Lecuter notes
Dynamic Programming and Stochastic Control
(SC42110)

Version 2024.05.23

Amin Sharifi Kolarijani

Peyman Mohajerin Esfahani

Delft Center for Systems and Control, Faculty of Mechanical Engineering, Delft University of Technology
Email address: m.a.sharifikolarijani@tudelft.nl

Delft Center for Systems and Control, Faculty of Mechanical Engineering, Delft University of Technology
Email address: p.mohajerinesfahani@tudelft.nl

Contents

Acknowledgement

This note, as it stands, is mostly based on the lectures for the course *Dynamic Programming and Stochastic Control (SC42110)* offered at the Delft Center for Systems and Control, TU Delft, the Netherlands. The main source of the course material and hence this note is Chapters 1 and 4 of Volume I of the book *Dynamic Programming and Optimal Control* by Dimitri Bertsekas. Another source of the course material is the slides of the lectures given by Daniel Kuhn for the course *Applied Probability and Stochastic Processes* offered in 2018 at EPFL, Switzerland.

The exercises of the course were originally designed by Anton Proskurnikov. The complete solutions to the exercises were first written by Arman Sharifi Kolarijani and later modified by Gyula Max. The current exercises and solutions are a modified version of those original exercises and solutions. Some of the graphics in the solutions were also first produced by Gyula Max and are reproduced in the current solution manual.

Amin Sharifi Kolarijani
Peyman Mohajerin Esfahani

CHAPTER 1

Introduction: Peak Shaving

In this chapter, we look at a motivating example of the class of optimal control problems that are the focus of this course. Through this example, we will elaborate on the main features of this problem class.

1.1. Peak shaving problem

Maintaining a balance between electricity generation (supply) and consumption (demand) is of utmost importance for the proper functioning of the power grid. However, the increasing number of consumers of electricity is putting a lot of pressure on the grid, potentially increasing the risk of blackouts/brownouts, particularly during peak consumption. To meet the peak demand, utility companies resort to less efficient but more expensive peaking generators. That is why they charge the consumers based on both their *total* and *peak* energy consumption over some period, say, a day. For example, the consumer's electricity cost may be

$$g_{\mathrm{e}} = \sum_{t=1}^{24} L_t \cdot \phi(t) + \left(\max_{t=1}^{24} L_t\right) \cdot \phi_{\text{peak}},$$

where L_t is the (average) load during the t-th hour, $\phi(t)$ is the energy fee of the t-th hour, and ϕ_{peak} is the fee associated with the peak load over the day.

Demand-side management is a collection of measures taken to increase the energy efficiency on the demand side. Peak shaving is one such measure that can be achieved by using a battery energy storage system (BESS). To be precise, this is done by integrating a battery for energy storage between the grid and the consumer; see Figure 1. The consumer can charge and discharge the battery in which case the actual load L_t imposed on the grid will be different from the demand D_t of the consumer. This can be particularly used for shifting the electricity consumption from periods of high demand to those of low demand and hence "shaving the peak" of the load on the grid; see Figure 2. To be a bit more precise, let us assume the consumer can put the battery in one of the three modes of charge, discharge, or idle (i.e., neither charge nor discharge) during each period t, represented by the control inputs $U_{t-1} = \alpha$, $U_{t-1} = -\alpha$ and $U_{t-1} = 0$, respectively, where $\alpha > 0$ is the fixed hourly rate of charge and discharge of the battery. Also, let C_t denote the state-of-charge of the battery, i.e, the amount of energy in the battery at time t, which can take any value between 0 and C. This gives us a very simple dynamics of the form

$$\begin{cases} C_{t+1} = \min\{\max\{C_t + U_t, 0\}, C\}, \\ L_{t+1} = (C_{t+1} - C_t) + D_{t+1}, \end{cases} \qquad t = 0, 1, \ldots, 23.$$

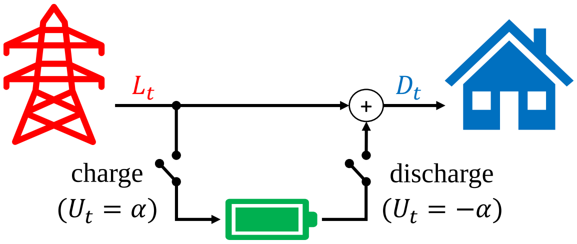


FIGURE 1. The battery energy storage system (BESS) integrated between the grid and the consumer.

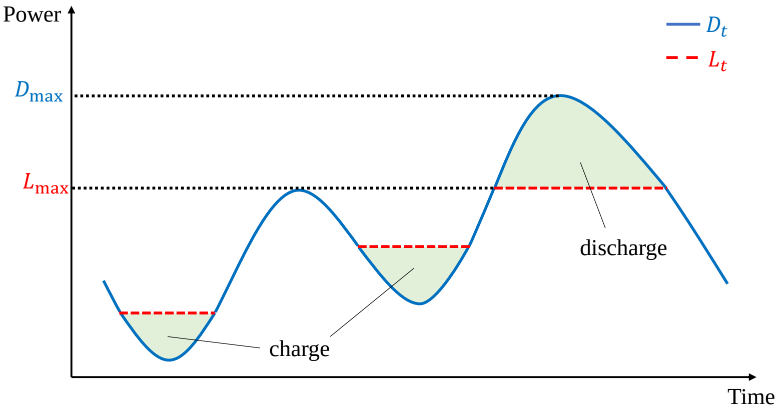


FIGURE 2. Peak shaving by shifting the electricity consumption from periods of high demand to those of low demand using BESS. D_t is the demand of the consumer and L_t is the actual load on the grid.

Moreover, in order the increase the lifetime of the battery, we want to avoid charging and discharging it as much as possible by introducing the cost

$$g_{\mathrm{b}} = \sum_{t=0}^{23} |U_t|.$$

The question is then when to charge and discharge the battery so that the total cost $g_{\mathrm{c}} + g_{\mathrm{b}}$ is minimized. So, we have the *objective* and the *decision variables* of our optimal control problem.

Let's see. So, our decision-making problem is an optimization problem. (What a surprise!) Then, why not simply apply one of the methods we know for solving a standard optimization problem? What is it that makes this problem different?

Well, here are the two main features of this decision-making problem: First, notice the *dynamic* nature of the problem. We are dealing with a dynamical system, and we have to make a *sequence* of decisions that influence its evolution. This means that we cannot look at these decisions in isolation. For example, if we decide to charge the battery at a particular time step, it leads to an increased load on the grid. This can potentially increase the peak load we have seen so far while giving us the ability to discharge the battery during a high-demand period in the future;

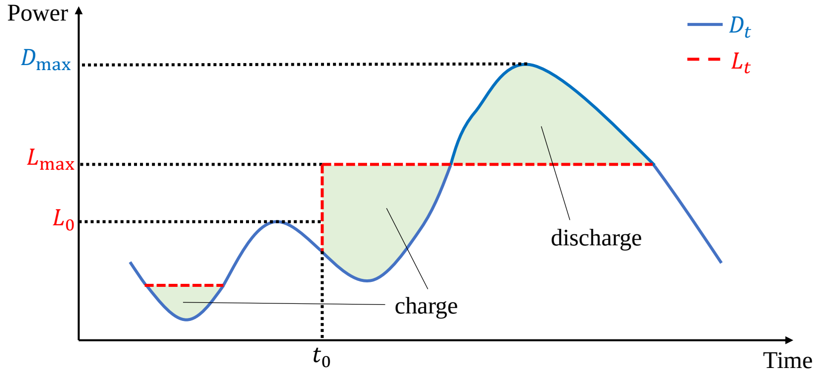


FIGURE 3. Dynamics in peak shaving via BESS. Observe that the decision to charge the battery at time t_0 leads to an increase in the peak load seen so far (from L_0 to $L_{\max}$). However, that allows for a discharge capacity in future steps that leads to a decrease in the peak load at the end of the period (from $D_{\max}$ to $L_{\max}$).

see Figure 3. So, we accept some cost now to avoid a potentially higher cost in the future. But, how hard this can be? Just look at the demand profile of the consumer, and charge the battery during the trough and discharge it during the peak; this is exactly what model predictive control (MPC) does!

Well, here comes the second feature of this decision-making problem which is the *uncertainty*. There is always some type of *stochasticity* that reveals itself in a sequential manner. Although we are not clueless about the evolution of the demand profile (e.g., its somewhat periodic nature), we can never be certain about it in advance because of stochasticity. That is, at any moment, the demand to be realized in the future steps can deviate from any deterministic prediction/model that we might come up with; see Figure 4. For this reason, we cannot use *open-loop* control laws for the entire horizon. Instead, we need to prepare ourselves for all possible states in the future by looking for *closed-loop* control laws that use the (history of) observed states for making decisions. This means that instead of a sequence of actions $(U_t)_{t=0}^{23} \in \{0, \pm\alpha\}^{24}$, we need to look for a sequence of functions $(\mu_t : \mathbb{X} \to \{0, \pm\alpha\})_{t=0}^{23}$ that map the state $X_t \in \mathbb{X}$ of the system to the proper control action U_t at time t. This is computationally much more demanding, if not impossible!

Is all hope lost? No. (Otherwise you wouldn't be reading these notes!) What comes to save us is the dynamic programming algorithm (DPA). DPA exploits the so-called dynamic programming principle to drastically reduce the complexity of finding an optimal control policy. Simply put, DPA solves the original problem for $t = 0, \ldots, 23$ by breaking it into 24 simpler problems for $t = 23$, $t = 22$, ..., $t = 0$, solved backward in time so that at each t we already know how to make optimal decisions from $t+1$ onward, and we only need to worry about the current decision!

1.2. What is *stochastic* control?

Let's see how this course adds to or differs from your previous knowledge of control. Consider a generic reference tracking problem in the presence of disturbance as depicted in Figure 5. In our peak shaving example above, you can think of

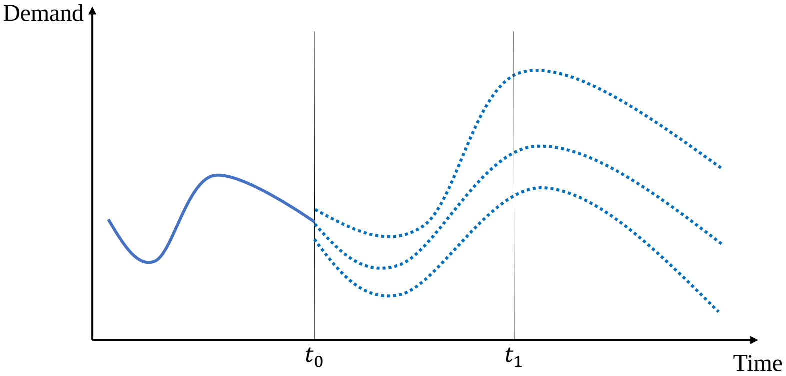


FIGURE 4. Uncertainty in peak shaving via BESS. At any moment t_0 in time, the demand to be realized in the future steps can deviate dramatically from any deterministic prediction/model. Therefore, committing to an *open-loop* control action U_{t_1} for time $t = t_1$ at time $t = 0$ does not make sense. Instead, we need to take that decision at time $t = t_1$ considering all the information that will be available till then (e.g., the peak load observed so far, the state of the charge of the battery, the realized demand).

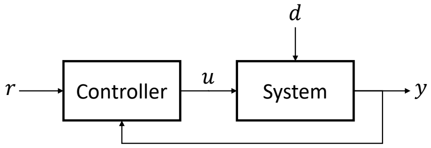


FIGURE 5. A generic reference tracking problem in the presence of disturbance.

the desired load profile as the reference signal r with the output y being the actual load on the grid, the consumer demand as the disturbance d, and the command signal to the battery (charge, discharge, or idle) as the control input u.

For the sake of the argument, assume you have an exact model of the disturbance, say, $D_t = \sin t$. In this case, you can use this model to design a controller that does what you want, given that you also have access to a perfect model of the system.[1]

Now, let's assume that all you know about the disturbance is that it is bounded, say, $-1 \leq D_t \leq 1$ for all t. With this limited amount of information, you most probably design a *robust* controller by taking into account the worst possible case.[2]

Finally, let's assume that we are somewhere in the middle. We do not know exactly what the disturbance is, but we do know more than mere lower and upper bounds on the disturbance. In particular, the extra information comes in the form of a probability distribution for the disturbance, say, $D_t = Z_t$, where Z_t's are independent and identically distributed (i.i.d.) with a triangular distribution over the interval $[-1, 1]$. This is where "stochastic control" comes into play by taking

[1] If you are in S&C MSc program, you have already designed such controllers in the courses "Control Theory " and/or "Model Predictive Control."

[2] If you are in S&C MSc program, you have already designed such controllers in the course "Robust Control."

advantage of this probabilistic information in the process of designing the controller. So, in a sense, stochastic control lies somewhere between model predictive control and robust control with the available information on the disturbance in the form of a probability distribution.

1.3. Course outline

In this course, we discuss such *dynamic decision-making problems under uncertainty.* To this end,

- we first discuss *Markov chains* (MCs) and *Markov decision processes* (MDPs) as the framework for the mathematical formulation of dynamic decision-making problems under uncertainty;
- we then introduce the *dynamic programming algorithm* (DPA) as a powerful tool for solving these decision-making problems;
- finally, we look at the application of the discussed framework and tool for solving three classic problems in control systems and operations research, namely, *linear quadratic regulation*, *portfolio optimization*, and *optimal stopping.*

1.4. General notations

We finish this chapter by introducing and fixing the related notations for the rest of this note.

Real numbers:

- $\mathbf{R}$ is the real line and $\mathbf{R}_+ \coloneqq \{x \in \mathbf{R} \;:\; x \geq 0\}$ denotes the non-negative reals.
- Intervals in $\mathbf{R}$ are described using the standard notation, e.g., $(a, b] \coloneqq \{x \in \mathbf{R} \;:\; a < x \leq b\}$ for $a, b \in \mathbf{R}$.

Natural numbers:

- $\mathbf{N} = \{1, 2, \ldots\}$ is the set of natural numbers, $\mathbf{N}_0 \coloneqq \mathbf{N} \cup \{0\} = \{0, 1, 2, \ldots\}$ denotes the non-negative integers, and $\mathbf{N}^\infty \coloneqq \mathbf{N} \cup \{\infty\}$ denotes the extended naturals.
- $[n] \coloneqq \{1, 2, \ldots, n\}$ for $n \in \mathbf{N}^\infty$ with the convention $[\infty] = \mathbf{N}$.

Given a set $\mathbb{X}$:

- $|\mathbb{X}|$ denotes the cardinality (i.e., number of elements) of $\mathbb{X}$,
- $2^{\mathbb{X}} = \{A \;:\; A \subset \mathbb{X}\}$ is the power set of $\mathbb{X}$ including all subsets of $\mathbb{X}$ (in particular, $\emptyset, \mathbb{X} \in 2^{\mathbb{X}}$),
- $\Delta(\mathbb{X})$ denotes the set of probability mass/density functions on $\mathbb{X}$ (i.e., $p \in \Delta(\mathbb{X})$ if $p : \mathbb{X} \to \mathbf{R}$, $p(x) \geq 0$ for all $x \in \mathbb{X}$, and $\int_{\mathbb{X}} p(x)\mathrm{d}x = 1$). Distributions $p \in \Delta(\mathbb{X})$ for $\mathbb{X} = [n]$ with $n \in \mathbf{N}^\infty$ are particularly treated as *row* vectors,
- $\mathbf{R}^{\mathbb{X}}$ denotes the class of functions $f : \mathbb{X} \to \mathbf{R}$

Distributions:

- $X \sim \mathcal{N}(\mu, \sigma^2)$ is a random variable with normal distribution with mean μ and variance σ^2,
- $X \sim \text{Uniform}[a, b]$ is a random variable with uniform distribution in the interval $[a, b] \in \mathbf{R}$,
- $X \sim \text{Binomial}(n, p)$ is a random variable with binomial distribution with parameters $n \in \mathbf{N}$ and $p \in (0, 1)$.

Vectors and matrices:

- $v(i)$ denotes the i-th entry of a row/column vector. Vectors are treated as *column* vectors unless specified otherwise (e.g., distributions $p \in \Delta([n])$ with $n \in \mathbf{N}^\infty$ and left eigenvectors of a square matrix which are *row* vectors).
- $e := (1, \ldots, 1)^\top$ is the all-one vector.
- $M(i, j)$ denotes the entry on row i and column j of the matrix $M \in \mathbf{R}^{n\times m}$.
- I is the identity matrix with $I(i, j) = 1$ if $i = j$ and $= 0$ otherwise.

Other notations:

- $|\alpha|$ is absolute value of a possibly complex number α.
- $[c]_+ = \max\{c, 0\}$ for $c \in \mathbf{R}$.
- $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient for $k \in \{0, 1, \ldots, n\}$ and $n \in \mathbf{N}$.

CHAPTER 2

Markov Chain

This chapter introduces Markov chains[1] as a class of stochastic processes with a very important property, namely, the Markov property.

2.1. Markov property

Let us start by defining a stochastic process:

DEFINITION 2.1 (Stochastic process). A discrete-time *stochastic process* is a family $(X_t)_{t\in\mathbf{N}_0}$ of random variables (r.v.'s) $X_0, X_1, X_2, \ldots$ taking values in the set $\mathbb{X}$ called the *state space* of the process. △

Note that the state spaces $\mathbb{X}_t$ of the r.v.'s X_t can be in general different from each other, in which case, we can take $\mathbb{X} = \bigcup_{t=1}^{\infty} \mathbb{X}_t$.

Stochastic processes are everywhere! Essentially, any time series (i.e., a variable taking values over time) you can think of is most probably a stochastic process.

EXAMPLE 2.2 (Stochastic process). Here are some examples:

- The daily closing price of a stock with $\mathbb{X} = \mathbf{R}_+$
- The yearly gross domestic product of a country with $\mathbb{X} = \mathbf{R}_+$
- The daily sales of a retail store with $\mathbb{X} = \mathbf{R}_+$
- The weekly temperature of a lake with $\mathbb{X} = \mathbf{R}$
- The yearly rate of unemployment with $\mathbb{X} = [0, 100]$
- The number of hits of a website every minute with $\mathbb{X} = \mathbf{N}_0$
- The demand in a power grid every minute with $\mathbb{X} = \mathbf{R}_+$
- The hourly occupancy level of a building with $\mathbb{X} = \mathbf{N}_0$ △

For a generic stochastic process, the distributions of the state X_t at different times $t \in \mathbf{N}_0$ can potentially depend on each other, without any restriction. The story is however different for a process with the *Markov property.* Informally, a stochastic process has the Markov property if its "future depends on the past only through the present." So, the only part of the history of the process that affects its future evolution is its *current state.* Put differently, if we want to know the distribution of the next state given the current state, there is *no additional information* in knowing the past history of the process. Here is the exact mathematical definition.

DEFINITION 2.3 (Markov chain). A discrete-time stochastic process $(X_t)_{t\in\mathbf{N}_0}$ with state space $\mathbb{X}$ has the *Markov property* if

$$\mathbb{P}(X_{t+1} \in A \mid X_0 = x_0, X_1 = x_1, \ldots, X_t = x_t) = \mathbb{P}(X_{t+1} \in A \mid X_t = x_t),$$

[1] Also known as Markov processes.

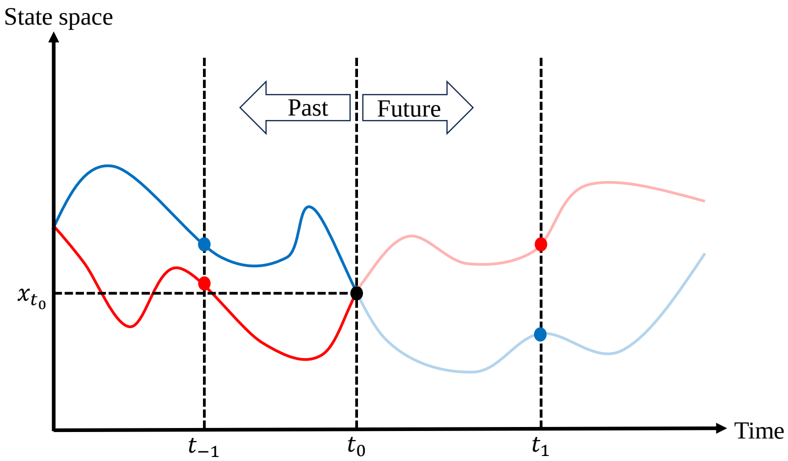


Figure 1. Two different realizations of a stochastic process starting from the same initial state. For an MC, when it comes to the evolution of the process for $t > t_0$, it does not matter how the process arrives at the realization x_{t_0} at $t = t_0$, and hence the two realizations are effectively equivalent.

for all time $t \in \mathbf{N}_0$, sequence of observed states $(x_0, \ldots, x_t) \in \mathbb{X}^{t+1}$,[2] and event $A \subset \mathbb{X}$. Such a stochastic process is called a *Markov chain* (MC). △

Figure 1 provides a schematic illustration of the Markov property. Shown are two different realizations of a stochastic process starting from the same initial state with $t = t_0$ being the current time and $X_{t_0} = x_{t_0}$ being the current state. Now, for a generic stochastic process, the future distribution of X_t for $t > t_0$ (e.g., the distribution of X_{t_1}) can potentially depend on the entire history of X_t for $t < t_0$ (e.g., the specific realization of $X_{t_{-1}}$). However, for an MC, the only thing that matters is the current state. More precisely, given that $X_{t_0} = x_{t_0}$, the future distribution of X_t for $t > t_0$ (e.g., the distribution of X_{t_1}) is independent of the history of X_t for $t < t_0$ (e.g., the specific realization of $X_{t_{-1}}$). That is, when it comes to the evolution of the process for $t \geq t_0$, it does not matter how the process arrives at $X_{t_0} = x_{t_0}$ at $t = t_0$, and hence the two realizations of Figure 1 are effectively equivalent. You can think of the current state of an MC as the state of a linear time-invariant (LTI) system in the sense that it encompasses all the relevant information from the past concerning the future evolution of the system. Let us look at some examples.

Example 2.4 (Stock price). Let X_t be the closing price of a stock on day $t \in \mathbf{N}$ and R_{t+1} be the rate of return of this stock from day t to day $t+1$ so that

$$X_{t+1} = X_t(1 + R_{t+1}), \quad t \in \mathbf{N}_0.$$

Assume that the rates of return are i.i.d. and normally distributed, i.e.,

$$R_t \sim \mathcal{N}(\mu, \sigma^2), \quad t \in \mathbf{N}$$

Note that X_{t+1} depends on the stock price history $X_1, X_2, \ldots, X_t$ only through the current price X_t. This is particularly because of the assumption that R_t's are

[2] Recall that $x_0, x_1, \ldots, x_t$ are numbers and not random variables.

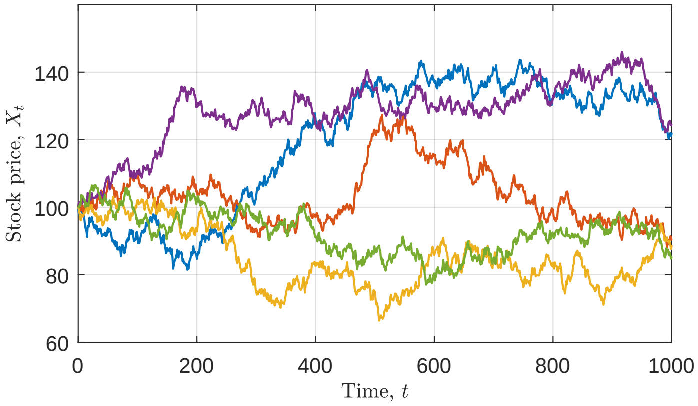


FIGURE 2. Five different realizations of the Markov chain $X_{t+1} = X_t(1 + R_{t+1})$ with $X_0 = 100$ and $R_t \sim \mathcal{N}(0, 1)$. See Example 2.4.

independent. This is indeed an MC with a continuous state space $\mathbb{X} = \mathbf{R}$. See Figure 2 for an example. △

EXAMPLE 2.5 (Coin flips). Consider a sequence of coin flips and define X_t as the number of heads observed up to (and including) time $t \in \mathbf{N}$ with $X_0 = 0$. Also, define

$$W_t = \begin{cases} 1 & \text{if the outcome at time } t \text{ is heads,} \\ 0 & \text{otherwise,} \end{cases} \qquad t \in \mathbf{N},$$

and note that W_t's are i.i.d. random variables (to be precise, W_t is a Bernoulli process). Then,

$$X_t = W_1 + W_2 + \ldots + W_t = X_{t-1} + W_t, \quad t \in \mathbf{N},$$

Thus, X_t depends on the past coin flips only through the most recent count X_{t-1}. Indeed, if we are only interested in the number of heads up to time t, we do not need the complete information on the sequence $W_1, W_2, \ldots, W_t$ of the result of coin flips. This is an MC with a discrete state space $\mathbb{X} = \mathbf{N}_0$ and some interesting properties, e.g.,

- $X_{t+1} \geq X_t$ for all $t \in \mathbf{N}$ (monotonicity)
- $X_t \leq t$ for all $t \in \mathbf{N}$

See Figure 3 for an example. Observe that, while X_{10} can take any value from 0 to 10 (corresponding to 10 consecutive tails and 10 consecutive heads, respectively), all five realizations shown in Figure 3 result in $5 \leq X_{10} \leq 7$. This has to do with the probability distribution of X_{10} which we will discuss later. △

Let us now look at a couple of classic examples from operations research.

EXAMPLE 2.6 ((s, S) inventory model). Consider an inventory of a certain product to be sold in response to a stochastic demand. Let X_t denote the inventory level *at the end of* period t and assume that we face a demand D_t *during* period t. A simple ordering policy, called the (s, S) model is as follows: *At the end of each period, order nothing as long as the inventory exceeds a level* $s \geq 0$*; otherwise,*

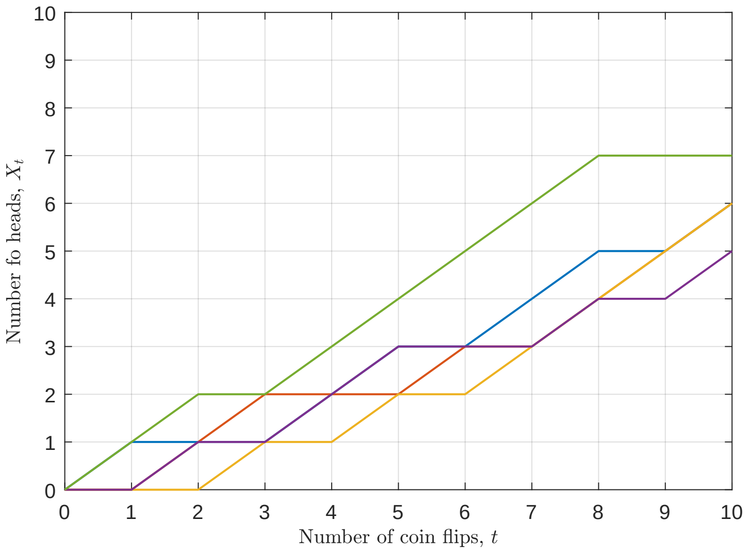


Figure 3. Five different realizations of the Markov chain $X_{t+1} = X_t + W_{t+1}$ with $X_0 = 0$ and W_t being an i.i.d. Bernoulli process with success probability $p = \frac{1}{2}$. See Example 2.5.

increase the inventory to a level $S > s$. See Figure 4 for an illustration. Using this ordering policy, we have:

- If the inventory level is less than or equal to s at the end of period t (i.e., $X_t \le s$), we increase the inventory level to S and then respond to the demand D_{t+1} during the period $t+1$ as much as the inventory allows. This leads to the inventory level $X_{t+1} = \max\{0, S - D_{t+1}\} =: [S - D_{t+1}]_+$ at the end of period $t+1$.
- If the inventory level is greater than s at the end of period t (i.e., $X_t > s$), we do not order anything and respond to the demand D_{t+1} during the period $t+1$ as much as our existing inventory from previous period allows. This leads to the inventory level $X_{t+1} = [X_t - D_{t+1}]_+$ at the end of period $t+1$.

Putting these together, the dynamics of the process is

$$X_{t+1} = \begin{cases} [S - D_{t+1}]_+ & \text{if } X_t \le s, \\ [X_t - D_{t+1}]_+ & \text{if } X_t > s. \end{cases}$$

Now, if the demands D_t are *independent*, then X_{t+1} depends on the past demands only through the current inventory level X_t and hence we have an MC. △

Example 2.7 (Queuing model). Consider a servicing system involving servers (e.g., a number of receptionists) and customers as shown in Figure 5. The customers arrive at the waiting room according to the stochastic process $(A_t)_{t\in\mathbf{N}}$, where A_t represents the number of customers arriving *during* period t. The servers can process D_t customers *during* period t.

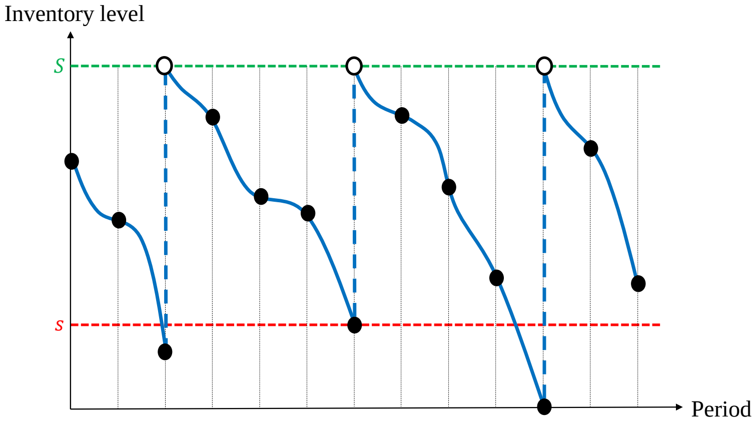


FIGURE 4. The (s, S) inventory model. See Example 2.6.

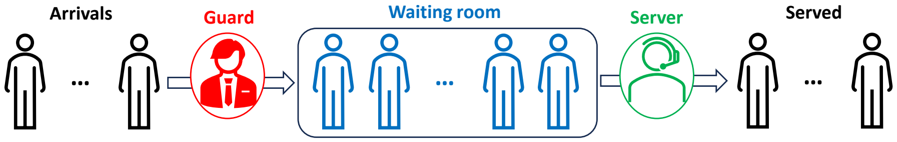


FIGURE 5. A queuing model. See Example 2.7.

