An Introduction to Reinforcement Learning

From theory to algorithms

Joon Kwon

November 26, 2023

Contents

1 Markov decision processes **4**
- 1.1 Definition . . . 4
- 1.2 Policies . . . 5
- 1.3 Induced probability distributions over histories . . . 5
- 1.4 Value functions . . . 7

2 Bellman operators & optimality **8**
- 2.1 Bellman operators . . . 8
- 2.2 Bellman equations . . . 10
- 2.3 Greedy policy . . . 11
- 2.4 Optimal value functions & policies . . . 12

3 Dynamic programming **14**
- 3.1 Value iteration for policy evaluation . . . 14
- 3.2 Value iteration for control . . . 14
- 3.3 Policy iteration for control . . . 14
- 3.4 Asynchronous value iteration . . . 14

4 Tabular reinforcement learning **15**
- 4.1 Asynchronous stochastic approximations . . . 15
- 4.2 Stochastic estimators of Bellman equations . . . 15
- 4.3 Policy evaluation . . . 15
- 4.4 Control . . . 15

5 Value function approximation **16**

6 Policy gradient **17**

7 Additional methods: actor-critic & model-based **18**

Foreword

As of Fall 2023, this document contains lecture notes from a course given in *Master 2 Mathématiques et intelligence artificielle* in *Université Paris–Saclay.* These are highly incomplete and constantly updated as the lectures are given.

Acknowledgements

These notes highly benefited from discussions with Sylvain Sorin, Erwan Le Pennec, the expertise of Jaouad Mourtada, and the encouragements from Liliane Bel.

Introduction

Chapter 1

Markov decision processes

For a finite set I, we denote $\Delta(I)$ the corresponding unit simplex in $\mathbb{R}^I$:

$$\Delta(I) = \left\{ x \in \mathbb{R}_+^I,\ \sum_{i \in I} x_i = 1 \right\}$$

and interpret it as set the probability distributions over I. For $i \in I$, the corresponding Dirac measure is denoted δ_i.

1.1 Definition

Definition 1.1.1. A *finite Markov Decision Process* (MDP) is a 4-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, p)$ where $\mathcal{S}, \mathcal{A}, \mathcal{R}$ are nonmepty finite sets and $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathcal{R} \to [0,1]$ is such that for all $s, a \in \mathcal{S} \times \mathcal{A}$,

$$\sum_{(s',r) \in \mathcal{S} \times \mathcal{R}} p(s,a,s',r) = 1.$$

The elements of $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{S}$ are respectively called *states*, *actions* and *rewards.*

From now on, we assume that a finite MDP is given. For fixed values $(s,a) \in \mathcal{S} \times \mathcal{A}$, $p(s,a,\,\cdot\,)$ defines a probability distribution on $\mathcal{S} \times \mathcal{R}$, which the following notation emphasizes:

$$p(s',r|s,a) = p(s,a,s',r), \quad (s,a,s',r) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathcal{R}.$$

Definition 1.1.2. Let $t \geqslant 1$. A *history of length* t is a finite sequence of the form

$$(s_0, a_0, r_1, s_1, a_1, r_2, s_2, a_2, \ldots, s_{t-1}, a_{t-1}, r_t, s_t) \in (\mathcal{S} \times \mathcal{A} \times \mathcal{R})^t \times \mathcal{S}.$$

By convention, a history of length 0 is an element $s_0 \in \mathcal{S}$. $\mathcal{H}^{(t)}$ denotes the set of histories of length t and $\mathcal{H}^\infty = (\mathcal{S} \times \mathcal{A} \times \mathcal{R})^{\mathbb{N}}$ the set of infinite histories.

1.2 Policies

Definition 1.2.1. A *policy* is a sequence of maps $\pi = (\pi_t)_{t\geqslant 0}$ where $\pi_t : \mathcal{H}^{(t)} \to \Delta(\mathcal{A})$. For each $t \geqslant 0$ and $h^{(t)} \in \mathcal{H}^{(t)}$, denote

$$\pi_t(a|h^{(t)}) := \pi_t(h^{(t)})_a.$$

Π denotes the set of all policies.

