Appendix A

Preliminaries for Probability Theory

Reinforcement learning heavily relies on probability theory. We next summarize some concepts and results frequently used in this book.

◇ *Random variable*: The term "variable" indicates that a random variable can take values from a set of numbers. The term "random" indicates that taking a value must follow a probability distribution.

A random variable is usually denoted by a capital letter. Its value is usually denoted by a lowercase letter. For example, X is a random variable, and x is a value that X can take.

This book mainly considers the case where a random variable can only take a finite number of values. A random variable can be a scalar or a vector.

Like normal variables, random variables have normal mathematical operations such as summation, product, and absolute value. For example, if X, Y are two random variables, we can calculate $X + Y$, $X + 1$, and XY.

◇ *A stochastic sequence* is a sequence of random variables.

One scenario we often encounter is collecting a stochastic sampling sequence $\{x_i\}_{i=1}^n$ of a random variable X. For example, consider the task of tossing a die n times. Let x_i be a random variable representing the value obtained for the ith toss. Then, $\{x_1, x_2, \ldots, x_n\}$ is a stochastic process.

It may be confusing to beginners why x_i is a random variable instead of a deterministic value. In fact, if the sampling sequence is {1,6,3,5,...}, then this sequence is not a stochastic sequence because all the elements are already determined. However, if we use a variable x_i to represent the values that can possibly be sampled, it is a random variable since x_i can take any value in $\{1, \ldots, 6\}$. Although x_i is a lowercase letter, it still represents a random variable.

◇ *Probability*: The notation $p(X = x)$ or $p_X(x)$ describes the probability of the random variable X taking the value x. When the context is clear, $p(X = x)$ is often written as $p(x)$ for short.

◇ *Joint probability*: The notation $p(X = x, Y = y)$ or $p(x, y)$ describes the probability of the random variable X taking the value x and Y taking the value y. One useful identity is as follows:

$$\sum_y p(x, y) = p(x).$$

◇ *Conditional probability*: The notation $p(X = x|A = a)$ describes the probability of the random variable X taking the value x given that the random variable A has already taken the value a. We often write $p(X = x|A = a)$ as $p(x|a)$ for short.

It holds that

$$p(x, a) = p(x|a)p(a)$$

and

$$p(x|a) = \frac{p(x, a)}{p(a)}.$$

Since $p(x) = \sum_a p(x, a)$, we have

$$p(x) = \sum_a p(x, a) = \sum_a p(x|a)p(a),$$

which is called the *law of total probability*.

◇ *Independence*: Two random variables are *independent* if the sampling value of one random variable does not affect the other. Mathematically, X and Y are independent if

$$p(x, y) = p(x)p(y).$$

Since $p(x, y) = p(x|y)p(y)$, the above equation implies

$$p(x|y) = p(x).$$

◇ *Conditional independence:* Let X, A, B be three random variables. X is said to be conditionally independent of A given B if

$$p(X = x|A = a, B = b) = p(X = x|B = b).$$

In the context of reinforcement learning, consider three consecutive states: s_t, s_{t+1}, s_{t+2}. Since they are obtained consecutively, s_{t+2} is dependent on s_{t+1} and also s_t. However, if s_{t+1} is already given, then s_{t+2} is conditionally independent of s_t. That is

$$p(s_{t+2}|s_{t+1}, s_t) = p(s_{t+2}|s_{t+1}).$$

This is also the memoryless property of Markov processes.

◇ *Law of total probability*: The law of total probability was already mentioned when we

introduced the concept of conditional probability. Due to its importance, we list it again below:

$$p(x) = \sum_y p(x, y)$$

and

$$p(x|a) = \sum_y p(x, y|a).$$

◇ *Chain rule* of conditional probability and joint probability. By the definition of conditional probability, we have

$$p(a, b) = p(a|b)p(b).$$

This can be extended to

$$p(a, b, c) = p(a|b, c)p(b, c) = p(a|b, c)p(b|c)p(c),$$

and hence $p(a, b, c)/p(c) = p(a, b|c) = p(a|b, c)p(b|c)$. The fact that $p(a, b|c) = p(a|b, c)p(b|c)$ implies the following property:

$$p(x|a) = \sum_b p(x, b|a) = \sum_b p(x|b, a)p(b|a).$$

◇ *Expectation/expected value/mean*: Suppose that X is a random variable and the probability of taking the value x is $p(x)$. The expectation, expected value, or mean of X is defined as

$$\mathbb{E}[X] = \sum_x p(x)x.$$

The linearity property of expectation is

$$\begin{aligned}\mathbb{E}[X + Y] &= \mathbb{E}[X] + \mathbb{E}[Y], \\ \mathbb{E}[aX] &= a\mathbb{E}[X].\end{aligned}$$

The second equation above can be trivially proven by definition. The first equation is proven below:

$$\begin{aligned}\mathbb{E}[X + Y] &= \sum_x \sum_y (x + y)p(X = x, Y = y) \\ &= \sum_x x \sum_y p(x, y) + \sum_y y \sum_x p(x, y) \\ &= \sum_x xp(x) + \sum_y yp(y) \\ &= \mathbb{E}[X] + \mathbb{E}[Y].\end{aligned}$$

Due to the linearity of expectation, we have the following useful fact:

$$\mathbb{E}\left[\sum_i a_i X_i\right] = \sum_i a_i \mathbb{E}[X_i].$$

Similarly, it can be proven that

$$\mathbb{E}[AX] = A\mathbb{E}[X],$$

where $A \in \mathbb{R}^{n\times n}$ is a deterministic matrix and $X \in \mathbb{R}^n$ is a random vector.

◇ *Conditional expectation*: The definition of conditional expectation is

$$\mathbb{E}[X|A=a] = \sum_x x p(x|a).$$

Similar to the law of total probability, we have the *law of total expectation*:

$$\mathbb{E}[X] = \sum_a \mathbb{E}[X|A=a]p(a).$$

The proof is as follows. By the definition of expectation, it holds that

$$\begin{aligned}
\sum_a \mathbb{E}[X|A=a]p(a) &= \sum_a \left[\sum_x p(x|a)x\right] p(a) \\
&= \sum_x \sum_a p(x|a)p(a)x \\
&= \sum_x \left[\sum_a p(x|a)p(a)\right] x \\
&= \sum_x p(x)x \\
&= \mathbb{E}[X].
\end{aligned}$$

The law of total expectation is frequently used in reinforcement learning.

Similarly, conditional expectation satisfies

$$\mathbb{E}[X|A=a] = \sum_b \mathbb{E}[X|A=a, B=b]p(b|a).$$

This equation is useful in the derivation of the Bellman equation. A hint of its proof is the chain rule: $p(x|a,b)p(b|a) = p(x,b|a)$.

Finally, it is worth noting that $\mathbb{E}[X|A=a]$ is different from $\mathbb{E}[X|A]$. The former is a value, whereas the latter is a random variable. In fact, $\mathbb{E}[X|A]$ is a function of the random variable A. We need rigorous probability theory to define $\mathbb{E}[X|A]$.

◇ *Gradient of expectation*: Let $f(X, \beta)$ be a scalar function of a random variable X and a deterministic parameter vector β. Then,

$$\nabla_\beta \mathbb{E}[f(X,\beta)] = \mathbb{E}[\nabla_\beta f(X,\beta)].$$

Proof: Since $\mathbb{E}[f(X,\beta)] = \sum_x f(x,\beta)p(x)$, we have $\nabla_\beta \mathbb{E}[f(X,\beta)] = \nabla_\beta \sum_x f(x,\beta)p(x) = \sum_x \nabla_\beta f(x,\beta)p(x) = \mathbb{E}[\nabla_\beta f(X,\beta)]$.

◇ *Variance, covariance, covariance matrix*: For a single random variable X, its *variance* is defined as $\text{var}(X) = \mathbb{E}[(X - \bar{x})^2]$, where $\bar{x} = \mathbb{E}[X]$. For two random variables X, Y, their *covariance* is defined as $\text{cov}(X,Y) = \mathbb{E}[(X - \bar{x})(Y - \bar{y})]$. For a random vector $X = [X_1, \ldots, X_n]^T$, the covariance matrix of X is defined as $\text{var}(X) \doteq \Sigma = \mathbb{E}[(X-\bar{x})(X-\bar{x})^T] \in \mathbb{R}^{n\times n}$. The ijth entry of Σ is $[\Sigma]_{ij} = \mathbb{E}[[X-\bar{x}]_i[X-\bar{x}]_j] = \mathbb{E}[(X_i - \bar{x}_i)(X_j - \bar{x}_j)] = \text{cov}(X_i, X_j)$. One trivial property is $\text{var}(a) = 0$ if a is deterministic. Moreover, it can be verified that $\text{var}(AX + a) = \text{var}(AX) = A\text{var}(X)A^T = A\Sigma A^T$.

