{9}$$

We say that a policy g^* is *optimal* if

$$J_T(g^*) = \min_g J_T(g), \tag{10}$$

where the minimum is over all policies admissible according to Definition 1.1.

1.1 Stochastic optimization

In order to understand the problem of optimal stochastic control, we first need to have a clear idea of the corresponding *static* problem, where the action has to be taken only once. This is the case of $T = 1$. We have a random state X taking values in $\mathcal{X}$. A policy is simply a function g from $\mathcal{X}$ into an action space $\mathcal{U}$, and the goal is to choose g to minimize $J_1(g) = \mathbf{E}[c(X, g(X))]$ over all g. The optimal policy is evidently given by

$$g^*(x) = \arg\min_{u\in\mathcal{U}} c(x, u). \tag{11}$$

Since we have assumed that both $\mathcal{X}$ and $\mathcal{U}$ are finite sets, there exists at least one optimal policy g^*. Moreover, we arrive at a simple but important relation:

$$\min_{g:\mathcal{X}\to\mathcal{U}} \mathbf{E}[c(X, g(X))] = \mathbf{E}\left[\min_{u\in\mathcal{U}} c(X, u)\right]. \tag{12}$$

The importance of this relation comes from the fact that, on the left-hand side of (12), the minimization is over the set of allall functions $g : \mathcal{X} \to \mathcal{U}$, which has cardinality $|\mathcal{U}|^{|\mathcal{X}|}$, while on the right-hand side the minimization is over the action space $\mathcal{U}$. Thus, even in the simplest case when there are only two available actions, the minimization on the left-hand side is over a set of size $2^{|\mathcal{X}|}$, while on the right-hand side we only need to minimize over the two possible actions for each state $x \in \mathcal{X}$.

Now let us consider something a bit more complicated: the case of *partial observations* (or *imperfect observations*). In this setting, instead of the state X, we observe some correlated random variable Y, which we assume to take values in some finite *observation space* $\mathcal{Y}$. Thus, any admissible policy is a function $g : \mathcal{Y} \to \mathcal{U}$, and the objective is to minimize the expected cost $\mathbf{E}[c(X, g(Y))]$ over all admissible policies. We then have the following:

Proposition 1.1 *A policy $g^* : \mathcal{Y} \to \mathcal{U}$ is optimal if and only if, for each observation $y \in \mathcal{Y}$,*

$$\mathbf{E}[c(X, g^*(Y))|Y = y] = \min_{u\in\mathcal{U}} \mathbf{E}[c(X, u)|Y = y], \tag{13}$$

where the expectation is with respect to the conditional distribution of the state X given the observation $Y = y$, which is computed according to the Bayes' rule:

$$P_{X|Y}(x|y) = \frac{P_{XY}(x, y)}{\sum_{x'\in\mathcal{X}} P_{XY}(x', y)} \tag{14}$$

if the denominator $P_Y(y)$ is nonzero.

Before giving the proof, we record the following analogue of Eq. (12), which is a simple consequence of Proposition 1.1:

$$\min_{g:\mathcal{Y}\to\mathcal{U}} \mathbf{E}[c(X, g(Y))] = \mathbf{E}\left[\min_{u\in\mathcal{U}} \mathbf{E}[c(X, u)|Y = y]\right]. \tag{15}$$

As before, note that the minimization on the left-hand side is over the set of all functions $g : \mathcal{X}\times\mathcal{Y} \to \mathcal{U}$, which has cardinality $|\mathcal{U}|^{|\mathcal{Y}|}$, while on the right-hand side, for each $y \in \mathcal{Y}$, the minimization is over the set of actions $\mathcal{U}$.