Definition 1.2.2. A policy $\pi = (\pi_t)_{t\geqslant 0}$ is

- *deterministic* if for each $t \geqslant 0$ and $h^{(t)} \in \mathcal{H}^{(t)}$, there exists $a \in \mathcal{A}$ such that $\pi_t(h^{(t)})$ is the Dirac distribution in a;
- *Markovian* if for each $t \geqslant 0$, π_t is constant in all its variables but the last: in other words for a fixed value $s_t \in \mathcal{S}$, the map $\pi_t(\,\cdot\,, s_t)$ is constant; π_t can then be represented as $\pi_t : \mathcal{S} \to \Delta(\mathcal{A})$;
- *stationary* if it is Markovian and if $\pi_t = \pi_0$ for all $t \geqslant 0$; π can then be represented as $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ and denoted $\pi(a|s) = \pi(s)_a$ for $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Denote Π_0 (resp. $\Pi_{0,d}$) the set of stationary policies (resp. stationary and deterministic policies). A stationary and deterministic policy can be represented as $\pi : \mathcal{S} \to \mathcal{A}$.

1.3 Induced probability distributions over histories

Proposition 1.3.1. *Let $\mu \in \Delta(\mathcal{S})$ and a policy π. There exists a unique probability measure $\mathbb{P}_{\mu,\pi}$ on $\mathcal{H}^\infty = (\mathcal{S} \times \mathcal{A} \times \mathcal{R})^{\mathbb{N}}$ (equipped with the product sigma-algebra) such that for all $T \geqslant 0$, $a_1, \ldots, a_T \in \mathcal{A}$, $s_0, \ldots, s_{T+1} \in \mathcal{S}$, and $r_1, \ldots, r_{T+1} \in \mathcal{R}$,*

$$\mathbb{P}_{\mu,\pi}\left(\prod_{t=0}^{T}(\{s_t\} \times \{a_t\} \times \{r_{t+1}\}) \times \{s_{T+1}\} \times (\mathcal{S} \times \mathcal{A} \times [0,1])^{\mathbb{N}}\right)$$
$$= \mu(s_0)\prod_{t=0}^{T}\pi_t(a_t|h^{(t)})p(s_{t+1}, r_{t+1}|s_t, a_t).$$

where for each $1 \leqslant t \leqslant T$, $h^{(t)} = (s_0, a_0, r_1, \ldots, s_{t-1}, a_{t-1}, r_t, s_t)$.

Sketch of proof. The above expression defines a value for each set of the form

$$\prod_{t=0}^{T}(\{s_t\} \times \{a_t\} \times \{r_{t+1}\}) \times \{s_{T+1}\} \times (\mathcal{S} \times \mathcal{A} \times \mathcal{R})^{\mathbb{N}}.$$

The map $\mathbb{P}_{\mu,\pi}$ can then be extended to so-called cylinder sets of the form

$$\prod_{t=0}^{T}(\mathcal{S}_t \times \mathcal{A}_t \times \mathcal{R}_{t+1}) \times \mathcal{S}_{T+1} \times (\mathcal{S} \times \mathcal{A} \times \mathcal{R})^{\mathbb{N}},$$

where $\mathcal{S}_0, \dots, \mathcal{S}_{T+1} \subset \mathcal{S}$, $\mathcal{A}_0, \dots, \mathcal{A}_T \subset \mathcal{A}$ and $\mathcal{R}_1, \dots, \mathcal{R}_{T+1} \subset \mathcal{R}$ by summing as follows:

$$\begin{aligned}
&\mathbb{P}_{\mu,\pi}\left(\prod_{t=0}^{T}(\mathcal{S}_t \times \mathcal{A}_t \times \mathcal{R}_{t+1}) \times \mathcal{S}_{T+1} \times (\mathcal{S} \times \mathcal{A} \times \mathcal{R})^{\mathbb{N}}\right) \\
&\quad = \sum_{\substack{s_0 \in \mathcal{S}_0 \\ \vdots \\ s_{T+1} \in \mathcal{S}_{T+1}}} \sum_{\substack{a_0 \in \mathcal{A}_0 \\ \vdots \\ a_T \in \mathcal{A}_T}} \sum_{\substack{r_1 \in \mathcal{R}_1 \\ \vdots \\ r_{T+1} \in \mathcal{R}_{T+1}}} \mu(s_0) \prod_{t=0}^{T} \pi_t(a_t | h^{(t)}) p(s_{t+1}, r_{t+1} | s_t, a_t).
\end{aligned}$$

$\mathbb{P}_{\mu,\pi}$ can then be seen to satisfy the assumptions of Kolmogorov's extension theorem which assures that $\mathbb{P}_{\mu,\pi}$ can be extended to a unique probability measure on $\mathcal{H}^{\infty}$. $\square$

Definition 1.3.2. Let $\mu \in \Delta(\mathcal{S})$ and $\pi \in \Pi$. $\mathbb{P}_{\mu,\pi}$ is called the *probability distribution over histories* induced by initial state distribution μ and policy π.