Let X_t denote by the total number of customers in the waiting room *at the end of* period t. Then, the dynamics of this process can be written as

$$\begin{aligned} X_{t+1} &= \begin{cases} 0 & \text{if } X_t + A_{t+1} \leq D_{t+1} \\ X_t + A_{t+1} - D_{t+1} & \text{otherwise} \end{cases} \\ &= [X_t + A_{t+1} - D_{t+1}]_+ . \end{aligned}$$

Therefore, X_t follows a Markov chain if the arrivals are *independent*.

Now, let us assume that a security guard controls the access to the waiting room: The guard observes the number of people in the room with a delay of 1 period (i.e., at time $t-1$) and if more than K people are in the waiting room, then any new arrivals during period $t+1$ are turned away. In this case, the number of people in the waiting room satisfies the recursion

$$X_{t+1} = f(X_t, X_{t-1}, A_{t+1}, D_{t+1}) = \begin{cases} [X_t + A_{t+1} - D_{t+1}]_+ & \text{if } X_{t-1} \leq K, \\ [X_t - D_{t+1}]_+ & \text{otherwise.} \end{cases}$$

Observe that in this case, X_t is *not* a Markov chain since X_{t+1} depends also on X_{t-1}. However, if we *increase the state dimension* and define $X'_t = (X_t, X_{t-1})$, then X'_t is again a Markov chain! Indeed, the dynamics of X'_t is given by

$$X'_{t+1} = (X_{t+1}, X_t) = \big(f(X_t, X_{t-1}, A_{t+1}, D_{t+1}), X_t\big) = f'(X'_t, A_{t+1}, D_{t+1}).$$

△

A couple of remarks are in order. First, the last example shows that *many stochastic processes can be converted to Markov chains by enlarging the state space.*

Indeed, as we have seen in the preceding example, for any memory parameter $m \geq 1$, the recursion

$$X_{t+1} = f(X_t, X_{t-1}, \dots, X_{t-m}),$$

can be transformed to

$$X'_{t+1} = f'(X'_t),$$

by including all the relevant memory in a new state variable

$$X'_t = (X_t, X_{t-1}, \dots, X_{t-m}).$$

The problem however is *the state space explosion* which has negative implications, computationally speaking. Therefore, the key to a good MC model is to control the size of the state space!

Second, as you can see, in all of the examples above, we managed to describe the MC by specifying a *recursion*

$$X_{t+1} = f(X_t, W_{t+1}),$$

that expresses X_{t+1} in terms of X_t and some *independent* disturbance process W_t. To be precise, for this recursion to describe an MC, we need the disturbance W_{t+1} to be *conditionally* independent *given* X_t. This characterization allows us to describe the evolution of the *state* and is commonly used for MC's with a *continuous* state space.[3] In the next section, we look at another way to describe the dynamics of an MC which is particularly handy when the state space is *countable*.

2.2. Transition probability matrix

In the rest of this chapter, we consider MC's with a *countable* (i.e., discrete) state space $\mathbb{X}$. Therefore, without loss of generality (w.l.o.g.), we assume that $\mathbb{X} = [n] := \{1, 2, \dots, n\}$ with $n \in \mathbf{N}^\infty := \mathbf{N} \cup \{\infty\}$.[4] In particular, deviating from our standard notation, we use i and j to denote generic elements of the state space to emphasize the fact that $\mathbb{X}$ is countable.[5]

Definition 2.8 (Transition probability matrix). For a Markov chain with countable state space $\mathbb{X} = [n]$ where $n \in \mathbf{N}^\infty$, we define the *transition probability matrix* $P_t \in [0,1]^{n \times n}$ with entries

$$P_t(i,j) = \mathbb{P}(X_{t+1} = j \mid X_t = i), \quad i, j \in \mathbb{X}, \tag{2.1}$$

as the collection of the probabilities of transitioning from state i at time t to state j at time $t+1$. △

An example of these transition probabilities for an MC with five states (i.e., $\mathbb{X} = \{1, 2, 3, 4, 5\}$) is shown in Figure 6. The transition probability matrix can generally be time-varying, however, for the rest of this chapter, we focus only on *time-homogeneous* MC's:

Definition 2.9 (Time-homogeneous Markov chain). A Markov chain is *time-homogeneous* if the transition probability matrices of the chain are independent of time, that is, $P_t = P_{t'}$ for all $t, t' \in \mathbf{N}_0$. △

[3] This is probably the characterization you are familiar with in control systems.

[4] $n = \infty$ implies $\mathbb{X} = \mathbf{N}$.

[5] You can think of state i as the i-th state x_i under some fixed labeling of elements of $\mathbb{X} = \{x_1, x_2, \dots, x_i, \dots\}$. To be precise, there is an underlying 1-to-1 mapping $\phi : \mathbb{X} \to \{1, 2, \dots, n\}$ with $n \in \mathbf{N}^\infty$ such that $i = \phi(x)$.

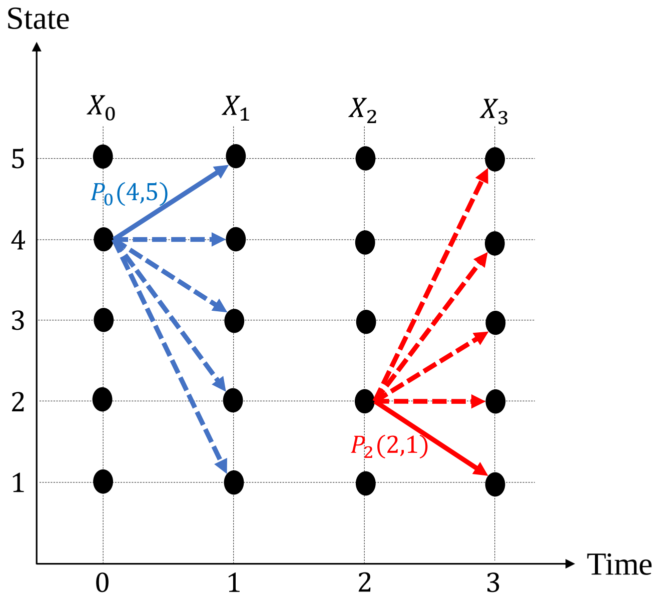


FIGURE 6. Examples of transition probabilities for an MC with five states.

Therefore, for a time-homogeneous MC, we can drop the subscript t and define a *single* transition probability matrix $P \in [0,1]^{n\times n}$ where

$$P(i,j) = \mathbb{P}(X_{t+1} = j \mid X_t = i), \quad \forall t \in \mathbf{N}_0,\ i,j \in \mathbb{X},$$

See Figure 7. Note that P is called a "matrix" even though $\mathbb{X}$ might be countably infinite. If $\mathbb{X}$ is infinite, then P has also infinitely many entries. We can also represent the transition probabilities *graphicallay*.

REMARK 2.10 (Graphical representation). An MC with countable state space $\mathbb{X}$ and transition probability matrix P can be equivalently characterized by a weighted directed graph with

(1) the set of nodes $\mathbb{V} = \mathbb{X}$, i.e., one node for each element of the state space;
(2) the set of directed edges $\mathbb{E} \subset \mathbb{V} \times \mathbb{V}$ where

$$(i,j) \in \mathbb{E} \Longleftrightarrow P(i,j) > 0, \quad \forall (i,j) \in \mathbb{V} \times \mathbb{V},$$

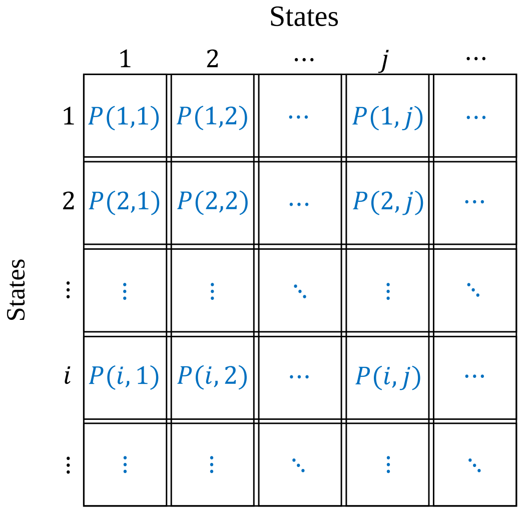
i.e., a directed edge from node i to node j iff $P(i,j) > 0$;
(3) the weight function $w : \mathbb{E} \to [0,1]$ where

$$w(i,j) = P(i,j), \quad \forall (i,j) \in \mathbb{E},$$

i.e., the directed edge from node i to node j is labeled by $P(i,j)$.

Note that the graph representation provides us with the exact same information as the transition probability matrix. Nevertheless, the graphical nature of this representation can sometimes help us identify specific properties of the system that can be exploited, e.g., symmetry in transitions. △

Let us look at a simple example.



States

States	1	2	⋯	j	⋯
1	$P(1,1)$	$P(1,2)$	⋯	$P(1,j)$	⋯
2	$P(2,1)$	$P(2,2)$	⋯	$P(2,j)$	⋯
⋮	⋮	⋮	⋱	⋮	⋱
i	$P(i,1)$	$P(i,2)$	⋯	$P(i,j)$	⋯
⋮	⋮	⋮	⋱	⋮	⋱

FIGURE 7. Transition probabilities matrix for a time-homogeneous MC with $P(i,j)$ being the probability of transitioning from state i to state j.

EXAMPLE 2.11 (Coin flips – continued). Consider the MC of Example 2.11 where X_t is the number of heads observed in a sequence of coin flips up to time $t \in \mathbf{N}$ with $X_0 = 0$. Note that the state space of this Markov chain is

$$\mathbb{X} = \mathbf{N}_0 = \{0, 1, 2, \ldots\}.$$

Assume that the coin is fair, i.e., the probability of the outcome being heads or tails is $\frac{1}{2}$. Therefore, if $X_t = x$ (i.e., the number of heads observed up to time t is x), then X_{t+1} can take the values x (if the outcome at time $t+1$ is tails) or $x+1$ (if the outcome at time $t+1$ is heads), both with probability $\frac{1}{2}$. Thus,

$$\mathbb{P}(X_{t+1} = y \mid X_t = x) = \begin{cases} \frac{1}{2} & \text{if } y = x+1, \\ \frac{1}{2} & \text{if } y = x, \\ 0 & \text{otherwise.} \end{cases}$$

The transition probability matrix of this Markov chain is then (with the i-th row/column of P corresponding to the state $x = i-1$ in $\mathbb{X}$)

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & \cdots \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & \cdots \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

We can also construct the directed graph corresponding to this Markov chain as shown in Figure 8. △

The transition probability matrix P is a *(row) stochastic matrix*, that is,

(1) P is non-negative: $P(i,j) \geq 0$ for all $i, j \in \mathbb{X}$.
(2) The rows of P sum to 1: $\sum_{j \in \mathbb{X}} P(i,j) = 1$ for all $i \in \mathbb{X}$.

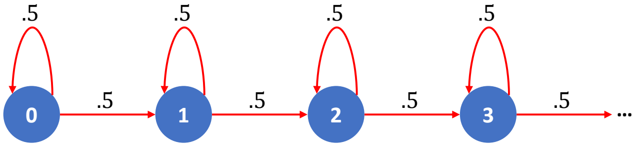


FIGURE 8. Graphical representation of the MC of Example 2.11. Circles are the nodes corresponding to the states and directed arcs are the edges corresponding to the transitions with the weights representing the probabilities.

Both properties above follow from the definition of P. To see this, note that $P(i,:) \in \Delta(\mathbb{X})$, that is, the i-th row of P is a p.m.f. on the space $\mathbb{X}$ which gives the probability distribution of the next state given that the current state is i. Stochastic matrices have the following interesting properties:

LEMMA 2.12 (Stochastic matrix). *Let $P \in \mathbf{R}^{n\times n}$ be a stochastic matrix. Then,*

(1) *the all-one vector $e = (1,\dots,1)^\top$ is a right eigenvector of P corresponding to the eigenvalue $\lambda = 1$, that is, $Pe = e$;*
(2) *all powers of P are also stochastic matrices, that is, P^t is a stochastic matrix for all $t \in \mathbf{N}$;*
(3) *all eigenvalues of P reside in the unit disc, that is, $|\lambda| \le 1$ for all eigenvalues λ of P.*

The second property above implies that the powers of P are also transition probability matrices. Indeed, these matrices describe the probabilities of multi-step transitions of the MC:

LEMMA 2.13 (Multi-step transition probability). *For each $s \in \mathbf{N}$, the entry (i,j) of P^s is the probability of transitioning from state i to state j in s time steps, that is,*

$$\mathbb{P}(X_{t+s} = j \mid X_t = i) = P^s(i,j), \quad \forall i,j \in \mathbb{X},\ s \in \mathbf{N},\ t \in \mathbf{N}_0.$$

PROOF. The proof is by induction. The base case of $s = 1$ holds by definition of the transition probability matrix P for time-homogeneous MC's. So, let us assume the statement holds for $s \in \mathbf{N}$. We need to show that it also holds for $s+1$. First observe that from the definition of the conditional and marginal probabilities, we have

$$\begin{aligned}
\mathbb{P}(X_{t+s+1} = j \mid X_t = i) &= \frac{\mathbb{P}(X_{t+s+1} = j,\ X_t = i)}{\mathbb{P}(X_t = i)} \\
&= \frac{\sum_{k\in\mathbb{X}} \mathbb{P}(X_{t+s+1} = j,\ X_{t+s} = k,\ X_t = i)}{\mathbb{P}(X_t = i)} \\
&= \frac{\sum_{k\in\mathbb{X}} \mathbb{P}(X_{t+s+1} = j \mid X_{t+s} = k,\ X_t = i)\ \mathbb{P}(X_{t+s} = k,\ X_t = i)}{\mathbb{P}(X_t = i)} \\
&= \sum_{k\in\mathbb{X}} \mathbb{P}(X_{t+s+1} = j \mid X_{t+s} = k,\ X_t = i)\ \mathbb{P}(X_{t+s} = k \mid X_t = i),
\end{aligned}$$

Using the Markov property, we have

$$\mathbb{P}(X_{t+s+1} = j \mid X_{t+s} = k,\ X_t = i) = \mathbb{P}(X_{t+s+1} = j \mid X_{t+s} = k) = P(k,j),$$

and the induction assumption implies that

$$\mathbb{P}(X_{t+s} = k \mid X_t = i) = P^s(i,k).$$

Hence,

$$\mathbb{P}(X_{t+s+1} = j \mid X_t = i) = \sum_{k \in \mathbb{X}} P^s(i,k)\ P(k,j) = P^{s+1}(i,j).$$

This completes the proof. □

The preceding result says that P^t can be considered as the "t-step" transition probability matrix, that is, the i-th row of P^t is the distribution of X_t given that $X_0 = i$. This implies that computing state distributions essentially boils down to computing the powers of P. This brings us to the following definition:

Definition 2.14 (Countable-state Markov chain). A countable-state MC is a tuple $(\mathbb{X}, P, p_0)$ describing the stochastic process $(X_t)_{t \in \mathbf{N}_0}$ with

(1) countable state space $\mathbb{X} = [n]$ where $n \in \mathbf{N}^\infty$, i.e., X_t is a random variable taking values in $\mathbb{X}$ for all $t \in \mathbf{N}_0$;
(2) transition probability matrix $P \in [0,1]^{n \times n}$, i.e., $\mathbb{P}(X_{t+1} = j | X_t = i) = P(i,j)$ for all $t \in \mathbf{N}_0$ and $i, j \in \mathbb{X}$;
(3) initial distribution $p_0 \in \Delta(\mathbb{X})$, i.e., $X_0 \sim p_0$ and hence $\mathbb{P}(X_0 = i) = p_0(i)$ for all $i \in \mathbb{X}$.

Indeed, a countable-state MC is fully characterized by the tuple $(\mathbb{X}, P, p_0)$. In particular, using the initial distribution p_0 and the transition probability matrix P, we can explicitly compute the probability distribution of any trajectory (i.e. the *joint* distribution of states) and the probability distribution of the states at any time (see also Figure 9):

Lemma 2.15 (Trajectory and state distribution). *Consider a countable-state MC defined by the triple $(\mathbb{X} = [n], P, p_0)$ where $n \in \mathbf{N}^\infty$.*

(1) Joint state distribution: *For each trajectory $(i_0, i_1, \ldots, i_t) \in \mathbb{X}^{t+1}$, we have*
$$\mathbb{P}(X_t = i_t, \ldots, X_0 = i_0) = p_0(i_0)\ P(i_0, i_1)\ \cdots\ P(i_{t-1}, i_t). \tag{2.2}$$
(2) State distribution: *Let $p_t \in \Delta(\mathbb{X})$ denote the state distribution at time t, that is, $p_t(i) = \mathbb{P}(X_t = i)$ for each $i \in \mathbb{X}$. We have*[6]
$$p_t = p_0 P^t. \tag{2.3}$$

Proof. For the first item, we can use the definition of conditional probability to write

$$\begin{aligned}\mathbb{P}(X_t = i_t, \ldots, X_0 = i_0) &= \mathbb{P}(X_{t-1} = i_t, \ldots, X_0 = i_0) \times \\ &\quad \mathbb{P}(X_t = i_t \mid X_{t-1} = i_{t-1}, \ldots, X_0 = i_0).\end{aligned}$$

Continuing this process, we have

$$\begin{aligned}\mathbb{P}(X_t = i_t, \ldots, X_0 = i_0) &= \mathbb{P}(X_0 = i_0) \times \\ &\quad \mathbb{P}(X_1 = i_1 \mid X_0 = i_0) \times \\ &\quad \mathbb{P}(X_2 = i_2 \mid X_1 = i_1,\ X_0 = i_0) \times \\ &\quad \cdots \times\end{aligned}$$

[6] Recall that distributions are treated as *row* vectors.

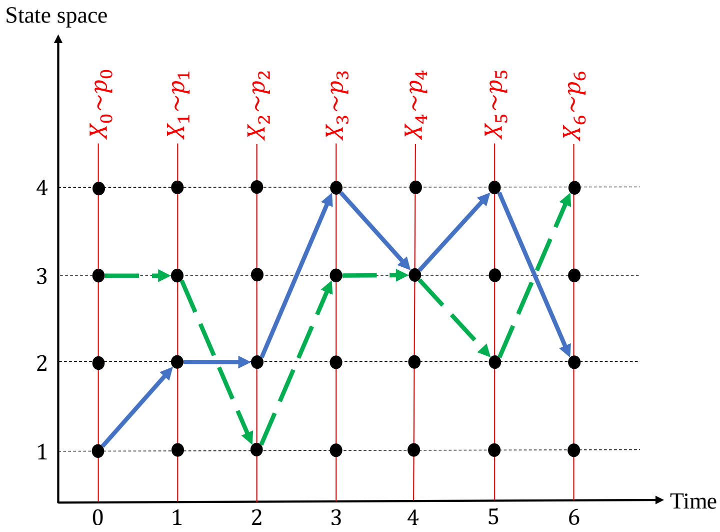


FIGURE 9. Trajectory and state distributions in an MC with four states over $t = 6$ steps. $p_t \in \Delta([4])$ is distribution of the states at time t. The two trajectories are $(1, 2, 2, 4, 3, 4, 2)$ shown by the solid blue line and $(3, 3, 1, 3, 3, 2, 4)$ shown by the dashed green line.

$$\mathbb{P}(X_t = i_t \mid X_{t-1} = i_{t-1}, \dots, X_0 = i_0).$$

Then, from the Markov property, it follows that

$$\begin{aligned}\mathbb{P}(X_t = i_t, \dots, X_0 = i_0) \;\; = \;\; & \mathbb{P}(X_0 = i_0)\times \\ & \mathbb{P}(X_1 = i_1 \mid X_0 = i_0)\times \\ & \mathbb{P}(X_2 = i_2 \mid X_1 = i_1,)\times \\ & \cdots\times \\ & \mathbb{P}(X_t = i_t \mid X_{t-1} = i_{t-1}).\end{aligned}$$

The result in (2.2) then follows from the definition of P and p_0.

The proof of the second item is by induction. The base case $t = 0$ holds trivially as $p_0 P^0 = p_0 I = p_0$, where I is the identity matrix. So, we assume that the statement holds for t and show that it also holds for $t + 1$. The law of total probability implies that for each $i \in \mathbb{X}$, we have

$$\begin{aligned} p_{t+1}(i) &= \mathbb{P}(X_{t+1} = i) \\ &= \sum_{j\in\mathbb{X}} \mathbb{P}(X_{t+1} = i \mid X_t = j)\ \mathbb{P}(X_t = j) \\ &= \sum_{j\in\mathbb{X}} P(j, i)\ p_t(j) = \sum_{j\in\mathbb{X}} P^\top(i, j)\ p_t(j). \end{aligned}$$

That is, $p_{t+1}^\top = P^\top p_t^\top$, or equivalently, $p_{t+1} = p_t P$. Then, using the induction assumption, we obtain

$$p_{t+1} = (p_0 P^t)P = p_0 P^{t+1}.$$

This completes the proof. □

Observe that the equality in (2.3) corresponds to the recursion $p_{t+1}^\top = P^\top p_t^\top$ for $t \in \mathbf{N}_0$. This is essentially an LTI system with state vector $p_t^\top$ and dynamics matrix $P^\top$. Therefore, the characterization of a *countable*-state MC via its *transition probability matrix* allows us to describe the evolution of its *state distribution*.[7] Let us put this result to use in a simple example.

EXAMPLE 2.16 (Coin flips – continued). Consider the MC of Examples 2.5 and 2.11 where X_t is the number of heads observed in a sequence of $t \in \mathbf{N}$ coin flips with $X_0 = 0$. Recall the state space

$$\mathbb{X} = \mathbf{N}_0 = \{0, 1, 2, \ldots\},$$

and the transition probability matrix (with the i-th row/column of P corresponding to the state $x = i - 1$ in $\mathbb{X}$)

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & \cdots \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & \cdots \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Also, observe that the provided initial condition corresponds to the initial distribution (with the i-th entry of p_0 corresponding to the state $x = i - 1$ in $\mathbb{X}$)

$$p_0(i) = \begin{cases} 1 & \text{if } i = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Now, let us focus on the probability of x heads in t coin flips, i.e., $\mathbb{P}(X_t = x)$. Considering the fact that X_t has a *binomial distribution*, we have[8]

$$\mathbb{P}(X_t = x) = \binom{t}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{t-x} = \binom{t}{x} \frac{1}{2^t}, \quad 0 \leq x \leq t.$$

Alternatively, we can compute this probability using Lemma 2.15 as follows

$$\begin{aligned} p_t = p_0 P^t \quad &= \begin{pmatrix} 1 & 0 & 0 & \cdots \end{pmatrix} P^t \\ = p_1 P^{t-1} &= \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & \cdots \end{pmatrix} P^{t-1} \\ = p_2 P^{t-2} &= \begin{pmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & 0 & \cdots \end{pmatrix} P^{t-2} \\ = p_3 P^{t-3} &= \begin{pmatrix} \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} & 0 & 0 & \cdots \end{pmatrix} P^{t-3} \\ = \cdots . \end{aligned}$$

Observe that the row vectors on the right-hand side of the equations above are exactly the p.m.f.'s of the binomial distribution. △

2.3. Limiting and invariant distributions

We now focus on the long-term behavior of a *finite*-state MC $(\mathbb{X} = [n], P, p_0)$ with $n \in \mathbf{N}$.[9] Recall that $\mathbb{X}$ is the state space, $P \in [0, 1]^{n \times n}$ is the transition probability matrix and $p_0 \in \Delta(\mathbb{X})$ is the initial distribution. To be precise, we are interested in the asymptotic behavior of $p_t \in \Delta(\mathbb{X})$, i.e., the state distribution at

[7] For MC's with a *continuous* state space, we have to specify the transition probability *kernel*, i.e, the conditional p.d.f. $p_{X_{t+1}|X_t}(\cdot|x_t)$ over the state space $\mathbb{X}$ for each $x_t \in \mathbb{X}$.

[8] $\binom{t}{x} = \frac{t!}{x!(t-x)!}$ is the binomial coefficient.

[9] The result of this section do *not* hold for *countable*-state MC's with $\mathbb{X} = \mathbf{N}$.

time t, as $t \to \infty$. This gives us information about the fraction of time the MC spends in each state as t becomes large.

Definition 2.17 (Limiting distribution). Consider a finite-state MC ($\mathbb{X} = [n], P, p_0$) where $n \in \mathbf{N}$. $p_\infty \in \Delta(\mathbb{X})$ is called the *limiting distribution* of the MC if $p_\infty = \lim\limits_{t\to\infty} p_t$ for any $p_0 \in \Delta(\mathbb{X})$.

That is, by definition, the limiting distribution is independent of the initial distribution. So, does this limiting distribution always exist? And, if it does, how can we compute it?

Let us first try to answer these questions using the dynamics of the state distribution given in Lemma 2.15. We know $p_t = p_0 P^t$. Therefore, one way to find the limiting distribution is to analyze the behavior of P^t as t becomes large. Indeed, we have:

Lemma 2.18 (Limiting distribution I). *$p_\infty \in \Delta(\mathbb{X})$ is the limiting distribution if and only if $\lim\limits_{t\to\infty} P^t = e \cdot p_\infty$, where $e := (1, \dots, 1)^\top$ is the all-one (column) vector.*

Proof. ($\Rightarrow$) Fix $i \in [n]$. Let $P^t(i,:)$ denote the i-th row of P^t. From Definition 2.17, we have $p_\infty = \lim_{t\to\infty} p_t = \lim_{t\to\infty} p_0 P^t$ for any $p_0 \in \Delta(\mathbb{X})$. So, set the i-th entry of p_0 to one and all other entries to zero (i.e., $p_0(j) = 1$ if $j = i$ and $= 0$ otherwise) to derive $p_\infty = \lim_{t\to\infty} P^t(i,:)$. Since the latter equality holds for all $i \in [n]$, we have $\lim_{t\to\infty} P^t = e \cdot p_\infty$.

($\Leftarrow$) Using the fact that $\lim_{t\to\infty} P^t = e \cdot p_\infty$, we can write

$$\lim_{t\to\infty} p_t = \lim_{t\to\infty} p_0 P^t = p_0 \cdot e \cdot p_\infty$$

Then, since $p_0 \cdot e = 1$ (recall that p_0 is a p.m.f.), we have $\lim_{t\to\infty} p_t = p_\infty$. □

From linear algebra, we know that the powers of a matrix can be computed using its Jordan form. In particular, if the matrix is diagonalizable, computing its power can be easily done using its eigenvalue decomposition:

Remark 2.19 (Powers of a diagonalizable matrix). Assume that the matrix $P \in \mathbf{R}^{n\times n}$ is diagonalizable. That is, there exist a *diagonal* matrix Λ and an *invertible* matrix R such that

$$P = R\Lambda R^{-1} = R \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} R^{-1}.$$

Above, $\lambda_1, \dots, \lambda_n$ are the *eigenvalues* of P, the columns of $R = (r_1, \dots, r_n)$ are the corresponding *right (column) eigenvectors*, and the rows of $R^{-1} = (q_1^\top, \dots, q_n^\top)^\top$ are the corresponding *left (row) eigenvectors.* That is, $Pr_i = \lambda_i r_i$ and $q_i P = \lambda_i q_i$ for each $i \in [n]$. (Note that $q_i,\ i \in [n]$, are *row* vectors). Then,

$$\begin{aligned} P^t &= (R\Lambda R^{-1})^t = R\Lambda \underbrace{R^{-1}R}_{=I} \Lambda R^{-1} \cdots R\Lambda R^{-1} = R\Lambda^t R^{-1} \\ &= R \begin{pmatrix} \lambda_1^t & & \\ & \ddots & \\ & & \lambda_n^t \end{pmatrix} R^{-1} = \sum_{i\in[n]} \lambda_i^t (r_i \cdot q_i). \end{aligned}$$

Hence, P^t can be easily computed using the diagonal form of P.

The computation of R and R^{-1}, which may be hard, can also be avoided! (The general case of a non-diagonalizable matrix is more intricate but can be handled using the Jordan form). In this regard, observe that $P^t = \sum_{i \in [n]} \lambda_i^t (r_i \cdot q_i)$ is a *linear combination* of $\lambda_1^t, \ldots, \lambda_n^t$. Indeed, we have

$$P^t = A_1 \lambda_1^t + \cdots A_n \lambda_n^t, \quad \forall t \in \mathbf{N}_0,$$

where $A_i \coloneqq r_i \cdot q_i \in \mathbf{R}^{n \times n}$. Therefore, by explicitly calculating P^t for $t = 0, \ldots, n-1$, we can interpret the preceding equation as a system of n^3 linear equations for the n^3 entries of the matrices $A_1, \ldots, A_n$.

Finally, we note that if P is also a stochastic matrix, then by Lemma 2.12, P has all its eigenvalues within the unit disc with at least one eigenvalue equal to 1 corresponding to all-one right eigenvector. That is, we can assume $1 = \lambda_1 \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$ w.l.o.g. so that $r_1 = e \coloneqq (1, \ldots, 1)^\top$ and q_1 are the right and left eigenvectors corresponding to $\lambda_1 = 1$, respectively. Therefore, we have $P^t = e \cdot q_1 + \sum_{i=2}^n \lambda_i^t (r_i \cdot q_i)$. △

But, is there any other way to find the limiting distribution other than computing the powers of P? Yes, using the *invariant distributions* of P!

Definition 2.20 (Invariant distribution). Consider a finite-state MC ($\mathbb{X} = [n], P, p_0$) where $n \in \mathbf{N}$. $\pi \in \Delta(\mathbb{X})$ is called a *invariant distribution* of the MC if $\pi = \pi P$, that is, π is a left eigenvector of P corresponding the eigenvalue 1 such that $\pi(i) \geq 0$ for all $i \in \mathbb{X}$ and $\sum_{i \in \mathbb{X}} \pi(i) = 1$.

The preceding definition implies that if $p_{t_0} = \pi$ for some $t_0 \in \mathbf{N}_0$, then $p_t = \pi$ for all $t \geq t_0$. That is, if the MC reaches the invariant distribution, it will stay in the invariant distribution (hence, the name "invariant"). You might have already guessed why we care about invariant distributions:

Lemma 2.21 (Limiting distribution II). *The limiting distribution is an invariant distribution.*

Proof. Let p_∞ be the limiting distribution of the MC. Then,

$$p_\infty = \lim_{t \to \infty} p_t = \lim_{t \to \infty} p_{t+1} = \lim_{t \to \infty} p_0 P^{t+1} = \underbrace{\left[\lim_{t \to \infty} p_0 P^t\right]}_{p_\infty} P = p_\infty P.$$

That is, p_∞ is an invariant distribution of the MC. □

Note that the reverse statement does *not* necessarily hold, i.e., an invariant distribution is not necessarily the limiting distribution. Indeed, the limiting distribution may not even exist, however, a finite-state MC has at least one invariant distribution.[10] To make matters worse, while an invariant distribution always exists, it does *not* need to be unique, that is, an MC can have multiple invariant distributions. Whether an invariant distribution is unique and is also the limiting distribution can be determined by finding the eigenvalues of P:

Lemma 2.22 (Invariant and limiting distributions). *Consider a finite-state MC* ($\mathbb{X} = [n], P, p_0$) *where* $n \in \mathbf{N}$. *Let* $\lambda_1, \ldots, \lambda_n$ *be the eigenvalues of* P *such*

[10] The existence of an invariant distribution can be shown using the *Brouwer Fixed Point Theorem* for general MC's and using the *Perron–Frobenius theorem* for a certain class of MC's called irreducible.

that $1 = \lambda_1 \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$. *Also, let* $\pi \in \Delta(\mathbb{X})$ *be an invariant distribution of the MC, i.e., a solution to* $\pi = \pi P \in \Delta(\mathbb{X})$.