Proof: For any $g : \mathcal{Y} \to \mathcal{U}$, we have

$$J(g) = \mathbf{E}[c(X, g(Y))] \tag{16}$$
$$= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_{XY}(x, y)c(x, g(y)) \tag{17}$$
$$= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_Y(y)P_{X|Y}(x|y)c(x, g(y)) \tag{18}$$
$$= \sum_{y\in\mathcal{Y}} P_Y(y) \underbrace{\sum_{x\in\mathcal{X}} P_{X|Y}(x|y)c(x, g(y))}_{=\mathbf{E}[c(X,g(Y))|Y=y]}, \tag{19}$$

or, more succinctly,

$$J(g) = \mathbf{E}[\mathbf{E}[c(X, g(Y))|Y]], \tag{20}$$

using the tower property of conditional expectation (or the law of iterated expectation). Evidently, $\mathbf{E}[c(X, g(Y))|Y = y] \geq \mathbf{E}[c(X, g^*(Y))|Y = y]$ for any g^* that satisfies (13), which implies that $J(g) \geq J(g^*)$. Conversely, if some g^* satisfies $J(g^*) \leq J(g)$ for all policies g, then

$$J(g^*) = \mathbf{E}[\mathbf{E}[c(X, g^*(Y))|Y]] \tag{21}$$
$$= \mathbf{E}\left[\min_{u\in\mathcal{U}} \mathbf{E}[c(X, u)|Y]\right] \tag{22}$$
$$\leq \mathbf{E}\left[\mathbf{E}[c(X, g(Y))|Y]\right] \tag{23}$$
$$= \mathbf{E}[c(X, g(Y))] \tag{24}$$
$$= J(g). \tag{25}$$

In particular, $g^*(y) = \arg\min_{u\in\mathcal{U}} \mathbf{E}[c(X, u)|Y = y]$ satisfies (13). ■

Finally, we state a simple but key result, which also helps to simplify the complexity of optimization problems that arise in stochastic control. A more general version of this result, which is valid in so-called *Borel spaces*, was obtained in 1964 by David Blackwell. To state it, consider the case when there is a random state X, we take an action U as a function of X and some additional random variable Y correlated with X, and incur the cost $c(X, U)$. Thus, admissible policies are functions of the form $g : \mathcal{X} \times \mathcal{Y} \to \mathcal{U}$, but the cost depends only on the state X and the action U. As before, we seek an optimal policy, i.e., the one that minimizes

$$J(g) := \mathbf{E}[c(X, g(X, Y))]. \tag{26}$$

It is tempting to think about Y as some side information that may help achieve smaller expected cost. Counter to intuition, however, this extra information turns out to be *irrelevant*:

Proposition 1.2 (Blackwell's principle of irrelevant information) *There is no loss of optimality if we restrict the minimization in (26) to policies that ignore Y, i.e.,*

$$\min_{g:\mathcal{X}\times\mathcal{Y}\to\mathcal{U}} \mathbf{E}[c(X, g(X, Y))] = \min_{g:\mathcal{X}\to\mathcal{U}} \mathbf{E}[c(X, g(X))]. \tag{27}$$

Proof: When all sets are finite, the proof is embarrassingly simple: Fix any admissible policy $g : \mathcal{X} \times \mathcal{Y} \to \mathcal{U}$ and observe that, for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$,

$$c(x, g(x, y)) \geq \min_{u \in \mathcal{U}} c(x, u). \tag{28}$$

Taking the expectation of both sides and invoking (7), we see that

$$\mathbf{E}[c(X, g(X, Y))] \geq \mathbf{E}\left[\min_{u \in \mathcal{U}} g(X, u)\right] = \min_{g : \mathcal{X} \to \mathcal{U}} \mathbf{E}[c(X, g(X))]. \tag{29}$$

Minimizing the left-hand side over all $g : \mathcal{X} \times \mathcal{Y} \to \mathcal{U}$ completes the proof. ■

1.2 Optimality of Markov policies

We are now ready to prove a key result about optimal stochastic control over a finite horizon. Consider the control problem specified in Eqs. (9) and (10). The minimization on the right-hand side of (10) is over the set of all policies admissible according to Def. 1.1. Each such policy g is a T-tuple of functions $(g_0, \ldots, g_{T-1})$, where g_t ia a function from $\mathcal{X}_0^t \times \mathcal{U}_0^{t-1}$ into $\mathcal{U}$.[1] There are $|\mathcal{U}|^{|\mathcal{X}|^{t+1} \cdot |\mathcal{U}|^t}$ such functions, so the complexity of the optimization problem that needs to be solved at each time step grows *doubly exponentially* in t! Indeed, in the simplest case when $|\mathcal{X}| = |\mathcal{U}| = 2$, we need to search over $2^{2^{t+1}2^t} = 2^{2^{2t+1}}$ functions at time step t.