Some useful facts are summarized below.

- Fact: $\mathbb{E}[(X - \bar{x})(Y - \bar{y})] = \mathbb{E}[XY] - \bar{x}\bar{y} = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.
  Proof: $\mathbb{E}[(X-\bar{x})(Y-\bar{y})] = \mathbb{E}[XY - X\bar{y} - \bar{x}Y + \bar{x}\bar{y}] = \mathbb{E}[XY] - \mathbb{E}[X]\bar{y} - \bar{x}\mathbb{E}[Y] + \bar{x}\bar{y} = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.
- Fact: $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ if X, Y are independent.
  Proof: $\mathbb{E}[XY] = \sum_x \sum_y p(x,y)xy = \sum_x \sum_y p(x)p(y)xy = \sum_x p(x)x \sum_y p(y)y = \mathbb{E}[X]\mathbb{E}[Y]$.
- Fact: $\text{cov}(X, Y) = 0$ if X, Y are independent.
  Proof: When X, Y are independent, $\text{cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] = 0$.

Appendix B

Measure-Theoretic Probability Theory

We now briefly introduce measure-theoretic probability theory, which is also called rigorous probability theory. We only present basic notions and results. Comprehensive introductions can be found in [96–98]. Moreover, measure-theoretic probability theory requires some basic knowledge of measure theory, which is not covered here. Interested readers may refer to [99].

The reader may wonder if it is necessary to understand measure-theoretic probability theory before studying reinforcement learning. The answer is yes if the reader is interested in rigorously analyzing the convergence of stochastic sequences. For example, we often encounter the notion of *almost sure* convergence in Chapter 6 and Chapter 7. This notion is taken from measure-theoretic probability theory. If the reader is not interested in the convergence of stochastic sequences, it is okay to skip this part.

Probability triples

A *probability triple* is fundamental for establishing measure-theoretic probability theory. It is also called a probability space or probability measure space. A probability triple consists of three ingredients.

- Ω: This is a set called the *sample space* (or outcome space). Any element (or point) in Ω, denoted as ω, is called an *outcome*. This set contains all the possible outcomes of a random sampling process.

  Example: When playing a game of dice, we have six possible outcomes $\{1, 2, 3, 4, 5, 6\}$. Hence, $\Omega = \{1, 2, 3, 4, 5, 6\}$.

- $\mathcal{F}$: This is a set called the *event space*. In particular, it is a σ-algebra (or σ-field) of Ω. The definition of a σ-algebra is given in Box B.1. An element in $\mathcal{F}$, denoted as A, is called an *event*. An *elementary event* refers to a single outcome in the sample space. An event may be an elementary event or a combination of multiple elementary events.

Example: Consider the game of dice. An example of an elementary event is "the number you get is i", where $i \in \{1, \ldots, 6\}$. An example of a nonelementary event is "the number you get is greater than 3". We care about such an event in practice because, for example, we can win the game if this event occurs. This event is mathematically expressed as $A = \{\omega \in \Omega : \omega > 3\}$. Since $\Omega = \{1, 2, 3, 4, 5, 6\}$ in this case, we have $A = \{4, 5, 6\}$.

- $\mathbb{P}$: This is a probability measure, which is a mapping from $\mathcal{F}$ to $[0, 1]$. Any $A \in \mathcal{F}$ is a set that contains some points in Ω. Then, $\mathbb{P}(A)$ is the measure of this set.

  Example: If $A = \Omega$, which contains all ω values, then $\mathbb{P}(A) = 1$; if $A = \emptyset$, then $\mathbb{P}(A) = 0$. In the game of dice, consider the event "the number you get is greater than 3". In this case, $A = \{\omega \in \Omega : \omega > 3\}$, and $\Omega = \{1, 2, 3, 4, 5, 6\}$. Then, we have $A = \{4, 5, 6\}$ and hence $\mathbb{P}(A) = 1/2$. That is, the probability of us rolling a number greater than 3 is $1/2$.

Box B.1: Definition of a σ-algebra

An *algebra* of Ω is a set of some subsets of Ω that satisfy certain conditions. A *σ-algebra* is a specific and important type of algebra. In particular, denote $\mathcal{F}$ as a σ-algebra. Then, it must satisfy the following conditions.

- $\mathcal{F}$ contains $\emptyset$ and Ω;
- $\mathcal{F}$ is closed under complements;
- $\mathcal{F}$ is closed under countable unions and intersections.

The σ-algebras of a given Ω are not unique. $\mathcal{F}$ may contain all the subsets of Ω, and it may also merely contain some of them as long as it satisfies the above three conditions (see the examples below). Moreover, the three conditions are not independent. For example, if $\mathcal{F}$ contains Ω and is closed under complements, then it naturally contains $\emptyset$. More information can be found in [96–98].

- Example: When playing the dice game, we have $\Omega = \{1, 2, 3, 4, 5, 6\}$. Then, $\mathcal{F} = \{\Omega, \emptyset, \{1, 2, 3\}, \{4, 5, 6\}\}$ is a σ-algebra. The above three conditions can be easily verified. There are also other σ-algebras such as $\{\Omega, \emptyset, \{1, 2, 3, 4, 5\}, \{6\}\}$. Moreover, for any Ω with finite elements, the collection of all the subsets of Ω is a σ-algebra.

Random variables

Based on the notion of probability triples, we can formally define random variables. They are called variables, but they are actually functions that map from Ω to $\mathbb{R}$. In particular,

a random variable assigns each outcome in Ω a numerical value, and hence it is a function: $X(\omega) : \Omega \to \mathbb{R}$.

Not all mappings from Ω to $\mathbb{R}$ are random variables. The formal definition of a random variable is as follows. A function $X : \Omega \to \mathbb{R}$ is a random variable if

$$A = \{\omega \in \Omega | X(\omega) \leq x\} \in \mathcal{F}$$

for all $x \in \mathbb{R}$. This definition indicates that X is a random variable only if $X(\omega) \leq x$ is an event in $\mathcal{F}$. More information can be found in [96, Section 3.1].

Expectation of random variables

The definition of the expectation of general random variables is sophisticated. Here, we only consider the special yet important case of simple random variables. In particular, a random variable is *simple* if $X(\omega)$ only takes a finite number of values. Let $\mathcal{X}$ be the set of all the possible values that X can take. A simple random variable is a function: $X(w) : \Omega \to \mathcal{X}$. It can be defined in a closed form as

$$X(\omega) \doteq \sum_{x \in \mathcal{X}} x \mathbb{1}_{A_x}(\omega),$$

where

$$A_x = \{\omega \in \Omega | X(\omega) = x\} \doteq X^{-1}(x)$$

and

$$\mathbb{1}_{A_x}(\omega) \doteq \begin{cases} 1, & \omega \in A_x, \\ 0, & \text{otherwise.} \end{cases} \tag{B.1}$$

Here, $\mathbb{1}_{A_x}(\omega)$ is an *indicator function* $\mathbb{1}_{A_x}(\omega) : \Omega \to \{0, 1\}$. If ω is mapped to x, the indicator function equals one; otherwise, it equals zero. It is possible that multiple ω's in Ω map to the same value in $\mathcal{X}$, but a single ω cannot be mapped to multiple values in $\mathcal{X}$.

With the above preparation, the expectation of a simple random variable is defined as

$$\mathbb{E}[X] \doteq \sum_{x \in \mathcal{X}} x \mathbb{P}(A_x), \tag{B.2}$$

where

$$A_x = \{\omega \in \Omega | X(\omega) = x\}.$$

The definition in (B.2) is similar to but more formal than the definition of expectation in the nonmeasure-theoretic case: $\mathbb{E}[X] = \sum_{x \in \mathcal{X}} xp(x)$.