(1) Uniqueness of invariant distribution: π *is unique if and only if* $\lambda_i \neq 1$ *for all* $i \in \{2, \ldots, n\}$, *that is,* $\lambda_1 = 1$ *is a simple eigenvalue of multiplicity* 1.

(2) Existence of limiting distribution: π *is the limiting distribution of the MC if and only if* $|\lambda_i| < 1$ *for all* $i \in \{2, \ldots, n\}$, *that is,* $\lambda_1 = 1$ *is the only eigenvalue on the unit circle. Moreover, the rate of convergence of the state distribution to the limiting distribution is* $t^m |\lambda_2|^t$ *with* $1 \leq m \leq n-1$,[11] *that is, there exists a constant* $C > 0$ *such that*

$$\lim_{t\to\infty} |P^t(i,j) - \pi(j)| \leq C \cdot t^m \cdot |\lambda_2|^t, \quad \forall i, j \in \mathbb{X}.$$

Some remarks are in order regarding the preceding result:

Remark 2.23 (Invariant and limiting distributions). The necessary and sufficient condition for the *uniqueness of invariant distribution* is equivalent to the MC having only one *recurrent class* and possibly some *transient states*: A *class* is a subset $\widehat{\mathbb{X}} \subset \mathbb{X}$ of states that communicate with each other in the sense that

$$\forall i, j \in \widehat{\mathbb{X}}, \ \exists t < \infty : \ \mathbb{P}(X_t = j | X_0 = i) > 0.$$

A state $i \in \mathbb{X}$ is *recurrent* if it is visited infinitely often given $X_0 = i$, that is,

$$\mathbb{P}(X_t = i \text{ for some } t \geq 1 \mid X_0 = x) = 1.$$

A state is called *transient* if it is not recurrent. Notice that these are class properties, i.e., if a state in a class is recurrent (respectively, transient), then all the states in that class are recurrent (respectively, transient). An MC with a single (recurrent) class is called *irreducible.*

The necessary and sufficient condition for the *existence of limiting distribution* is equivalent to the MC having only one *aperiodic* recurrent class and possibly some transient states: A state $i \in \mathbb{X}$ is said to be of period t if any return to state i occurs in multiples of t time steps, i.e.,

$$t := \gcd\{t \in \mathbf{N} : \ \mathbb{P}(X_t = i | X_0 = i) > 0\}.$$

A state is called aperiodic if it is of period 1. For example, any state with a self-transition (i.e., $P(i,i) > 0$) is aperiodic. An MC is called aperiodic if all its states are aperiodic. Notice that periodicity is a class property which means that, for example, in an irreducible MC, all the states have the same periodicity. △

Let us look at the invariant and limiting distributions of the simplest MC.

Example 2.24 (The simplest MC). Consider the simplest MC with only two states as shown in Figure 10. The probability transition matrix of this MC is

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}, \quad \alpha, \beta \in [0,1].$$

The invariant distribution $\pi = (\pi(1), \pi(2)) \in \Delta([2])$ of P is is the solution to

$$\pi = \pi P, \quad \pi(1) + \pi(2) = 1, \quad \pi(1), \pi(2) \geq 0.$$

Solving the preceding equations, we derive

[11] To be precise, $m+1$ is the size of the largest Jordan block corresponding to λ_2 in the Jordan normal form of P.

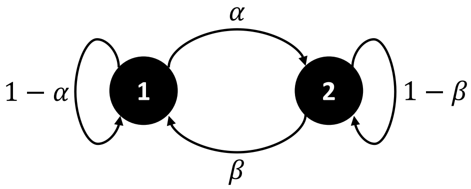


FIGURE 10. The simplest MC with two states.

(1) If $\alpha + \beta = 0$ (i.e., $\alpha = \beta = 0$), then $\pi = (\gamma, 1-\gamma)$ is a solution for any $\gamma \in [0,1]$, that is, the MC has infinitely many invariant distributions.
(2) If $\alpha + \beta \neq 0$, then there exists a unique solution $\pi = (\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta})$.

Note that while the invariant distribution always exists, it need not be unique. More importantly, even a unique invariant distribution does not imply that it is also the limiting distribution. This is indeed the case in this example. So, let us look at the eigenvalues of P, i.e., $\lambda_1 = 1$ and $\lambda_2 = 1 - \alpha - \beta$:

(1) If $\alpha + \beta = 0$, then $\lambda_1 = \lambda_2 = 1$, i.e., both eigenvalues of P are at 1. Then, by Lemma 2.22, the invariant distribution is not unique which we have already seen in case (1) above. Indeed, in this case, we have $P = I$, the identity matrix, and hence $p_t = p_0 P^t = p_0$ for all $t \in \mathbf{N}_0$. That is, the initial distribution is preserved at all times and in particular $\lim_{t\to\infty} p_t = p_0$ which depends on the initial distribution.
(2) If $\alpha + \beta \neq 0$, then P has two distinct eigenvalues with the eigenvalue $\lambda_1 = 1$ being simple. Then, by Lemma 2.22, the invariant distribution is unique which has already been computed in case (2) above. However, this unique invariant distribution may or may not be the limiting distribution:
(2a) If $\alpha + \beta = 2$ (i.e., $\alpha = \beta = 1$), then $\lambda_2 = -1$ is also on the unit circle. Thus, by Lemma 2.22, $\pi = (\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta})$ is not the limiting distribution. Indeed, in this case, we have

$$\lim_{t\to\infty} P^t = \lim_{t\to\infty} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}^t = \begin{cases} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} & \text{if } t \text{ is even,} \\ \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} & \text{if } t \text{ is odd.} \end{cases}$$

That is, the limit does not exist. For instance, if the MC starts in the state $X_0 = 1$, it will keep switching between the two states, i.e, $X_1 = 2,\ X_3 = 1,\ X_4 = 2, X_5 = 1, \ldots$.
(2a) If $\alpha + \beta < 2$ then $|\lambda_2| = |1 - \alpha - \beta| < 1$ and $\lambda_1 = 1$ is the only eigenvalue on the unit circle. Thus, by Lemma 2.22, $\pi = (\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta})$ is indeed the limiting distribution. Observe that in this case, we also have

$$P = R \begin{pmatrix} 1 & 0 \\ 0 & \lambda_2 \end{pmatrix} R^{-1} \quad \text{where} \quad R = \begin{pmatrix} 1 & \frac{-\alpha}{\alpha+\beta} \\ 1 & \frac{\beta}{\alpha+\beta} \end{pmatrix}.$$

and hence

$$\lim_{t\to\infty} P^t = \lim_{t\to\infty} \begin{pmatrix} \frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta}\lambda_2^t & \frac{\alpha}{\alpha+\beta} - \frac{\alpha}{\alpha+\beta}\lambda_2^t \\ \frac{\beta}{\alpha+\beta} - \frac{\beta}{\alpha+\beta}\lambda_2^t & \frac{\alpha}{\alpha+\beta} + \frac{\beta}{\alpha+\beta}\lambda_2^t \end{pmatrix} = e \cdot \pi.$$

Notice the rate of convergence which is exponential (i.e., $|\lambda_2|^t$). △

2.4. Markov chain with reward

In this final section, we briefly introduce the extension of the Markov chain with reward, a.k.a. Markov reward process (MRP). An MRP involves associating *rewards* to different states of the chain which are related to some kind of a "performance" measure that we are interested in. Mathematically, we define a reward function $r : \mathbb{X} \to \mathbf{R}$ (or, equivalently, a reward *column* vector $r \in \mathbf{R}^{\mathbb{X}}$) with $r(i)$ being the reward received when the process is in state $i \in \mathbb{X}$. Then, an MRP can be characterized by the tuple $(\mathbb{X}, P, r, p_0)$ with the extra element compared to an MC being the reward function r.

Example 2.25 (Queuing model – continued). Consider the queuing model of Example 2.7, where X_t is the number of customers in a queue at time t. By assigning a reward 1 to states $x > 0$ and a reward 0 to the state $x = 0$, we can use the reward signal to determine if the server is "busy." △

Having specified a reward function, one can focus on various statistics based on that reward function. A common measure is the (limiting) expected reward:

Lemma 2.26 (Expected reward). *Consider a finite-state MRP* $(\mathbb{X} = [n], P, r, p_0)$ *where* $n \in \mathbf{N}$*. The* expected reward at time t *is given by*

$$\mathbb{E}(r(X_t)) = p_0 P^t r.$$

Moreover, assuming the limiting distribution p_∞ *exists, the* limiting expected reward *is given by*

$$\lim_{t\to\infty} \mathbb{E}(r(X_t)) = p_\infty \cdot r.$$

Proof. The results follow from Lemma 2.15 and Definition 2.17. □

CHAPTER 3

Dynamic Programming

In this chapter, we look at controlled stochastic processes by introducing Markov decision processes and their optimal control problem. We then look at the dynamic programming algorithm as the standard method for solving these problems.

3.1. Markov decision process

We started our journey with the recursion $X_{t+1} = f(X_t, W_{t+1})$ with the state variable X_t, taking values in the state space $\mathbb{X}$, and the disturbance variable W_t with a given distribution, which was the source of stochasticity. We now make things a lot more interesting by adding the "control" input (a.k.a. action, decision) variable U_t, taking values in the action space $\mathbb{U}$, to the dynamics and considering the recursion

$$X_{t+1} = f(X_t, U_t, W_{t+1}). \tag{3.1}$$

This is called a *controlled* stochastic process. The variable U_t, as the name suggests, can be manipulated for the purpose of steering the process in a desired fashion. Our goal is to find "the best" control actions to do so. But, best for what? Of course, we need to specify an objective. To that end, for a given finite *planning horizon* $T \in \mathbf{N}$, we define

- the *running* (a.k.a. stage) cost function $g : \mathbb{X} \times \mathbb{U} \to \mathbf{R}$ such that $g(x, u)$ is the cost of taking the action u in state x at time $t \in \{0, 1, \ldots, T-1\}$, and
- the *terminal* cost function $G : \mathbb{X} \to \mathbf{R}$ such that $G(x)$ is the cost of the process being in state x at the end of planning horizon $t = T$.

Let us assume that the system is initially in some state $X_0 = x \in \mathbb{X}$. Given these functions, the cost of the state sequence $(X_t)_{t=0}^{T} = (X_0 = x, X_1, \ldots, X_T)$ in response to the control sequence $(U_t)_{t=0}^{T-1} = (U_0, \ldots, U_{t-1})$ is derived by adding up the costs over the trajectory, i.e,

$$\sum_{t=0}^{T-1} g(X_t, U_t) + G(X_T).$$

See Figure 1. Note that this cost is a *random* variable. So, if we wish to minimize such a random cost, we need to work with a statistic of this cost. We choose to work with the *expected* accumulated cost

$$\mathbb{E}\left(\sum_{t=0}^{T-1} g(X_t, U_t) + G(X_T) \;\middle|\; X_0 = x \right).$$

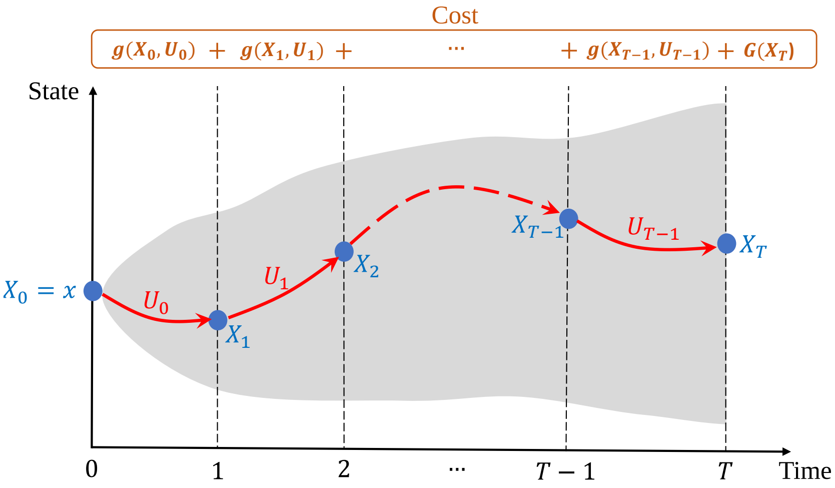


FIGURE 1. The cost of the state trajectory $(X_0, \ldots, X_T)$ in response to the action trajectory $(U_0, U_1 \ldots, U_{T-1})$ from the initial state $X_0 = x$. The optimal control problem involves finding a control policy with the minimum expected cost.

The objective is then to find a control sequence that minimizes this cost, that is,[1]

$$\min_{(U_t)_{t=0}^{T-1} \in \mathbb{U}^T} \mathbb{E}\left(\sum_{t=0}^{T-1} g(X_t, U_t) + G(X_T) \,\middle|\, X_0 = x \right).$$

This is indeed a *sequential* decision-making problem under *uncertainty*. In particular, it does not make sense to look for an *open-loop* control law, i.e., a predetermined sequence of control actions $U_t = u_t^\star \in \mathbb{U}$ for $t \in \{0, 1, \ldots, T-1\}$. The main reason is stochasticity which prevents us from being certain about the state of the system in the future. By committing to the open-loop control sequence $(u_0^\star, \ldots, u_{T-1}^\star)$, we are disregarding the actual realization of the states over the planning horizon in the sense that each action $u_t^\star$ should be optimal for any possible state $X_t \in \mathbb{X}$ in which the process can end up in time t. But, why do that when we can make decisions sequentially as the state of the system is revealed. So, U_0 should be taken at time $t = 0$ based on the state $X_0 = x$ of the system at that time. Similarly, we should decide about U_1 based on the information up to time $t = 1$, and so on. This means that we need to look for *closed-loop* control laws

$$U_t = \mu_t(X_t, U_{t-1}, X_{t-1}, \ldots, U_0, X_0) \in \mathbb{U}, \quad t \in \{0, 1, \ldots, T-1\},$$

called a *(control) policy*, that allows us to take actions in response to the particular trajectory visited up time t. OK! This is catastrophic computationally speaking: We need to find functions that map each possible trajectory

$$(x_t, u_{t-1}, x_{t-1}, \ldots, u_0, x_0) \in \mathbb{X} \times (\mathbb{U} \times \mathbb{X})^t,$$

to an action $u_t \in \mathbb{U}$ for each $t \in \{0, 1, \ldots, T-1\}$![2]

[1] Strictly speaking, this minimization and the similar ones that follow are subject to the dynamics (3.1) for the given probability distributions of W_t's.

[2] The situation is worse than that! Actually, the control policy $\mu_0, \mu_1, \ldots, \mu_{T-1}$ can in general be *stochastic* with $\mu_t(\cdot | X_t, U_{t-1}, X_{t-1}, \ldots, U_0, X_0) \in \Delta(\mathbb{U})$ being a *distribution* over the action space $\mathbb{U}$. Using this policy, we take the action $U_t \sim \mu_t$, i.e., an action chosen at random with

This is why we focus on a specific class of controlled stochastic processes in which the disturbances W_{t+1} are *conditionally independent given* X_t *and* U_t. Such a process is called a *Markov decision process* (MDP). In particular, the process is called an MDP because the only relevant information at each time t concerning the future distribution of the state is the current state-action pair (X_t, U_t). Luckily, the optimal policies of MDPs are much simpler:

THEOREM 3.1. *An MDP has a* Markov (i.e., memory-less) *optimal policy of the form* $U_t = \mu_t(X_t)$ *for all* $t \in \{0, 1, \ldots, T-1\}$.

That is, there exists an optimal policy in the form of simple state-feedback laws that depend only on the current state X_t. This should not be surprising. After all, we are talking about MDPs in which at time t, the future evolution of the state only depends on the current state X_t and action U_t. So, the best action $U_t^\star$ that we can take at time t in order to minimize the expected cost in the future should only depend on the current state X_t. We note that under a Markov policy, an MDP reduces to a Markov chain. To summarize, in order to find an optimal Markov policy, we need to solve the optimization problem

$$\min_{(\mu_t:\mathbb{X}\to\mathbb{U})_{t=0}^{T-1}} \mathbb{E}\left(\sum_{t=0}^{T-1} g\big(X_t, \mu_t(X_t)\big) + G(X_T) \;\middle|\; X_0 = x\right)$$

Observe that even with a Markov policy, looking for closed-loop control laws (i.e., functions) is still much more computationally demanding compared to open-loop control laws. To illustrate, for a finite MDP with n states and m actions ($n, m \in \mathbf{N}$), given the initial state $X_0 = x_0$, the total number of open-loop control laws is m^T, while the total number of closed-loop control laws is $m^{1+(T-1)n}$. Let us illustrate this through an example: Say, $n = 4$, $m = 2$, and $T = 10$. An MDP of this size should be definitely manageable! But these numbers mean $2^{10} \approx 10^3$ open-loop control laws and $2^{37} \approx 10^{11}$ closed-loop control laws! We will revisit this example later in this chapter.

Here comes the formal definition of the optimal control problem of an MDP:

DEFINITION 3.2 (Stochastic optimal control). Consider a Markov decision process (MDP) with

- planning horizon $T \in \mathbf{N}$,
- state space $\mathbb{X}$ and action space $\mathbb{U}$,
- dynamics
$$X_{t+1} = f(X_t, U_t, W_{t+1}), \quad t = 0, 1, \ldots, T-1, \tag{3.2}$$
where $X_t \in \mathbb{X}$ is the state process, $U_t \in \mathbb{U}$ is the control process, and $W_t \in \mathbb{W}$ is the disturbance process assumed to be conditionally independent given X_t and U_t with probability distribution $p_{W_{t+1}|X_t,U_t} \in \Delta(\mathbb{W})$,
- running cost $g : \mathbb{X} \times \mathbb{U} \to \mathbf{R}$ and terminal cost $G : \mathbb{X} \to \mathbf{R}$.

Consider the class

$$\mathcal{C} = \left\{\mu \ : \ \mu = (\mu_t : \mathbb{X} \to \mathbb{U})_{t=0}^{T-1}\right\},$$

distribution μ_t in response to the trajectory visited up time t. However, we do not need to worry about this because, in the class of problems we discuss, the existence of a *deterministic* optimal policy is theoretically guaranteed.

of all *admissible*, deterministic, Markov control policies. The expected cost of a policy $\mu \in \mathcal{C}$ from the initial state $X_0 = x \in \mathbb{X}$ is then

$$J_0^\mu(x) \coloneqq \mathbb{E}\left(\sum_{t=0}^{T-1} g\big(X_t, \mu_t(X_t)\big) + G(X_T) \;\middle|\; X_0 = x\right).$$

The problem of interest is to find an *optimal policy* $\mu^\star = (\mu_t^\star)_{t=0}^{T-1}$ that minimizes this cost, i.e.,[3]

$$J_0^{\mu^\star}(x) = \min_{\mu \in \mathcal{C}} J_0^\mu(x), \quad \forall x \in \mathbb{X}.$$

with $J_0^\star \coloneqq J_0^{\mu^\star} : \mathbb{X} \to \mathbf{R}$ being the *optimal cost-to-go* (a.k.a. value function) at $t = 0$. △

Some remarks are in order regarding the setup described above. First, understanding the sequential flow of information is very important: At each time t, we observe the state X_t, we take the control action $U_t = \mu_t(X_t)$ and then move to the next state $X_{t+1} = f(X_t, U_t, W_{t+1})$. In particular, the disturbance W_{t+1} is only revealed after our commitment to the control action U_t. This is exactly the reason for our somewhat unconventional notation for the time index of the disturbance in the dynamics of the system. (You are most probably familiar with the standard notation $X_{t+1} = f(X_t, U_t, W_t)$).

Second, similar to Markov chains, the dynamics of an MDP can be alternatively described via the transition probabilities. For example, for a finite MDP with state space $\mathbb{X} = [n] \coloneqq \{1, \dots, n\}$ and action space $\mathbb{U} = [m]$ with $n, m \in \mathbf{N}$, we can define a transition probability matrix P_k for each action $k \in \mathbb{U}$ with entries

$$P_k(i,j) = \mathbb{P}(X_{t+1} = j \mid X_t = i,\ U_t = k), \quad \forall (i,j,k) \in \mathbb{X} \times \mathbb{X} \times \mathbb{U},$$

that is, $P_k(i,j)$ is the probability of the transition to state j from state i under the control action k. Then, the family $\{P_k\}_{k \in \mathbb{U}}$ fully describes the dynamics of the MDP. Indeed, this description is equivalent to the recursion[4]

$$X_{t+1} = f(X_t, U_t, W_{t+1}) = W_{t+1},$$

where the process $(W_t)_{t=1}^T$ takes values in $\mathbb{X}$ and is (conditionally) independent with distribution

$$p_{W_{t+1}|X_t,U_t}(w|x,u) = P_u(x,w), \quad \forall (x,u,w) \in \mathbb{X} \times \mathbb{U} \times \mathbb{X}.$$

Finally, we note that while we focus on minimizing costs, the same formulation and results hold for the case when our goal is to maximize rewards. This is commonly the case in the reinforcement learning literature. Just use reward functions instead of cost functions and maximization instead of minimization.

Let us look at an example:

Example 3.3 (Machine replacement). We want to rent a machine to use in a production line over a fixed number of periods of operation. The problem is that the state of the machine can deteriorate after each period to the extent that it can break and stop operating. We can take the machine out of the line at the beginning of each period and carry out some maintenance to avoid that. Of course,

[3] Strictly speaking, we have to use infimum because the minimum may not be attained. In these notes, we are going to assume that the minimum is attained in all the provided optimization problems.

[4] We hope that this justifies our choice of time indices!

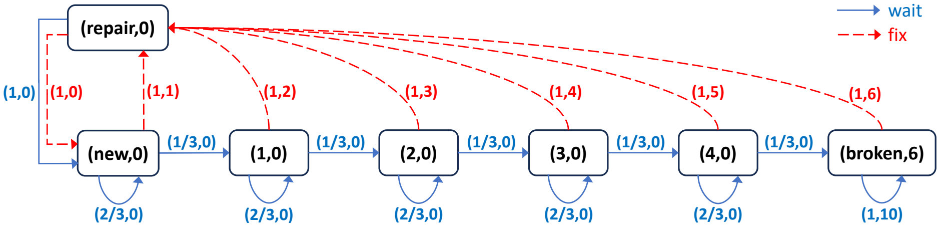


Figure 2. The graphical representation of the machine replacement problem of Example 3.3. The rectangles show the states with the label (x, G) denoting the state x and the corresponding terminal cost G. The arcs show the transitions with the weight (p, g) denoting the probability p and the running cost g of the corresponding transition.

if the machine is in a better condition, the cost of maintenance will be lower. We have to return the machine at the end and if it is broken when we return it, we incur a large cost.

To be more precise, the machine has states 'repair', 'new', '1', '2', '3', '4', and 'broken', ordered by deteriorating operating conditions. In each period, we can either 'fix' the machine at some cost or 'wait'. In the 'broken' state, the decision to 'wait' incurs a large penalty. The graphical representation of the system is shown in Figure 2 including the probability and the running cost of the transitions between states and the terminal cost of each state. There is no terminal cost unless the machine is 'broken' in which case we incur a large penalty of 6. The question we are interested in is *when we should repair the machine to minimize the expected costs if the machine is initially at state '1' and used for 10 periods.*

Let us look at this question closely. First note that the planning horizon is really important for this question to have a non-trivial answer. For example, if we were to use the machine for only 3 periods, the answer would be clear: Never fix it! Because the worst thing that can happen is for the machine to end up in the state '4' at time $t = 3$. But with a planning horizon of 10, there is a good chance that the machine 'breaks' down at some point. Second, let us again see why open-loop control is meaningless for this problem. Say, we commit to an open-loop control law at time $t = 0$ that says we have to 'wait' at time $t = 4$. But there is a non-zero probability ($\frac{1}{3^4}$, to be exact) that the machine is 'broken' at that time! 'Waiting' clearly cannot be the optimal decision if that happens to be the case. What if the open-loop policy says that we have to 'fix' at time $t = 4$? Again, the machine can be still in state '1' at time $t = 4$ with a non-zero probability. Should we really 'fix' the machine if that happens to be the case? The problem is that we cannot be sure about the state of the machine in the future and it does not make sense to commit to any future action in advance. We should indeed take into account the particular realization of the state and make decisions *sequentially* based on the most current state of the machine. That is exactly closed-loop control.

The first step is the standard formulation of the problem:

- Planning horizon: $T = 10$
- State space: $\mathbb{X} = \{\text{repair}, \text{new}, 1, 2, 3, 4, \text{broken}\}$
- Action space: $\mathbb{U} = \{\text{wait}, \text{fix}\}$

- Dynamics: In this case, it is easier to work with transition probability matrices of each action as follows:

$P_{\text{fix}}(x_t, x_{t+1})$

$x_t \backslash x_{t+1}$	repair	new	1	2	3	4	broken
repair	0	1	0	0	0	0	0
new/1/2/3/4/broken	1	0	0	0	0	0	0

$P_{\text{wait}}(x_t, x_{t+1})$

$x_t \backslash x_{t+1}$	repair	new	1	2	3	4	broken
repair	0	1	0	0	0	0	0
new	0	2/3	1/3	0	0	0	0
1	0	0	2/3	1/3	0	0	0
2	0	0	0	2/3	1/3	0	0
3	0	0	0	0	2/3	1/3	0
4	0	0	0	0	0	2/3	1/3
broken	0	0	0	0	0	0	1

- Running cost:

$g(x, u)$

$u \backslash x$	repair	new	1	2	3	4	broken
wait	0	0	0	0	0	0	10
fix	0	1	2	3	4	5	6

- Terminal cost: $G(x) = \begin{cases} 6 & \text{if } x = \text{broken}, \\ 0 & \text{otherwise.} \end{cases}$

Now that the problem data are specified, we can write down the optimal control problem: We are looking for the optimal control policy that minimizes the expected cost if the machine is initially at state '1' and used for 10 periods, i.e.,

$$J_0^\star(1) = \min_{(\mu_t : \mathbb{X} \to \mathbb{U})_{t=0}^{9}} \mathbb{E}\left(\sum_{t=0}^{9} g\big(X_t, \mu_t(X_t)\big) + G(X_{10}) \;\middle|\; X_0 = 1 \right).$$

We will come back to the solution of this problem in the next section. △

Let us emphasize that modeling, i.e., formulating the optimal control problem, is one of the most difficult parts of real applications (and exams). Among different elements of the problem formulation, proper identification of the state variable is of utmost importance: It should contain all the necessary and sufficient information required for the process to be Markovian.

3.2. Dynamic programming algorithm

Let us recall the problem that we are trying to solve:

$$J_0^\star(x) := \min_{(\mu_k : \mathbb{X} \to \mathbb{U})_{k=0}^{T-1}} \mathbb{E}\left(\sum_{k=0}^{T-1} g\big(X_k, \mu_k(X_k)\big) + G(X_T) \;\middle|\; X_0 = x \right), \quad \forall x \in \mathbb{X}. \quad (\mathcal{P}_0)$$

We denote this problem by $\mathcal{P}_0$. As we discussed, this is a difficult problem since we are optimizing in the space of functions, that is, Markov policies $(\mu_k)_{k=0}^{T-1}$ that map states to actions at each stage of the problem. So, how can we solve this problem? Well, the trick is to use the *dynamic programming algorithm* and solve the problem by *backward induction*! To explain this idea, let us also define the *tail* problem $\mathcal{P}_t$

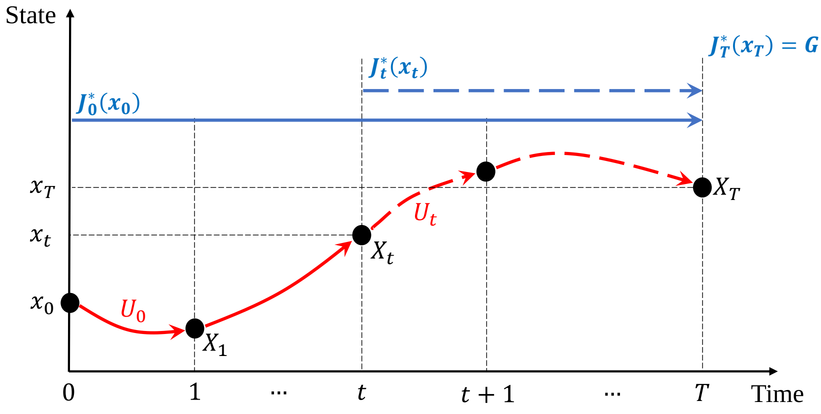


FIGURE 3. The original problem $\mathcal{P}_0$ starting from $k = 0$ and the corresponding *tail* problem $\mathcal{P}_t$ starting from $k = t$.

whereby we minimize the *cost-to-go starting from state* $X_t = x$ *at some time* $t \geq 0$ *to time* T:

$$J_t^\star(x) := \min_{(\mu_k : \mathbb{X} \to \mathbb{U})_{k=t}^{T-1}} \mathbb{E}\left(\sum_{k=t}^{T-1} g\big(X_k, \mu_k(X_k)\big) + G(X_T) \;\middle|\; X_t = x \right), \quad \forall x \in \mathbb{X}. \quad (\mathcal{P}_t)$$

Notice that the only difference is in the starting point; solving $\mathcal{P}_t$ gives us the optimal expected cost-to-go from time t onward; see Figure 3. Also, observe that, by definition, $\mathcal{P}_{t=0}$ is indeed the original problem and $\mathcal{P}_{t=T}$ corresponds to $J_T^\star(x) = G(x)$ for all $x \in \mathbb{X}$.

The important point to notice is that the solutions of the original problem $\mathcal{P}_0$ and the corresponding *tail* problem $\mathcal{P}_t$ are related. Indeed, by solving $\mathcal{P}_0$ and finding the optimal policies for the entire horizon $\{0, \ldots, T\}$, we also solve $\mathcal{P}_t$ for the tail $\{t, \ldots, T\}$. This is known as the *principal of optimally*:

PRINCIPLE OF OPTIMALITY (Bellman, 1957). *If the policy* $(\mu_k^\star)_{k=0}^{T-1}$ *is optimal in the problem* $\mathcal{P}_0$*, then the* tail policy $(\mu_k^\star)_{k=t}^{T-1}$ *is optimal in the* tail problem $\mathcal{P}_t$ *for all* $t \in \{0, 1, \ldots, T-1\}$.