We introduce some additional notation. Let $\mu \in \Delta(\mathcal{S})$ and $\pi \in \Pi$. We use $\mathbb{E}_{\mu,\pi}[\,\cdot\,]$ as a shorthand for

$$\mathbb{E}_{(S_0, A_0, R_1, \dots) \sim \mathbb{P}_{\mu,\pi}}[\,\cdot\,].$$

If μ is the Dirac in some state $s \in \mathcal{S}$, we write $\mathbb{P}_{s,\pi}$ (resp. $\mathbb{E}_{s,\pi}[\,\cdot\,]$) instead of $\mathbb{P}_{\delta_s,\pi}$ (resp. $\mathbb{E}_{\delta_s,\pi}[\,\cdot\,]$).

Definition 1.3.3. Let $(s, a) \in \mathcal{S} \times \mathcal{A}$, $\pi = (\pi_t)_{t \geqslant 0}$ a policy and $\pi' = (\pi'_t)_{t \geqslant 0}$ defined as

$$\begin{aligned}
\pi'_0(s) &= \delta_a, \\
\pi'_0(s') &= \pi_0(s') \quad \text{for } s' \neq s \\
\pi'_t &= \pi_t \quad \text{for } t \geqslant 1.
\end{aligned}$$

Then, $\mathbb{P}_{s,\pi'}$ is called the probability distribution induced by initial state s, initial action a, and policy π, and is denoted $\mathbb{P}_{s,a,\pi}$.

For $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $\pi \in \Pi$, we also introduce the shorthand

$$\mathbb{E}_{s,a,\pi}[\,\cdot\,] := \mathbb{E}_{(S_0, A_0, R_1, \dots) \sim \mathbb{P}_{s,a,\pi}}[\,\cdot\,].$$

1.4 Value functions

Definition 1.4.1. (i) A *state-value function* (aka *V-function*) is a function $v : \mathcal{S} \to \mathbb{R}$ or equivalently a vector $v = (v(s))_{s\in\mathcal{S}} \in \mathbb{R}^{\mathcal{S}}$.

(ii) An *action-value function* (aka *Q-function*) is a function $q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ or equivalently a vector $q = (q(s,a))_{(s,a)\in\mathcal{S}\times\mathcal{A}} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$.

Proposition 1.4.2. *Let $(R_t)_{t\geqslant 1}$ be a sequence of random variables with values in $\mathcal{R}$ and $\gamma \in (0,1)$. Then, the series $\sum_{t\geqslant 1} \gamma^{t-1} R_t$ converges almost-surely, and its sum is integrable. Moreover,*

$$\mathbb{E}\left[\sum_{t=1}^{+\infty} \gamma^{t-1} R_t\right] = \sum_{t=1}^{+\infty} \gamma^{t-1} \mathbb{E}\left[R_t\right].$$

Proof. $\mathcal{R}$ being a finite subset of $\mathbb{R}$, it holds that $\max_{r\in\mathcal{R}} |r| < +\infty$. Then,

$$\left|\gamma^{t-1} R_t\right| \leqslant \gamma^{t-1} \max_{r\in\mathcal{R}} |r|, \quad \text{a.s.}$$

The result follows the dominated convergence theorem. □

Definition 1.4.3. Let $\pi \in \Pi$ and $\gamma \in (0,1)$.

(i) The *state-value function of policy* π with discount factor γ is defined as

$$v_\pi^{(\gamma)}(s) = \mathbb{E}_{s,\pi}\left[\sum_{t=1}^{+\infty} \gamma^{t-1} R_t\right], \quad s \in \mathcal{S}.$$

(ii) The *action-value function of policy* π with discount factor γ is defined as

$$q_\pi^{(\gamma)}(s,a) = \mathbb{E}_{s,a,\pi}\left[\sum_{t=1}^{+\infty} \gamma^{t-1} R_t\right], \quad (s,a) \in \mathcal{S} \times \mathcal{A}.$$

We may denote $v_\pi = v_\pi^{(\gamma)}$ and $q_\pi = q_\pi^{(\gamma)}$ when γ is clear from the context.

Chapter 2

Bellman operators & optimality

We assume that $\gamma \in (0,1)$ in given. The image of an element $x \in X$ by a map $F : X \to Y$ will often be denoted Fx instead of $F(x)$.