As a demonstrative example, we next calculate the expectation of the indicator func-

tion in (B.1). It is notable that the indicator function is also a random variable that maps Ω to $\{0, 1\}$ [96, Proposition 3.1.5]. As a result, we can calculate its expectation. In particular, consider the indicator function $\mathbb{1}_A$ where A denotes any event. We have

$$\mathbb{E}[\mathbb{1}_A] = \mathbb{P}(A).$$

To prove that, we have

$$\begin{aligned}\mathbb{E}[\mathbb{1}_A] &= \sum_{z\in\{0,1\}} z\mathbb{P}(\mathbb{1}_A = z) \\ &= 0 \cdot \mathbb{P}(\mathbb{1}_A = 0) + 1 \cdot \mathbb{P}(\mathbb{1}_A = 1) \\ &= \mathbb{P}(\mathbb{1}_A = 1) \\ &= \mathbb{P}(A).\end{aligned}$$

More properties of indicator functions can be found in [100, Chapter 24].

Conditional expectation as a random variable

While the expectation in (B.2) maps random variables to a specific value, we next introduce a conditional expectation that maps random variables to another random variable.

Suppose that X, Y, Z are all random variables. Consider three cases. First, a conditional expectation like $\mathbb{E}[X|Y = 2]$ or $\mathbb{E}[X|Y = 5]$ is specific *number*. Second, $\mathbb{E}[X|Y = y]$, where y is a variable, is a *function* of y. Third, $\mathbb{E}[X|Y]$, where Y is a random variable, is a function of Y and hence also a *random variable*. Since $\mathbb{E}[X|Y]$ is also a random variable, we can calculate, for example, its expectation.

We next examine the third case closely since it frequently emerges in the convergence analyses of stochastic sequences. The rigorous definition is not covered here and can be found in [96, Chapter 13]. We merely present some useful properties [101].

Lemma B.1 (Basic properties)**.** *Let X, Y, Z be random variables. The following properties hold.*

(a) $\mathbb{E}[a|Y] = a$, *where a is a given number.*

(b) $\mathbb{E}[aX + bZ|Y] = a\mathbb{E}[X|Y] + b\mathbb{E}[Z|Y]$.

(c) $\mathbb{E}[X|Y] = \mathbb{E}[X]$ *if X, Y are independent.*

(d) $\mathbb{E}[Xf(Y)|Y] = f(Y)\mathbb{E}[X|Y]$.

(e) $\mathbb{E}[f(Y)|Y] = f(Y)$.

(f) $\mathbb{E}[X|Y, f(Y)] = \mathbb{E}[X|Y]$.

(g) *If* $X \geq 0$, *then* $\mathbb{E}[X|Y] \geq 0$.

(h) *If* $X \geq Z$, *then* $\mathbb{E}[X|Y] \geq \mathbb{E}[Z|Y]$.

Proof. We only prove some properties. The others can be proven similarly.

To prove $\mathbb{E}[a|Y] = a$ as in (a), we can show that $\mathbb{E}[a|Y = y] = a$ is valid for any y that Y can possibly take. This is clearly true, and the proof is complete.

To prove the property in (d), we can show that $\mathbb{E}[Xf(Y)|Y = y] = f(Y = y)\mathbb{E}[X|Y = y]$ for any y. This is valid because $\mathbb{E}[Xf(Y)|Y = y] = \sum_x xf(y)p(x|y) = f(y)\sum_x xp(x|y) = f(y)\mathbb{E}[X|Y = y]$. □

Since $\mathbb{E}[X|Y]$ is a random variable, we can calculate its expectation. The related properties are presented below. These properties are useful for analyzing the convergence of stochastic sequences.

Lemma B.2. *Let X, Y, Z be random variables. The following properties hold.*

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

(b) $\mathbb{E}\big[\mathbb{E}[X|Y, Z]\big] = \mathbb{E}[X]$.

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

Proof. To prove the property in (a), we need to show that $\mathbb{E}\big[\mathbb{E}[X|Y = y]\big] = \mathbb{E}[X]$ for any y that Y can possibly take. To that end, considering that $\mathbb{E}[X|Y]$ is a function of Y, we denote it as $f(Y) = \mathbb{E}[X|Y]$. Then,

$$\begin{aligned}
\mathbb{E}\big[\mathbb{E}[X|Y]\big] = \mathbb{E}\big[f(Y)\big] &= \sum_y f(Y = y)p(y) \\
&= \sum_y \mathbb{E}[X|Y = y]p(y) \\
&= \sum_y \left(\sum_x xp(x|y)\right) p(y) \\
&= \sum_x x \sum_y p(x|y)p(y) \\
&= \sum_x x \sum_y p(x, y) \\
&= \sum_x xp(x) \\
&= \mathbb{E}[X].
\end{aligned}$$

The proof of the property in (b) is similar. In particular, we have

$$\mathbb{E}\big[\mathbb{E}[X|Y, Z]\big] = \sum_{y,z} \mathbb{E}[X|y, z]p(y, z) = \sum_{y,z}\sum_x xp(x|y, z)p(y, z) = \sum_x xp(x) = \mathbb{E}[X].$$

The proof of the property in (c) follows immediately from property (e) in Lemma B.1. That is because $\mathbb{E}[X|Y]$ is a function of Y. We denote this function as $f(Y)$. It then follows that $\mathbb{E}[\mathbb{E}[X|Y]|Y] = \mathbb{E}[f(Y)|Y] = f(Y) = \mathbb{E}[X|Y]$. □

Definitions of stochastic convergence

One main reason why we care about measure-theoretic probability theory is that it can rigorously describe the convergence properties of stochastic sequences.

Consider the stochastic sequence $\{X_k\} \doteq \{X_1, X_2, \dots, X_k, \dots\}$. Each element in this sequence is a random variable defined on a triple $(\Omega, \mathcal{F}, \mathbb{P})$. When we say $\{X_k\}$ *converges* to a random variable X, we should be careful since there are different types of convergence as shown below.

◇ *Sure convergence:*

Definition: $\{X_k\}$ converges *surely* (or *everywhere* or *pointwise*) to X if

$$\lim_{k\to\infty} X_k(\omega) = X(\omega), \quad \text{for all } \omega \in \Omega.$$

It means that $\lim_{k\to\infty} X_k(\omega) = X(\omega)$ is valid for *all* points in Ω. This definition can be equivalently stated as

$$A = \Omega \quad \text{where} \quad A = \left\{\omega \in \Omega : \lim_{k\to\infty} X_k(\omega) = X(\omega)\right\}.$$

◇ *Almost sure convergence:*

Definition: $\{X_k\}$ converges *almost surely* (or *almost everywhere* or *with probability 1* or *w.p.1*) to X if

$$\mathbb{P}(A) = 1 \quad \text{where} \quad A = \left\{\omega \in \Omega : \lim_{k\to\infty} X_k(\omega) = X(\omega)\right\}. \tag{B.3}$$

It means that $\lim_{k\to\infty} X_k(\omega) = X(\omega)$ is valid for *almost all* points in Ω. The points, for which this limit is invalid, form a set of zero measure. For the sake of simplicity, (B.3) is often written as

$$\mathbb{P}\left(\lim_{k\to\infty} X_k = X\right) = 1.$$

Almost sure convergence can be denoted as $X_k \xrightarrow{a.s.} X$.

◇ *Convergence in probability:*

Definition: $\{X_k\}$ converges *in probability* to X if for any $\epsilon > 0$,

$$\lim_{k\to\infty} \mathbb{P}(A_k) = 0 \quad \text{where} \quad A_k = \{\omega \in \Omega : |X_k(\omega) - X(\omega)| > \epsilon\}. \tag{B.4}$$

For simplicity, (B.4) can be written as

$$\lim_{k\to\infty} \mathbb{P}(|X_k - X| > \epsilon) = 0.$$

The difference between convergence in probability and (almost) sure convergence is as follows. Both sure convergence and almost sure convergence first evaluate the convergence of every point in Ω and then check the measure of these points that converge. By contrast, convergence in probability first checks the points that satisfy $|X_k - X| > \epsilon$ and then evaluates if the measure will converge to zero as $k \to \infty$.

◇ *Convergence in mean:*

Definition: $\{X_k\}$ converges *in the r-th mean* (or *in the L^r norm*) to X if

$$\lim_{k\to\infty} \mathbb{E}[|X_k - X|^r] = 0.$$

The most frequently used cases are $r = 1$ and $r = 2$. It is worth mentioning that convergence in mean is not equivalent to $\lim_{k\to\infty} \mathbb{E}[X_k - X] = 0$ or $\lim_{k\to\infty} \mathbb{E}[X_k] = \mathbb{E}[X]$, which indicates that $\mathbb{E}[X_k]$ converges but the variance may not.