Let us think about this principle and what it says. It does make a lot of sense! If the tail policy $(\mu_k^\star)_{k=t}^{T-1}$ is *not* an optimal solution to $\mathcal{P}_t$, then while applying $(\mu_k^\star)_{k=0}^{T-1}$ for solving $\mathcal{P}_0$, when we reach to the step $k = t$, we should *not* follow $(\mu_k^\star)_{k=t}^{T-1}$ for the following steps! Let us illustrate this through an example: Imagine that you want to take the train from Delft to Amsterdam to go to a concert with one of your friends who lives in Leiden. You consider all possible routes to get to Amsterdam and find the 'best' one for yourself. Interestingly, the 'best' route for you happens to make a stop at Leiden. Now, when you arrive at Leiden, you see your friend waiting at the station and not getting on your train. But then, when you arrive in Amsterdam and get out of the station, you notice that your friend is already there! So, was your plan really the 'best' plan?

So, how can we use this principle? We are going to use it the other way around. That is, we are going the use the fact that the optimal solution to $\mathcal{P}_t$ is the tail of the optimal solution to $\mathcal{P}_0$. Or, more generally, for $t' > t$, the solution to $\mathcal{P}_{t'}$ gives us the tail of the solution to $\mathcal{P}_t$. Simply put, based on this principle, if we know how to make the best decisions from the next stage onward, all we need to worry about is the current decision. This means we can use backward induction to solve these problems from $t = T-1$ to $t = 0$ till we have the solution to the original problem $\mathcal{P}_0$. This is called the *dynamic programming algorithm*:

ALGORITHM 3.4 (Dynamic programming algorithm). *The dynamic programming algorithm (DPA) is as follows:*
(1) Initialization: *Set $t = T$ and the* value function $J_T : \mathbb{X} \to \mathbf{R}$ *as*

$$J_T(x) = G(x), \quad \forall x \in \mathbb{X}.$$

(2) Backward iteration: *Count backwards from $t = T-1$ to $t = 0$ and find the* value function $J_t : \mathbb{X} \to \mathbf{R}$ *by solving*[5]

$$\begin{aligned} J_t(x) &= \min_{u \in \mathbb{U}} \big\{ g(x,u) + \mathbb{E}\left(J_{t+1}(X_{t+1}) \;\middle|\; X_t = x, U_t = u \right) \big\} \\ &= \min_{u \in \mathbb{U}} \big\{ g(x,u) + \mathbb{E}\left(J_{t+1}\big(f(x,u,W_{t+1})\big) \;\middle|\; X_t = x, U_t = u \right) \big\} \\ &= \min_{u \in \mathbb{U}} \big\{ g(x,u) + \sum_{x_+ \in \mathbb{X}} P_u(x,x_+) J_{t+1}(x_+) \big\}, \quad \forall x \in \mathbb{X}, \end{aligned}$$

and set the policy $\mu_t : \mathbb{X} \to \mathbb{U}$ *to be the corresponding minimizer, i.e.,*

$$\mu_t(x) = \operatorname*{argmin}_{u \in \mathbb{U}} \big\{ g(x,u) + \mathbb{E}\left(J_{t+1}(X_{t+1}) \;\middle|\; X_t = x, U_t = u \right) \big\}, \quad \forall x \in \mathbb{X}.$$

△

Note that each iteration t of DPA involves finding the *function* $J_t : \mathbb{X} \to \mathbf{R}$, that is, the optimal cost-to-go at time t for *all* $x \in \mathbb{X}$. Here is the formal statement and proof of the fact that DPA outputs the optimal costs-to-go and policy:

LEMMA 3.5 (Dynamic programming algorithm). *The DP Algorithm 3.4 outputs the optimal costs-to-go and an optimal policy, that is, $(J_t)_{t=0}^{T-1} = (J_t^\star)_{t=0}^{T-1}$ and $(\mu_t)_{t=0}^{T-1} = (\mu_t^\star)_{t=0}^{T-1}$.*

PROOF. We prove by *induction on t* that $J_t(x) = J_t^*(x)$ for all $x \in \mathbb{X}$:[6]
Base step ($t = T$): By construction, we have

$$J_T(x) = G(x) = J_T^*(x), \quad \forall x \in \mathbb{X}.$$

Induction step ($t \in \{T-1, \dots, 0\}$): Assume that $J_{t+1}(x) = J_{t+1}^*(x)$ for all $x \in \mathbb{X}$ (induction hypothesis) and recall the problem $\mathcal{P}_t$ given by

$$J_t^*(x) = \min_{(\mu_k)_{k=t}^{T-1}} \mathbb{E}\left(\sum_{k=t}^{T-1} g\big(X_k, \mu_k(X_k)\big) + G(X_T) \;\middle|\; X_t = x \right), \quad \forall x \in \mathbb{X}. \tag{3.3}$$

Fix $x \in \mathbb{X}$ and observe that

$$J_t^*(x) = \min_{\mu_t(x),\ (\mu_k)_{k=t+1}^{T-1}} \mathbb{E}\left(g\big(x, \mu_t(x)\big) + \sum_{k=t+1}^{T-1} g\big(X_k, \mu_k(X_k)\big) + G(X_T) \;\middle|\; X_t = x \right)$$

[5] The third equality is under the assumption that the state space $\mathbb{X}$ is finite.

[6] It is of utmost importance that you understand the concept of proof by induction and how to apply it as it will be used extensively in the rest of these notes.

$$= \min_{\mu_t(x),\ (\mu_k)_{k=t+1}^{T-1}} g\big(x, \mu_t(x)\big) + \mathbb{E}\left(\sum_{k=t+1}^{T-1} g\big(X_k, \mu_k(X_k)\big) + G(X_T) \;\middle|\; X_t = x\right),$$

where for the last equality we used the fact that $g\big(X_t, \mu_t(X_t)\big) = g\big(x, \mu_t(x)\big)$ is not random given that $X_t = x$. Then, by total law of expectation (a.k.a. tower rule),[7] we have

$$\begin{aligned} J_t^*(x) = \min_{\mu_t(x),\ (\mu_k)_{k=t+1}^{T-1}} & \Big\{ g\big(x, \mu_t(x)\big) \\ & + \mathbb{E}\Big(\mathbb{E}\Big[\sum_{k=t+1}^{T-1} g\big(X_k, \mu_k(X_k)\big) + G(X_T) \mid X_{t+1}\Big] \Bigm| X_t = x\Big)\Big\}. \end{aligned}$$

Above, the outer expectation is w.r.t. X_{t+1} and the inner expectation is w..r.t. the tail state-action trajectory $(U_{t+1}, X_{t+2}, U_{t+2}, \ldots, X_T)$. Now, observe that $g\big(x, \mu_t(x)\big)$ and the outer expectation (w.r.t. X_{t+1}) does not affect $(\mu_k)_{k=t+1}^{T-1}$ and hence we can split the minimization and write

$$\begin{aligned} J_t^*(x) = \min_{\mu_t(x)} & \Big\{ g\big(x, \mu_t(x)\big) \\ & + \mathbb{E}\Big(\min_{(\mu_k)_{k=t+1}^{T-1}} \mathbb{E}\Big[\sum_{k=t+1}^{T-1} g\big(X_k, \mu_k(X_k)\big) + G(X_T) \mid X_{t+1}\Big] \Bigm| X_t = x\Big)\Big\} \end{aligned}$$

What we managed to show so far is the principle of optimality: The optimal solution $(\mu_k^\star)_{k=t+1}^{T-1}$ to the tail problem

$$J_{t+1}^*(X_{t+1}) = \min_{(\mu_k)_{k=t+1}^{T-1}} \mathbb{E}\Big[\sum_{k=t+1}^{T-1} g\big(X_k, \mu_k(X_k)\big) + G(X_T) \mid X_{t+1}\Big]$$

constitutes the tail portion of the optimal solution to (3.3). Finally, we replace the control law $\mu_t(x)$ with the control input u to derive

$$\begin{aligned} J_t^*(x) =& \min_u \big\{ g(x,u) + \mathbb{E}\big(J_{t+1}^*(X_{t+1}) \mid X_t = x,\ U_t = u\big)\big\} \\ =& \min_u \big\{ g(x,u) + \mathbb{E}\big(J_{t+1}(X_{t+1}) \mid X_t = x,\ U_t = u\big)\big\} \\ =& J_t(x) \end{aligned}$$

where the last two equations follow from the induction hypothesis and the definition of J_t, respectively. This completes the proof of the equality $(J_t)_{t=0}^{T-1} = (J_t^\star)_{t=0}^{T-1}$. Replacing min by argmin, $J_t^\star$ by $\mu_t^\star$, and J_t by μ_t in the above arguments, one also has $(\mu_t)_{t=0}^{T-1} = (\mu_t^\star)_{t=0}^{T-1}$. □

Let us use DPA to solve the machine replacement problem:

Example 3.6 (Machine replacement – continued). Consider again the problem described in Example 3.3. We are looking for the optimal policy on when to repair a machine initially in state '1' used for 10 periods. We can solve this problem using DPA. The algorithm in particular involves working our way backward in time and filling the entries of the following table:

[7] $\mathbb{E}(Y) = \mathbb{E}\big(\mathbb{E}(Y|Z)\big)$

$J_t(x)/\mu_t(x)$							
$t\backslash x$	repair	new	1	2	3	4	broken
10	$*$	$*$	$*$	$*$	$*$	$*$	$*$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
0	$*$	$*$	$*$	$*$	$*$	$*$	$*$

Initialization: We set $t = T = 10$ and

$$J_{10}(x) = G(x) = \begin{cases} 6 & \text{if } x = \text{broken}, \\ 0 & \text{otherwise.} \end{cases}$$

Backward iteration: For $t = T - 1 = 9$ to $t = 0$, we set

$$\begin{aligned} J_t(x) &= \min_{u \in \mathbb{U}} \left\{ g(x,u) + \mathbb{E}\left(J_{t+1}(X_{t+1}) \;\middle|\; X_t = x, U_t = u \right) \right\} \\ &= \min_{u \in \mathbb{U}} \Big\{ g(x,u) + \sum_{x_+ \in \mathbb{X}} P_u(x, x_+) J_{t+1}(x_+) \Big\}, \quad \forall x \in \mathbb{X}. \end{aligned}$$

So, for $t = 9$, we have

$$J_9(\text{repair}) = J_{10}(\text{new}) = 0, \qquad \mu_9(\text{repair}) = \text{wait}$$

$$J_9(\text{new}) = \min\{\underbrace{1 + J_{10}(\text{repair})}_{\text{fix}}, \underbrace{(2/3)J_{10}(\text{new}) + (1/3)J_{10}(1)}_{\text{wait}}\} = 0, \qquad \mu_9(\text{new}) = \text{wait}$$

$$\vdots \qquad\qquad \vdots$$

$$J_9(4) = \min\{\underbrace{5 + J_{10}(\text{repair})}_{\text{fix}}, \underbrace{(2/3)J_{10}(4) + (1/3)J_{10}(\text{broken})}_{\text{wait}}\} = 2, \qquad \mu_9(4) = \text{wait}$$

$$J_9(\text{broken}) = \min\{\underbrace{6 + J_{10}(\text{repair})}_{\text{fix}}, \underbrace{10 + J_{10}(\text{broken})}_{\text{wait}}\} = 6, \qquad \mu_9(\text{broken}) = \text{fix}$$

And, for $t = 8$, we have

$$J_8(\text{repair}) = J_9(\text{new}) = 0, \qquad \mu_8(\text{repair}) = \text{wait}$$

$$J_8(\text{new}) = \min\{\underbrace{1 + J_9(\text{repair})}_{\text{fix}}, \underbrace{(2/3)J_9(\text{new}) + (1/3)J_9(1)}_{\text{wait}}\} = 0, \qquad \mu_8(\text{new}) = \text{wait}$$

$$\vdots \qquad\qquad \vdots$$

$$J_8(3) = \min\{\underbrace{4 + J_9(\text{repair})}_{\text{fix}}, \underbrace{(2/3)J_9(3) + (1/3)J_9(4)}_{\text{wait}}\} = 0.67, \qquad \mu_8(3) = \text{wait}$$

$$J_8(4) = \min\{\underbrace{5 + J_9(\text{repair})}_{\text{fix}}, \underbrace{(2/3)J_9(4) + (1/3)J_9(\text{broken})}_{\text{wait}}\} = 3.33, \qquad \mu_8(4) = \text{wait}$$

$$J_8(\text{broken}) = \min\{\underbrace{6 + J_9(\text{repair})}_{\text{fix}}, \underbrace{10 + J_9(\text{broken})}_{\text{wait}}\} = 6, \qquad \mu_8(\text{broken}) = \text{fix}$$

Continuing in the same manner for $t = 7, 6, \ldots, 0$, we can find $J_t(x)$ for all t and x and the corresponding policy $\mu_t(x)$ as follows ('w' stands for 'wait' and and '**f**' stands for 'fix')

$J_t(x)/\mu_t(x)$							
$t\backslash x$	repair	new	1	2	3	4	broken
10	0.00/-	0.00/-	0.00/-	0.00/-	0.00/-	0.00/-	6.00/-
9	0.00/w	0.00/w	0.00/w	0.00/w	0.00/w	2.00/w	6.00/**f**
8	0.00/w	0.00/w	0.00/w	0.00/w	0.67/w	3.33/w	6.00/**f**
7	0.00/w	0.00/w	0.00/w	0.22/w	1.56/w	4.22/w	6.00/**f**
6	0.00/w	0.00/w	0.07/w	0.67/w	2.44/w	4.81/w	6.00/**f**
5	0.00/w	0.02/w	0.27/w	1.26/w	3.23/w	5.00/**f**	6.00/**f**
4	0.02/w	0.11/w	0.60/w	1.92/w	3.82/w	5.00/**f**	6.00/**f**
3	0.11/w	0.27/w	1.04/w	2.55/w	4.02/**f**	5.02/**f**	6.02/**f**
2	0.27/w	0.53/w	1.54/w	3.04/w	4.11/**f**	5.11/**f**	6.11/**f**
1	0.53/w	0.87/w	2.04/w	3.27/**f**	4.27/**f**	5.27/**f**	6.27/**f**
0	0.87/w	1.26/w	2.45/w	3.53/**f**	4.53/**f**	5.53/**f**	6.53/**f**

Therefore, if the machine is initially in state '1', the minimal expected cost under the optimal policy is given by $J_0(1) = 2.45$.

So, how do we use the derived optimal policy? Well, we simply apply it in the forward (in time) run of the system. As an illustration, Figure 4 shows the result of the 1000 forward simulations of the system from initial state $X_0 = 1$ under the optimal policy. Observe that, due to stochasticity, the state of the system at $t = 1$ can be either $X_1 = 1$ or $X_1 = 2$. In the first case, the optimal policy tells us to 'wait' ($\mu_1(1) = \text{wait}$), while in the second case, we should 'fix' the machine ($\mu_1(2) = \text{fix}$). Moreover, note that the actual cost of the individual realized trajectories can be different from $J_0(1) = 2.45$. For example, in almost half of the simulations reported in Figure 4, the cost of the realized trajectory is zero. The empirical *average* of the cost of the realized trajectories on the other hand is close to $J_0(1) = 2.45$ since each element $J_t(x)$ of the table above is indeed the optimal *expected* cost-to-go from state x at time t.

Finally, an interesting observation is that the optimal policy is in the form of a threshold which requires the time-state pairs at the bottom-left corner of the table above to be 'fixed.'[8] How do you think the optimal policy would look like for planning horizon $T = 20$? What about $T = 100$? Will this threshold behavior in the optimal policy prevail? $\triangle$

Now that we understand the mechanics of the DPA, let us discuss some straightforward extensions. Indeed, the provided DPA can be easily extended to handle more general problems than the one provided in Definition 3.2. In particular,

- the running cost $g = g_t$ and the dynamics $f = f_t$ can be *time-varying* and hence change from one step to another: all we need to do in this case is to use the corresponding cost and dynamics at each iteration of DPA;

[8] To be honest, the model that we considered in this example is too simple and quite far from reality. But, as George Box once said, "All models are wrong, some are useful." The usefulness of our model is indeed in this powerful insight about the threshold behavior in the optimal policy.

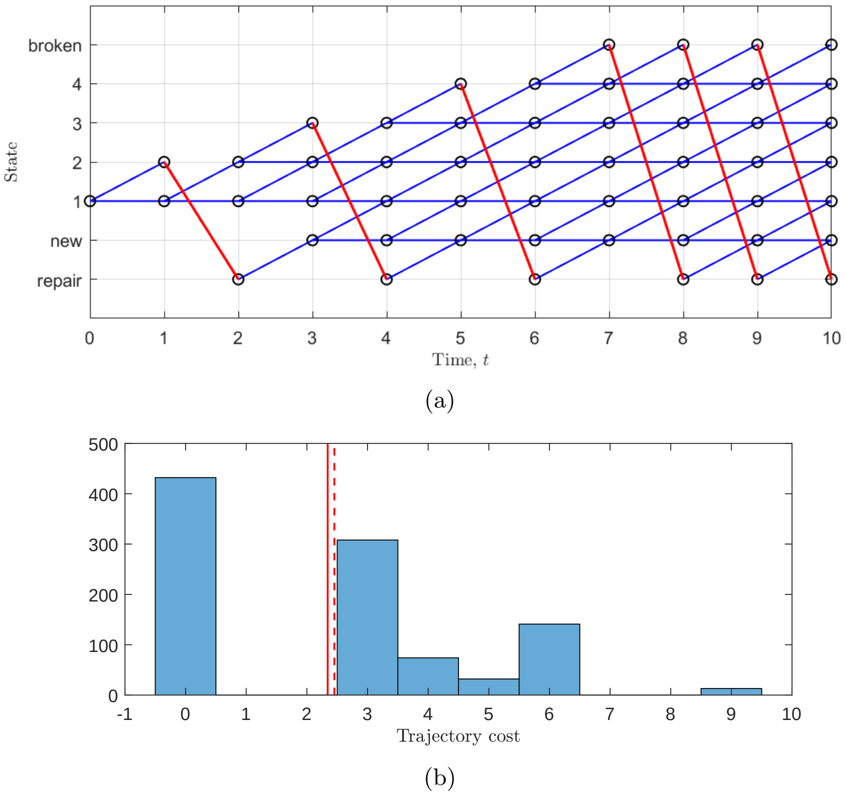


FIGURE 4. Forward simulation of the machine replacement problem of Example 3.6 form initial state $X_0 = 1$ under the optimal policy: (a) 1000 realizations of the state trajectory – the blue and red lines correspond to the actions 'wait' and 'fix', respectively; (b) the histogram of the cost of the realized trajectories – the solid red line shows the average of the 1000 runs and the dashed red line shows $J_0(1)$.

- the running cost $g = g(x, u, w)$ can be *stochastic* and depend on the disturbance variable w: in this case, we need to consider the *expected value of the running cost* in the DPA;
- there can be *state-dependent input constraints*, that is, the action space $\mathbb{U} = \mathbb{U}(x)$ can depend on the state x of the system, and hence any control policy must be *admissible* in the sense that $\mu_t(x) \in \mathbb{U}(x)$ for all x and t: in this case, the minimization over the input at each iteration should also consider the state-dependent constraints.[9]

[9] Alternatively, we might be able to define the action space to be $\mathbb{U} = \bigcup_{x\in\mathbb{X}} \mathbb{U}(x)$ and modify the dynamics and the costs such that the state-dependent input constraints are automatically satisfied. This is actually the case in Example 3.3. Observe that the action 'fix' in the state 'repair' does not make any sense, so what we have done is that we "copied" the admissible action 'wait' and just labeled it with 'fix'. Similarly, the only admissible action in the state 'broken' is 'fix' as we cannot simply leave the machine to stay 'broken'. Nevertheless, we did consider the action 'wait' for the state 'broken', but with a very large cost to make sure that it cannot be optimal. These are the two standard ways that allow us to take the action space to be $\mathbb{U} = \bigcup_{x\in\mathbb{X}} \mathbb{U}(x)$ in the presence of state-dependent input constraints.

These extensions lead to the following iteration in the DP algorithm:

$$J_t(x) = \min_{u \in \mathbb{U}(x)} \mathbb{E}\left(g_t(x, u, W_{t+1}) + J_{t+1}\big(f_t(x, u, W_{t+1})\big) \;\middle|\; X_t = x, U_t = u\right), \quad \forall x \in \mathbb{X}.$$

We finish this section with the following remark.

Remark 3.7 (Bellman operator). Each iteration of the DPA can be characterized by the functional equation

$$J_t = \mathcal{T}[J_{t+1}] \iff J_t(x) = \mathcal{T}[J_{t+1}](x), \ \forall x \in \mathbb{X},$$

where $\mathcal{T} : \mathbf{R}^{\mathbb{X}} \to \mathbf{R}^{\mathbb{X}}$ is the so-called *Bellman (a.k.a. DP) operator*, i.e., a mapping from the space $\mathbf{R}^{\mathbb{X}}$ of real-valued functions on $\mathbb{X}$ to itself, defined by

$$\mathcal{T}[J](x) \coloneqq \min_{u \in \mathbb{U}} \left\{ g(x, u) + \gamma \mathbb{E}\left(J(X_+) \;\middle|\; X = x, U = u\right)\right\}, \quad \forall x \in \mathbb{X},$$

with $\gamma = 1$, where $X_+ = f(X, U, W_+)$ is the next state (at $t+1$) of the system given that the current state and action (at t) are X and U, respectively. The Bellman operator is particularly of interest in the *γ-discounted, infinite-horizon* setting in which one aims to minimize $\mathbb{E}\big(\sum_{t=0}^{\infty} \gamma^t g(X_t, U_t)\big)$, with $\gamma \in (0, 1)$ being the so-called discount factor. Indeed, in this case, one can show that the value function (i.e., the optimal expected γ-discounted, infinite-horizon cost) is the fixed-point of the Bellman operator, i.e., $J^\star = \mathcal{T} J^\star$. △

3.3. The curse of dimensionality

In this final section, we discuss the computational complexity of DPA. Let us first point out that DPA provides a *sequential* procedure for finding an optimal policy $(\mu_t^\star)_{t=0}^{T-1}$ whereby we first find $\mu_{T-1}^\star$ and then $\mu_{T-2}^\star$ and so on up to $\mu_0^\star$. This reduces the size of the search space significantly compared to the case in which we try to find all of these state-feedback laws *at once*. To illustrate, for a finite MDP with n states and m actions, given the initial state $X_0 = x$, the number of (sequential) closed-loop control laws in DPA is $m + (T-1)m^n$ with a *linear* dependence on the horizon T of the problem. Compare this with the total number $m^{1+(T-1)n}$ of closed-loop control laws with an *exponential* dependence on T. Revisiting our example with $n = 4$, $m = 2$, and $T = 10$, the size of the search space reduces to $2 + 9 \cdot 2^4 = 146$ from $2^{37} \approx 10^{11}$ closed-loop control laws! However, as you can see, the size of the search space still increases *exponentially* with the number n of the states.

Let us now take a closer look at the *per-iteration* complexity of DPA. To this end, observe that each iteration of DPA involves three important and nested operations:

- (functional equation) we need to find the value function for each state $x \in \mathbb{X}$;
- (minimization) we need to solve a minimization over the action $u \in \mathbb{U}$ for a given $x \in \mathbb{X}$;
- (expectation) we need to compute the expected value of the next state w.r.t. the disturbance $W \in \mathbb{W}$ for a given $(x, u) \in \mathbb{X} \times \mathbb{U}$.

To illustrate, for a finite MDP with n states, m actions, and $p \leq n$ possible values for disturbance (that is, the maximum number of states that are reachable from each state-action pair), the per-iteration time complexity of DPA is $O(nmp)$:[10] For each $x \in \mathbb{X}$, we need to enumerate over $u \in \mathbb{U}$ to find the minimum of an objective that requires computing the expectation over $W \in \mathbb{W}$.

In case any of the variables (state, action, or disturbance) are *continuous*, the situation can (and will!) get significantly complicated. For continuous-state MDPs, the functional equation $J_t = \mathcal{T}[J_{t+1}]$ becomes *infinite*-dimensional: each iteration of DPA involves solving an infinite number of minimization problems. So, either this equation leads to a closed-form analytic solution, or, we have to use *finite-dimensional approximations*, i.e., a parameterization J_{θ_t} of the value functions with $\theta_t \in \mathbf{R}^q$ being the parameter. For example, in the case of *linear* parameterization, we have $J_{\theta_t}(x) = \theta_t^\top \phi(x)$ for each $x \in \mathbb{X}$, where $\phi : \mathbb{X} \to \mathbf{R}^q$ is the vector of basis functions (a.k.a. features). To find the parameters, one can form a regression problem that minimizes the "Bellman error" over a *finite* number of points in the state space, e.g., for $t = T-1, \dots, 0$, and $\{x_i\}_{i \in [N]} \subset \mathbb{X}$, we set

$$\theta_t = \operatorname*{argmin}_{\theta \in \mathbf{R}^q} \sum_{i \in [N]} \Big| J_\theta(x_i) - \underbrace{\min_{u \in \mathbb{U}} \big\{ g(x_i, u) + \mathbb{E}\big(J_{\theta_{t+1}}(X_{t+1}) \bigm| X_t = x, U_t = u \big) \big\}}_{= \mathcal{T}[J_{\theta_{t+1}}](x_i)} \Big|^2.$$

The procedure described above is called *fitted value iteration.* While this procedure reduces the original infinite-dimensional problem to a finite-dimensional one, we still need to compute $\mathcal{T}[J_{\theta_{t+1}}](x_i)$ by solving a minimization problem with an expectation operation in the objective. Now, if the action space is also continuous, the minimization problem can easily become a non-convex, difficult problem. This may even leave us with no choice but to discretize the action space and enumerate over it. This is actually one of the main reasons why a large body of dynamic programming (and reinforcement learning) literature only considers finite action spaces. Similarly, if the disturbance is a continuous r.v., the expectation is essentially an integration that is again known to be a numerically demanding operation.

To summarize, when it comes to MDPs with continuous state, action, and disturbance, there is a good chance that the exact implementation of DPA is simply impossible and we have to use some form of *approximate dynamic programming.* A common approximation technique is the discretization (or abstraction) of the original MDP. This essentially boils down to a proper partitioning of the state and action space and then applying the standard DPA to the resulting finite MDP. The problem however is that for a "good" approximation, the size (i.e., the number of partitions) of the finite MDP must increase *exponentially* with the dimension (i.e., the number) of the state and action variables. To illustrate, consider a dynamical system with a d-dimensional state variable $x \in \mathbf{R}^d$; e.g., a simple inverted pendulum has two state variables (angle and angular velocity) with $d = 2$. Then, if you decide to partition *each* state in N bins, the total number of regions in the entire state space will be N^d. This is known as the "curse of dimensionality," which is arguably the most important drawback of DPA.

[10] The big O notation means that the time complexity (simply put, the number of binary operations) does not grow faster than *nmp*. To be precise, for two functions $f, g : \mathbf{R}_+ \to \mathbf{R}_+$, we write $f(x) = O\big(g(x)\big)$ if there exist $x_0, M \in \mathbf{R}_+$ such that $f(x) \leq Mg(x)$ for all $x \geq x_0$.

CHAPTER 4

Applications

In this chapter, we look at the application of the dynamic programming algorithm (DPA) for solving classic problems in control systems and operation research. The setup in each problem allows us to derive analytic solutions for the value function and the optimal policy based on the problem data.

4.1. Linear quadratic control

In this section, we look at a classic problem from control, namely, the optimal stochastic control of a linear dynamical system with quadratic state and input costs, i.e., the *linear quadratic* (LQ) optimal control problem (a.k.a. LQR where 'R' stands for regulation). To simplify the exposition, we start with one-dimensional systems with a single state variable and a single action variable.

4.1.1. One-dimensional problem. Consider the linear dynamical system with state variable $X_t \in \mathbb{X} = \mathbf{R}$, action variable $U_t \in \mathbb{U} = \mathbf{R}$, and disturbance variable $W_t \in \mathbb{W} = \mathbf{R}$, described by the recursion

$$X_{t+1} = aX_t + U_t + W_{t+1}, \quad t = 0, 1, 2, \ldots, \tag{4.1}$$

where $a \in \mathbf{R}$ is a 'memory' coefficient. Let the initial state of the system be $X_0 = x \in \mathbf{R}$. Assume that $W_1, W_2, \ldots$ is a sequence of i.i.d. random variables with finite first and second moments such as Gaussian or uniform (with bounded support). For simplicity, we are going to assume $\mathbb{E}(W_t) = 0$ and $\operatorname{var}(W_t) = 1$ for all t. Also, consider quadratic stage and terminal costs given by

$$g(x, u) = qx^2 + ru^2, \quad G(x) = qx^2, \tag{4.2}$$

where $q, r > 0$, over the planning horizon $T \in \mathbf{N}$. Such a cost implies that our goal is to keep the state X_t close to 0 with control actions U_t that are close to 0.

Let us first take a look at the autonomous system with $U_t = 0$ for all t. In this case, we have

$$\begin{cases} X_{t+1} = aX_t + W_{t+1}, & t = 0, 1, 2, \ldots, \\ X_0 = x. \end{cases} \tag{4.3}$$

Therefore, for all $t = 0, 1, \ldots$, we have[1]

$$\mathbb{E}(X_{t+1}) = a\mathbb{E}(X_t) + \mathbb{E}(W_{t+1}) = a\mathbb{E}(X_t) = \cdots = a^{t+1}\mathbb{E}(X_0) = a^{t+1}x,$$

and[2]

$$\operatorname{var}(X_{t+1}) = a^2 \operatorname{var}(X_t) + \operatorname{var}(W_{t+1}) = a^2 \operatorname{var}(X_t) + 1 = \cdots$$

[1] For two r.v.'s X and Y, we have $\mathbb{E}(aX + Y) = a\mathbb{E}(X) + \mathbb{E}(Y)$.

[2] For two *independent* r.v.'s X and Y, we have $\operatorname{var}(aX + Y) = a^2 \operatorname{var}(X) + \operatorname{var}(Y)$.

