Appendix A

Preliminaries to Probability Theory

Reinforcement learning (RL) 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 in a set of numbers or events. The term "random" indicates that taking a value must follow a probability distribution.

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

  In this book, we mainly consider the case where a random variable can only take a finite number of values. A random variable can be a scalar or vector.

  Like normal variables, random variables have normal mathematical operations or functions 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.

- *Stochastic sequence* is a sequence of random variables.

  One scenario that we often encounter is that we collect 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 in the ith toss, then $\{x_1, x_2, \dots, x_n\}$ 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 value that can be possibly sampled, it is a random variable since x_i can take any values in $\{1, \dots, 6\}$. Although x_i is a lowercase letter here, it still represents the 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 $p(x)$ in 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 of x and, in the meantime, Y taking the value of y. One useful identity is

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

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

  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 of each other if the sampling value of one random variable does not affect the other. Mathematically, X and Y are independent of each other if

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

  Another equivalent definition is

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

  The above two definitions are equivalent because because $p(x, y) = p(x|y)p(y)$, which implies $p(x|y) = p(x)$ if and only if $p(x, y) = p(x)p(y)$.

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

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

  The intuitive interpretation is as follows. X and A are conditionally independent given B if and only if, given knowledge of whether $B = b$ occurs, the knowledge of whether $X = x$ occurs provides no information on the likelihood of $A = a$ occurring, and vice versa.

  In the context of RL, 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 to 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 is mentioned when we introduce 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

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

It can be further 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 useful fact about conditional probability:

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

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

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

The linearity property of expectation is

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

The second equation above can be 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 fixed matrix and $X \in \mathbb{R}^n$ is a random vector.

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

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

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, the right-hand side equals

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

The law of total expectation is frequently used in RL.

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 the 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. To well define $\mathbb{E}[X|A]$, we need the rigorous probability theory.

– *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 $\mathbb{E}[f(X,\beta)]=\sum_x f(x,a)p(x)$, we have $\nabla_\beta \mathbb{E}[f(X,\beta)]=\nabla_\beta \sum_x f(x,a)p(x)=\sum_x \nabla_\beta f(x,a)p(x)=\mathbb{E}[\nabla_\beta f(X,\beta)]$.

– *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 $X=[X_1,\dots,X_n]^T$, 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

Preliminaries to Gradient Descent

B.1 Basics

Mean value theorem

1) The statement of the theorem:

   Suppose x, y are two real variables and $f(\cdot)$ is a differentiable scalar function.

   - Scalar case: If x, y are *scalars* in $\mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}$, then there exists $z \in [x, y]$ such that
   $$f(y) - f(x) = f'(z)(y - z),$$
   where $f'(z)$ is the derivative of f at z.
   - Vector case: If x, y are *vectors* in $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$, then there exists $z = cx + (1 - c)y$ with $c \in [0, 1]$ such that
   $$f(y) - f(x) = \nabla f(z)^T (y - x), \tag{B.1}$$
   where $\nabla f(z)$ is the gradient of f parameterized at z.

2) Generalization:

   Consider $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ as a new function. Applying (B.1) in the mean value theorem to this new function gives
   $$\nabla f(y) - \nabla f(x) = \nabla^2 f(z)(y - x), \tag{B.2}$$
   where $z = cx + (1 - c)y$ with $c \in [0, 1]$, and $\nabla^2 f \in \mathbb{R}^{n \times n}$ is the *Hessian matrix*.

3) A related and useful result:

   Suppose $x, y \in \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$. Then we have
   $$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(z)(y - x), \tag{B.3}$$

where

$$z = cx + (1-c)y$$

with $c \in [0,1]$. This is the Talyor expansion or quadratic expansion of multivariate functions [39, Section 9.1.2].

Furthermore, suppose that $\nabla^2 f$ satisfies $\nabla^2 f(z) \preceq LI_n$, where L is a positive constant and I_n is the $n \times n$ identity matrix. Then, (B.3) implies

$$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|^2.$$

Similarly, if $\nabla^2 f(z) \succeq \ell I_n$ with ℓ as a positive constant, then (B.3) implies

$$f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{\ell}{2}\|y-x\|^2.$$

Convexity

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

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

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

- Evaluation conditions:

  How to evaluate if a function is convex or not? Two evaluation conditions are given below.

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

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

    When x is scalar, $\nabla f(x)$ is the slope of the tangent line of $f(x)$ at x. If we fix x and consider different y, the geometric interpretation of this inequality 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 a convex function if

    $$H \doteq \nabla^2 f(x) \geq 0, \quad \text{for all } x, \tag{B.4}$$

where $H = \nabla^2 f(x)$ is the Hessian matrix, which provides a simple way to evaluate convex functions.

– Degree of convexity:

Given a convex function, it is often of interest how strong the convexity of the function is. Hessian matrix is a useful tool to describe the degree of convexity of a function. If H is close to rank deficient at a point, then the function is *flat* around that point and hence of *weak convexity.* Otherwise, if the minimum singular value of H 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, as shown later.