However, as we will see shortly, there is no loss of optimality if we restrict the search to a much smaller class of policies:

Definition 1.2 *A policy $g = (g_0, g_1, \ldots, g_{T-1})$ is* Markov *if, for each t, the action $u_t = g_t(x_0^t, u_0^{t-1})$ depends only on the most recent state x_t. In other words, a Markov policy g is a T-tuple $(g_0, \ldots, g_{T-1})$ of functions $g_t : \mathcal{X} \to \mathcal{U}$, $t \in \{0, \ldots, T-1\}$. We will denote the collection of all T-step Markov policies by $\mathcal{M}_T$.*

The reason for using the term 'Markov' will become clear later on. For now, we note that, if we limit our attention to Markov policies, then the complexity of the optimization problem that needs to be solved at each time step does not depend on time. More precisely, we have the following result:

Theorem 1.1 (Optimality of Markov policies) *Consider the stochastic optimal control problem over a finite horizon T, specified according to Eqs. (9) and (10). Then*

$$\min_{g} J_T(g) = \min_{g \in \mathcal{M}_T} J_T(g). \tag{30}$$

The proof of Theorem 1.1 relies crucially on three structural features of the problem:

1. Independent primitive random variables: the initial state X_0 and the stochastic disturbances $W_0, W_1, \ldots$ are mutually independent.

[1] From now on, given a set $\mathcal{A}$, we will denote by $\mathcal{A}_0^t$ the set of all tuples $a_0^t = (a_0, \ldots, a_t)$ with entries from $\mathcal{A}$.

2. The Markov dynamics: according to the state update rule (5), the next state is a function of the current state, current action, and a stochastic disturbance.

3. Additive costs: the total expected cost of a policy is the sum of expected state-action costs at each time step.

These three features will allow us to exploit Blackwell's principle of irrelevant information and argue that, at each time step t, the past states X_0^{t-1} and past actions U_0^{t-1} are irrelevant, and the action U_t can be selected based solely on the current state X_t.

The proof will be by induction. The base cases, for $T = 2$ and $T = 3$, will be established in the next section. The overall idea of the proof comes from a 1979 paper by Hans Witsenhausen.

1.2.1 Preparation: two-step lemma and three-step lemma

Lemma 1.1 (Two-step lemma) *If $T = 2$, then Markov policies are optimal:*

$$\min_g J_2(g) = \min_{g \in \mathcal{M}_2} J_2(g). \tag{31}$$

Proof: Fix an arbitrary 2-step policy $g = (g_0, g_1)$ with some $g_0 : \mathcal{X} \to \mathcal{U}$ and $g_1 : \mathcal{X}_0^1 \times \mathcal{U} \to \mathcal{U}$. Note that g_0 already has the Markov structure, since U_0 depends only on X_0. We will show that there exists a function $\bar{g}_1 : \mathcal{X} \to \mathcal{U}$, such that

$$J_2(g_0, \bar{g}_1) \leq J_2(g_0, g_1). \tag{32}$$

To that end, we first write

$$\begin{aligned} J_2(g) &= \mathbf{E}^g[c_0(X_0, U_0) + c_1(X_1, U_1)] &(33)\\ &= \mathbf{E}^g[c_0(X_0, g_0(X_0))] + \mathbf{E}^g[c_1(X_1, g_1(X_0^1, U_0))]. &(34)\end{aligned}$$

Since the first term is not affected by g_1, it suffices to show that there exists some $\bar{g}_1 : \mathcal{X} \to \mathcal{U}$, such that

$$\mathbf{E}^{g_0, \bar{g}_1}[c_1(X_1, \bar{g}_1(X_1))] \leq \mathbf{E}^{g_0, g_1}[c_1(X_1, g_1(X_0^1, U_0))]. \tag{35}$$

Indeed, such a function exists by Blackwell's principle of irrelevant information applied to $X = X_1$ and $Y = (X_0, U_0)$. ■