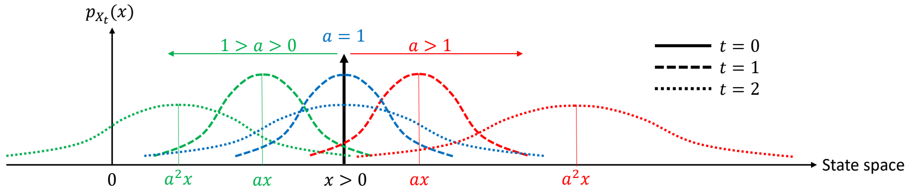


FIGURE 1. The evolution of the dynamics (4.3) for three different values of a and $W_t \sim \mathcal{N}(0,1)$ for all t.

$$= a^{2(t+1)}\,\mathrm{var}(X_0) + \sum_{k=0}^{t} a^{2k} = \begin{cases} t+1 & \text{if } |a| = 1, \\ \frac{1-a^{2(t+1)}}{1-a^2} & \text{otherwise.} \end{cases}$$

Observe that the variance of X_t increases in time regardless of the value of $a \neq 0$. In particular,

1. if $|a| < 1$, then $\mathbb{E}(X_t) \to 0$ and $\mathrm{var}(X_t) \to \frac{1}{1-a^2}$ as $t \to \infty$;
2. if $|a| = 1$, then $\mathbb{E}(X_t) = \pm x$ for all t but $\mathrm{var}(X_t) \to \infty$ as $t \to \infty$;[3]
3. if $|a| > 1$, then $|\mathbb{E}(X_t)| \to \infty$ and $\mathrm{var}(X_t) \to \infty$ as $t \to \infty$.

As you can see, it is only in the first case above that the system is stable in the sense that the state process X_t has finite first and second moments asymptotically. See Figure 1 for an illustration.

Recall that our goal is to solve the following problem

$$J_0^\star(x) = \min_{(\mu_t:\mathbf{R}\to\mathbf{R})_{t=0}^{T-1}} \mathbb{E}\left(\sum_{t=0}^{T-1} qX_t^2 + r\mu_t(X_t)^2 + qX_T^2 \;\middle|\; X_0 = x \right), \quad x \in \mathbf{R}.$$

Let us deploy DPA to solve this stochastic optimal control problem:

- **Initialization** at $t = T$:
$$J_T(x) = G(x) = qx^2, \quad \forall x \in \mathbf{R}.$$
- **Backward iteration** for $t = T-1, T-2, \dots, 0$:
$$\begin{aligned} J_t(x) &= \min_{u\in\mathbf{R}} \left\{ g(x,u) + \mathbb{E}\big(J_{t+1}(X_{t+1}) \;\big|\; X_t = x,\ U_t = u\big) \right\} \\ &= \min_{u\in\mathbf{R}} \left\{ qx^2 + ru^2 + \mathbb{E}\big(J_{t+1}(ax+u+W_{t+1}) \;\big|\; X_t = x,\ U_t = u\big) \right\} \\ &= \min_{u\in\mathbf{R}} \left\{ qx^2 + ru^2 + \mathbb{E}\big(J_{t+1}(ax+u+W_{t+1})\big) \right\}, \quad \forall x \in \mathbf{R}, \end{aligned}$$
where for the last equality we used the fact that W_t's are indepedendent.

Note that this is an MDP with *continuous* state and action spaces. The good news however is that this problem can be solved (almost) *analytically*! Let's see how: For the **first iteration** $t = T-1$, since W_t's are assumed to be zero mean with unit variance, we have

$$\begin{aligned} J_{T-1}(x) &= \min_{u\in\mathbf{R}} \left\{ qx^2 + ru^2 + \mathbb{E}\big(q(ax+u+W_T)^2\big) \right\}. \\ &= \min_{u\in\mathbf{R}} \left\{ qx^2 + ru^2 + q(ax+u)^2 + \mathbb{E}(2q(ax+u)W_T + qW_T^2) \right\} \\ &= \min_{u\in\mathbf{R}} \left\{ qx^2 + ru^2 + q(ax+u)^2 + q \right\} \end{aligned}$$

[3] This is called a *random walk*.

$$= q(1+a^2)x^2 + q + \min_{u\in\mathbf{R}} \left\{(q+r)u^2 + 2aqxu\right\}, \quad \forall x \in \mathbf{R}$$

Observe that the objective of the preceding minimization problem is a convex quadratic function of u since $q+r>0$ by the assumptions. Hence, using the first-order optimality condition, we obtain

$$\begin{aligned} & 0 = \frac{\partial}{\partial u}\big((q+r)u^2 + 2aqxu\big)\Big|_{u=u^\star} \\ \iff & 0 = (q+r)u^\star + aqx \\ \iff & u^\star = \mu_{T-1}(x) = -\frac{aq}{q+r}x, \quad \forall x \in \mathbf{R}, \end{aligned}$$

and

$$\begin{aligned} J_{T-1}(x) &= q(1+a^2)x^2 + q + (q+r)u^{\star 2} + 2aqxu^\star \\ &= \left(q + \frac{a^2 rq}{q+r}\right)x^2 + q, \quad \forall x \in \mathbf{R}. \end{aligned}$$

Observe that

1. the optimal policy μ_{T-1} is *linear* in x, and,
2. more importantly, the value function J_{T-1} is *quadratic* in x similar to J_T.

These observations signal the possibility of all value functions being quadratic (and hence the optimal policy being a linear state feedback). This is indeed the case and can be proven by induction:

For the *base step* at $t = T$, we have that $J_T(x) = qx^2$ is quadratic in x with $q > 0$. Also, for the *induction step* at $t < T$, we assume (induction hypothesis)

$$J_{t+1}(x) = k_{t+1}x^2 + c_{t+1}, \quad \forall x \in \mathbf{R},$$

for some $k_{t+1} > 0$ and $c_{t+1} \in \mathbf{R}$. Then, using DPA, and following similar arguments as above, we have

$$\begin{aligned} J_t(x) &= \min_{u\in\mathbf{R}} \left\{qx^2 + ru^2 + \mathbb{E}\big(k_{t+1}(ax+u+W_{t+1})^2 + c_{t+1}\big)\right\} \\ &= (q + a^2k_{t+1})x^2 + k_{t+1} + c_{t+1} + \min_{u\in\mathbf{R}} \left\{(k_{t+1}+r)u^2 + 2ak_{t+1}xu\right\}, \quad \forall x \in \mathbf{R}. \end{aligned}$$

Observe that the objective of the preceding minimization problem is again a convex quadratic function of u since $k_{t+1} + r > 0$ by the assumptions. Hence, using the first-order optimality condition, we can obtain the optimal policy

$$u^\star = \mu_t(x) = -\frac{ak_{t+1}}{k_{t+1}+r}x =: \ell_t x, \quad \forall x \in \mathbf{R},$$

and the optimal cost-to-to

$$J_t(x) = \left(q + \frac{a^2 rk_{t+1}}{k_{t+1}+r}\right)x^2 + k_{t+1} + c_{t+1} =: k_t x^2 + c_t, \quad \forall x \in \mathbf{R}.$$

at time t. That is, J_t is also quadratic in x and in particular

$$k_t = q + \frac{a^2 rk_{t+1}}{k_{t+1}+r} > 0.$$

This completes the proof by induction. Hence, the two properties listed above hold for all t, i.e., we have

$$J_t(x) = k_t x^2 + c_t, \quad \mu_t(x) = \ell_t x, \quad \forall x \in \mathbf{R},$$

where k_t, c_t, and ℓ_t can be computed *recursively* as follows

$$\begin{array}{lll}
k_T = q, & c_T = 0, & (\text{no } \ell_T); \\
k_{T-1} = q + \frac{a^2 r k_T}{k_T + r}, & c_{T-1} = k_T + c_T, & \ell_{T-1} = -\frac{a k_T}{k_T + r}; \\
\vdots & \vdots & \vdots \\
k_t = q + \frac{a^2 r k_{t+1}}{k_{t+1} + r}, & c_t = k_{t+1} + c_{t+1} = \sum_{s=t+1}^{T} k_s, & \ell_t = -\frac{a k_{t+1}}{k_{t+1} + r} x; \\
\vdots & \vdots & \vdots \\
k_0 = q + \frac{a^2 r k_1}{k_1 + r}, & c_0 = \sum_{s=1}^{T} k_s, & \ell_0 = -\frac{a k_1}{k_1 + r} x.
\end{array}$$

To summarize, for $t \in \{0, 1, \dots, T-1\}$, we have

$$J_t(x) = k_t x^2 + \sum_{s=t+1}^{T} k_s, \quad \mu_t(x) = -\frac{a k_{t+1}}{k_{t+1} + r} x, \quad \forall x \in \mathbf{R},$$

where

- (Initialization) $k_T = q$,
- (Backward iteration) $k_t = \mathcal{F}(k_{t+1}) \coloneqq q + \frac{a^2 r k_{t+1}}{k_{t+1} + r}$ for $t = T-1, \dots, 0$.

See Figure 2 for an illustration. What we observe to be the case is that under linear dynamics and quadratic stage cost, each iteration of DPA preserves the quadratic shape of the value function. This is exactly what we proved by induction. As a result, we managed to transform the *infinite*-dimensional functional equation $J_t(x) = \mathcal{T}[J_{t+1}](x)$, $\forall x \in \mathbf{R}$, in each iteration of DPA to a *one*-dimensional equation $k_t = \mathcal{F}(k_{t+1})$ in terms of the parameter k_t of the value function $J_t(x) = k_t x^2 + \sum_{s=t+1}^{T} k_s$. We emphasize that in this problem, the chosen quadratic parameterization of the value functions leads to *exact* solutions to iterations of DPA.

The procedure we used above essentially involves (1) 'guessing' a proper shape (i.e., parameterization) for the value functions, and (2) using induction to show that the iterations of DPA preserve this shape and meanwhile solve the problem analytically to derive a recursive formula for the parameters. How can we 'guess' the particular shape of the value functions that are preserved through iterations? Well, there is no general recipe; we can use the problem data, e.g., the terminal cost, or try to solve a couple of iterations of DPA analytically. In the following sections, we look at other examples of problems that have this kind of property which allows for such analytic solutions. Let us emphasize however that in real-world applications this is often not case and we must use some form of approximation.

Let us now look at the asymptotic behavior of the system as the planning horizon T tends to ∞. Mathematically, this is equivalent to fixing $T \in \mathbf{N}$ and letting $t \to -\infty$ in the solution we derived above. Thus, we need to analyze the asymptotic behavior of the recursion $k_t = \mathcal{F}(k_{t+1})$ as $t \to -\infty$, initialized by $k_T = q$. This is nothing but consecutive applications of the mapping $\mathcal{F}$. So, if the sequence $(k_T, k_{T-1}, \dots)$ converges, it must converge to a *fixed-point* $k_* = \mathcal{F}(k_*)$ of the mapping $\mathcal{F}$. The fixed-point equation (a.k.a. the *algebraic Riccati equation*) is

$$k_*^2 + (r - q - a^2 r) k_* - qr = 0,$$

which has exactly two solutions, one positive and one negative. Taking a closer look at the mapping $\mathcal{F}(k) = q + \frac{a^2 r k}{k + r}$, we see that

- $\mathcal{F}(0) = q > 0$ and $\lim_{k \to \infty} \mathcal{F}(0) = q + a^2 r$;

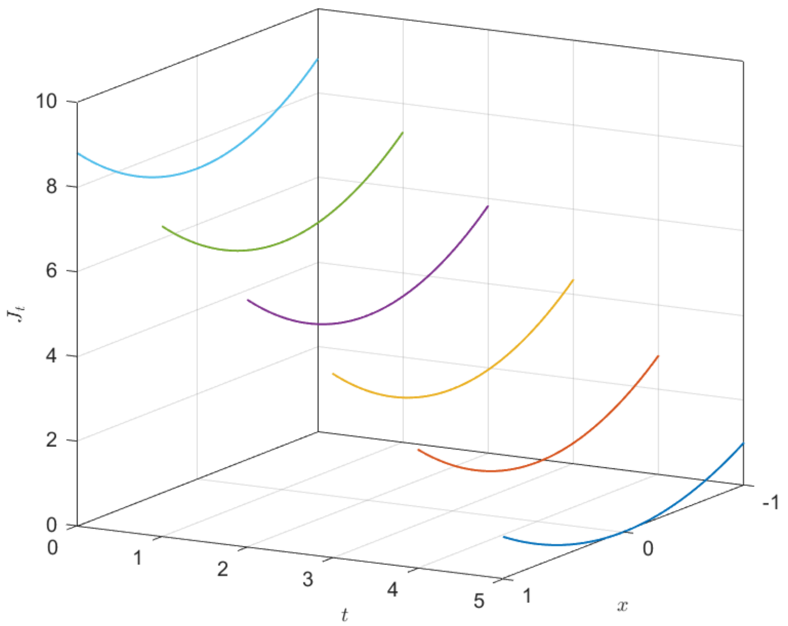


Figure 2. The value function for linear dynamics (4.1) with $a = 0.9$ and quadratic costs (4.2) with $q = r = 1$ over the planning horizon $T = 5$.

- $\mathcal{F}(k)$ is increasing and concave for $k > -r$ since

$$\mathcal{F}'(k) = a^2r^2(k+r)^{-2} > 0 \quad \text{and} \quad \mathcal{F}''(k) = -2a^2r^2(k+r)^{-3} < 0.$$

These properties imply that the sequence $(k_T, k_{T-1}, \ldots)$ indeed converges to the *positive* fixed-point $k_* > 0$ of $\mathcal{F}$; see Figure 3 for an illustration.

The corresponding optimal policy as $k_t \to k_*$ is then

$$\mu_*(x) = -\frac{ak_*}{k_* + r}, \qquad \forall x \in \mathbf{R}. \tag{4.4}$$

In particular, observe that the preceding policy is a *stationary* policy that does *not* depend on time t explicitly. We note that this stationary policy is a good approximation of the optimal policy even for finite t (the reason is the relatively fast convergence of k_t to k_*). Under this stationary policy, the system obeys the recursion

$$X_{t+1} = aX_t + U_t + W_{t+1} = \left(a - \frac{ak_*}{k_* + r}\right)X_t + W_{t+1} = \underbrace{\frac{ar}{k_* + r}}_{=:b} X_t + W_{t+1},$$

and hence

$$\mathbb{E}(X_t) = b^t x \quad \text{and} \quad \operatorname{var}(X_t) = \frac{1 - b^{2t}}{1 - b^2}.$$

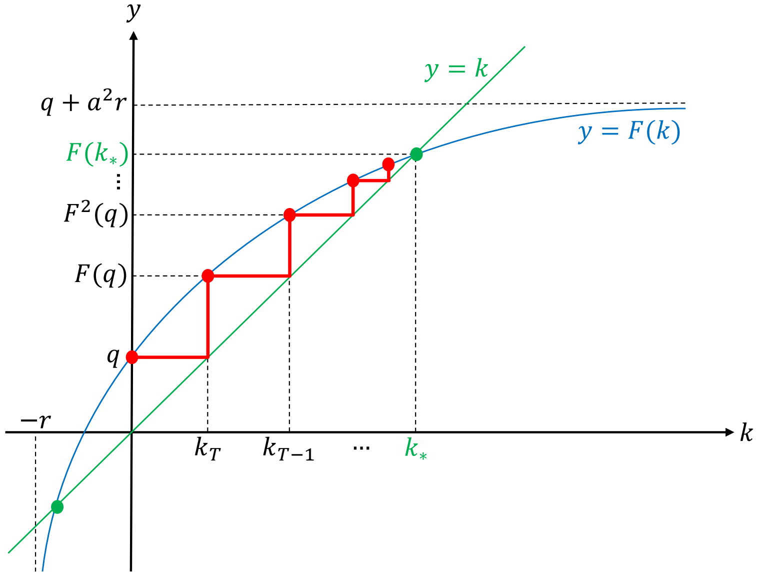


FIGURE 3. Convergence of the recursion $k_t = \mathcal{F}(k_{t+1})$ initialized by $k_T = q$ to the fixed-point $k_* = \mathcal{F}(k_*) > 0$ as $t \to -\infty$.

Now, observe that since $k_* = \mathcal{F}(k_*) = q + \frac{a^2 r k_*}{k_* + r} = q + abk_*$, we further have

$$|b| = \begin{cases} |a| \left| \frac{r}{k_* + r} \right| < 1 & \text{for } |a| < 1, \\ \left| \frac{k_* - q}{a k_*} \right| = \frac{1}{|a|} \left| 1 - \frac{q}{k_*} \right| < 1 & \text{for } |a| \geq 1. \end{cases}$$

Therefore, $\mathbb{E}(X_t) \to 0$ and $\operatorname{var}(X_t) \to 1/(1 - b^2)$ as $t \to \infty$, i.e., asymptotically, the state process has a zero mean and a bounded variance irrespective of a. That is, the stationary policy (4.4) is stabilizing even for $|a| \geq 1$.

4.1.2. Multi-dimensional problem. In this subsection, we extend the results of the previous subsection to generic problems with linear dynamics with quadratic costs. So, let us consider a linear (time-varying) dynamics

$$X_{t+1} = A_t X_t + B_t U_t + W_{t+1}, \quad t = 0, 1, 2, \ldots, \tag{4.5}$$

where $X_t \in \mathbb{X} = \mathbf{R}^n$ is the state variable, $U_t \in \mathbb{U} = \mathbf{R}^m$ is the control variable, and $W_t \in \mathbb{W} = \mathbf{R}^n$ is the disturbance variable which is assumed to be independent with zero mean and finite covariance. Above, $A_t \in \mathbf{R}^{n \times n}$ and $B_t \in \mathbf{R}^{n \times m}$ are the state and inupt matrices at time t, respectively. Also, consider quadratic (time-varying) stage and terminal costs given by

$$\begin{aligned} &g_t(x, u) = x^\top Q_t x + u^\top R_t u \quad \text{for } t = 0, 1, \ldots, T-1, \\ &G(x) = x^\top Q_T x, \end{aligned} \tag{4.6}$$

where $Q_T, Q_t \in \mathbf{R}^{n \times n}$ are positive semi-definite, $R_t \in \mathbf{R}^{m \times m}$ are positive definite, and $T \in \mathbf{N}$ is the planning horizon. The goal is to solve the following problem

$$J_0^\star(x) = \min_{(\mu_t : \mathbf{R}^n \to \mathbf{R}^m)_{t=0}^{T-1}} \mathbb{E}\left(\sum_{t=0}^{T-1} X_t^\top Q_t X_t + \mu_t(X_t)^\top R_t \mu_t(X_t) + X_T^\top Q_T X_T \,\middle|\, X_0 = x \right),$$

for each $x \in \mathbf{R}^n$. The solution to this stochastic optimal control problem can be derived using similar arguments to the ones provided in the previous subsection. We summarize the results in the following lemma:

Lemma 4.1 (Linear quadratic control). *Consider the stochastic optimal control problem with dynamics* (4.5) *and costs* (4.6) *where the disturbance W_t is assumed to be independent with zero mean and finite covariance. The value functions are of quadratic form*

$$J_t(x) = x^\top K_t x + c_t, \quad \forall x \in \mathbf{R},$$

where c_t's are some constants and K_t's are positive semi-definite matrices given by the discrete-time Riccati equation

$$\begin{aligned} &K_T = Q_T, \\ &K_t = \mathcal{F}_t(K_{t+1}), \quad t = T-1, \dots, 0. \end{aligned} \tag{4.7}$$

where

$$\mathcal{F}_t(K) \coloneqq Q_t + A_t^\top \big(K - KB_t(B_t^\top K B_t + R_t)^{-1} B_t^\top K\big) A_t.$$

Moreover, the optimal policy is a linear state feedback of the form

$$\mu_t(x) = L_t x, \quad \forall x \in \mathbf{R},$$

where the matrix gains L_t are given by

$$L_t = -(B_t^\top K_{t+1} B_t + R_t)^{-1} B_t^\top K_{t+1} A_t, \quad t = T-1, \dots, 0.$$

Once again, by letting $t \to -\infty$ in the solution provided in the preceding lemma, we can analyze the asymptotic behavior of the system for a long planning horizon T. To this end, we need to assume the dynamics and costs are *time-invariant*, i.e.,

$$A_t = A, \quad B_t = B, \quad Q_t = Q, \quad R_t = R, \quad \forall t,$$

such that the pair (A, B) is *controllable* and the pair (A, Q) is *observable*.[4] Under these assumptions, the recursion (4.7) converges, i.e., $\lim_{t \to -\infty} K_t = K_*$ such that K_* is the unique positive semi-definite solution to the algebraic Riccati equation

$$K_* = \mathcal{F}(K_*) \coloneqq Q + A^\top \big(K_* - K_* B(B^\top K_* B + R)^{-1} B^\top K_*\big) A.$$

Moreover, the corresponding stationary optimal policy

$$\mu_*(x) = L_* x = -(B^\top K_* B + R)^{-1} B^\top K_* A x,$$

is stabilizing in the sense that the poles of the closed-loop system matrix $A + L_* B$ reside (strictly) inside the unit circle. This means that the state process, under the control policy μ_*, has a zero mean and a finite covariance asymptotically.

You might have noticed that the optimal policy derived above is the same as the solution to the *deterministic* LQR problem:

Remark 4.2 (Certainty equivalence). In the class of problems considered in this section, we get the exact same optimal policy if we replace the disturbance W_t by its expected value $\mathbb{E}(W_t) = 0$ and solve the resulting *deterministic* optimal control problem. This is because the problem setup leads to the optimal policy depending on W_t only through its expected value (observe however that the covariance of the disturbance affects the constant terms c_t in the value functions $J_t(x) = x^\top K_t x + c_t$). This is the so-called *certainty equivalence principle*. △

[4] The pair (A, B) is controllable if and only if the matrix $(B, AB, A^2B, \dots, A^{n-1}B)$ has full row rank. The pair (A, Q) is observable if and only if $(A^\top, Q^\top)$ is controllable.

4.2. Optimal portfolio selection

In this section, we look at a classic problem from operations research involving portfolio optimization. Let us start by providing a mathematical definition of the problem.

4.2.1. Portfolio selection and utility function. Suppose that the financial *market* offers $n \in \mathbf{N}$ different risky assets to invest in. Let $W_{t,i} \in \mathbb{W}_i = \mathbf{R}_+$ denote the random *total return* of the asset $i \in [n]$ over the t-th period; that is, each euro invested in asset i at time $t-1$ will be worth $W_{t,i}$ euro at time t. Also, denote $W_t = (W_{t,1}, W_{t,2}, \dots, W_{t,n}) \in \mathbb{W} = \mathbf{R}^n_+$ to be the random vector of all total returns, assumed to be *independent*.

Consider an *investor* with the total available budget (i.e., wealth) $X_t \in \mathbb{X} = \mathbf{R}_+$ euro at time t. The investor decides to allocate a percentage $U_{t,i}$ of their wealth X_t to each asset $i \in [n]$. That is, the investor allocates $X_{t,i} = X_t \cdot U_{t,i} \geq 0$ euro to asset i at time t. Then, $U_t = (U_{t,1}, \dots, U_{t,n}) \in \mathbb{U} = \Delta_n := \Delta([n])^5$ is the *column* vector of investor's *portfolio* weights. In particular, using the return $W_{t+1,i}$ of asset i from time t to $t+1$, the investor's wealth at time $t+1$ with the chosen portfolio U_t at time t is given by

$$X_{t+1} = \sum_{i \in [n]} X_t \cdot U_{t,i} \cdot W_{t+1,i} = X_t \cdot U_t^\top W_{t+1} =: f(X_t, U_t, W_{t+1}). \tag{4.8}$$

Of course, the goal of an investor is to maximize their wealth by selecting 'the best' portfolio. For that, the investor needs a *metric for ranking random wealth levels*, that is, a *utility function* $G : \mathbb{X} \to \mathbf{R}$ which assigns a utility index to each possible wealth level. Using this utility function, alternative random wealth levels can be ranked by evaluating their expected utility values: random wealth variable X is preferred over Y if $\mathbb{E}\big(G(X)\big) > \mathbb{E}\big(G(Y)\big)$.

Utility functions vary among decision-makers, depending on their financial environment and their *risk tolerance*. A common characteristic of the utility function is that it is an *increasing* function (i.e, higher wealth corresponds to higher utility). The simplest utility function is $G(x) = x$ which implies ranking by only using the expected values. But, is this really the best choice? Notice that we are comparing the *random* wealth levels with different *distributions* as opposed to deterministic values. Let us illustrate this through an example:

Example 4.3 (Utility function and risk tolerance). Assume that we have two alternatives for future wealth:

(1) Alternative 1: we obtain x_1 with probability α or x_2 with probability $1-\alpha$ for some $\alpha \in (0,1)$;
(2) Alternative 2: we obtain $x^* = \alpha x_1 + (1-\alpha)x_2$ with *certainty*.

To grasp the difference between the two alternatives, consider the following three cases with $\alpha = 0.5$:

- Case 1: $x_1 =$ €0, $x_2 =$ €10, and hence $x^* =$ €5;
- Case 2: $x_1 =$€-5, $x_2 =$ €15, and hence $x^* =$ €5;
- Case 3: $x_1 =$ €0, $x_2 =$ €1000, and hence $x^* =$ €500.

[5] Recall that $\Delta([n]) = \{u \in \mathbf{R}^n \ : \ u_i \geq 0, \ \forall i \in [n], \ \sum_{i \in [n]} u_i = 1\}$ is the set of all p.m.f.'s defined on the set $[n]$.

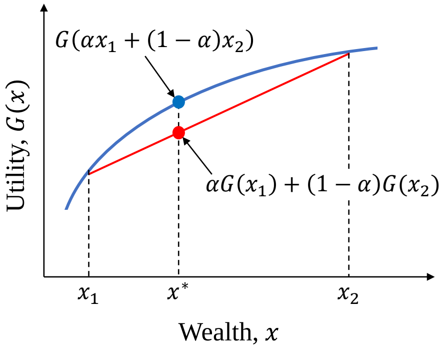

$G(\alpha x_1 + (1-\alpha)x_2)$
$\alpha G(x_1) + (1-\alpha)G(x_2)$
Utility, $G(x)$
x_1
x^*
x_2
Wealth, x

Figure 4. Risk aversion with concave utility function; see Example 4.3.

In all the cases above, both alternatives have the *same expected wealth* x^*. Therefore, based on the utility function $G(x) = x$ neither alternative is better than the other. This is why individuals using this utility function are called *risk neutral*. Are you neutral in all of the three cases above? Most probably not. That is why we should look for better utility functions!

Let us also compare the two alternatives using the utility function $G(x) = -e^{-x}$. Then, the expected utility of the first alternative is

$$\alpha G(x_1) + (1-\alpha)G(x_2) = -\alpha e^{-x_1} - (1-\alpha)e^{-x_2},$$

while the expected utility of the second alternative is

$$G(x^*) = G(\alpha x_1 + (1-\alpha)x_2) = -e^{-\alpha x_1 - (1-\alpha)x_2}.$$

Since $G(x) = -e^{-x}$ is a concave function,[6] the second (i.e., risk-free) alternative is preferred. Actually, the risk-free alternative will be the preferred one for any (strictly) concave utility function; see Figure 4. That is why a *concave* utility function is called *risk averse*. Conversely, for a *risk-seeking* investor, the utility function will be *convex* which leads to the first alternative above being the preferred one. △

Some of the most common risk-averse utility functions are (see Figure 5):

- *Exponential*: $G(x) = -e^{-ax}$ for some $a > 0$;
- *Logarithmic*: $G(x) = \ln(x)$ – defined only for $x > 0$;
- *Power*: $G(x) = \frac{x^{1-\gamma}}{1-\gamma}$ for some $\gamma \geq 0$, $\gamma \neq 1$ – defined only for $x > 0$ (respectively, $x \geq 0$) if $\gamma > 1$ (respectively, $\gamma < 1$);
- *Quadratic*: $G(x) = x - bx^2$ for some $b > 0$ – increasing only for $x \leq \frac{1}{2b}$.

Observe that among all the functions listed above the *power* utility with $\gamma > 1$ and the *logarithmic* utility have the property that $\lim_{x\to 0^+} G(x) = -\infty$ and hence they can guard the investor against "bankruptcy."

[6] A function $G : [a, b] \to \mathbf{R}$ is *concave* if for any $\alpha \in [0, 1]$ and for any $x, y \in [a, b]$ there holds

$$G\big(\alpha x + (1-\alpha)y\big) \geq \alpha G(x) + (1-\alpha)G(y).$$

That is, the straight line drawn between two points on the function must lie below or on the function itself.

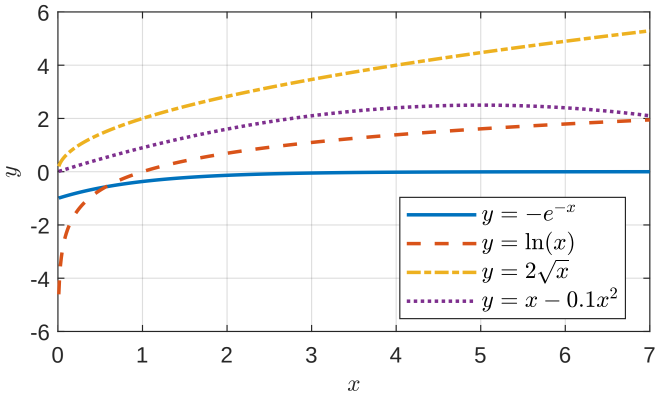


FIGURE 5. Examples of concave utility functions.

With the dynamics and the objective at hand, we can define the corresponding stochastic optimal control problem: Consider an investor with *initial wealth* $X_0 = x \in \mathbb{X}$ who is aiming to maximize the expected *utility* $G(X_T)$ of their terminal wealth X_T after $T \in \mathbf{N}$ periods of investment. Such an investor must solve the dynamic portfolio selection problem

$$J_0^\star(x) = \max_{(\mu_t:\mathbb{X}\to\mathbb{U})_{t=0}^{T-1}} \mathbb{E}\big(G(X_T) \mid X_0 = x\big). \tag{4.9}$$

Notice that there is no running reward in this problem, i.e., $g(x,u) = 0$ for all x, u.

4.2.2. Optimal portfolio with power utility. In this section, we use DPA to solve the dynamic portfolio selection problem (4.9) with *power* utility function $G(x) = \frac{x^{1-\gamma}}{1-\gamma}$ for a given $\gamma \geq 0,\ \gamma \neq 1$. As usual, we try to carry out the first one or two iterations of DPA and examine the results to see if the shape the value function and the optimal policy with respect to the state variable x follows a certain pattern over the iterations. So, let's deploy DPA to solve this problem:

- **Initialization** at $t = T$:

$$J_T(x) = G(x) = \frac{x^{1-\gamma}}{1-\gamma}, \quad \forall x \in \mathbf{R}_+.$$[7]

- **Backward iteration** for $t = T-1, T-2, \dots, 0$:

$$\begin{aligned} J_t(x) &= \max_{u\in\Delta_n} \big\{g(x,u) + \mathbb{E}\big(J_{t+1}(X_{t+1}) \mid X_t = x,\ U = u\big)\big\} \\ &= \max_{u\in\Delta_n} \big\{\mathbb{E}\big(J_{t+1}(x \cdot u^\top W_{t+1}) \mid X_t = x,\ U_t = u\big)\big\} \\ &= \max_{u\in\Delta_n} \big\{\mathbb{E}\big(J_{t+1}(x \cdot u^\top W_{t+1})\big)\big\}, \quad \forall x \in \mathbf{R}_+. \end{aligned}$$

where for the last equality we used the fact that W_t's are independent.