◇ *Convergence in distribution:*

Definition: The *cumulative distribution function* of X_k is defined as $\mathbb{P}(X_k \leq a)$ where $a \in \mathbb{R}$. Then, $\{X_k\}$ converges to X *in distribution* if the cumulative distribution function converges:

$$\lim_{k\to\infty} \mathbb{P}(X_k \leq a) = \mathbb{P}(X \leq a), \quad \text{for all } a \in \mathbb{R}.$$

A compact expression is

$$\lim_{k\to\infty} \mathbb{P}(A_k) = \mathbb{P}(A),$$

where

$$A_k \doteq \{\omega \in \Omega : X_k(\omega) \leq a\}, \quad A \doteq \{\omega \in \Omega : X(\omega) \leq a\}.$$

The relationships between the above types of convergence are given below:

almost sure convergence $\Rightarrow$ convergence in probability $\Rightarrow$ convergence in distribution

convergence in mean $\Rightarrow$ convergence in probability $\Rightarrow$ convergence in distribution

Almost sure convergence and convergence in mean do not imply each other. More information can be found in [102].

Appendix C

Convergence of Sequences

We next introduce some results about the convergence of deterministic and stochastic sequences. These results are useful for analyzing the convergence of reinforcement learning algorithms such as those in Chapters 6 and 7.

We first consider deterministic sequences and then stochastic sequences.

C.1 Convergence of deterministic sequences

Convergence of monotonic sequences

Consider a sequence $\{x_k\} \doteq \{x_1, x_2, \ldots, x_k, \ldots\}$ where $x_k \in \mathbb{R}$. Suppose that this sequence is deterministic in the sense that x_k is not a random variable.

One of the most well-known convergence results is that a nonincreasing sequence with a lower bound converges. The following is a formal statement of this result.

Theorem C.1 (Convergence of monotonic sequences)**.** *If the sequence $\{x_k\}$ is nonincreasing and bounded from below:*

- *Nonincreasing: $x_{k+1} \leq x_k$ for all k;*
- *Lower bound: $x_k \geq \alpha$ for all k;*

then x_k converges to a limit, which is the infimum of $\{x_k\}$, as $k \to \infty$.

Similarly, if $\{x_k\}$ is *nondecreasing* and bounded from above, then the sequence is convergent.

Convergence of nonmonotonic sequences

We next analyze the convergence of *nonmonotonic* sequences.

To analyze the convergence of nonmonotonic sequences, we introduce the following useful operator [103]. For any $z \in \mathbb{R}$, define

$$z^+ \doteq \begin{cases} z, & \text{if } z \geq 0, \\ 0, & \text{if } z < 0, \end{cases}$$
$$z^- \doteq \begin{cases} z, & \text{if } z \leq 0, \\ 0, & \text{if } z > 0. \end{cases}$$

It is obvious that $z^+ \geq 0$ and $z^- \leq 0$ for any z. Moreover, it holds that

$$z = z^+ + z^-$$

for all $z \in \mathbb{R}$.

To analyze the convergence of $\{x_k\}$, we rewrite x_k as

$$\begin{aligned} x_k &= x_k - x_{k-1} + x_{k-1} - x_{k-2} + \cdots - x_2 + x_2 - x_1 + x_1 \\ &= \sum_{i=1}^{k-1} (x_{i+1} - x_i) + x_1 \\ &\doteq S_k + x_1, \end{aligned} \tag{C.1}$$

where $S_k \doteq \sum_{i=1}^{k-1}(x_{i+1} - x_i)$. Note that S_k can be decomposed as

$$S_k = \sum_{i=1}^{k-1} (x_{i+1} - x_i) = S_k^+ + S_k^-,$$

where

$$S_k^+ = \sum_{i=1}^{k-1} (x_{i+1} - x_i)^+ \geq 0, \qquad S_k^- = \sum_{i=1}^{k-1} (x_{i+1} - x_i)^- \leq 0.$$

Some useful properties of S_k^+ and S_k^- are given below.

- $\{S_k^+ \geq 0\}$ is a nondecreasing sequence since $S_{k+1}^+ \geq S_k^+$ for all k.
- $\{S_k^- \leq 0\}$ is a nonincreasing sequence since $S_{k+1}^- \leq S_k^-$ for all k.
- If S_k^+ is bounded from above, then S_k^- is bounded from below. This is because $S_k^- \geq -S_k^+ - x_1$ due to the fact that $S_k^- + S_k^+ + x_1 = x_k \geq 0$.

With the above preparation, we can show the following result.

Theorem C.2 (Convergence of nonmonotonic sequences)**.** *For any nonnegative sequence*

$\{x_k \geq 0\}$, *if*

$$\sum_{k=1}^{\infty}(x_{k+1} - x_k)^+ < \infty, \tag{C.2}$$

then $\{x_k\}$ *converges as* $k \to \infty$.

Proof. First, the condition $\sum_{k=1}^{\infty}(x_{k+1} - x_k)^+ < \infty$ indicates that $S_k^+ = \sum_{i=1}^{k-1}(x_{i+1} - x_i)^+$ is bounded from above for all k. Since $\{S_k^+\}$ is nondecreasing, the convergence of $\{S_k^+\}$ immediately follows from Theorem C.1. Suppose that S_k^+ converges to S_*^+.

Second, the boundedness of S_k^+ implies that S_k^- is bounded from below since $S_k^- \geq -S_k^+ - x_1$. Since $\{S_k^-\}$ is nonincreasing, the convergence of $\{S_k^-\}$ immediately follows from Theorem C.1. Suppose that S_k^- converges to S_*^-.

Finally, since $x_k = S_k^+ + S_k^- + x_1$, as shown in (C.1), the convergence of S_k^+ and S_k^- implies that $\{x_k\}$ converges to $S_*^+ + S_*^- + x_1$. □

Theorem C.2 is more general than Theorem C.1 because it allows x_k to increase as long as the increase is damped as in (C.2). In the monotonic case, Theorem C.2 still applies. In particular, if $x_{k+1} \leq x_k$, then $\sum_{k=1}^{\infty}(x_{k+1} - x_k)^+ = 0$. In this case, (C.2) is still satisfied and the convergence follows.

We next consider a special yet importance case. Suppose that $\{x_k \geq 0\}$ is a nonnegative sequence satisfying

$$x_{k+1} \leq x_k + \eta_k.$$

When $\eta_k = 0$, we have $x_{k+1} \leq x_k$, meaning that the sequence is monotonic. When $\eta_k \geq 0$, the sequence is *not* monotonic because x_{k+1} may be greater than x_k. Nevertheless, we can still ensure the convergence of the sequence under some mild conditions. The following result is an immediate corollary of Theorem C.2.

Corollary C.1. *For any nonnegative sequence* $\{x_k \geq 0\}$, *if*

$$x_{k+1} \leq x_k + \eta_k$$

and $\{\eta_k \geq 0\}$ *satisfies*

$$\sum_{k=1}^{\infty}\eta_k < \infty,$$

then $\{x_k \geq 0\}$ *converges.*

Proof. Since $x_{k+1} \leq x_k + \eta_k$, we have $(x_{k+1} - x_k)^+ \leq \eta_k$ for all k. Then, we have

$$\sum_{k=1}^{\infty}(x_{k+1} - x_k)^+ \leq \sum_{k=1}^{\infty}\eta_k < \infty.$$

As a result, (C.2) is satisfied and the convergence follows from Theorem C.2. □

C.2 Convergence of stochastic sequences

We now consider stochastic sequences. While various definitions of stochastic sequences have been given in Appendix B, how to determine the convergence of a given stochastic sequence has not yet been discussed. We next present an important class of stochastic sequences called *martingales*. If a sequence can be classified as a martingale (or one of its variants), then the convergence of the sequence immediately follows.

Convergence of martingale sequences

◇ Definition: A stochastic sequence $\{X_k\}_{k=1}^{\infty}$ is called a *martingale* if $\mathbb{E}[|X_k|] < \infty$ and

$$\mathbb{E}[X_{k+1}|X_1, \ldots, X_k] = X_k \tag{C.3}$$

almost surely for all k.