The lower and upper bounds of H play important roles in characterizing the function convexity.

– Lower bound of H: A function is called *strictly convex* if $\nabla^2 f(x) \succeq \ell I_n$ where $\ell > 0$ for all x. This definition of strong convexity is to set a positive lower bound for the second-order derivative of f. An equivalent first-order condition is $(x - y)^T(\nabla f(x) - \nabla f(y)) \geq m\|x - y\|^2$ for all x, y. This first-order condition is equivalent to $\nabla^2 f(x) \succeq \ell I_n$. The proof is based on the mean value theorem in (B.2) and omitted here.

– Upper bound of H: If H is bounded from above so that $\nabla^2 f(x) \leq L I_n$, then the change of the first order derivative $\nabla f(x)$ could not be arbitrarily fast; or equivalently, the function could not be arbitrarily convex at some points.

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

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

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

then $\nabla^2 f(x) \preceq L I_n$ for all x. Here, $\|\cdot\|$ is a vector norm.

Proof. Following (B.2), we have $\nabla f(x) - \nabla f(y) = \nabla^2 f(z)(x - y)$. It follows from the Lipschitz property that

$$\|\nabla f(x) - \nabla f(y)\| = \|\nabla^2 f(z)(x - y)\| \leq L\|x - y\|.$$

Let $v \doteq x - y$. Then we have

$$\|\nabla^2 f(z) v\| \leq L v \Rightarrow \left\| \nabla^2 f(z) \frac{v}{\|v\|} \right\| \leq L. \tag{B.5}$$

By definition, the maximum singular value of $\nabla^2 f(z)$ is

$$\sigma_{\max}(\nabla^2 f(z)) \doteq \max_{v,\|v\|=1} \|\nabla^2 f(z)v\|.$$

Since (B.5) is valid for all v, we have $\sigma_{\max}(\nabla^2 f(z)) \leq L$. Since $\nabla^2 f(z)$ is positive semi-definite, we know $\nabla^2 f(z) \preceq LI_n$. Since z is determined by x, y which can be arbitrarily selected, the theorem is proven. □

B.2 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 well-known gradient-descent algorithm is

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), \quad k = 0, 1, 2, \ldots \tag{B.6}$$

where α_k is a positive coefficient that may be fixed or time-varying. Here, $\alpha_k$is called the *step size* or *learning rate*.

Why can this algorithm solve the optimization problem? We next first given some intuition. The rigorous proofs will be given later.

1) Direction of change: $\nabla f(x_k)$ is a vector, which points to a 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.

2) 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 is close to the optimum x^* where $\nabla f(x^*) = 0$, the magnitude of $\nabla f(x_k)$, which is $\|\nabla f(x_k)\|$, is small. In this case, the update of x_k is slow, which is reasonable because we do not want to update x too aggressively to 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 of x_k is fast. This is also reasonable because we hope the estimate could get close to the optimum as fast as possible.

   - Step size α_k:

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

- How to select α_k? The selection of α_k should depend on the degree of convexity of $f(x_k)$. Here, the degree of convexity can be described by the Hessian matrix $\nabla^2 f$. 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* convex around the optimum (the degree of convexity is weak), then the step size could be large so that x_k could approach the optimum fast. The above intuition will be verified in the following convergence analysis.

Convergence analysis

To prove the convergence of the gradient-descent algorithm in (B.6), we need to make some assumptions.

- $f(x)$ is convex and twice differentiable: The mathematical condition given by this assumption is
$$\nabla^2 f(x) \succeq 0.$$
- $\nabla f(x)$ is Lipschitz continuous with constant L: Following Lemma B.1, the mathematical condition given by this assumption is
$$\nabla^2 f(x) \preceq LI_n.$$
This assumption requires that the first derivative of the function could not change arbitrarily fast.

We next give two proofs to show that (B.6) converges.

The First Proof

Recall

$$f(y) = f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T\nabla^2 f(z)(y-x)$$

as shown in (B.3). Here, z is a convex combination of x, y. Since $\nabla^2 f(z) \preceq LI_n \Rightarrow \|\nabla^2 f(z)\| \leq L$, we have

$$\begin{aligned} f(y) &\leq f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}\|\nabla^2 f(z)\|\|y-x\|^2 \\ &\leq f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|^2. \end{aligned} \tag{B.7}$$

Replacing y, x in (B.7) by x_{k+1}, x_k, respectively, gives

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \nabla f(x_k)^T(-\alpha_k \nabla f(x_k)) + \frac{L}{2}\|\alpha_k \nabla f(x_k)\|^2 \\ &= f(x_k) - \alpha_k\|\nabla f(x_k)\|^2 + \frac{\alpha_k^2 L}{2}\|\nabla f(x_k)\|^2 \\ &= f(x_k) - \underbrace{\alpha_k\left(1 - \frac{\alpha_k L}{2}\right)}_{\eta_k}\|\nabla f(x_k)\|^2. \end{aligned} \tag{B.8}$$

We next show that if we select

$$0 < \alpha_k < \frac{2}{L}, \tag{B.9}$$

then the sequence $\{f(x_k)\}_{k=1}^{\infty}$ converges to $f(x*)$ where $\nabla f(x^*) = 0$. First, (B.9) implies that $\eta_k > 0$. It then follows from (B.9) that $f(x_{k+1}) \leq f(x_k)$. Second, we know that $f(x_k)$ is bounded from below by $f(x^*)$, which is the minimum value. As a result, the sequence converges as $k \to \infty$ according to the monotone convergence theorem. Suppose the limit of the sequence is f^*. Then, taking limit on both sides of (B.8) 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.