For the **first iteration** $t = T-1$, we have

$$J_{T-1}(x) = \max_{u\in\Delta_n} \left\{\mathbb{E}\left(\frac{(x\cdot u^\top W_T)^{1-\gamma}}{1-\gamma}\right)\right\}$$

[7] Strictly speaking, for $\gamma > 1$, the state space is the positive reals, i.e., $\mathbb{X} = \mathbf{R}_+ \setminus \{0\}$.

$$= \max_{u\in\Delta_n} \left\{ x^{1-\gamma} \cdot \mathbb{E}\left(\frac{(u^\top W_T)^{1-\gamma}}{1-\gamma} \right) \right\}, \quad \forall x \in \mathbf{R}_+.$$

Then, since $x \geq 0$, we have

$$J_{T-1}(x) = x^{1-\gamma} \cdot \max_{u\in\Delta_n} \mathbb{E}\left(\frac{(u^\top W_T)^{1-\gamma}}{1-\gamma} \right), \quad \forall x \in \mathbf{R}_+.$$

The optimal portfolio is then[8]

$$\mu_{T-1}(x) = u^\star_{T-1} = \operatorname*{argmax}_{u\in\Delta_n} \mathbb{E}\left(\frac{(u^\top W_T)^{1-\gamma}}{1-\gamma} \right), \quad \forall x \in \mathbf{R}_+,$$

which is a *constant* vector and *independent of the state* x. Then,

$$J_{T-1}(x) = \beta_{T-1} \cdot \frac{x^{1-\gamma}}{1-\gamma}, \quad \forall x \in \mathbf{R}_+,$$

where

$$\beta_{T-1} = \mathbb{E}\left(({u^\star_{T-1}}^\top W_T)^{1-\gamma} \right).$$

That is, J_{T-1} is just a (scaled) power utility. We note that $\beta_{T-1} \geq 0$ since $u^\star_{T-1} \in \Delta_n$ and $W_T \in \mathbf{R}^n_+$. Using this observation, we hypothesize that $J_t(x) = \beta_t \cdot \frac{x^{1-\gamma}}{1-\gamma}$ for some $\beta_t \geq 0$ for all $t = 0, \ldots, T$. The proof is of course by induction as follows.

For the *base step* at $t = T$, we clearly have that $J_T(x) = \beta_T \cdot \frac{x^{1-\gamma}}{1-\gamma}$ with $\beta_T = 1 \geq 0$. For the *induction step* at $t < T$, we assume (induction hypothesis)

$$J_{t+1}(x) = \beta_{t+1} \cdot \frac{x^{1-\gamma}}{1-\gamma}, \quad \forall x \in \mathbf{R}_+,$$

for some $\beta_{t+1} \geq 0$. Then, using DPA, and following similar arguments as above, we have

$$\begin{aligned} J_t(x) &= \max_{u\in\Delta_n} \left\{ \mathbb{E}\big(J_{t+1}(x \cdot u^\top W_{t+1})\big) \right\} \\ &= \max_{u\in\Delta_n} \left\{ \mathbb{E}\left(\beta_{t+1} \cdot \frac{(x \cdot u^\top W_{t+1})^{1-\gamma}}{1-\gamma} \right) \right\} \\ &= \beta_{t+1} \cdot x^{1-\gamma} \cdot \max_{u\in\Delta_n} \left\{ \mathbb{E}\left(\frac{(u^\top W_{t+1})^{1-\gamma}}{1-\gamma} \right) \right\}, \quad \forall x \in \mathbf{R}_+. \end{aligned}$$

The optimal portfolio is then *independent of the state* x and given by

$$\mu_t(x) = u^\star_t = \operatorname*{argmax}_{u\in\Delta_n} \mathbb{E}\left(\frac{(u^\top W_{t+1})^{1-\gamma}}{1-\gamma} \right), \quad \forall x \in \mathbf{R}_+. \tag{4.10}$$

This means that the optimal portfolio does not depend on the total amount of money that we decide to invest in the market, which to some extent makes sense. Moreover,

$$J_t(x) = \beta_t \cdot \frac{x^{1-\gamma}}{1-\gamma}, \quad \forall x \in \mathbf{R}_+,$$

where

$$\beta_t = \beta_{t+1} \cdot \mathbb{E}\left(({u^\star_t}^\top W_{t+1})^{1-\gamma} \right) \geq 0,$$

[8] We are assuming that the maximum is attained.

This completes the proof by induction. To summarize, for all $t \in \{0, 1, \ldots, T-1\}$, we have

$$J_t(x) = \beta_t \cdot \frac{x^{1-\gamma}}{1-\gamma}, \ \mu_t(x) = u_t^\star = \operatorname*{argmax}_{u \in \Delta_n} \mathbb{E}\left(\frac{(u^\top W_{t+1})^{1-\gamma}}{1-\gamma} \right), \quad \forall x \in \mathbf{R}_+,$$

where

$$\beta_T = 1 \quad \text{and} \quad \beta_t = \alpha_t \beta_{t+1} = \Pi_{s=t}^{T-1} \alpha_s \ \text{ for } \ t = T-1, \ldots, 0.$$

with

$$\alpha_t = \mathbb{E}\left(({u_t^\star}^\top W_{t+1})^{1-\gamma} \right) \geq 0.$$

We finish this subsection with two remarks on two special cases of the setup described above.

Remark 4.4 (Linear utility). A particular case of interest is $\gamma = 0$ which corresponds to the *linear* utility function $G(x) = x$. In this case, the optimal portfolio $\mu_t(x) = u_t^\star$ at time t is given by

$$u_t^\star(i) = \begin{cases} 1 & \text{if } i = i_t^\star, \\ 0 & \text{otherwise,} \end{cases} \qquad \text{where} \quad i_t^\star = \operatorname*{argmax}_{i \in [n]} \mathbb{E}(W_{t+1,i}).$$

That is, the optimal portfolio is the one that invests all the available budget in the asset with the highest expected return during the t-th period. Correspondingly, we have $J_t(x) = \left(\Pi_{s=t}^{T-1} \mathbb{E}(W_{s+1, i_t^\star}) \right) \cdot x$ for $t \in \{0, 1, \ldots, T-1\}$. △

Remark 4.5 (Myopic policy). If the total returns $(W_t)_{t=1}^T$ are also *identically distributed* and hence i.i.d., the optimization problem in (4.10) will be also *independent of time* t. That is, the optimal portfolio is a fixed-mix policy of the form $\mu_t(x) = u^\star$ for all $x \in \mathbf{R}_+$ and all $t = 0, \ldots, T-1$, where $u^\star$ is the optimal solution of the (static) optimization problem

$$u^\star = \operatorname*{argmax}_{u \in \Delta_n} \mathbb{E}\left(\frac{(u^\top W_1)^{1-\gamma}}{1-\gamma} \right).$$

This means that optimizing only over *one* period would result in the same portfolio for the multi-period problem. We thus say that the optimal policy is *myopic*. Moreover, the corresponding value functions are of the from $J_t(x) = \alpha^{T-t} \cdot \frac{x^{1-\gamma}}{1-\gamma}$ for $x \in \mathbf{R}_+$, where $\alpha = \mathbb{E}\left(({u^\star}^\top W_1)^{1-\gamma} \right) \geq 0$. △

4.2.3. Optimal portfolio with logarithmic utility. In this section, we look at the dynamic portfolio selection problem (4.9) with *logarithmic* utility function $G(x) = \ln(x)$. This problem also has an analytic solution that can be derived using similar arguments to the ones provided in the previous subsection. We summarize the results in the following lemma for the case of i.i.d. total returns:

Lemma 4.6 (Optimal portfolio with logarithmic utility). *Consider the dynamic portfolio selection problem* (4.9) *with* logarithmic *utility function* $G(x) = \ln(x)$. *Assume that the the total returns* $(W_t)_{t=1}^T$ *are* i.i.d. *Then, the optimal value functions are given by* $J_t(x) = \alpha \cdot (T-t) + \ln(x)$ *for all* $x > 0$ *and all* $t \in \{0, \ldots, T\}$, *where*

$$\alpha = \max_{u \in \Delta_n} \mathbb{E}\left(\ln(u^\top W_1) \right). \tag{4.11}$$

Moreover, the optimal policy is a fixed-mix policy of the form $\mu_t(x) = u^\star$ *for all* $x > 0$ *and all* $t \in \{0, \ldots, T-1\}$, *where* $u^\star$ *is an optimal solution of the (static) optimization problem* (4.11).

Let us now look at a simple but insightful example of portfolio selection to see the importance of the proper choice of the utility function.

Example 4.7 (Logarithmic utility in a binomial market). Consider a simple model of a financial market with only two assets, whose returns in each period are determined by flipping a fair coin as follows

	Heads		Tails	
	Asset 1	Asset 2	Asset 1	Asset 2
Net return	+40%	−20%	−30%	+15%

Thus, the asset returns are

$$W_t = \begin{cases} \begin{pmatrix} 1.40 \\ 0.80 \end{pmatrix} & \text{with probability } \frac{1}{2}, \\ \begin{pmatrix} 0.70 \\ 1.15 \end{pmatrix} & \text{with probability } \frac{1}{2}, \end{cases}$$

and i.i.d. for $t = 1, \ldots, T$. By Lemma 4.6, for $G(x) = \ln(x)$, the optimal portfolio is then myopic and can be found by solving

$$\begin{aligned} u^\star &= \operatorname*{argmax}_{u \in \Delta_2} \mathbb{E}\left(\ln(u^\top W_t)\right) \\ &= \operatorname*{argmax}_{u=(a,b)^\top} \left\{ \frac{1}{2} \ln(1.4a + 0.8b) + \frac{1}{2} \ln(0.7a + 1.15b) \right\} \\ &\qquad \text{s.t.} \quad a \geq 0,\ b \geq 0,\ a + b = 1 \\ &= \operatorname*{argmax}_{u=(a,1-a)^\top} \{\ln(0.8 + 0.6a) + \ln(1.15 - 0.45a)\ :\ 0 \leq a \leq 1\} \end{aligned}$$

The optimal solution is then one of the critical points of the objective function within the interval $[0, 1]$, i.e., the two boundary points $a = 0$ and $a = 1$, and the point at which the derivative vanishes, that is,

$$\begin{aligned} 0 &= \frac{\partial}{\partial a}\big(\ln(0.8 + 0.6a) + \ln(1.15 - 0.45a)\big) \\ \iff 0 &= \frac{0.6}{0.8 + 0.6a} - \frac{0.45}{1.15 - 0.45a} \\ \iff a &= \frac{11}{18}. \end{aligned}$$

The optimal solution is indeed $a^\star = 11/18$ and hence $u^\star = (a^\star,\ 1 - a^\star)^\top$. We note that $a = 1$ corresponds to the optimal policy for the linear utility functions $G(x) = x$; see Remark 4.4. Indeed, the expected logarithmic and linear utilities per period and unit of investment for the corresponding three policies $u = (a, 1 - a)^\top$ can be computed as

a	1	0	11/18
$\mathbb{E}(\ln(u^\top W_t))$	−1.01	−4.17	+1.03
$\mathbb{E}(u^\top W_t)$	1.05	0.975	1.0208

Figure 6 shows the result of 100 simulations of investment in this market starting from $X_0 = 1$ and using three policies:

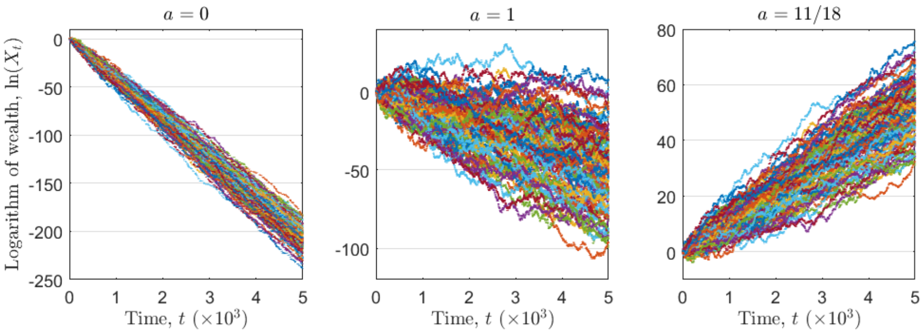


FIGURE 6. Logarithm of the wealth in 100 simulations of the binomial market of Example 4.7 starting from $X_0 = 1$ and using three different myopic policies $u_t = (a,\ 1-a)$ for all t. $a = 1$ and $a = 11/18$ are the optimal policies for linear and logarithmic utilities, respectively.

- $a = 0$, i.e., investing all the available budget in Asset 2 at each time step;
- $a = 1$, i.e., investing all the available budget in Asset 1 at each time step, which is the optimal policy for linear utility $G(x) = x$;
- $a = 11/18$, i.e., investing all the available budget in a fixed mix of Assets 1 and 2 at each time step, which is the optimal policy for logarithmic utility $G(x) = \ln(x)$.

The results show that investing only in Asset 2 is an absolute mistake; this is indeed expected since the expected net return of this asset is $\frac{1}{2}(-20) + \frac{1}{2}(+15) = -2.5\% < 0$. What is surprising in the simulation results is that with $a = 1$, in most of the random realizations of the market, we end up losing money! This means that investing only in Asset 1 (i.e., the *optimal* policy for linear utility) also seems to be a mistake even though the expected net return of this asset is $\frac{1}{2}(+40) + \frac{1}{2}(-30) = +2.5\% > 0$. On the other hand, the optimal policy for logarithmic utility (i.e., $a = 11/18$) is indeed performing well and leads to an increase in the wealth in all the runs for $t \geq 2 \times 10^3$. It seems that this policy is making money from losing assets! How can that be?

The answer is in the "distribution" of wealth under these policies and how relying only on the expected values can be very misleading. Let X_T^{lin} and X_T^{log} denote the random wealth after T periods of investment under the fixed policies $u_t = u^{\star,\text{lin}} = (1,0)^\top$ and $u_t = u^{\star,\text{log}} = (11/18, 7/18)^\top$ for all $t \in 0, 1, \ldots, T-1$, respectively, starting from initial wealth $X_0 = 1$. Then,

$$\begin{aligned} X_T^{\text{lin}} &= X_{T-1} \cdot u_{T-1}^\top W_T = X_{T-2} \cdot u_{T-2}^\top W_{T-1} \cdot u_{T-1}^\top W_T = \cdots \\ &= X_0 \cdot (\Pi_{t=1}^T\ u_{t-1}^\top W_t) = \Pi_{t=1}^T Y_t^{\text{lin}} \end{aligned}$$

where

$$Y_t^{\text{lin}} = (u^{\star,\text{lin}})^\top W_t = \begin{cases} 1.4 & \text{with probability } \frac{1}{2}, \\ 0.7 & \text{with probability } \frac{1}{2}. \end{cases}$$

is an i.i.d. random variable. Therefore, X_T^{lin} is a discrete r.v. with at most $T+1$ possible values given by

$$X_T^{\text{lin}} = (1.4)^S (0.7)^{T-S}, \tag{4.12}$$

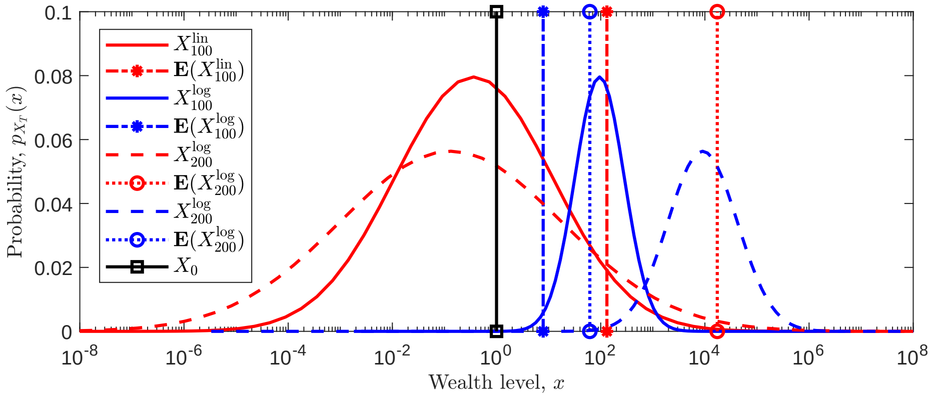


FIGURE 7. The distribution of X_T^{lin} and $X_T^{\log}$ and their expected values for $T = 100$ and $T = 200$. $X_0 = 1$ is the initial investment.

where $S \sim \text{Binomial}(T, \frac{1}{2})$ with

$$\mathbb{P}\left(S = s\right) = \frac{\binom{T}{s}}{2^T}, \quad s \in \{0, 1, \dots, T\}.$$

Similarly, we have

$$X_T^{\log} = \Pi_{t=1}^T Y_t^{\log}$$

where

$$Y_t^{\log} = (u^{\star,\log})^\top W_t = \begin{cases} \frac{21}{18} & \text{with probability } \frac{1}{2}, \\ \frac{16.9}{18} & \text{with probability } \frac{1}{2}. \end{cases}$$

is an i.i.d. random variable and hence

$$X_T^{\log} = \left(\frac{21}{18}\right)^S \left(\frac{15.75}{18}\right)^{T-S}, \tag{4.13}$$

where $S \sim \text{Binomal}(T, \frac{1}{2})$. Note that the vertical cross-sections of the plots in Figure 6 at each time s corresponds to 100 realizations of the distributions above for $T = s$. Figure 7 shows the distribution of X_T^{lin} and $X_T^{\log}$ and their expected values for two different values of T. Observe that when it comes to the expected values, X_T^{lin} outperforms $X_T^{\log}$. This is indeed what we expect to be the case as the optimal *linear* policy aims to maximize $\mathbb{E}(X_T)$. However, we can see that as T increases, the distribution of X_T^{lin} becomes flatter with its (right) tail moving towards extremely large values of x and its peak moving toward $x = 0$ such that the majority of its mass falls in the area with $x \leq X_0 = 1$. Indeed, the large expected value of X_T^{lin} is exactly because of the highly rare events in the tail of its distribution with extremely large values of x, while the typical behavior of X_T^{lin} captured by its peak and majority of mass is not profitable as we have also observed to be the case in Figure 6. On the other hand, we observe that the typical behavior of $X_T^{\log}$ to be profitable, captured by its peak and majority of mass falling in the area with $x \geq X_0 = 1$.

To be more concrete about the discussion above, let us compute the objects of interest rigorously: The expected value of X_T^{lin} can be computed as

$$\mathbb{E}(X_T^{\text{lin}}) = \mathbb{E}\big(\Pi_{t=1}^T Y_t^{\text{lin}}\big) = \mathbb{E}(Y_t^{\text{lin}})^T = (1.05)^T,$$

where for the second equality we used the fact that Y_t^{lin}'s are i.i.d. Similarly, for the expected value of $X_T^{\log}$, we have

$$\mathbb{E}(X_T^{\log}) = \mathbb{E}\big(\Pi_{t=1}^T Y_t^{\log}\big) = \mathbb{E}(Y_t^{\log})^T = (1.0208)^T,$$

Therefore, we indeed have

$$\frac{\mathbb{E}(X_T^{\text{lin}})}{\mathbb{E}(X_T^{\log})} = \left(\frac{1.05}{1.0208}\right)^T > 1, \quad \forall T \in \mathbf{N},$$

with the ratio growing exponentially fast as $t \to \infty$.

Next, let us consider the peaks of the distribution of X_T^{lin} and $X_T^{\log}$, that is, the most probable wealth level under optimal *linear* and *logarithmic* policies, respectively. Using (4.12) and (4.13) and because $S \sim \text{Binomial}(T, \frac{1}{2})$ has a binomial distribution, the most probable wealth corresponds to $S = T/2$ if T is even, and $S = (T \pm 1)/2$ otherwise. Therefore, we have

$$\begin{aligned}\underset{x}{\operatorname{argmax}}\, \mathbb{P}(X_T^{\text{lin}} = x) &\leq (1.4)^{\frac{T+1}{2}}(0.7)^{\frac{T-1}{2}} \\ &= \sqrt{2}\cdot(0.98)^{\frac{T-1}{2}} < 1, \quad \forall T \geq 37,\end{aligned}$$

and

$$\begin{aligned}\underset{x}{\operatorname{argmax}}\, \mathbb{P}(X_T^{\log} = x) &\geq (21/18)^{\frac{T-1}{2}}(15.75/18)^{\frac{T+1}{2}} \\ &= \sqrt{0.75}\cdot(1.0208)^{\frac{T-1}{2}} > 1, \quad \forall T \geq 15.\end{aligned}$$

In particular, observe that

$$\underset{x}{\operatorname{argmax}}\, \mathbb{P}(X_T^{\text{lin}} = x) \to 0 \quad \text{and} \quad \underset{x}{\operatorname{argmax}}\, \mathbb{P}(X_T^{\log} = x) \to \infty \quad \text{as} \quad T \to \infty.$$

Finally, let us consider the probability of ending up with *no loss* under the optimal *linear* policy $u^{\star,\text{lin}}$ after T periods of investment, that is, $\mathbb{P}(X_T^{\text{lin}} \geq X_0)$ assuming initial wealth $X_0 = 1$. Using (4.12), we have

$$\begin{aligned}\mathbb{P}(X_T^{\text{lin}} \geq 1) &= \mathbb{P}\left((1.4)^S(0.7)^{T-S} \geq 1\right) \\ &= \mathbb{P}\left(S\ln(1.4) + (T-S)\ln(0.7) \geq 0\right) \\ &= \mathbb{P}\left(\frac{S}{T} \geq \frac{\ln(10/7)}{\ln(2)}\right).\end{aligned}$$

Observe that

$$\frac{1}{2} < c = \frac{\ln(10/7)}{\ln(2)} < 1,$$

and hence (using the fact that $S \sim \text{Binomial}(T, \frac{1}{2})$ has a binomial distribution)

$$\mathbb{P}(X_T^{\text{lin}} \geq 1) = \mathbb{P}\left(\frac{S}{T} \geq c\right) = \frac{\sum_{s \geq cT}\binom{T}{s}}{2^T} \leq \frac{(1-c)T\cdot\binom{T}{cT}}{2^T}$$

Now, using Stirling's bounds

$$e^{\frac{1}{12n+1}}\,\sqrt{2\pi n}\,(n/e)^n < n! < e^{\frac{1}{12n}}\,\sqrt{2\pi n}\,(n/e)^n, \quad \forall n \in \mathbf{N},$$

one can show that

$$\binom{T}{cT} \leq \frac{e}{\sqrt{2\pi c(1-c)T}\cdot\left(c^c\cdot(1-c)^{(1-c)}\right)^T}, \quad \forall c \in [0,1]. \tag{4.14}$$

Therefore,

$$\mathbb{P}(X_T^{\text{lin}} \geq 1) \leq \frac{e\sqrt{1-c}}{\sqrt{2\pi c}} \cdot \frac{\sqrt{T}}{\left(2 \cdot c^c \cdot (1-c)^{(1-c)}\right)^T} \leq \frac{1.0533\sqrt{T}}{(1.0004)^T}.$$

This means that $\mathbb{P}(X_T^{\text{lin}} \geq 1) \to 0$ exponentially as $T \to \infty$, that is, the probability of making a profit under the optimal *linear* policy approaches zeros for large values of T. On the other hand, for the probability of ending up with *no profit* under the optimal *logarithmic* policy $u^{\star,\log}$ after T periods of investment with initial wealth $X_0 = 1$, we can use (4.13) to write

$$\begin{aligned}\mathbb{P}(X_T^{\log} \leq 1) &= \mathbb{P}\left((21/18)^S (15.75/18)^{T-S} \leq 1\right)\\ &= \mathbb{P}\left(S \ln(21/18) + (T-S)\ln(15.75/18) \leq 0\right)\\ &= \mathbb{P}\left(\frac{S}{T} \leq \frac{\ln(18/15.75)}{\ln(21/15.75)}\right),\end{aligned}$$

where

$$0 < d = \frac{\ln(18/15.75)}{\ln(21/15.75)} < \frac{1}{2}.$$

Hence,

$$\mathbb{P}(X_T^{\log} \leq 1) = \mathbb{P}\left(\frac{S}{T} \leq d\right) = \frac{\sum_{s \leq dT} \binom{T}{s}}{2^T} \leq \frac{dT \cdot \binom{T}{dT}}{2^T}.$$

Then, using the bound in (4.14), we have

$$\mathbb{P}(X_T^{\log} \leq 1) \leq \frac{e\sqrt{d}}{\sqrt{2\pi(1-d)}} \cdot \frac{\sqrt{T}}{\left(2 \cdot d^d \cdot (1-d)^{(1-d)}\right)^T} \leq \frac{0.9737\sqrt{T}}{(1.0057)^T}.$$

This means that $\mathbb{P}(X_T^{\log} \leq 1) \to 0$ exponentially fast as $T \to \infty$, that is, the probability of loss under the optimal *logarithmic* policy approaches zeros for large values of T. △

4.3. Optimal stopping

In this section, we look at a new class of problems that allow the decision-maker the "stop" the process prematurely while (possibly) incurring some cost. That is, the decision-maker is not obliged to continue the process till the end of the planning horizon T and has the option to terminate the process at some earlier step $t < T$ at a certain cost. As we will see in this section, a classic example of such problems is when an investor is trying to maximize their profit by selling or buying an asset in the market. Indeed, the decision to sell or buy means that the investor made a transaction and "stopped" interacting with the market for that particular asset. Another example is the decision to follow a particular treatment, e.g., a surgical operation, for a patient based on their symptoms and physiological recordings. Of course, any operation has its risks and costs and the question is when the patient should accept those risks and costs and go through the operation.

4.3.1. DPA with stopping. Recall the optimal control problem of a system with dynamics

$$X_{t+1} = f(X_t, U_t, W_{t+1}), \quad t = 0, 1, \dots, T-1,$$

with running cost $g : \mathbb{X} \times \mathbb{U} \to \mathbf{R}$ and terminal cost $G : \mathbb{X} \to \mathbf{R}$, where T is the planning horizon, $X_t \in \mathbb{X}$ is the state process, $U_t \in \mathbb{U}$ is the control process, and $W_t \in \mathbb{W}$ is the disturbance process (assumed to be conditionally independent given X_t and U_t with a given probability distribution). Now, assume that we are *not* obliged to continue the process till the end of the planning horizon. Instead, we have the option to "stop" the process at any time $t < T$ while incurring the *stopping* cost $h_t : \mathbb{X} \to \mathbf{R}$, besides the running cost during the first t periods. That is, if we decide to stop the process at some time $t = S \in \{0, 1, \dots, T-1\}$, the total cost is

$$\mathbb{E}\left(\sum_{t=0}^{S-1} g(X_t, U_t) + h_S(X_S)\right),$$

while in case we decide to continue the process till the end of the planning horizon T, the total cost is as before, i.e.,

$$\mathbb{E}\left(\sum_{t=0}^{T-1} g(X_t, U_t) + G(X_T)\right).$$

In other words, the decision to stop the process at time $t = S$ is equivalent to reducing the planning horizon from T to S and replacing the terminal cost G with the "termination" cost h_S. Therefore, we must also decide the *stopping time* S besides potentially the control inputs U_t. The objective is again to minimize the expected total cost as described above. Here is the formal definition of the optimal stopping problem.

Definition 4.8 (Optimal stopping). Consider an MDP described by the dynamics

$$X_{t+1} = f(X_t, U_t, W_{t+1}), \quad t = 0, 1, \dots, T-1,$$

over the planning horizon T, where $X_t \in \mathbb{X}$ is the state process, $U_t \in \mathbb{U}$ is the control process, and $W_t \in \mathbb{W}$ is the disturbance process assumed to be conditionally independent given X_t and U_t with a given probability distribution. Also, let $g : \mathbb{X} \times \mathbb{U} \to \mathbf{R}$ be the running cost so that $g(x, u)$ is the cost of taking the control action u in state x, $G : \mathbb{X} \to \mathbf{R}$ be the terminal cost so that $G(x)$ is the cost of being in state x at end of the horizon $t = T$, and $h_t : \mathbb{X} \to \mathbf{R}$ be the stopping cost so that $h_t(x)$ is the cost of stopping the process in state x at time $t \in \{0, 1, \dots, T-1\}$. The problem of interest is to find optimal stopping time $S \in \{0, 1, \dots, T-1\}$ and control sequence $(U_t)_{t=0}^{S-1} \in \mathbb{U}^S$ that minimizes the expected accumulated cost, i.e.,

$$\min_{\substack{(U_t)_{t=0}^{S-1} \in \mathbb{U}^S \\ S \in \{0,1,\dots,T\}}} \mathbb{E}\left(\sum_{t=0}^{S-1} g_t(X_t, U_t) + H_S(X_S) \;\middle|\; X_0 = x\right),$$

where

$$H_S(x) = \begin{cases} h_S(x) & \text{if } S < T, \\ G(x) & \text{if } S = T, \end{cases} \qquad \forall x \in \mathbb{X}.$$

△

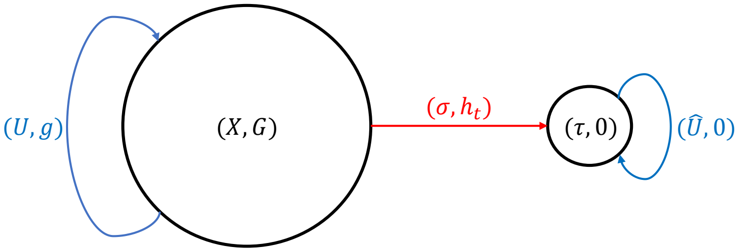


FIGURE 8. The graphical representation of modified state space, action space, dynamics, and costs for the optimal stopping problem. The circles show the states with the label (x, G) denoting the state x and the corresponding terminal cost G. The arcs show the transitions with the label (u, g) denoting the action u and the running cost g of the corresponding transition.

Observe that we cannot directly apply DPA to solve this problem. To be able to do so, we need to modify different elements of the problem so that it fits the standard setting of the stochastic optimal control problem. Let's do that then.