Here, $\mathbb{E}[X_{k+1}|X_1, \ldots, X_k]$ is a random variable rather than a deterministic value. The term "almost surely" in the second condition is due to the definition of such expectations. In addition, $\mathbb{E}[X_{k+1}|X_1, \ldots, X_k]$ is often written as $\mathbb{E}[X_{k+1}|\mathcal{H}_k]$ for short where $\mathcal{H}_k = \{X_1, \ldots, X_k\}$ represents the "history" of the sequence. $\mathcal{H}_k$ has a specific name called a *filtration*. More information can be found in [96, Chapter 14] and [104].

◇ Example: An example that can demonstrate martingales is *random walk*, which is a stochastic process describing the position of a point that moves randomly. Specifically, let X_k denote the position of the point at time step k. Starting from X_k, the expectation of the next position X_{k+1} equals X_k if the mean of the one-step displacement is zero. In this case, we have $\mathbb{E}[X_{k+1}|X_1, \ldots, X_k] = X_k$ and hence $\{X_k\}$ is a martingale.

A basic property of martingales is that

$$\mathbb{E}[X_{k+1}] = \mathbb{E}[X_k]$$

for all k and hence

$$\mathbb{E}[X_k] = \mathbb{E}[X_{k-1}] = \cdots = \mathbb{E}[X_2] = \mathbb{E}[X_1].$$

This result can be obtained by calculating the expectation on both sides of (C.3) based on property (b) in Lemma B.2.

While the expectation of a martingale is constant, we next extend martingales to submartingales and supermartingales, whose expectations vary monotonically.

◇ Definition: A stochastic sequence $\{X_k\}$ is called a *submartingale* if it satisfies $\mathbb{E}[|X_k|] < \infty$ and

$$\mathbb{E}[X_{k+1}|X_1, \ldots, X_k] \geq X_k \tag{C.4}$$

for all k.

Taking the expectation on both sides of (C.4) yields $\mathbb{E}[X_{k+1}] \geq \mathbb{E}[X_k]$. In particular, the left-hand side leads to $\mathbb{E}[\mathbb{E}[X_{k+1}|X_1, \ldots, X_k]] = \mathbb{E}[X_{k+1}]$ due to property (b) in Lemma B.2. By induction, we have

$$\mathbb{E}[X_k] \geq \mathbb{E}[X_{k-1}] \geq \cdots \geq \mathbb{E}[X_2] \geq \mathbb{E}[X_1].$$

Therefore, the expectation of a submartingale is nondecreasing.

It may be worth mentioning that, for two random variables X and Y, $X \leq Y$ means $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$. It does not mean the maximum of X is less than the minimum of Y.

◇ Definition: A stochastic sequence $\{X_k\}$ is called a *supermartingale* if it satisfies $\mathbb{E}[|X_k|] < \infty$ and

$$\mathbb{E}[X_{k+1}|X_1, \ldots, X_k] \leq X_k \tag{C.5}$$

for all k.

Taking expectation on both sides of (C.5) gives $\mathbb{E}[X_{k+1}] \leq \mathbb{E}[X_k]$. By induction, we have

$$\mathbb{E}[X_k] \leq \mathbb{E}[X_{k-1}] \leq \cdots \leq \mathbb{E}[X_2] \leq \mathbb{E}[X_1].$$

Therefore, the expectation of a supmartingale is nonincreasing.

The names "submartingale" and "supmartingale" are standard, but it may not be easy for beginners to distinguish them. Some tricks can be employed to do so. For example, since "supermartingale" has a letter "p" that points down, its expectation decreases; since submartingale has a letter "b" that points up, its expectation increases [104].

A supermartingale or submartingale is comparable to a deterministic monotonic sequence. While the convergence result for monotonic sequences has been given in Theorem C.1, we provide a similar convergence result for martingales as follows.

Theorem C.3 (Martingale convergence theorem)**.** *If $\{X_k\}$ is a submartingale (or supermartingale), then there is a finite random variable X such that $X_k \to X$ almost surely.*

The proof is omitted. A comprehensive introduction to martingales can be found in [96, Chapter 14] and [104].

Convergence of quasimartingale sequences

We next introduce quasimartingales, which can be viewed as a generalization of martingales since their expectations are not monotonic. They are comparable to nonmonotonic deterministic sequences. The rigorous definition and convergence results of quasimartingales are nontrivial. We merely list some useful results.

The event A_k is defined as $A_k \doteq \{\omega \in \Omega : \mathbb{E}[X_{k+1} - X_k|\mathcal{H}_k] \geq 0\}$, where $\mathcal{H}_k = \{X_1, \ldots, X_k\}$. Intuitively, A_k indicates that X_{k+1} is greater than X_k in expectation. Let $\mathbb{1}_{A_k}$ be an indicator function:

$$\mathbb{1}_{A_k} = \begin{cases} 1, & \mathbb{E}[X_{k+1} - X_k|\mathcal{H}_k] \geq 0, \\ 0, & \mathbb{E}[X_{k+1} - X_k|\mathcal{H}_k] < 0. \end{cases}$$

The indicator function has a property that

$$1 = \mathbb{1}_A + \mathbb{1}_{A^c}$$

for any event A where A^c denotes the complementary event of A. As a result, it holds for any random variable that

$$X = \mathbb{1}_A X + \mathbb{1}_{A^c} X.$$

Although quasimartingales do not have monotonic expectations, their convergence is still ensured under some mild conditions as shown below.

Theorem C.4 (Quasimartingale convergence theorem)**.** *For a nonnegative stochastic sequence $\{X_k \geq 0\}$, if*

$$\sum_{k=1}^{\infty} \mathbb{E}[(X_{k+1} - X_k)\mathbb{1}_{A_k}] < \infty,$$

then $\sum_{k=1}^{\infty} \mathbb{E}[(X_{k+1} - X_k)\mathbb{1}_{A_k^c}] > -\infty$ and there is a finite random variable such that $X_k \to X$ almost surely as $k \to \infty$.

Theorem C.4 can be viewed as an analogy of Theorem C.2, which is for nonmonotonic deterministic sequences. The proof of this theorem can be found in [105, Proposition 9.5]. Note that X_k here is required to be nonnegative. As a result, the boundedness of $\sum_{k=1}^{\infty} \mathbb{E}[(X_{k+1} - X_k)\mathbb{1}_{A_k}]$ implies the boundedness of $\sum_{k=1}^{\infty} \mathbb{E}[(X_{k+1} - X_k)\mathbb{1}_{A_k^c}]$.

Summary and comparison

We finally summarize and compare the results for deterministic and stochastic sequences.

◇ Deterministic sequences:

- Monotonic sequences: As shown in Theorem C.1, if a sequence is monotonic and bounded, then it converges.
- Nonmonotonic sequences: As shown in Theorem C.2, given a nonnegative sequence, even if it is nonmonotonic, it can still converge as long as its variation is damped in the sense that $\sum_{k=1}^{\infty}(x_{k+1}-x_k)^+ < \infty$.

◇ Stochastic sequences:

- Supermartingale/submartingale sequences: As shown in Theorem C.3, the expectation of a supermartingale or submartingale is monotonic. If a sequence is a supermartingale or submartingale, then the sequence converges almost surely.
- Quasimartingale sequences: As shown in Theorem C.4, even if a sequence's expectation is nonmonotonic, it can still converge as long as its variation is damped in the sense that $\sum_{k=1}^{\infty}\mathbb{E}[(X_{k+1}-X_k)\mathbf{1}_{\mathbb{E}[X_{k+1}-X_k|\mathcal{H}_k]>0}] < \infty$.

The above properties are summarized in Table C.1.

Variants of martingales	**Monotonicity of** $\mathbb{E}[X_k]$
Martingale	Constant: $\mathbb{E}[X_{k+1}] = \mathbb{E}[X_k]$
Submartingale	Increasing: $\mathbb{E}[X_{k+1}] \geq \mathbb{E}[X_k]$
Supermartingale	Decreasing: $\mathbb{E}[X_{k+1}] \leq \mathbb{E}[X_k]$
Quasimartingale	Non-monotonic

Table C.1: Summary of the monotonicity of different variants of martingales.

Appendix D

Preliminaries for Gradient Descent

We next present some preliminaries for the gradient descent method, which is one of the most frequently used optimization methods. The gradient descent method is also the foundation for the stochastic gradient descent method introduced in Chapter 6.