If we consider the simplest case where $\alpha_k = \alpha$ is constant, then

$$\eta_k = \eta = \alpha\left(1 - \frac{\alpha L}{2}\right).$$

If $0 < \alpha < 2/L$, then $\eta > 0$ and hence $\|\nabla f(x_k)\|^2 \to 0$ as $k \to \infty$. As a result, $\nabla f(x_k)$ converges to zero, indicating that the minimum is achieved.

The above proof is inspired by [40].

The inequality in (B.9) provides valuable insights in how α_k should be selected. First, the step size should be sufficiently small so that it is bounded from above. The upper bound is determined by the convexity of the function. If the function is flat (L is small), the step size could be large; otherwise, if the function is strongly convex (L is large), then the step size must be sufficiently small.

The Second Proof

Another way to prove the convergence of (B.6) is based on the contraction mapping theorem, which is a general method to analyze the convergence of sequences. We can rewrite (B.6) as

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) \doteq v(x_k).$$

If we can show that $v(x) : \mathbb{R}^n \to \mathbb{R}^n$ is a contraction mapping, then $\{x_k\}$ converges to the fixed-point satisfying $x^* = v(x^*)$. This is the idea of the proof. The details are given below.

Consider the simplest case where $\alpha_k = \alpha$ is constant. For any x, y we have

$$\begin{aligned} v(x) - v(y) &= x - \alpha\nabla f(x) - y + \alpha\nabla f(y) \\ &= x - y - \alpha(\nabla f(x) - \nabla f(y)) \\ &= x - y - \alpha\nabla^2 f(z)(x - y) \\ &= (I_n - \alpha\nabla^2 f(z))(x - y), \end{aligned}$$

where z is a convex combination of x, y according to the mean value theorem in (B.2). As a result,

$$\|v(x) - v(y)\| \leq \|I_n - \alpha\nabla^2 f(z)\|\|x - y\|. \tag{B.10}$$

Assume $\ell I_n \preceq \nabla^2 f(z)$ which means $f(x)$ is strictly convex. Then, we have $\ell I_n \preceq \nabla^2 f(z) \preceq L I_n$ and hence

$$(1 - \alpha L)I_n \preceq I_n - \alpha\nabla^2 f(z) \preceq (1 - \alpha\ell)I_n.$$

If we select

$$0 < \alpha < \frac{1}{L},$$

then

$$0 \prec (1 - \alpha L)I_n \preceq I_n - \alpha\nabla^2 f(z) \preceq (1 - \alpha\ell)I_n \prec I_n.$$

Hence, $I_n - \alpha\nabla^2 f(z)$ is positive definite and in the meantime $\|I_n - \alpha\nabla^2 f(z)\| \leq 1 - \alpha\ell < 1$. As a result, (B.10) becomes

$$\|v(x) - v(y)\| \leq (1 - \alpha\ell)\|x - y\|.$$

Since $0 < \alpha < 1/L$, we have $0 < 1 - \alpha\ell < 1$. As a result, the above inequality indicates that $v(x)$ is a contraction mapping. It then follows from the contraction mapping theorem that x converges to the fixed point x^* satisfying $v(x^*) = x^*$. It can be verified that $\nabla f(x^*) = 0$. The proof is complete.

The second proof requires the function to be *strongly convex* such that $\ell I_n \preceq \nabla^2 f(z)$, whereas the first proof does not. The second proof is inspired by [41, Lemma 3]. More information about convex optimization can be found in [39].

Bibliography

[1] M. Pinsky and S. Karlin, *An introduction to stochastic modeling (3rd Edition).* A-cademic 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.* MIT press, 2018.

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

[5] 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.

[6] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

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

[8] 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.

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

[10] A. Dvoretzky, "On stochastic approximation," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1956.

[11] 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.

[12] L. Bottou, "Online learning and stochastic approximations," *Online learning in neural networks*, vol. 17, no. 9, p. 142, 1998.

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

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

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

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

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

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

[19] 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.

[20] R. A. Horn and C. R. Johnson, *Matrix analysis.* Cambridge university press, 2012.

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

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

[23] 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.

[24] B. Scherrer, "Should one compute the temporal difference fix point or minimize the bellman residual? the unified oblique projection view,"

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

[26] 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.

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

[28] 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.

[29] 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.

[30] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[31] 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, PMLR, 2020.

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

[33] C. D. Meyer, *Matrix analysis and applied linear algebra.* SIAM, 2000.

[34] 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.

[35] 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.

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

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

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

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

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

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