Lemma 1.2 (Three-step lemma) *Suppose $T = 3$, and let a 3-step policy $g = (g_0, g_1, g_2)$ be given, where $g_2 : \mathcal{X} \to \mathcal{U}$ is Markov. Then there exists a function $\bar{g}_1 : \mathcal{X} \to \mathcal{U}$, such that*

$$J_3(g_0, \bar{g}_1, g_2) \leq J_3(g_0, g_1, g_2). \tag{36}$$

Proof: The expected cost of g is given by

$$\begin{aligned} J_3(g) &= J_3(g_0, g_1, g_2) &(37)\\ &= \mathbf{E}^g[c_0(X_0, g_0(X_0)) + c_1(X_1, g_1(X_0^1, U_1)) + c_2(X_2, g_2(X_2))]. &(38)\end{aligned}$$

As before, g_1 does not affect the first term. In contrast with the two-step case, however, g_1 affects both the second term through $U_1 = g_1(X_0^1, U_0)$ and the third term through $X_2 = f_2(X_1, U_1, W_1)$. Therefore, we need to proceed carefully. Using the law of iterated expectation, we have

$$\mathbf{E}^g[c_1(X_1, U_1) + c_2(X_2, U_2)] = \mathbf{E}^g[\mathbf{E}^g[c_1(X_1, U_1) + c_2(X_2, U_2)|X_1, U_1]]. \tag{39}$$

We claim that the conditional expectation $\mathbf{E}^g[c_1(X_1, U_1) + c_2(X_2, U_2)|X_1, U_1]$ does not depend on g_1, i.e., there exists some function $h : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$, such that, for any $g_0' : \mathcal{X} \to \mathcal{U}$ and $g_1' : \mathcal{X}_0^1 \times \mathcal{U} \to \mathcal{U}$ and with g_2 fixed,

$$\mathbf{E}^{g_0', g_1', g_2}[c_1(X_1, U_1) + c_2(X_2, U_2)|X_1, U_1] = h(X_1, U_1). \tag{40}$$

This property is known as *policy independence of conditional expectations*: the conditional expectation of future costs given the state-action pair at time 1 does not depend on the portion of the policy that resulted in that state-action pair. To prove (40), we first show that the conditional distribution of X_2 given $X_1 = x_1$ and $U_1 = u_1$ does not depend on g_0, g_1. Indeed, from (5) it follows that

$$\begin{aligned}
\mathbf{P}^g[X_2 = x_2|X_1 = x_1, U_1 = u_1] &= \mathbf{P}[f_1(X_1, U_1, W_1) = x_2|X_1 = x_1, U_1 = u_1] &(41)\\
&= \mathbf{P}[f_1(x_1, u_1, W_1) = x_2] &(42)\\
&=: P^{u_1}(x_2|x_1), &(43)
\end{aligned}$$

and the latter is determined by the probability distribution of W_2 and by the values of x_1, x_2, u_1. Therefore, for any g_0', g_1', g_2, we have

$$\begin{aligned}
&\mathbf{E}^{g_0', g_1', g_2}[c_1(X_1, U_1) + c_2(X_2, U_2)|X_1 = x_1, U_1 = u_1]\\
&\quad= \sum_{x_2 \in \mathcal{X}} \mathbf{P}^{g_0', g_1'}[X_2 = x_2|X_1 = x_1, U_1 = u_1] c_2(x_2, g_2(x_2)) &(44)\\
&\quad= \sum_{x_2 \in \mathcal{X}} P^{u_1}(x_2|x_1) c_2(x_2, g_2(x_2)) &(45)\\
&\quad=: h(x_1, u_1). &(46)
\end{aligned}$$

Note that $h(\cdot)$ depends on g_2, but not on g_0' or g_1'. Therefore, using (39) and (40), we can write

$$\begin{aligned}
\mathbf{E}^g[c_1(X_1, U_1) + c_2(X_2, U_2)] &= \mathbf{E}^g[h(X_1, U_1)] &(47)\\
&= \mathbf{E}[h(X_1, g_1(X_0, X_1, U_0))]. &(48)
\end{aligned}$$