Convexity

◇ Definitions:

- Convex set: Suppose that $\mathcal{D}$ is a subset of $\mathbb{R}^n$. This set is *convex* if $z \doteq cx + (1-c)y \in \mathcal{D}$ for any $x, y \in \mathcal{D}$ and any $c \in [0, 1]$.
- Convex function: Suppose $f : \mathcal{D} \to \mathbb{R}$ where $\mathcal{D}$ is convex. Then, the function $f(x)$ is *convex* if

$$f(cx + (1-c)y) \leq cf(x) + (1-c)f(y)$$

for any $x, y \in \mathcal{D}$ and $c \in [0, 1]$.

◇ Convex conditions:

- First-order condition: Consider a function $f : \mathcal{D} \to \mathbb{R}$ where $\mathcal{D}$ is convex. Then, f is convex if [106, 3.1.3]

$$f(y) - f(x) \geq \nabla f(x)^T (y - x), \quad \text{for all } x, y \in \mathcal{D}. \tag{D.1}$$

When x is a scalar, $\nabla f(x)$ represents the slope of the tangent line of $f(x)$ at x. The geometric interpretation of (D.1) is that the point $(y, f(y))$ is always located above the tangent line.

- Second-order condition: Consider a function $f : \mathcal{D} \to \mathbb{R}$ where $\mathcal{D}$ is convex. Then, f is convex if

$$\nabla^2 f(x) \succeq 0, \quad \text{for all } x \in \mathcal{D},$$

where $\nabla^2 f(x)$ is the Hessian matrix.

◇ Degree of convexity:

Given a convex function, it is often of interest how strong its convexity is. The Hessian matrix is a useful tool for describing the degree of convexity. If $\nabla^2 f(x)$ is close to rank deficiency at a point, then the function is *flat* around that point and hence *weakly convex*. Otherwise, if the minimum singular value of $\nabla^2 f(x)$ is positive and large, the function is *curly* around that point and hence *strongly convex*. The degree of convexity influences the step size selection in gradient descent algorithms.

The lower and upper bounds of $\nabla^2 f(x)$ play an important role in characterizing the function convexity.

- Lower bound of $\nabla^2 f(x)$: A function is called *strongly convex* or *strictly convex* if $\nabla^2 f(x) \succeq \ell I_n$, where $\ell > 0$ for all x.
- Upper bound of $\nabla^2 f(x)$: If $\nabla^2 f(x)$ is bounded from above so that $\nabla^2 f(x) \preceq L I_n$, then the change in the first-order derivative $\nabla f(x)$ cannot be arbitrarily fast; equivalently, the function cannot be arbitrarily convex at a point.

  The upper bound can be implied by a Lipschitz condition of $\nabla f(x)$, as shown below.

  Lemma D.1. *Suppose that f is a convex function. If $\nabla f(x)$ is Lipschitz continuous with a constant L so that*

  $$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \textit{for all } x, y,$$

  then $\nabla^2 f(x) \preceq L I_n$ for all x. Here, $\|\cdot\|$ denotes the Euclidean norm.

Gradient descent algorithms

Consider the following optimization problem:

$$\min_x f(x)$$

where $x \in \mathcal{D} \subseteq \mathbb{R}^n$ and $f : \mathcal{D} \to \mathbb{R}$. The gradient descent algorithm is

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), \quad k = 0, 1, 2, \ldots \tag{D.2}$$

where α_k is a positive coefficient that may be fixed or time-varying. Here, α_k is called the *step size* or *learning rate*. Some remarks about (D.2) are given below.

◇ Direction of change: $\nabla f(x_k)$ is a vector that points in the direction along which $f(x_k)$ *increases* the fastest. Hence, the term $-\alpha_k \nabla f(x_k)$ changes x_k in the direction along which $f(x_k)$ *decreases* the fastest.

◇ Magnitude of change: The magnitude of the change $-\alpha_k \nabla f(x_k)$ is jointly determined by the step size α_k and the magnitude of $\nabla f(x_k)$.

- Magnitude of $\nabla f(x_k)$:

  When x_k is close to the optimum x^* where $\nabla f(x^*) = 0$, the magnitude $\|\nabla f(x_k)\|$ is small. In this case, the update process of x_k is slow, which is reasonable because we do not want to update x too aggressively and miss the optimum.

  When x_k is far from the optimum, the magnitude of $\nabla f(x_k)$ may be large, and hence the update process of x_k is fast. This is also reasonable because we hope that the estimate can approach the optimum as quickly as possible.

- Step size α_k:

  If α_k is small, the magnitude of $-\alpha_k \nabla f(x_k)$ is small, and hence the convergence process is slow. If α_k is too large, the update process of x_k is aggressive, which leads to either fast convergence or divergence.

  How to select α_k? The selection of α_k should depend on the degree of convexity of $f(x_k)$. If the function is *curly* around the optimum (the degree of convexity is strong), then the step size α_k should be small to guarantee convergence. If the function is *flat* around the optimum (the degree of convexity is weak), then the step size could be large so that x_k can quickly approach the optimum. The above intuition will be verified in the following convergence analysis.

Convergence analysis

We next present a proof of the convergence of the gradient descent algorithm in (D.2). That is to show x_k converges to the optimum x^* where $\nabla f(x^*) = 0$. First of all, we make some assumptions.

◇ Assumption 1: $f(x)$ is strongly convex such that

$$\nabla^2 f(x) \succeq \ell I,$$

where $\ell > 0$.

◇ Assumption 2: $\nabla f(x)$ is Lipschitz continuous with a constant L. This assumption implies the following inequality according to Lemma D.1:

$$\nabla^2 f(x) \preceq L I_n.$$

The convergence proof is given below.

Proof. For any x_{k+1} and x_k, it follows from [106, Section 9.1.2] that

$$f(x_{k+1}) = f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{1}{2}(x_{k+1} - x_k)^T \nabla^2 f(z_k)(x_{k+1} - x_k), \quad \text{(D.3)}$$

where z_k is a convex combination of x_k and x_{k+1}. Since it is assumed that $\nabla^2 f(z_k) \preceq LI_n$, we have $\|\nabla^2 f(z_k)\| \leq L$. (D.3) implies

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) + \frac{1}{2}\|\nabla^2 f(z_k)\|\|x_{k+1} - x_k\|^2 \\ &\leq f(x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2. \end{aligned}$$

Substituting $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$ into the above inequality yields

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \nabla f(x_k)^T(-\alpha_k \nabla f(x_k)) + \frac{L}{2}\|\alpha_k \nabla f(x_k)\|^2 \\ &= f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \frac{\alpha_k^2 L}{2}\|\nabla f(x_k)\|^2 \\ &= f(x_k) - \underbrace{\alpha_k \left(1 - \frac{\alpha_k L}{2}\right)}_{\eta_k} \|\nabla f(x_k)\|^2. \end{aligned} \tag{D.4}$$

We next show that if we select

$$0 < \alpha_k < \frac{2}{L}, \tag{D.5}$$

then the sequence $\{f(x_k)\}_{k=1}^{\infty}$ converges to $f(x^*)$ where $\nabla f(x^*) = 0$. First, (D.5) implies that $\eta_k > 0$. Then, (D.4) implies that $f(x_{k+1}) \leq f(x_k)$. Therefore, $\{f(x_k)\}$ is a nonincreasing sequence. Second, since $f(x_k)$ is always bounded from below by $f(x^*)$, we know that $\{f(x_k)\}$ converges as $k \to \infty$ according to the monotone convergence theorem in Theorem C.1. Suppose that the limit of the sequence is f^*. Then, taking the limit on both sides of (D.4) gives

$$\begin{aligned} &\lim_{k\to\infty} f(x_{k+1}) \leq \lim_{k\to\infty} f(x_k) - \lim_{k\to\infty} \eta_k \|\nabla f(x_k)\|^2 \\ &\Leftrightarrow f^* \leq f^* - \lim_{k\to\infty} \eta_k \|\nabla f(x_k)\|^2 \\ &\Leftrightarrow 0 \leq -\lim_{k\to\infty} \eta_k \|\nabla f(x_k)\|^2. \end{aligned}$$

Since $\eta_k \|\nabla f(x_k)\|^2 \geq 0$, the above inequality implies that $\lim_{k\to\infty} \eta_k \|\nabla f(x_k)\|^2 = 0$. As a result, x converges to x^* where $\nabla f(x^*) = 0$. The proof is complete. The above proof is inspired by [107]. □

The inequality in (D.5) provides valuable insights into how α_k should be selected. If the function is flat (L is small), the step size can be large; otherwise, if the function is strongly convex (L is large), then the step size must be sufficiently small to ensure convergence. There are also many other ways to prove the convergence such as the contraction mapping theorem [108, Lemma 3]. A comprehensive introduction to convex optimization can be found in [106].