To start, we introduce a new state τ to the state space which represents the case in which the process has been already terminated. We also introduce the control action σ to the control space which represents the decision to stop the process. So, our new state and action spaces are $\widehat{\mathbb{X}} = \mathbb{X} \cup \{\tau\}$ and $\widehat{\mathbb{U}} = \mathbb{U} \cup \{\sigma\}$, respectively. Observe that when we decide to stop the process by taking the control action $U_t = \sigma$ in some state $X_t \neq \tau$ in time t, we incur the cost $h_t(X_t)$, and the process enters the state $X_{t+1} = \tau$ and *stays there till the end of the planning horizon without incurring any further cost*; see Figure 8. We now need to modify the dynamics and the costs to account for the new state τ and the new action σ:

- for the dynamics, we have
$$X_{t+1} = \widehat{f}(X_t, U_t, W_{t+1}) = \begin{cases} \tau & \text{if } X_t = \tau \text{ or } U_t = \sigma, \\ f(X_t, U_t, W_{t+1}) & \text{otherwise,} \end{cases} \tag{4.15}$$
for $t = 0, 1, \ldots, T-1$;
- for the running cost, we have
$$\widehat{g}_t(x, u) = \begin{cases} 0 & \text{if } x = \tau, \\ h_t(x) & \text{if } x \neq \tau \text{ and } u = \sigma, \\ g(x, u) & \text{otherwise,} \end{cases}$$
for $t = 0, 1, \ldots, T-1$;
- for the terminal cost, we have
$$\widehat{G}(x) = \begin{cases} 0 & \text{if } x = \tau, \\ G(x) & \text{if } x \neq \tau. \end{cases}$$

Observe that, mathematically, in this modified version of the problem, one can take any of the actions $u \in \widehat{\mathbb{U}}$ when the process is state τ. However, regardless of that action, the next state is again τ and the stage cost is 0. So, effectively, once the process enters the state τ, there is nothing to do.

The corresponding stochastic optimal control problem is then given by

$$J_0^\star(x) = \min_{(\mu_t : \widehat{\mathbb{X}} \to \widehat{\mathbb{U}})_{t=0}^{T-1}} \mathbb{E}\left(\sum_{t=1}^{T-1} \widehat{g}_t\big(X_t, \mu_t(X_t)\big) + \widehat{G}(X_T) \;\middle|\; X_0 = x \right), \quad \forall x \in \widehat{\mathbb{X}}.$$

Notice that the preceding minimization problem will be solved subject to the *modified* dynamics (4.15). Solving this problem indeed amounts to solving our original optimal stopping problem. Indeed, the optimal policy $(\mu_t^\star : \widehat{\mathbb{X}} \to \widehat{\mathbb{U}})_{t=0}^{T-1}$ provides us with the states $x \in \mathbb{X}$ for which the optimal decision is to stop for each time t.

Let us now employ DPA to solve this problem. Using the modified dynamics and costs, the value functions can be derived by DPA as follows

- Initialization for $t = T$: for each $x \in \widehat{\mathbb{X}}$

$$J_T(x) = \widehat{G}(x) = \begin{cases} 0 & \text{if } x = \tau, \\ G(x) & \text{if } x \neq \tau. \end{cases} \tag{4.16}$$

- Backward iteration for $t = T-1, \ldots, 0$: for each $x \in \widehat{\mathbb{X}}$

$$\begin{aligned} J_t(x) &= \min_{u \in \widehat{\mathbb{U}}} \left\{ \widehat{g}_t(x,u) + \mathbb{E}\left(J_{t+1}(X_{t+1}) \;\middle|\; X_t = x,\ U_t = u \right) \right\} \\ &= \begin{cases} J_{t+1}(\tau) & \text{if } x = \tau, \\ \displaystyle\min_{u \in \widehat{\mathbb{U}}} \left\{ \widehat{g}_t(x,u) + \mathbb{E}\left(J_{t+1}(X_{t+1}) \;\middle|\; X_t = x,\ U_t = u \right) \right\} & \text{if } x \neq \tau. \end{cases} \end{aligned} \tag{4.17}$$

Observe that for the state τ, the value function is independent of the control action. That is, any action $u \in \widehat{\mathbb{U}}$ is optimal when the process is in state τ. This is just a by-product of our mathematical formulation. In practice, once the process is stopped, we no longer make any decisions. Moreover, we have:

Lemma 4.9 (Value of τ). *$J_t(\tau) = 0$ for all $t \in \{0, 1, \ldots, T\}$.*

Proof. The proof is by induction. For the base case, we clearly have $J_T(\tau) = 0$ by (4.16). Now, assume $J_{t+1}(\tau) = 0$ for $t < T$. Then using (4.17), we have $J_t(\tau) = \mathbb{E}\big(J_{t+1}(\tau)\big) = 0$. This completes the proof. □

Therefore, the cost-to-go from the state τ is also zero. This is indeed expected considering our modifications in the dynamics and costs: the state τ means that the process has been already stopped, in which case, we do not incur any further cost, and regardless of the control action u, the process stays in the state τ till the end of the planning horizon.

Finally, let us use the fact that the value of τ is always zero to simplify the backward iteration in (4.17) for $x \neq \tau$. For each $x \in \mathbb{X}$, using the modified dynamics and costs, we have

$$\begin{aligned} J_t(x) &= \min_{u \in \{\sigma\} \cup \mathbb{U}} \left\{ \widehat{g}_t(x,u) + \mathbb{E}\left(J_{t+1}(X_{t+1}) \;\middle|\; X_t = x,\ U_t = u \right) \right\} \\ &= \min \left\{ h_t(x) + \mathbb{E}\big(J_{t+1}(\tau)\big),\ K_t(x) \right\} \\ &= \min \left\{ h_t(x),\ K_t(x) \right\}, \end{aligned}$$

where

$$K_t(x) = \min_{u \in \mathbb{U}} \left\{ g(x,u) + \mathbb{E}\left(J_{t+1}\big(f(x,u,W_{t+1})\big) \;\middle|\; X_t = x,\ U_t = u \right) \right\}. \tag{4.18}$$

The corresponding optimal policy is then of the form

$$\mu_t(x) = \begin{cases} \sigma & \text{if } h_t(x) \leq K_t(x), \\ \nu_t(x) & \text{otherwise.} \end{cases}$$

where

$$\nu_t(x) = \operatorname*{argmin}_{u \in \mathbb{U}} \left\{ g(x,u) + \mathbb{E}\left(J_{t+1}\big(f(x,u,W_{t+1})\big) \;\middle|\; X_t = x, U_t = u \right) \right\}. \tag{4.19}$$

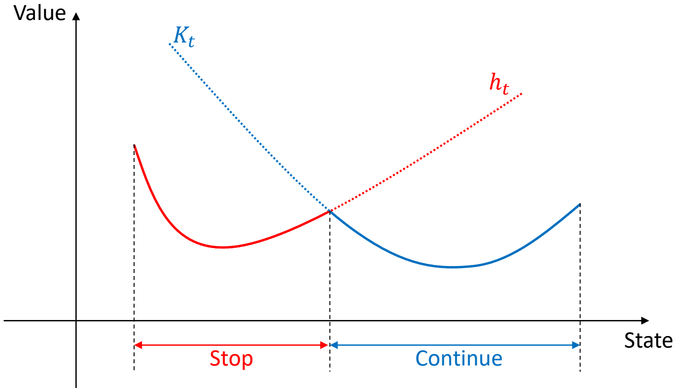


FIGURE 9. The optimal stopping policy based on the comparison between the cost K_t of continuing the process and the cost $h_t$of stopping the process.

In words, at each $t < T$, we first perform a "standard" backward iteration based on (4.18) and (4.19) to find the optimal cost-to-go $K_t(x)$ and policy $\nu_t(x)$ *for the case of continuing the process at time t in state x*. Then, if the cost of continuing the process (i.e, $K_t(x)$) is greater than the cost of stopping the process (i.e, $h_t(x)$), we set $J_t(x) = h_t(x)$ with the optimal decision being to stop the process (i.e., $\mu_t(x) = \sigma$). Otherwise, we proceed as usual by setting $J_t(x) = K_t(x)$ and $\mu_t(x) = \nu_t(x)$. See Figure 9. Try to reflect on this procedure and see if it makes sense! We shall call this procedure *DPA with stopping*.

Observe that since the value and optimal policy for the termination state τ are clear (and not relevant!), we can simply discard it in our DPA and only focus on the states $x \in \mathbb{X}$. To summarize, we showed that the following algorithm solves the optimal stopping problem.

ALGORITHM 4.10 (DPA with stopping). *The DPA with stopping is as follows:*
(1) Initialization: *Set $t = T$ and the* value function $J_T : \mathbb{X} \to \mathbf{R}$ *as*

$$J_T(x) = G(x), \quad \forall x \in \mathbb{X}.$$

(2) Backward iteration: *Count backwards from $t = T - 1$ to $t = 0$ and find the* value function $J_t : \mathbb{X} \to \mathbf{R}$ *by solving*

$$J_t(x) = \min\left\{h_t(x),\ \min_{u \in \mathbb{U}} Q_t(x,u)\right\}, \quad \forall x \in \mathbb{X},$$

and set the policy $\mu_t : \mathbb{X} \to \mathbb{U} \cup \{\text{stop}\}$ *to be the corresponding minimizer, i.e.,*

$$\mu_t(x) = \begin{cases} \text{stop} & \text{if } J_t(x) = h_t(x), \\ \operatorname*{argmin}_{u \in \mathbb{U}} Q_t(x,u) & \text{otherwise}, \end{cases} \qquad \forall x \in \mathbb{X},$$

where

$$\begin{aligned} Q_t(x,u) &:= g(x,u) + \mathbb{E}\left(J_{t+1}(X_{t+1}) \;\middle|\; X_t = x, U_t = u\right) \\ &= g(x,u) + \mathbb{E}\left(J_{t+1}\big(f(x,u,W_{t+1})\big) \;\middle|\; X_t = x, U_t = u\right), \quad \forall (x,u) \in \mathbb{X} \times \mathbb{U}. \end{aligned}$$

△

Observe that by setting $h_t(x) = +\infty$ for all $x \in \mathbb{X}$ and $t \in \{0, \ldots, T-1\}$, i.e., removing the option to stop the process prematurely, we end up with the standards DP Algorithm 3.4 as expected. Moreover, we note that the same extensions concerning time-varying dynamics and running cost, stochastic running cost, and state-dependent input constraints described for DP Algorithm 3.4 can be applied to the DP Algorithm 4.10 with stopping.

Remark 4.11 (Modeling with or without τ). Upon recognizing that a given problem is an optimal stopping problem, we have two options:

(1) We can consider the modified state space $\widehat{\mathbb{X}} = \mathbb{X} \cup \{\tau\}$ and develop the modified dynamics $\widehat{f}$, running cost $\widehat{g}_t$ and terminal cost $\widehat{G}$ based on the problem data as described above. Note that, in this case, the stopping cost must be incorporated in the modified running cost. We then use the (standard) DP Algorithm 3.4 to solve this problem.

(2) We can ignore the termination state τ and use the original state space $\mathbb{X}$ and develop the dynamics f, running cost g, and terminal cost G as if we cannot stop the process. In this case, the possibility of stopping the process is modeled by the stopping cost h_t. We then use the DP Algorithm 4.10 with stopping to solve this problem. $\triangle$

4.3.2. Asset selling. Assume that you want to sell an asset (e.g., a piece of land) sometime between now ($t = 0$) and a terminal time ($t = T$). At each time $t \in \{1, \ldots, T\}$, you receive a *random offer* $W_t \in \mathbf{R}_+$. The offers are i.i.d. with a given p.d.f. $p_{W_t}(w) = \rho(w)$ for $w \geq 0$. You can either *accept* or *reject* the offers for $t < T$. However, the offer W_T at the terminal time *must be accepted* if all prior offers were rejected. If you accept the offer at time $t < T$, the money earned is invested at a *fixed per-period interest rate* $r > 0$ until the terminal time $t = T$. Notice that the rejected offers are *not renewed.* The objective is to find a policy for accepting and rejecting offers that *maximizes the revenue at time T*.

The first step is to recognize that this is indeed an optimal stopping problem: The decision to sell corresponds to stopping the process and "leaving the market." The next step is to model this problem as an optimal stopping problem.

The state variable X_t should contain all the relevant information that we need to make the decision to accept or reject an offer, which is nothing but the offer itself. Therefore, the state space is $\mathbb{X} = \mathbf{R}_+$ since the offers are non-negative. The action space is $\mathbb{U} \cup \{\text{accept}\}$, where 'accept' corresponds to 'stop' and $\mathbb{U} = \{\text{reject}\}$ corresponds to the standard action space according to Definition 4.8. The disturbance is also the offered price W_t at time t. Therefore, the dynamics are

$$X_{t+1} = f(X_t, \text{reject}, W_{t+1}) = W_{t+1}, \quad t = 0, 1, \ldots, T-1,$$

with $X_0 = 0$ being a *fictitious null offer.* For the running cost and terminal reward, we have

$$g(x, \text{reject}) = 0, \quad G(x) = x, \quad \forall x \in \mathbf{R}_+,$$

because if we reject an offer X_t at time $t < T$, we receive no money; and, the last offer X_T must be accepted if the asset has not yet been sold. Note that for determining the dynamics f, running cost g, and the terminal cost G, we are essentially ignoring the fact that this is an optimal stopping problem. What remains to be determined is the stopping reward h_t. Recall that if we decide to accept an offer X_t at time $t < T$, we will invest it at the fixed interest rate $r > 0$ until the terminal time $t = T$, which amounts to $(1+r)^{T-t}X_t$ by the end of the horizon. Thus,

$$h_t(x) = (1+r)^{T-t}x, \quad \forall x \in \mathbf{R}_+.$$

By defining the function $H_t(\cdot) = h_t(\cdot)$ for $t < T$ and $= G(\cdot)$ for $t = T$ as in Definition 4.8, the objective is to find an optimal stopping policy that maximizes the expected accumulated reward, i.e.,

$$\max_{\substack{(U_t)_{t=0}^{S-1} \in \mathbb{U}^S \\ S \in \{0,1,\dots,T\}}} \mathbb{E}\left(\sum_{t=0}^{S-1} g_t(X_t, U_t) + H_S(X_S) \;\middle|\; X_0 = 0 \right).$$

We now use DP Algorithm 4.10 with stopping to solve this problem. For *initialization* at $t = T$, we have

$$J_T(x) = G(x) = x, \quad \forall x \in \mathbf{R}_+.$$

The *backward iteration* for $t = T-1, \dots, 0$, then gives us

$$\begin{aligned} J_t(x) &= \max\left\{h_t(x),\ \max_{u \in \mathbb{U}} Q_t(x,u)\right\} \\ &= \max\Bigg\{(1+r)^{T-t}x, \\ &\qquad\qquad g(x, \text{reject}) + \mathbb{E}\left(J_{t+1}\big(f(x, \text{reject}, W_{t+1})\big) \;\middle|\; X_t = x, U_t = u\right)\Bigg\} \\ &= \max\left\{(1+r)^{T-t}x,\ \mathbb{E}\big(J_{t+1}(W_{t+1})\big)\right\} \\ &= (1+r)^{T-t} \cdot \max\left\{x,\ \frac{\mathbb{E}\big(J_{t+1}(W_{t+1})\big)}{(1+r)^{T-t}}\right\}, \quad \forall x \in \mathbf{R}_+, \end{aligned}$$

where for the last equality we used the fact that $1 + r > 0$. Then, defining

$$\alpha_t = \frac{\mathbb{E}\big(J_{t+1}(W_{t+1})\big)}{(1+r)^{T-t}}. \tag{4.20}$$

we have

$$J_t(x) = (1+r)^{T-t} \cdot \max\{x,\ \alpha_t\}, \quad \forall x \in \mathbf{R}_+, \tag{4.21}$$

with the corresponding optimal policy

$$\mu_t(x) = \begin{cases} \text{accept} & \text{if } x \geq \alpha_t, \\ \text{reject} & \text{otherwise,} \end{cases} \quad \forall x \in \mathbf{R}_+.$$

That is, the optimal policy is in the form of a *threshold*: Accept the offer x at time t if $x \geq \alpha_t$ and reject it if $x < \alpha_t$. So, let us compute these thresholds. For $t = T-1$, we have

$$\alpha_{T-1} = \frac{\mathbb{E}\big(J_T(W_T)\big)}{1+r} = \frac{\mathbb{E}(W_T)}{1+r} = \frac{1}{1+r} \cdot \int_{\mathbf{R}_+} w\ \rho(w)\ \mathrm{d}w.$$

For $t < T-1$, we can write

$$\begin{aligned} \alpha_t &= \frac{\mathbb{E}\big(J_{t+1}(W_{t+1})\big)}{(1+r)^{T-t}} && \text{[using (4.20)]} \\ &= \frac{\mathbb{E}\big((1+r)^{T-(t+1)} \cdot \max\{W_{t+1},\ \alpha_{t+1}\}\big)}{(1+r)^{T-t}} && \text{[using (4.21)]} \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{1+r} \cdot \mathbb{E}(\max\{W_{t+1},\ \alpha_{t+1}\}) \\
&= \frac{1}{1+r} \cdot \int_{\mathbf{R}_+} \max\{w,\ \alpha_{t+1}\}\ \rho(w)\ \mathrm{d}w \\
&= \frac{\alpha_{t+1}}{1+r} \int_{-\infty}^{\alpha_{t+1}} \rho(w)\ \mathrm{d}w + \frac{1}{1+r} \int_{\alpha_{t+1}}^{\infty} w\ \rho(w)\ \mathrm{d}w,
\end{aligned}
$$

where the p.d.f. ρ is extended by setting $\rho(w) = 0$ for $w < 0$ (recall that W_t is a non-negative r.v. taking values in $\mathbf{R}_+$). To summarize, the value functions and the optimal policy are of the form

$$
\begin{cases} J_t(x) = (1+r)^{T-t} \cdot \max\{x,\ \alpha_t\}, \\ \mu_t(x) = \text{accept iff } x \geq \alpha_t, \end{cases} \qquad \forall x \in \mathbf{R}_+,\ \forall t \in \{0, 1, \ldots, T\},
$$

where the thresholds α_t can be computed recursively by

$$
\alpha_T = 0 \quad \text{and} \quad \alpha_t = \mathcal{F}(\alpha_{t+1}) \ \text{ for } \ t = T-1, \ldots, 0,
$$

where

$$
\mathcal{F}(\alpha) \coloneqq \frac{\alpha}{1+r} \int_{-\infty}^{\alpha} \rho(w)\ \mathrm{d}w + \frac{1}{1+r} \int_{\alpha}^{\infty} w\ \rho(w)\ \mathrm{d}w.
$$

Note that the initialization $\alpha_T = 0$ leads to the terminal action $\mu_T(x) =$ accept for all $x \in \mathbf{R}_+$ (i.e., sell whatever the terminal offer X_T is), the value function $J_T(x) = x$ for all $x \in \mathbf{R}_+$, and $\alpha_{T-1} = \frac{\mathbb{E}(W_T)}{1+r}$, which are all the same as above.

An interesting observation is that the threshold decreases as the deadline approaches:

Lemma 4.12. *$\alpha_t \geq \alpha_{t+1}$ for all $t \in \{0, 1, \ldots, T-1\}$.*

Proof. The proof is by induction. For the base case $t = T-1$, we clearly have $\alpha_{T-1} = \frac{\mathbb{E}(W_T)}{1+r} \geq 0 = \alpha_T$ since W_T is a non-negative r.v. Let us now assume $\alpha_{t+1} \geq \alpha_{t+2}$ for some $t < T-1$ (induction hypothesis). Then,

$$
\begin{aligned}
\alpha_t &= \frac{1}{1+r} \cdot \mathbb{E}(\max\{W_{t+1},\ \alpha_{t+1}\}) \\
&\geq \frac{1}{1+r} \cdot \mathbb{E}(\max\{W_{t+1},\ \alpha_{t+2}\}) && \text{[induction hypothesis]} \\
&= \frac{1}{1+r} \cdot \mathbb{E}(\max\{W_{t+2},\ \alpha_{t+2}\}) && [W_t \text{ is i.i.d.}] \\
&= \alpha_{t+1}.
\end{aligned}
$$

This completes the proof. □

This means that if an offer is acceptable at time t, it should also be acceptable at time $t+1$; see Figure 10. This is indeed expected based on intuition: As we get closer to the deadline, there is *less chance for improvement* so we have to lower our threshold of acceptance.

Finally, let us consider the behavior of the optimal policy for a long time horizon, i.e., as $T \to \infty$. Equivalently, we can look at the behavior of the backward recursion $\alpha_t = \mathcal{F}(\alpha_{t+1})$ as $t \to -\infty$. Using the property $\alpha_t \geq \alpha_{t+1}$, it can be seen that the sequence α_t converges to a fixed point $\bar{\alpha} = \mathcal{F}(\bar{\alpha})$ as $t \to -\infty$; see Figure 10. This implies that for $T \to \infty$, the optimal policy can be approximated by the stationary policy specified by the acceptance threshold $\bar{\alpha}$, i.e., accept the

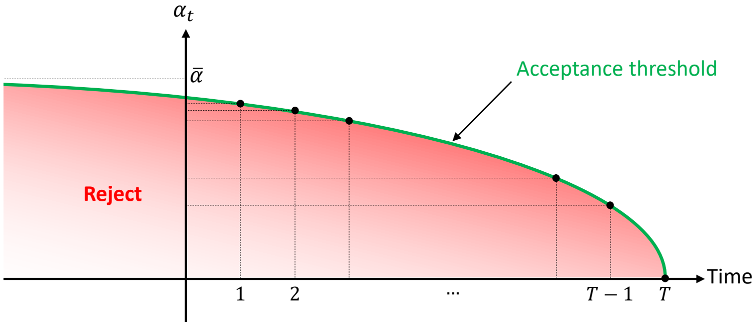


FIGURE 10. The behavior of the acceptance threshold in the asset selling problem. The threshold decreases as the deadline $t = T$ approaches. For long horizons (i.e, $T \to \infty$), the optimal policy can be approximated by the stationary acceptance threshold $\bar{\alpha}$.

offer X_t if and only if $X_t \geq \bar{\alpha}$. As an example, if $W_t \sim \text{Uniform}[0,1]$, i.e., $\rho(w) = 1$ for $w \in [0,1]$ and $= 0$ otherwise, we have

$$\begin{aligned}\mathcal{F}(\alpha) &= \frac{\alpha}{1+r}\int_{-\infty}^{\alpha}\rho(w)\ \mathrm{d}w + \frac{1}{1+r}\int_{\alpha}^{\infty} w\ \rho(w)\ \mathrm{d}w \\ &= \begin{cases} \frac{1}{2(1+r)} & \text{if } \alpha < 0, \\ \frac{\alpha}{1+r}\cdot\alpha + \frac{1}{1+r}\cdot\frac{1}{2}\cdot(1-\alpha^2) = \frac{1+\alpha^2}{2(1+r)} & \text{if } \alpha \in [0,1], \\ \frac{\alpha}{1+r}\cdot\alpha = \frac{\alpha^2}{1+r} & \text{if } \alpha > 1. \end{cases}\end{aligned}$$

One can check that $\mathcal{F}$ has only one fixed point in the interval $[0,1]$. Then, the fixed point $\bar{\alpha}$ can be computed as

$$\bar{\alpha} \in [0,1] \text{ and } \bar{\alpha} = \frac{1+\bar{\alpha}^2}{2(1+r)} \implies \bar{\alpha} = 1 + r - \sqrt{(1+r)^2 - 1}.$$

Observe that as $r \to 0$, we have $\bar{\alpha} \to 1$, that is, we should wait until we see a high offer close to the upper bound 1. On the other hand, as $r \to \infty$, we have $\bar{\alpha} \to 0$, that is, we should immediately accept any non-zero offer. Are these policies justified considering these extreme values of r and T?

4.3.3. Purchase with correlated prices. Assume that a fixed quantity of some raw material is needed within T periods of time. The prices X_t of the material can be described by a linear system driven by independent disturbances. That is, given some initial price $X_0 = x \geq 0$ at time $t = 0$, we have

$$X_{t+1} = \lambda X_t + W_{t+1}, \quad t = 0, 1, \ldots, T-1,$$

where $\lambda \in [0,1)$ is a memory coefficient, while the disturbances $(W_t)_{t=1}^T$ are i.i.d. and take only *positive* values with a given p.d.f. $p_{W_t}(w) = \rho(w)$ for $w > 0$.

We can decide to buy at today's price or to wait a period during which the price can go up or down due to the fluctuations of the market. The objective is to *minimize the expected price of the purchase*. Once again, observe that this is indeed an optimal stopping problem: The decision to buy corresponds to stopping the process and "leaving the market."

Let us now model this problem as an optimal stopping problem according to Definition 4.8. The state variable $X_t \in \mathbf{R}_+$ and the disturbance variable $W_t \in \mathbf{R}_+$

are already defined. The action space is $\mathbb{U} \cup \{\text{buy}\}$, where 'buy' corresponds to 'stop' and $\mathbb{U} = \{\text{wait}\}$ corresponds to the standard action space. For determining the dynamics f, running cost g, and the terminal cost G, we ignore the option to stop the process. Therefore, the dynamics are

$$X_{t+1} = f(X_t, \text{wait}, W_{t+1}) = \lambda X_t + W_{t+1}, \quad t = 0, 1, \dots, T-1,$$

and the costs are

$$g(x, \text{wait}) = 0, \quad G(x) = x, \quad \forall x \in \mathbf{R}_+.$$

For the stopping cost, observe that if we decide to buy at any time $t < T$, we incur the cost $h_t(X_t) = X_t$ with X_t being the current price of the material. Thus,

$$h_t(x) = x, \quad \forall x \in \mathbf{R}_+.$$

We now use DPA with stopping to solve this problem:

- *Initialization* at $t = T$:

$$J_T(x) = G(x) = x, \quad \forall x \in \mathbf{R}_+.$$

- *Backward iteration* for $t = T-1, \dots, 0$:

$$\begin{aligned} J_t(x) &= \min\big\{h_t(x),\ \min_{u \in \mathbb{U}} Q_t(x,u)\big\} \\ &= \min\big\{x,\ g(x, \text{wait}) + \mathbb{E}\left(J_{t+1}\big(f(x, \text{wait}, W_{t+1})\big) \;\middle|\; X_t = x, U_t = u\right)\big\} \\ &= \min\big\{x,\ \mathbb{E}\big(J_{t+1}(\lambda x + W_{t+1})\big)\big\}, \quad \forall x \in \mathbb{X}, \end{aligned} \tag{4.22}$$

with the corresponding optimal policy

$$\mu_t(x) = \begin{cases} \text{buy} & \text{if } x \le \mathbb{E}\big(J_{t+1}(\lambda x + W_{t+1})\big), \\ \text{wait} & \text{otherwise,} \end{cases} \qquad \forall x \in \mathbf{R}_+,$$

Therefore, the optimal policy is to buy at the price x at time t if and only if

$$x \le \mathbb{E}\big(J_{t+1}(\lambda x + W_{t+1})\big).$$

Note that the right-hand side of this inequality also depends on x, and hence, unlike the asset selling problem, it is not yet clear that this condition can be equivalently characterized by a threshold policy $x \le \alpha_t$ for some $\alpha_t \in \mathbf{R}_+$. For $t = T-1$, this is indeed the case since we have

$$\begin{aligned} J_{T-1}(x) &= \min\{x,\ \mathbb{E}\big(J_T(\lambda x + W_T)\big)\} \\ &= \min\{x,\ \mathbb{E}(\lambda x + W_T)\} \\ &= \min\{x,\ \lambda x + \mathbb{E}(W_T)\} \\ &= \min\{x,\ \lambda x + \bar{w}\}, \end{aligned}$$

where $\bar{w} \coloneqq \mathbb{E}(W_t) > 0$, and hence

$$\mu_{T-1}(x) = \text{buy} \quad \text{iff} \quad x \le \lambda x + \bar{w} \quad \text{iff} \quad x \le \alpha_{T-1} = \frac{\bar{w}}{1-\lambda}.$$

See Figure 11. In what follows, we will show that this is also the case for $t < T-1$.

Lemma 4.13. *$J_t(x) \le J_{t+1}(x)$ for all $x \in \mathbf{R}_+$ and all $t \in \{0, 1, \dots, T-1\}$.*

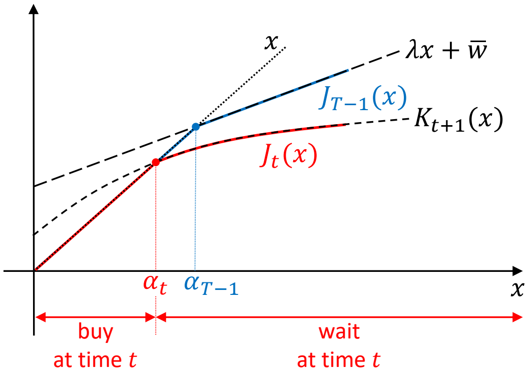


FIGURE 11. The value function and the buying threshold at time $T-1$ and a generic time $t < T-1$. Observe that the equation $x = K_{t+1}(x)$ has exactly one solution which determines the the buying threshold α_t at time t.