Since $U_0 = g_0(X_0)$, the underlying probability distribution depends only on g_0. Now we may invoke Blackwell's principle with $X = X_1$ and $Y = (X_0, U_0)$ to conclude that there exists a function $\bar{g}_1 : \mathcal{X} \to \mathcal{U}$, such that

$$\mathbf{E}^{g_0, \bar{g}_1, g_2}[c_1(X_1, U_1) + c_2(X_2, U_2)] \le \mathbf{E}^{g_0, g_1, g_2}[c_1(X_1, U_1) + c_2(X_2, U_2)]. \tag{49}$$

This implies that $J_3(g_0, \bar{g}_1, g_2) \le J_3(g_0, g_1, g_2)$. ■

Now, a careful examination of the above two proofs shows that these results hold even if the state and the action spaces change from time step to time step. This will be important in the next section.

1.3 The proof of Theorem 1.1

We now proceed as follows. If $T = 1$, the result is obvious. If $T = 2$, optimality of Markov strategies follows from the two-step lemma. Now suppose that $T \geq 3$ and let an admissible policy $g = (g_0, \ldots, g_{T-1})$ be given. We will first use the two-step lemma to argue that we can replace $g_{T-1} : \mathcal{X}_0^{T-1} \times \mathcal{U}_0^{T-2} \to \mathcal{U}$ with $\bar{g}_{T-1} : \mathcal{X} \to \mathcal{U}$, such that the modified policy $\bar{g} := (g_0, \ldots, g_{T-2}, \bar{g}_{T-1})$ will have smaller expected cost: $J_T(\bar{g}) \leq J_T(g)$. Once this is done, we will replace all the other g_t's by $\bar{g}_t$'s with Markov structure by repeated invocation of the three-step lemma.

We can convert the original T-step problem into a two-step problem as follows: The state at time $T-1$ is a function of the tuples X_0^{T-2}, U_0^{T-2} of past states and actions, as well as of the disturbance W_{T-2}. The action $U_{T-1} = g_{T-1}(X_0^{T-1}, U_0^{T-2})$ is a function of the past states X_0^{T-2}, the current state X_{T-1}, and the past actions U_0^{T-2}. If we merge the time steps $t \in \{0, \ldots, T-2\}$ into a single time-step, then we face a two-step problem with stepwise costs $\sum_{t=0}^{T-2} c_t$ and c_{T-1}. The two-step lemma then allows us to replace $g_{T-1} : \mathcal{X}_0^{T-1} \times \mathcal{U}_0^{T-2} \to \mathcal{U}$ with another function $\bar{g}_{T-1} : \mathcal{X} \to \mathcal{U}$ that generates the action U_{T-1} as a function of the current state X_{T-1} only.

We now have reduced the problem to the setting where $g = (g_0, \ldots, g_{T-2}, g_{T-1})$ is such that g_{T-1} depends only on the state X_{T-1}. By the three-step lemma, we can now replace $g_{T-2} : \mathcal{X}_0^{T-2} \times \mathcal{U}_0^{T-3} \to \mathcal{U}$ with $\bar{g}_{T-2} : \mathcal{X} \to \mathcal{U}$, such that the resulting policy $\bar{g} := (g_0, \ldots, g_{T-3}, \bar{g}_{T-2}, g_{T-1})$ has smaller expected cost: $J_T(\bar{g}) \leq J_T(g)$. Inductively, for each $s \in \{1, \ldots, T-2\}$, we consider a three-step problem, where we merge all time steps $r \in \{0, \ldots, s-1\}$ into step 0, treat time step s as step 1, and merge all time steps $r \in \{s+1, \ldots, T-1\}$ into step 2. The stepwise costs are aggregated accordingly. Then, since the policy at time step 2 is Markov by the inductive assumption, the three-step lemma guarantees that the policy at time step 1, given by $g_s : \mathcal{X}_0^s \times \mathcal{U}_0^{s-1} \to \mathcal{U}$, can be replaced by another function $\bar{g}_s : \mathcal{X} \to \mathcal{U}$, such that the new overall policy has smaller expected cost.