Bibliography

[1] M. Pinsky and S. Karlin, *An introduction to stochastic modeling (3rd Edition)*. Academic Press, 1998.

[2] M. L. Puterman, *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[3] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction (2nd Edition)*. MIT Press, 2018.

[4] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge University Press, 2012.

[5] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-dynamic programming*. Athena Scientific, 1996.

[6] H. K. Khalil, *Nonlinear systems (3rd Edition)*. Patience Hall, 2002.

[7] G. Strang, *Calculus*. Wellesley-Cambridge Press, 1991.

[8] A. Besenyei, "A brief history of the mean value theorem," 2012. Lecture notes.

[9] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *International Conference on Machine Learning*, vol. 99, pp. 278–287, 1999.

[10] R. E. Bellman, *Dynamic programming*. Princeton University Press, 2010.

[11] R. E. Bellman and S. E. Dreyfus, *Applied dynamic programming*. Princeton University Press, 2015.

[12] J. Bibby, "Axiomatisations of the average and a further generalisation of monotonic sequences," *Glasgow Mathematical Journal*, vol. 15, no. 1, pp. 63–65, 1974.

[13] A. S. Polydoros and L. Nalpantidis, "Survey of model-based reinforcement learning: Applications on robotics," *Journal of Intelligent & Robotic Systems*, vol. 86, no. 2, pp. 153–173, 2017.

[14] T. M. Moerland, J. Broekens, A. Plaat, and C. M. Jonker, "Model-based reinforcement learning: A survey," *Foundations and Trends in Machine Learning*, vol. 16, no. 1, pp. 1–118, 2023.

[15] F.-M. Luo, T. Xu, H. Lai, X.-H. Chen, W. Zhang, and Y. Yu, "A survey on model-based reinforcement learning," *arXiv:2206.09328*, 2022.

[16] X. Wang, Z. Zhang, and W. Zhang, "Model-based multi-agent reinforcement learning: Recent progress and prospects," *arXiv:2203.10603*, 2022.

[17] M. Riedmiller, R. Hafner, T. Lampe, M. Neunert, J. Degrave, T. Wiele, V. Mnih, N. Heess, and J. T. Springenberg, "Learning by playing solving sparse reward tasks from scratch," in *International Conference on Machine Learning*, pp. 4344–4353, 2018.

[18] J. Ibarz, J. Tan, C. Finn, M. Kalakrishnan, P. Pastor, and S. Levine, "How to train your robot with deep reinforcement learning: Lessons we have learned," *The International Journal of Robotics Research*, vol. 40, no. 4-5, pp. 698–721, 2021.

[19] S. Narvekar, B. Peng, M. Leonetti, J. Sinapov, M. E. Taylor, and P. Stone, "Curriculum learning for reinforcement learning domains: A framework and survey," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 7382–7431, 2020.

[20] C. Szepesvári, *Algorithms for reinforcement learning*. Springer, 2010.

[21] A. Maroti, "RBED: Reward based epsilon decay," *arXiv:1910.13701*, 2019.

[22] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[23] W. Dabney, G. Ostrovski, and A. Barreto, "Temporally-extended epsilon-greedy exploration," *arXiv:2006.01782*, 2020.

[24] H.-F. Chen, *Stochastic approximation and its applications*, vol. 64. Springer Science & Business Media, 2006.

[25] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

[26] J. Venter, "An extension of the Robbins-Monro procedure," *The Annals of Mathematical Statistics*, vol. 38, no. 1, pp. 181–190, 1967.

[27] D. Ruppert, "Efficient estimations from a slowly convergent Robbins-Monro process," tech. rep., Cornell University Operations Research and Industrial Engineering, 1988.

[28] J. Lagarias, "Euler's constant: Euler's work and modern developments," *Bulletin of the American Mathematical Society*, vol. 50, no. 4, pp. 527–628, 2013.

[29] J. H. Conway and R. Guy, *The book of numbers.* Springer Science & Business Media, 1998.

[30] S. Ghosh, "The Basel problem," *arXiv:2010.03953*, 2020.

[31] A. Dvoretzky, "On stochastic approximation," in *The Third Berkeley Symposium on Mathematical Statistics and Probability*, 1956.

[32] T. Jaakkola, M. I. Jordan, and S. P. Singh, "On the convergence of stochastic iterative dynamic programming algorithms," *Neural Computation*, vol. 6, no. 6, pp. 1185–1201, 1994.

[33] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear estimation.* Prentice Hall, 2000.

[34] C. K. Chui and G. Chen, *Kalman filtering.* Springer, 2017.

[35] G. A. Rummery and M. Niranjan, *On-line Q-learning using connectionist systems.* Technical Report, Cambridge University, 1994.

[36] H. Van Seijen, H. Van Hasselt, S. Whiteson, and M. Wiering, "A theoretical and empirical analysis of Expected Sarsa," in *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, pp. 177–184, 2009.

[37] M. Ganger, E. Duryea, and W. Hu, "Double Sarsa and double expected Sarsa with shallow and deep learning," *Journal of Data Analysis and Information Processing*, vol. 4, no. 4, pp. 159–176, 2016.

[38] C. J. C. H. Watkins, *Learning from delayed rewards.* PhD thesis, King's College, 1989.

[39] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[40] T. C. Hesterberg, *Advances in importance sampling.* PhD Thesis, Stanford University, 1988.

[41] H. Hasselt, "Double Q-learning," *Advances in Neural Information Processing Systems*, vol. 23, 2010.

[42] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *AAAI Conference on Artificial Intelligence*, vol. 30, 2016.

[43] C. Dann, G. Neumann, and J. Peters, "Policy evaluation with temporal differences: A survey and comparison," *Journal of Machine Learning Research*, vol. 15, pp. 809–883, 2014.

[44] J. Clifton and E. Laber, "Q-learning: Theory and applications," *Annual Review of Statistics and Its Application*, vol. 7, pp. 279–301, 2020.

[45] B. Jang, M. Kim, G. Harerimana, and J. W. Kim, "Q-learning algorithms: A comprehensive classification and applications," *IEEE Access*, vol. 7, pp. 133653–133667, 2019.

[46] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine Learning*, vol. 3, no. 1, pp. 9–44, 1988.

[47] G. Strang, *Linear algebra and its applications (4th Edition)*. Belmont, CA: Thomson, Brooks/Cole, 2006.

[48] C. D. Meyer and I. Stewart, *Matrix analysis and applied linear algebra*. SIAM, 2023.

[49] M. Pinsky and S. Karlin, *An introduction to stochastic modeling*. Academic Press, 2010.

[50] M. G. Lagoudakis and R. Parr, "Least-squares policy iteration," *The Journal of Machine Learning Research*, vol. 4, pp. 1107–1149, 2003.

[51] R. Munos, "Error bounds for approximate policy iteration," in *International Conference on Machine Learning*, vol. 3, pp. 560–567, 2003.

[52] A. Geramifard, T. J. Walsh, S. Tellex, G. Chowdhary, N. Roy, and J. P. How, "A tutorial on linear function approximators for dynamic programming and reinforcement learning," *Foundations and Trends in Machine Learning*, vol. 6, no. 4, pp. 375–451, 2013.

[53] B. Scherrer, "Should one compute the temporal difference fix point or minimize the Bellman residual? the unified oblique projection view," in *International Conference on Machine Learning*, 2010.

[54] D. P. Bertsekas, *Dynamic programming and optimal control: Approximate dynamic programming (Volume II)*. Athena Scientific, 2011.

[55] S. Abramovich, G. Jameson, and G. Sinnamon, "Refining Jensen's inequality," *Bulletin mathématique de la Société des Sciences Mathématiques de Roumanie*, pp. 3–14, 2004.

[56] S. S. Dragomir, "Some reverses of the Jensen inequality with applications," *Bulletin of the Australian Mathematical Society*, vol. 87, no. 2, pp. 177–194, 2013.