2.1 Bellman operators

Definition 2.1.1. Let π be a stationary policy. We define the following operators.

(i) $D^{(\gamma)} : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ as

$$(D^{(\gamma)}v)(s,a) = \sum_{(s',r)\in\mathcal{S}\times\mathcal{R}} p(s',r|s,a)(r+\gamma v(s')), \quad s\in\mathcal{S},\ a\in\mathcal{A}.$$

(ii) $E_\pi : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}}$ as

$$(E_\pi q)(s) = \sum_{a\in\mathcal{A}} \pi(s|a)q(s,a), \quad s\in\mathcal{S}.$$

(iii) $E_* : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}}$ as

$$(E_* q)(s) = \max_{a\in\mathcal{A}} q(s,a), \quad s\in\mathcal{S}.$$

(iv) $B_\pi^{(V,\gamma)} = E_\pi \circ D^{(\gamma)}$ (Bellman expectation operator for state-value functions)

(v) $B_*^{(V,\gamma)} = E_* \circ D^{(\gamma)}$ (Bellman optimality operator for state-value functions)

(vi) $B_\pi^{(Q,\gamma)} = D^{(\gamma)} \circ E_\pi$ (Bellman expectation operator for action-value functions)

(vii) $B_*^{(Q,\gamma)} = D^{(\gamma)} \circ E_*$ (Bellman optimality operator for action-value functions)

We will use lighter notation $D, E_\pi, E_*, B_\pi, B_*$ as soon as context prevents confusion. The following expressions follow from the definitions.

Proposition 2.1.2 (Explicit expression of Bellman operators)**.** *Let $v \in \mathbb{R}^{\mathcal{S}}$, $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, and π a stationary policy. Then, the following expressions hold.*

$$(B_\pi v)(s) = \sum_{(a,s',r)\in\mathcal{A}\times\mathcal{S}\times\mathcal{R}} \pi(a|s)p(s',r|s,a)\left(r+\gamma v(s')\right), \quad s \in \mathcal{S},$$

$$(B_* v)(s) = \max_{a\in\mathcal{A}} \sum_{(s',r)\in\mathcal{S}\times\mathcal{R}} p(s',r|s,a)(r+\gamma v(s')), \quad s \in \mathcal{S},$$

$$(B_\pi q)(s,a) = \sum_{(s',r,a')\in\mathcal{S}\times\mathcal{R}\times\mathcal{A}} p(s',r|s,a)\left(r+\gamma\pi(a'|s')q(s',a')\right), \quad (s,a)\in\mathcal{S}\times\mathcal{A},$$

$$(B_* q)(s,a) = \sum_{(s',r)\in\mathcal{S}\times\mathcal{R}} p(s',r|s,a)\left(r+\gamma\max_{a'\in\mathcal{A}} q(s',a')\right), \quad (s,a)\in\mathcal{S}\times\mathcal{A}.$$

Proof. Immediate from the definitions. □

Proposition 2.1.3. *Let $v \in \mathbb{R}^{\mathcal{S}}$, $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, $s \in \mathcal{S}$, $a \in \mathcal{A}$ and π a stationary policy. Then,*

$$\begin{aligned}(B_\pi v)(s) &= \mathbb{E}_{s,\pi}\left[R_1+\gamma v(S_1)\right]\\ (B_\pi q)(s,a) &= \mathbb{E}_{s,a,\pi}\left[R_1+\gamma q(S_1,A_1)\right].\end{aligned}$$

Proof. Using the explicit expression from Proposition 2.1.2 and the definition of the probability measure $\mathbb{P}_{s,\pi}$ (see Proposition 1.3.1), we write

$$\begin{aligned}(B_\pi v)(s) &= \sum_{(a,s',r)\in\mathcal{A}\times\mathcal{S}\times\mathcal{R}} \pi(a|s)p(s',r|s,a)(r+\gamma v(s'))\\ &= \sum_{\substack{a\in\mathcal{A}\\ r\in\mathcal{R}}} \mathbb{P}_{s,\pi}\left(\{s\}\times\{a\}\times\{r\}\times\{s'\}\times(\mathcal{S}\times\mathcal{A}\times\mathcal{R})^{\mathbb{N}}\right)\\ &\qquad\times(r+\gamma v(s'))\\ &= \mathbb{E}_{s,\pi}\left[R_1+\gamma v(S_1)\right].\end{aligned}$$

The expression for $B_\pi q$ is proved similary. □

Definition 2.1.4. Let $d, n \geqslant 1$ integers. A map $F : \mathbb{R}^d \to \mathbb{R}^n$ is *monotone* if for all $x, x' \in \mathbb{R}^d$, $x \leqslant x'$ implies $Fx \leqslant Fx'$, where the inequalities are to be understood component-wise.