PROOF. The proof is as usual by induction. For the base case $t = T-1$, we have

$$J_{T-1}(x) = \min\{x,\ \lambda x + \bar{w}\} \le x = J_T(x), \quad \forall x \in \mathbf{R}_+.$$

Let us now assume $J_{t+1}(x) \le J_{t+2}(x)$ for all $x \in \mathbf{R}_+$ and some $t < T-1$ (induction hypothesis). Then, for each $x \in \mathbf{R}_+$, we have

$$\begin{aligned} J_t(x) &= \min\left\{x,\ \mathbb{E}\big(J_{t+1}(\lambda x + W_{t+1})\big)\right\} \\ &\le \min\left\{x,\ \mathbb{E}\big(J_{t+2}(\lambda x + W_{t+1})\big)\right\} && \text{[induction hypothesis]} \\ &= \min\left\{x,\ \mathbb{E}\big(J_{t+2}(\lambda x + W_{t+2})\big)\right\} && [W_t\text{'s are i.i.d.}] \\ &= J_{t+1}(x) \end{aligned}$$

This completes the proof. □

LEMMA 4.14. *$J_t(x)$ is a non-negative, increasing, concave function in x and $\mathbb{E}\big(J_t(W_t)\big) > 0$ for all $t \in \{0, 1, \dots, T\}$.*

PROOF. The proof is again by induction. Let $\mathbb{J}$ denote the class of non-negative, increasing, and concave functions defined on $\mathbf{R}_+$. For the base case $t = T$, we have $J_T(x) = x$, which is clearly in the class $\mathbb{J}$. Also, $\mathbb{E}\big(J_T(W_T)\big) = \mathbb{E}(W_T) = \bar{w} > 0$ by the assumption that W_t takes only positive values. Let us now assume that $J_{t+1} \in \mathbb{J}$ and $\mathbb{E}\big(J_{t+1}(W_{t+1})\big) > 0$ (induction hypothesis). First, observe that

$$\begin{aligned} & J_{t+1}(x) \in \mathbb{J} \\ \Longrightarrow\ & J_{t+1}(\lambda x + w) \in \mathbb{J}, \quad \forall w > 0 \\ \Longrightarrow\ & \mathbb{E}\big(J_{t+1}(\lambda x + W_{t+1})\big) \in \mathbb{J} \\ \Longrightarrow\ & \min\left\{x,\ \mathbb{E}\big(J_{t+1}(\lambda x + W_{t+1})\big)\right\} \in \mathbb{J} \\ \Longrightarrow\ & J_t(x) \in \mathbb{J}. \end{aligned}$$

Above, we used the fact that if $J, \tilde{J} \in \mathbb{J}$, then $\theta J + (1-\theta)\tilde{J} \in \mathbb{J}$ for all $\theta \in [0,1]$, and $\min\{J, \tilde{J}\} \in \mathbb{J}$. Also, we have

$$\mathbb{E}\big(J_t(W_t)\big) = \mathbb{E}\Big(\min\left\{W_t,\ \mathbb{E}\big(J_{t+1}(\lambda W_t + W_{t+1})\big)\right\}\Big)$$

$$\begin{aligned} &\geq \mathbb{E}\big(\min\big\{W_t,\ \mathbb{E}\big(J_{t+1}(W_{t+1})\big)\big\}\big) && [W_t > 0 \text{ and } J_{t+1} \text{ is increasing}] \\ &> 0 && [W_t > 0 \text{ and induction hypothesis}] \end{aligned}$$

This completes the proof. □

The preceding lemmas imply that, for each $t \in \{0, 1, \dots, T-1\}$, the function $K_{t+1}(x) = \mathbb{E}\big(J_{t+1}(\lambda x + W_{t+1})\big)$ is a non-negative, increasing, concave function for $x \in \mathbf{R}_+$ such that $K_{t+1}(0) > 0$ and $K_{t+1}(x) \leq \lambda x + \bar{w}$ for all $x \in \mathbf{R}_+$. This, in turn, implies that

$$x \leq \mathbb{E}\big(J_{t+1}(\lambda x + W_{t+1})\big) \iff x \leq \alpha_t,$$

where α_t is the *unique* solution of the equation

$$x = \mathbb{E}\big(J_{t+1}(\lambda x + W_{t+1})\big).$$

See Figure 11. Finally, we note that Lemma 4.13 also implies that

$$\alpha_t \leq \alpha_{t+1}, \quad \forall t \in \{0, 1, \dots, T-2\},$$

that is, the threshold price increases as the deadline approaches.

4.3.4. Marriage problem. [9] [10] A *bride(-to-be)* is presented with a number T of *bachelors* to choose from to marry. The bachelors, if all presented together, can be *ranked* by the bride from the best to the worst unambiguously. However, they are presented one at a time and in a random order such that all the possible $T! = T \times (T-1) \times \cdots \times 1$ orders are equally likely. At each time, as a bachelor is presented, the bride can either *accept* the presented bachelor, in which case the process stops, or *reject* him. To make that decision, the bride can only use the *apparent* rank of the presented bachelor within the group of already presented bachelors. The bride *must*(!) marry one of the presented bachelors, so she must accept the last bachelor, if presented. Note that the decision made at each time is irrevocable, that is, the bride can never go back and choose a previously rejected bachelor who, in retrospect, turns out to be the best. The bride wishes to maximize *the probability of accepting the best of all T bachelors.*

It should be clear to you that this is indeed an optimal stopping problem. Before formulating this problem as a standard optimal stopping problem following Definition 4.8, let us compute the reward (i.e., the probability of accepting the best of all T bachelors) of a generic *open-loop* policy that decides, in advance, to accept the bachelor presented at a fixed time $t \in \{1, \dots, T\}$. Considering the fact that all the possible $T!$ orders are equally likely, the probability that the best bachelor is presented at a certain time t is $\frac{1}{T}$, for each $t \in \{1, \dots, T\}$. So, any open-loop policy has an expected reward of $\frac{1}{T}$. In what follows, we want to see if the bride can increase this probability by using the sequentially revealed information about the *apparent* rank of the presented bachelors.

The state variable X_t of the process should contain all the information available for making a decision which is exactly the *apparent rank of the bachelor presented at time t within the group of already presented bachelors up to time t.* See Figure 12 for an illustration. Therefore, $\mathbb{X}_t = [t] = \{1, \dots, t\}$ which is *time-dependent.* Observe that what we just described is also the disturbance variable W_t at time t and hence

[9] Disclaimer: The use of gender-specific terms in this section is solely to simplify the exposition. No offense or discrimination is intended.

[10] This problem has many names; another common name is *the secretary problem.*

$\mathbb{W}_t = [t]$. Moreover, since all the possible $T!$ orders of the bachelors are equally likely, the apparent rank of the bachelor presented at time t can be any of numbers $1, 2, \dots, t$, with equal probability, i.e.,

$$p_{W_t}(w) = \frac{1}{t}, \quad \forall w \in \mathbb{W}_t = [t].$$

Observe that W_t's are *independent* but *not identically distributed.* The action space is clearly $\mathbb{U} \cup \{\text{accept}\}$, where 'accept' corresponds to 'stop' and $\mathbb{U} = \{\text{reject}\}$ corresponds to the standard action space according to Definition 4.8. Based on the discussion above on the state and disturbance variables, one can see that the dynamics are given by

$$X_{t+1} = f(X_t, \text{reject}, W_{t+1}) = W_{t+1}, \quad t = 0, 1, \dots, T-1,$$

with $X_0 = 1$ being a *fictitious bachelor* presented at time $t = 0$. For the running and terminal rewards, we have

$$g_t(x, \text{reject}) = 0 \quad \forall x \in [t], \quad G(x) = \begin{cases} 1 & \text{if } x = 1, \\ 0 & \text{if } x \in \{2, 3, \dots, T\}. \end{cases}$$

To see this, let us ignore the option to stop the process for a moment. Then, the bride must marry the last bachelor that is presented to her at time T knowing the *absolute* rank X_T of the last bachelor among all T bachelors. This means that there is no running reward because all the bachelors presented at time $t < T$ are rejected (the only way to receive a reward is to accept a bachelor!). On the other hand, the probability of having the best bachelor at the terminal time is exactly 1 if $X_T = 1$ and 0 otherwise as given by terminal reward G. The next ingredient is the stopping reward h_t. Note that $h_t(x)$ is the probability of accepting the best of all T bachelors given that we decide to accept the bachelor presented at time t with apparent rank $x \in [t]$. Clearly, $h_t(x) = 0$ of $x \neq 1$: if the bachelor presented at time t is not the best among the bachelors presented up to time t, then he cannot be the best among all T bachelors. Let us next consider the case $x = 1$ corresponding to the bachelor presented at time t being the best among the bachelors presented up to time t and denote this event by B_t. Also, let A_t denote the event that the bachelor presented at time t is the best among all T bachelors and observe that $A_t \subset B_t$. Then, by definition, we have

$$h_t(1) = \mathbb{P}(A_t \mid B_t) = \frac{\mathbb{P}(A_t \cap B_t)}{\mathbb{P}(B_t)} = \frac{\mathbb{P}(A_t)}{\mathbb{P}(B_t)} = \frac{1/T}{1/t} = \frac{t}{T},$$

where we used the fact that $\mathbb{P}(A_t) = 1/T$ and $\mathbb{P}(B_t) = 1/t$ based on the assumption that any ordering of the T bachelors is equally likely. Therefore, we have

$$h_t(x) = \begin{cases} t/T & \text{if } x = 1, \\ 0 & \text{if } x \in \{2, 3, \dots, t\}. \end{cases}$$

Here comes the last element of modeling: Observe that the running, terminal, and stopping rewards are the same for all $x \neq 1$. Also, in the dynamics, the probability distribution of the next state does not depend on the current state. This means that as far as the value functions and the optimal policy are concerned, these states are the same. Therefore, we can group all the states $x \neq 1$ together and treat them as a single state $\neg 1$. Note that this also makes sense intuitively considering the objective to be maximized. The bride aims to maximize the probability of accepting *the best of all T bachelors.* Thus, if, at time t, she is presented by a bachelor of *apparent* rank $x \neq 1$ (i.e., who is not the best among the bachelors that the bride has seen so

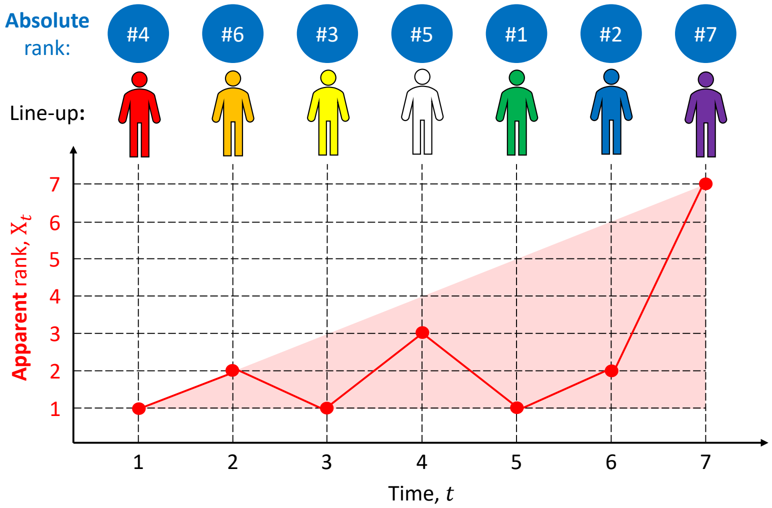


FIGURE 12. An illustration of the evolution of the state variable X_t for the marriage problem with $T = 7$ bachelors. The blue-filled circles in the top show the ordering of the bachelors in this example with $\#i$ being the *absolute* rank of the bachelor presented at each time within the whole group of T bachelors – note that the bride does not have access to this information. X_t is the *apparent* rank of the bachelor presented at time t within the group of already presented bachelors up time t – this is the information available to the bride. The red-shaded area shows the state space $\mathbb{X}_t = [t]$ at time t. Observe that the bachelor presented at time $t = 3$ is better than the first two bachelors presented at times $t = 1$ and $t = 2$ and hence its apparent rank is $X_3 = 1$.

far), then she should immediately reject him because he cannot be the best of all T bachelors. In other words, for the bride, it should not make any difference what the apparent rank of the presented bachelor is, unless it is $x = 1$.

Based on the discussion above, we can formulate the marriage problem as an optimal stopping problem following Definition 4.8 with

- state and disturbance spaces $\{1, \neg 1\}$ and action space $\{\text{reject}, \text{accept}\}$ where the action 'accept' corresponds to 'stopping' the process;
- dynamics

$$X_t = W_t = \begin{cases} 1 & \text{with probability } 1/t, \\ \neg 1 & \text{with probability } (t-1)/t, \end{cases} \qquad t = 1, 2. \dots, T;$$

- rewards

$$g(x, \text{reject}) = 0, \quad x \in \{1, \neg 1\},$$
$$G(x) = \begin{cases} 1 & \text{if } x = 1, \\ 0 & \text{if } x = \neg 1, \end{cases}$$
$$h_t(x) = \begin{cases} t/T & \text{if } x = 1, \\ 0 & \text{if } x = \neg 1, \end{cases} \qquad t = 1, 2. \dots, T-1.$$

We now use DP Algorithm 4.10 with stopping to solve this problem:

- *Initialization* at $t = T$:

$$J_T(x) = G(x) = \begin{cases} 1 & \text{if } x = 1, \\ 0 & \text{if } x = \neg 1. \end{cases}$$

- *Backward iteration* for $t = T-1, \ldots, 1$:

$$\begin{aligned} J_t(x) &= \max\left\{h_t(x),\ g_t(x, \text{reject}) + \mathbb{E}\left(J_{t+1}(X_{t+1}) \;\middle|\; X_t = x, U_t = u\right)\right\} \\ &= \begin{cases} \max\left\{\frac{t}{T},\ \frac{1}{t+1}J_{t+1}(1) + \frac{t}{t+1}J_{t+1}(\neg 1)\right\} & \text{if } x = 1, \\ \max\left\{0,\ \frac{1}{t+1}J_{t+1}(1) + \frac{t}{t+1}J_{t+1}(\neg 1)\right\} & \text{if } x = \neg 1, \end{cases} \end{aligned}$$

and thus

$$\begin{cases} J_t(\neg 1) &= \frac{1}{t+1}J_{t+1}(1) + \frac{t}{t+1}J_{t+1}(\neg 1), \\ J_t(1) &= \max\{\frac{t}{T},\ J_t(\neg 1)\}, \end{cases} \tag{4.23}$$

with the corresponding optimal policy

$$\begin{cases} \mu_t(\neg 1) &= \text{reject}, \\ \mu_t(1) &= \text{accept iff } \frac{t}{T} \geq J_t(\neg 1). \end{cases}$$

By now, it shouldn't be surprising to you that $\mu_t(\neg 1) =$ reject, i.e., the bride should *only* consider accepting a bachelor if he is the best among all the bachelors that are presented so far with apparent rank $x = 1$. The necessary and sufficient condition for accepting such a bachelor is $\frac{t}{T} \geq J_t(\neg 1)$. As an example, for $T = 7$ bachelors, the value function and the optimal policy can be computed as follows:

t	1	2	3	4	5	7	7
$J_t(1)$	0.41	0.41	0.43	0.57	0.71	0.85	1
$J_t(\neg 1)$	-	0.41	0.40	0.35	0.26	0.14	0
$\mu_t(1)$	reject	reject	accept	accept	accept	accept	-
$\mu_t(\neg 1)$	-	reject	reject	reject	reject	reject	-

See also Figure 13. Observe that the optimal policy is in the form of a threshold such that the bride must accept a bachelor who is best out of t (i.e., with apparent rank $X_t = 1$) for all $t \geq t^* = 3$. That is, the bride must reject the first two presented bachelors (even if the second presented bachelor is better than the first presented bachelor) and then, from the third presented bachelor onward, accept the one that is better than all the bachelors she has seen so far. Also, observe that $J_1(1) = 0.41 > 1/7 = 0.14$, that is, the derived policy indeed outperforms any open-loop policy by increasing the probability of accepting the best out of all T bachelors.

The threshold behavior we observed is not specific to the preceding example and is generic as shown by the following lemma.

Lemma 4.15. *If it is optimal to accept a bachelor who is* the best out of $t-1$, *then it is also optimal to accept one who is* the best out of t.

Proof. We prove the contra-position: If it is *not* optimal to accept a bachelor who is *the best out of* t, then it is *not* optimal to accept one who is *the best out of* $t-1$. So, let us assume that $\mu_t(1) =$ reject, i.e., the t-th bachelor (with apparent rank $x = 1$) is rejected, and hence $\frac{t}{T} \leq J_t(\neg 1) = J_t(1)$. Then, using the first

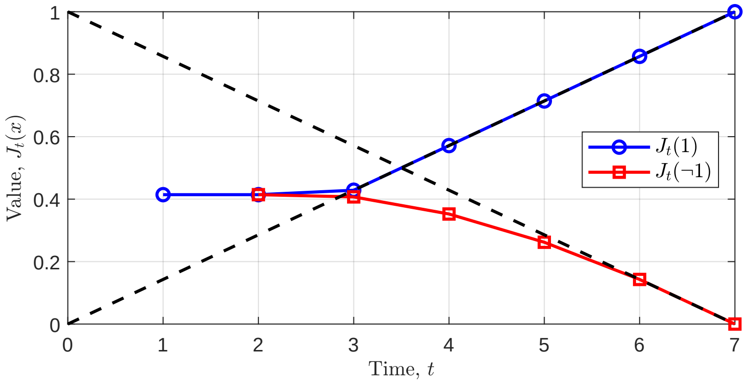


FIGURE 13. An example of the value function $J_t(x)$ in the marriage problem for $T = 7$ bachelors. For $t \geq t^* = 3$, the optimal policy is to accept a bachelor who is best out of t.

equation in the iteration (4.23), we have

$$J_{t-1}(\neg 1) = \frac{1}{t} J_t(1) + \frac{t-1}{t} \cdot J_t(\neg 1) = J_t(\neg 1) \geq \frac{t}{T} > \frac{t-1}{T}.$$

Hence, by the second equation in the iteration (4.23),

$$J_{t-1}(1) = \max\left\{\frac{t-1}{T},\ J_{t-1}(\neg 1)\right\} = J_{t-1}(\neg 1),$$

with the corresponding optimal policy $\mu_{t-1}(1) = $ reject, i.e., the $(t-1)$-th bachelor (with apparent rank $x = 1$) must also be rejected. □

The preceding result indeed implies that the optimal policy is to accept a bachelor who is best out of t if $t \geq t^*$. Let us now compute this threshold t^*: By definition, we have

$$\frac{t}{T} < J_t(\neg 1) = J_t(1), \quad \forall t < t^*,$$
$$J_t(\neg 1) \leq \frac{t}{T} = J_t(1), \quad \forall t \geq t^*.$$

Therefore, t^* is the *smallest* $t \in \{1, 2, \ldots, T\}$ for which $J_t(\neg 1) \leq \frac{t}{T}$ holds true; see Figure 13. Also, using the DPA above and the fact that $J_t(1) = \frac{t}{T}$ for $t \geq t^*$, we have the recursion

$$J_T(\neg 1) = 0,$$
$$J_t(\neg 1) = \frac{J_{t+1}(1)}{t+1} + \frac{t \cdot J_{t+1}(\neg 1)}{t+1} = \frac{1}{T} + \frac{t \cdot J_{t+1}(\neg 1)}{t+1}, \quad t = T-1, T-2, \ldots, t^*.$$

Defining $V_t \coloneqq \frac{T}{t} \cdot J_t(\neg 1)$, the above recursion reduces to

$$V_T = 0,$$
$$V_t = \frac{1}{t} + V_{t+1}, \quad t = T-1, T-2, \ldots, t^*.$$

The solution of the preceding recursion can be computed explicitly as

$$V_t = \frac{1}{t} + \frac{1}{t+1} + \cdots + \frac{1}{T-1}, \quad t \in \{t^*, \ldots, T-1\}.$$

Recall that t^* is the *smallest* $t \in \{1, 2, \ldots, T\}$ for which

$$J_t(\neg 1) \leq \frac{t}{T} \iff V_t = \frac{T}{t} \cdot J_t(\neg 1) \leq 1.$$

That is,

$$t^* = \min \left\{ t \in \{1, \ldots, T\} \; : \; \sum_{s=t}^{T-1} \frac{1}{s} \leq 1 \right\}.$$

Then, using the approximation

$$1 \approx \frac{1}{t^*} + \frac{1}{t^*+1} + \cdots + \frac{1}{T-1} \approx \int_{t^*}^{T} \frac{1}{t} \, \mathrm{d}t = \log \frac{T}{t^*},$$

we have $t^* \approx T/e$. In summary, we have found that the optimal policy is to *reject a fraction* $1/e \approx 0.37$ *of the bachelors and then accept any subsequent bachelor of apparent rank 1.* Note that the corresponding probability of accepting the best of all T bachelors under this policy (i.e., $J_1(1)$) is approximately $1/e$.

APPENDIX A

Preliminaries

In this chapter, we briefly review some preliminaries that are needed for this course.

A.1. Probability

A.1.1. Random variables. Consider a random variable (r.v.) X taking values in the "space" $\mathbb{X} \subset \mathbf{R}$. We use $\mathbb{P}(A) = \mathbb{P}(X \in A)$ to denote the probability of a (measurable) subset (a.k.a. event) A of $\mathbb{X}$, that is, the probability of a realization of X taking a value in A. In particular, we have $\mathbb{P}(\mathbb{X}) = 1$ and $\mathbb{P}(\emptyset) = 0$. We use $x \in \mathbb{X}$ to denote a generic value that the random variable X can take.

Example A.1 (Fair die). Consider a fair die with $\mathbb{X} = \{1, \dots, 6\}$ and probability *mass* function (p.m.f.) $p(x) = 1/6$ for all $x \in \mathbb{X}$ so that

$$\mathbb{P}(A) = \sum_{x \in \mathbb{X}} p(x) \mathbb{1}_A(x),$$

where $\mathbb{1}_A : \mathbb{X} \to \{0, 1\}$ is the indicator function of the event $A \subset \mathbb{X}$, that is,

$$\mathbb{1}_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases}$$

For instance, the probability that an odd number shows up when we roll the die is $\mathbb{P}(\{1, 3, 5\}) = 3/6 = 1/2$. This is an example of a *discrete* space $\mathbb{X}$. △

Example A.2 (Uniform r.v.). Consider the interval $\mathbb{X} = [0, 1]$ with probability *density* function (p.d.f.) $p(x) = 1$ for all $x \in \mathbb{X}$ so that

$$\mathbb{P}(A) = \int_{\mathbb{X}} p(x)\ \mathbb{1}_A(x)\ \mathrm{d}x = \int_A p(x)\ \mathrm{d}x$$

for all (measurable) $A \subset \mathbb{X}$. For instance, we have $\mathbb{P}([a, b]) = b - a$ for $b \geq a \geq 0$. This is an example of a *continuous* space $\mathbb{X}$. △

Hereafter, we focus on discrete r.v.'s while noting that the same definitions and results also hold for continuous r.v.'s.

A.1.2. Marginalization and conditioning. Consider two *discrete* r.v.'s $X \in \mathbb{X}$ and $y \in \mathbb{Y}$ with the *joint* probability distribution

$$p(x, y) = \mathbb{P}(X = x, Y = y), \quad \forall (x, y) \in \mathbb{X} \times \mathbb{Y}.$$

Using the joint distribution of these r.v.'s, we can compute the following probabilities (see Figure 1):

- The *marginal* probability of X is

$$p_X(x) = \mathbb{P}(X = x) = \sum_{y \in \mathbb{Y}} p(x, y), \quad \forall x \in \mathbb{X}.$$

- The *conditional* probability of X given Y is

$$p_{X|Y}(x|y) = \mathbb{P}(X = x \mid Y = y) = \frac{p(x,y)}{p_Y(y)}, \quad \forall (x,y) \in \mathbb{X} \times \mathbb{Y},\ p_Y(y) \neq 0,$$

  where p_Y is the marginal probability of Y.

Observe that both marginal and conditional probabilities above are probability distributions over $\mathbb{X}$, however, the conditional one is actually a function of $y \in \mathbb{Y}$. In particular, if the two r.v.'s are *independent*, the two probabilities coincide, i.e.,

$$p_{X|Y}(x|y) = p_X(x), \quad \forall (x,y) \in \mathbb{X} \times \mathbb{Y}.$$

or, equivalently,

$$p(x,y) = p_X(x) \cdot p_Y(y), \quad \forall (x,y) \in \mathbb{X} \times \mathbb{Y},$$

that is, the knowledge of Y gives us no information about X. From the probabilities provided above, we can immediately derive the following results:

LEMMA A.3 (Law of total probability). *We have*

$$p_X(x) = \sum_{y \in \mathbb{Y}} p_{X|Y}(x|y) \cdot p_Y(y), \quad \forall x \in \mathbb{X}.$$

LEMMA A.4 (Bayes' rule). *We have*

$$p_{X|Y}(x|y) \cdot p_Y(y) = p_{Y|X}(y|x) \cdot p_X(x), \quad \forall (x,y) \in \mathbb{X} \times \mathbb{Y}.$$

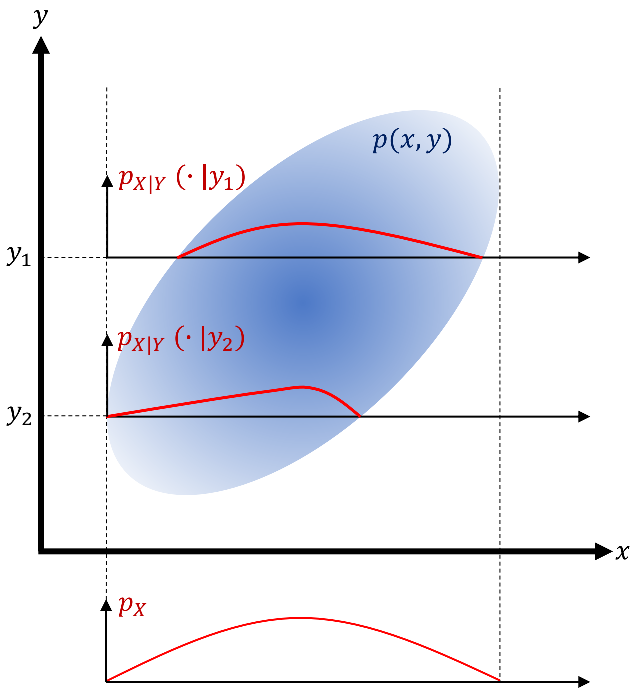


FIGURE 1. Marginal and conditional probabilities.

A.1.3. Expectation and variance. Once again, consider two *discrete* r.v.'s $X \in \mathbb{X}$ and $y \in \mathbb{Y}$ with the *joint* probability distribution

$$p(x, y) = \mathbb{P}(X = x, Y = y), \quad \forall (x, y) \in \mathbb{X} \times \mathbb{Y}.$$

and let

$$p_X(x) = \mathbb{P}(X = x) = \sum_{y \in \mathbb{Y}} p(x, y), \quad \forall x \in \mathbb{X}.$$

be the marginal probability of X. We have the following definitions.

- The *expectation* of X is
$$\mathbb{E}(X) = \sum_{x \in \mathbb{X}} x \cdot p_X(x).$$
- The *variance* of X is
$$\mathrm{var}(X) = \mathbb{E}\big((X - \mathbb{E}(X))^2\big) = \mathbb{E}(X^2) - \mathbb{E}(X)^2.$$
- The *conditional* expectation of X given Y is
$$\mathbb{E}(X|Y = y) = \sum_{x \in \mathbb{X}} x \cdot p_{X|Y}(x|y), \quad \forall y \in \mathbb{Y}.$$

Observe that, by definition, for any function $f : \mathbf{R} \to \mathbf{R}$, we have

$$\mathbb{E}\big(f(X)\big) = \sum_{x \in \mathbb{X}} f(x) \cdot p_X(x).$$

Also of importance are the following results concerning the expectation operation:

Lemma A.5 (Linearity of expectation). *We have*

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

Lemma A.6 (Law of total expectation). *We have*

$$\mathbb{E}(X) = \mathbb{E}\big(\mathbb{E}(X|Y)\big).$$

A.2. Optimization

Given a function $f : \mathbf{R}^n \to \mathbf{R}$ and a set $\mathbb{X} \in \mathbf{R}^n$, consider the following optimization problem

$$\text{minimize } f(x) \text{ subject to } x \in \mathbb{X},$$

which aims to find the minimum value that f can take with its argument x being in the set $\mathbb{X}$. Other standard notations for such an optimization problem are

$$\begin{array}{rl} \min & f(x) \\ \text{s.t.} & x \in \mathbb{X}. \end{array}$$

or

$$\min_{x \in \mathbb{X}} f(x).$$

Note the equivalence

$$\min_{x \in \mathbb{X}} f(x) = -\max_{x \in \mathbb{X}} \; -f(x),$$

which allows us to easily transform a minimization problem into a maximization problem and vice versa. Moreover, we use the notation

$$x^\star = \operatorname*{argmin}_{x \in \mathbb{X}} f(x),$$

to denote the optimal solution (minimizer) so that

$$f(x) \geq f(x^\star), \quad \forall x \in \mathbb{X}.$$

Strictly speaking, one needs to first show that the minimum is attained and the optimal solution is unique to use the notations above:

- If the minimum is not attained, the 'correct' notation is $\inf_{x \in \mathbb{X}} f(x)$.
- If the minimum is attained, but the minimizer is not unique, the 'correct' notation is $x^\star \in \operatorname{argmin}_{x \in \mathbb{X}} f(x)$.

A sufficient condition for the attainment of the minimum is that f is continuous and $\mathbb{X}$ is compact (i.e., closed and bounded).[1] A sufficient condition for the uniqueness of the optimal solution is that f is strictly convex. A classic example is the minimization of a strictly convex quadratic function, i.e.,

$$\operatorname*{argmin}_{x \in \mathbf{R}^n} \{x^\top Q x + q^\top x\} = -\frac{1}{2} Q^{-1} q, \tag{A.1}$$

where $Q \in \mathbf{R}^{n \times n}$ is *positive definite* and $q \in \mathbf{R}^n$.

We note that, in these notes, we use 'min/max' (as opposed to 'inf/sup') for defining optimization problems and use '=' (as opposed to '$\in$') for minimizers, noticing that there is an underlying assumption that the extremum is attained and the minimizer is unique.

A.3. Mathematical induction

Mathematical induction is a neat trick for showing that a statement $\mathcal{S}_k$ is true for all $k \in [n]$ where $n \in \mathbf{N}^\infty$, without actually going through the trouble of exhausting the index set $[n]$. The classic example is the equality

$$\mathcal{S}_k : \; 1 + 2 + \ldots + k = \frac{k(k+1)}{2}, \quad \forall k \in \mathbf{N}.$$

How should one go about proving this? Well, observe that for $k = 1$ (*the base case*), we clearly have $1 = \frac{1 \cdot (1+1)}{2}$ and hence $\mathcal{S}_1$ is indeed true. Next, let us assume $\mathcal{S}_k$ is true, i.e., $1 + 2 + \ldots + k = \frac{k(k+1)}{2}$, for some $k \in \mathbf{N}$ (*induction hypothesis*). Now, observe that

$$1 + 2 + \ldots + k + (k+1) = \frac{k(k+1)}{2} + (k+1) = \frac{(k+1)(k+2)}{2},$$

where for the first equality, we used the induction hypothesis. This means that if $\mathcal{S}_k$ is true, then $\mathcal{S}_{k+1}$ is true (*induction step*) for all $k \in \mathbf{N}$. Combining this result with the base case, we can indeed conclude that $\mathcal{S}_k$ is true for all $k \in \mathbf{N}$.

What we have just explained are the two steps of *proof by induction*: to show that a statement $\mathcal{S}_k$ is true for all $k \in [n]$, it suffices to show that

- *Base case*: $\mathcal{S}_k$ is true for $k = 1$.
- *Induction step*: if $\mathcal{S}_k$ is true for some $k < n$ (induction hypothesis), then $\mathcal{S}_{k+1}$ is true.

This method is the most important tool that we use in this course for proving formal statements concerning dynamic programming algorithm and its applications.

[1] Weierstrass extreme value theorem.