[57] S. J. Bradtke and A. G. Barto, "Linear least-squares algorithms for temporal difference learning," *Machine Learning*, vol. 22, no. 1, pp. 33–57, 1996.

[58] K. S. Miller, "On the inverse of the sum of matrices," *Mathematics Magazine*, vol. 54, no. 2, pp. 67–72, 1981.

[59] S. A. U. Islam and D. S. Bernstein, "Recursive least squares for real-time implementation," *IEEE Control Systems Magazine*, vol. 39, no. 3, pp. 82–85, 2019.

[60] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[61] J. Fan, Z. Wang, Y. Xie, and Z. Yang, "A theoretical analysis of deep Q-learning," in *Learning for Dynamics and Control*, pp. 486–489, 2020.

[62] L.-J. Lin, *Reinforcement learning for robots using neural networks*. 1992. Technical report.

[63] J. N. Tsitsiklis and B. Van Roy, "An analysis of temporal-difference learning with function approximation," *IEEE Transactions on Automatic Control*, vol. 42, no. 5, pp. 674–690, 1997.

[64] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," *Advances in Neural Information Processing Systems*, vol. 12, 1999.

[65] P. Marbach and J. N. Tsitsiklis, "Simulation-based optimization of Markov reward processes," *IEEE Transactions on Automatic Control*, vol. 46, no. 2, pp. 191–209, 2001.

[66] J. Baxter and P. L. Bartlett, "Infinite-horizon policy-gradient estimation," *Journal of Artificial Intelligence Research*, vol. 15, pp. 319–350, 2001.

[67] X.-R. Cao, "A basic formula for online policy gradient algorithms," *IEEE Transactions on Automatic Control*, vol. 50, no. 5, pp. 696–699, 2005.

[68] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, vol. 8, no. 3, pp. 229–256, 1992.

[69] J. Peters and S. Schaal, "Reinforcement learning of motor skills with policy gradients," *Neural Networks*, vol. 21, no. 4, pp. 682–697, 2008.

[70] E. Greensmith, P. L. Bartlett, and J. Baxter, "Variance reduction techniques for gradient estimates in reinforcement learning," *Journal of Machine Learning Research*, vol. 5, no. 9, 2004.

[71] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *International Conference on Machine Learning*, pp. 1928–1937, 2016.

[72] M. Babaeizadeh, I. Frosio, S. Tyree, J. Clemons, and J. Kautz, "Reinforcement learning through asynchronous advantage actor-critic on a GPU," *arXiv:1611.06256*, 2016.

[73] T. Degris, M. White, and R. S. Sutton, "Off-policy actor-critic," *arXiv:1205.4839*, 2012.

[74] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *International Conference on Machine Learning*, pp. 387–395, 2014.

[75] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv:1509.02971*, 2015.

[76] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International Conference on Machine Learning*, pp. 1861–1870, 2018.

[77] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, and P. Abbeel, "Soft actor-critic algorithms and applications," *arXiv:1812.05905*, 2018.

[78] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International Conference on Machine Learning*, pp. 1889–1897, 2015.

[79] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv:1707.06347*, 2017.

[80] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International Conference on Machine Learning*, pp. 1587–1596, 2018.

[81] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[82] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[83] Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, and J. Wang, "Mean field multi-agent reinforcement learning," in *International Conference on Machine Learning*, pp. 5571–5580, 2018.

[84] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, *et al.*, "Grandmaster level in StarCraft II using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[85] Y. Yang and J. Wang, "An overview of multi-agent reinforcement learning from game theoretical perspective," *arXiv:2011.00583*, 2020.

[86] S. Levine and V. Koltun, "Guided policy search," in *International Conference on Machine Learning*, pp. 1–9, 2013.

[87] M. Janner, J. Fu, M. Zhang, and S. Levine, "When to trust your model: Model-based policy optimization," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[88] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," in *International Conference on Machine Learning*, pp. 449–458, 2017.

[89] M. G. Bellemare, W. Dabney, and M. Rowland, *Distributional Reinforcement Learning*. MIT Press, 2023.

[90] H. Zhang, D. Liu, Y. Luo, and D. Wang, *Adaptive dynamic programming for control: algorithms and stability*. Springer Science & Business Media, 2012.

[91] F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis, "Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers," *IEEE Control Systems Magazine*, vol. 32, no. 6, pp. 76–105, 2012.

[92] F. L. Lewis and D. Liu, *Reinforcement learning and approximate dynamic programming for feedback control*. John Wiley & Sons, 2013.

[93] Z.-P. Jiang, T. Bian, and W. Gao, "Learning-based control: A tutorial and some recent results," *Foundations and Trends in Systems and Control*, vol. 8, no. 3, pp. 176–284, 2020.

[94] S. Meyn, *Control systems and reinforcement learning*. Cambridge University Press, 2022.

[95] S. E. Li, *Reinforcement learning for sequential decision and optimal control*. Springer, 2023.

[96] J. S. Rosenthal, *First look at rigorous probability theory (2nd Edition)*. World Scientific Publishing Company, 2006.

[97] D. Pollard, *A user's guide to measure theoretic probability*. Cambridge University Press, 2002.

[98] P. J. Spreij, "Measure theoretic probability," *UvA Course Notes*, 2012.

[99] R. G. Bartle, *The elements of integration and Lebesgue measure*. John Wiley & Sons, 2014.

[100] M. Taboga, *Lectures on probability theory and mathematical statistics (2nd Edition)*. CreateSpace Independent Publishing Platform, 2012.

[101] T. Kennedy, "Theory of probability," 2007. Lecture notes.

[102] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge University Press, 2000.

[103] L. Bottou, "Online learning and stochastic approximations," *Online Learning in Neural Networks*, vol. 17, no. 9, p. 142, 1998.

[104] D. Williams, *Probability with martingales*. Cambridge University Press, 1991.

[105] M. Métivier, *Semimartingales: A course on stochastic processes*. Walter de Gruyter, 1982.

[106] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[107] S. Bubeck *et al.*, "Convex optimization: Algorithms and complexity," *Foundations and Trends in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015.

[108] A. Jung, "A fixed-point of view on gradient methods for big data," *Frontiers in Applied Mathematics and Statistics*, vol. 3, p. 18, 2017.

Symbols

In this book, a matrix or a random variable is represented by capital letters. A vector, a scalar, or a sample is represented by a lowercase letter. The mathematical symbols that are frequently used in this book are listed below.

$=$	equality
$\approx$	approximation
$\doteq$	equality by definition
$\geq$, $>$, $\leq$, $<$	elementwise comparison
$\in$	is an element of
$\|\cdot\|_2$	Euclidean norm of a vector or the corresponding induced matrix norm
$\|\cdot\|_\infty$	maximum norm of a vector or the corresponding induced matrix norm
ln	natural logarithm
$\mathbb{R}$	set of real numbers
$\mathbb{R}^n$	set of n-dimensional real vectors
$\mathbb{R}^{n\times m}$	set of all $n \times m$-dimensional real matrices
$A \succeq 0$ ($A \succ 0$)	matrix A is positive semidefinite (definite)
$A \preceq 0$ ($A \succ 0$)	matrix A is negative semidefinite (definite)
$\|x\|$	absolute value of real scalar x
$\|\mathcal{S}\|$	number of elements in set $\mathcal{S}$
$\nabla_x f(x)$	gradient of scalar function $f(x)$ with respect to vector x. It may be written as $\nabla f(x)$ for short.
$[A]_{ij}$	element in the ith row and jth column of matrix A
$[x]_i$	ith element of vector x
$X \sim p$	p is the probability distribution of random variable X.
$p(X = x)$, $\Pr(X = x)$	probability of $X = x$. They are often written as $p(x)$ or $\Pr(x)$ for short.
$p(x\|y)$	conditional probability
$\mathbb{E}_{X\sim p}[X]$	expectation or expected value of random variable X. It is often written as $\mathbb{E}[X]$ for short when the distribution of X is clear.

$\mathrm{var}(X)$	variance of random variable X
$\arg\max_x f(x)$	maximizer of function $f(x)$
$\mathbf{1}_n$	vector of all ones. It is often written as $\mathbf{1}$ for short when its dimension is clear.
I_n	$n \times n$-dimensional identity matrix. It is often written as I for short when its dimensions are clear.

Index