Proposition 2.1.5. *Operators $D, E_\pi, B_\pi^{(V)}, B_\pi^{(Q)}$ are affine with nonnegative coefficients. E_π is moreover linear. In particular, they are monotone.*

Proof. Immediate from the definitions. □

Proposition 2.1.6. *Let $v \in \mathbb{R}^{\mathcal{S}}$, $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Then,*

(i) $(E_* q)(s,a) = \sup_{\pi\in\Pi_0} (E_\pi q)(s,a) = \sup_{\pi\in\Pi_{0,d}} (E_\pi q)(s,a)$,

(ii) $(B_* v)(s) = \sup_{\pi\in\Pi_0} (B_\pi v)(s) = \sup_{\pi\in\Pi_{0,d}} (B_\pi v)(s)$,

(iii) $(B_* q)(s,a) = \sup_{\pi\in\Pi_0} (B_\pi q)(s,a) = \sup_{\pi\in\Pi_{0,d}} (B_\pi q)(s,a)$.

Proof. (i) is an easy consequence from the definition of E_*. Then (ii) and (iii) follow using the monotonicity from Proposition 2.1.5. □

2.2 Bellman equations

Definition 2.2.1. Let X be a set and $F : X \to X$. An element $x \in X$ is a *fixed point* of F is $Fx = x$.

Theorem 2.2.2 (Banach's fixed point theorem)**.** *Let $0 \leqslant \gamma < 1$, (X, d) a complete metric space, and $F : X \to X$ a γ-Lipschitz map (with respect to distance d). Then, F has a unique fixed point $x_* \in X$ and for all sequence $(x_k)_{k\geqslant 0}$ satisfying $x_{k+1} = Fx_k$ ($k \geqslant 0$), it holds that*

$$d(x_k, x_*) \leqslant \gamma^k d(x_0, x_*), \quad k \geqslant 0,$$

and thus $x_k \longrightarrow x_*$ as $k \to +\infty$.

Proposition 2.2.3. *Let π be a stationary policy. With respect to the norms $\|\cdot\|_\infty$ in $\mathbb{R}^{\mathcal{S}}$ and $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$,*

(i) $D^{(\gamma)}$ *is γ-Lipschitz*

(ii) E_π *is 1-Lipschitz*

(iii) E_* *is 1-Lipschitz*

(iv) $B_\pi^{(V,\gamma)}$, $B_*^{(V,\gamma)}$, $B_\pi^{Q,\gamma}$ *and* $B_*^{(Q,\gamma)}$ *are γ-Lipschitz and admit unique fixed points.*

Proof. TODO □

Proposition 2.2.4. *Let π be a stationary policy. Then,*

(i) $v_\pi = E_\pi q_\pi$,

(ii) $q_\pi = Dv_\pi$,

(iii) v_π *is the unique fixed point of* $B_\pi^{(V)}$*, meaning the unique solution to the Bellman expectation equation for state-value functions.*

(iv) q_π *is the unique fixed point of* $B_\pi^{(Q)}$*, meaning the unique solution to the Bellman expectation equation for action-value functions.*

Proof. TODO: consequence of the definitions. □

2.3 Greedy policy

Definition 2.3.1. A stationary and deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$ is

(i) a *greedy policy* with respect to an action-value function $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ if for all $s \in \mathcal{S}$,

$$\pi(s) \in \operatorname*{Arg\,max}_{a\in\mathcal{A}} q(s, a),$$

where Arg max denotes the set of maximizers.

(ii) a *greedy policy* with respect to an state-value function $v \in \mathbb{R}^{\mathcal{S}}$ if $\pi \in \Pi_g[Dv]$.

$\Pi_g[q]$ denotes the set of greedy policies with respect to q and $\Pi_g[v]$ is a shorthand for $\Pi_g[Dv]$.

Proposition 2.3.2. *For* $v \in \mathbb{R}^{\mathcal{S}}$ *(resp.* $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$*),* $\Pi_g[v]$ *(resp.* $\Pi_g[q]$*) is nonempty.*

Proof. The set of actions $\mathcal{A}$ being finite (and nonempty), $\operatorname{Arg\,max}_{a\in\mathcal{A}} q(s, a)$ is nonempty, and the result follows. □

Notation $\pi_g[q]$ (resp. $\pi_g[v]$) denotes any element from $\Pi_g[q]$ (resp. $\Pi_g[v]$).

Proposition 2.3.3. *Let* $v \in \mathbb{R}^{\mathcal{S}}$ *and* $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$*. Then,*

(i) $E_* q = E_{\pi_g[q]} q$,

(ii) $B_* q = B_{\pi_g[q]} q$.

(iii) $B_* v = B_{\pi_g[v]} v$,

Proof. TODO □

2.4 Optimal value functions & policies

Definition 2.4.1. Let $\gamma \in (0,1)$. The *optimal state-value* and *actions-value functions* with respect to discount factor γ are respectively defined as

$$\begin{aligned} v_*^{(\gamma)}(s) &= \sup_{\pi\in\Pi} v_\pi^{(\gamma)}(s), \quad s \in \mathcal{S}, \\ q_*^{(\gamma)}(s,a) &= \sup_{\pi\in\Pi} q_\pi^{(\gamma)}(s,a), \quad (s,a) \in \mathcal{S}\times\mathcal{A}. \end{aligned}$$

As soon as discount factor γ is clear from the context, we may simply use notation v_* and q_*.

Definition 2.4.2. A policy π_* is *optimal* if $v_{\pi_*} = v_*$.

Theorem 2.4.3. *Let v_0 and q_0 the unique fixed points of $B_*^{(V)}$ and $B_*^{(Q)}$ respectively. Then, $\Pi_g[v_0] = \Pi_g[q_0]$ and for π_g in the latter set,*

(i) $v_* = v_0 = v_{\pi_g}$,

(ii) $q_* = q_0 = q_{\pi_g}$,

(iii) $v_* = E_* q_*$,

(iv) $q_* = D v_*$.

Remark 2.4.4. Some important takeaways from the above theorem are the following:

- v_* (resp. q_*) is the unique fixed point of $B_*^{(V)}$ (resp. $B_*^{(Q)}$), meaning the unique solution to the Bellman expectation equation for state-value function (resp. action-value function);
- there exists a stationary and deterministic optimal policy.

Proof. Let us first prove that $q_0 = Dv_0$ and $v_0 = E_* q_0$. Indeed,

$$Dv_0 = DB_*v_0 = DE_*Dv_0 = B_*(Dv_0),$$

therefore, Dv_0 is the unique fixed point of B_*, in other words $q_0 = Dv_0$. Then,

$$E_*q_0 = E_*Dv_0 = B_*v_0 = v_0.$$

Therefore, $\Pi_g[v_0] = \Pi_g[Dv_0] = \Pi_g[q_0]$. We recall that a set of greedy policies is never empty, as stated in Proposition 2.3.2.

Let $\pi_g \in \Pi_g[v_0]$. Then using the property of greedy policies from Proposition 2.3.3, $v_0 = B_*v_0 = B_{\pi_g}v_0$ and $q_0 = B_*q_0 = B_{\pi_g}q_0$. Value functions v_0 and q_0 are therefore the unique fixed points of $B_{\pi_g}^{(V)}$ and $B_{\pi_g}^{(Q)}$, respectively. In other words $v_0 = v_{\pi_g}$ and $q_0 = q_{\pi_g}$, by Proposition 2.2.4.

Therefore, $v_0 = v_{\pi_g} \leqslant \sup_{\pi \in \Pi_{0,d}} v_\pi$ because $\pi_g \in \Pi_{0,d}$ by definition, and similarly $q_0 \leqslant \sup_{\pi \in \Pi_{0,d}} q_\pi$.

Let us now prove that $v_0 \geqslant \sup_{\pi \in \Pi} v_\pi$. Let $\pi = (\pi_t)_{t \geqslant 0}$ be any policy, $s \in \mathcal{S}$, and consider random variables $(S_0, A_0, R_1, S_2, A_2, R_3, \dots) \sim \mathbb{P}_{s,\pi}$. Then for each $t \geqslant 0$,

$$\begin{aligned}
v_0(S_t) = (B_* v_0)(S_t) &= \max_{a \in \mathcal{A}} \sum_{(s',r) \in \mathcal{S} \times \mathbb{R}} p(s', r | s, a)(r + \gamma v_0(s')) \\
&\geqslant \sum_{(s',r) \in \mathcal{S} \times \mathbb{R}} p(s', r | S_t, A_t)(r + \gamma v_0(s')) \\
&= \mathbb{E}\left[R_{t+1} + \gamma v_0(S_{t+1}) \,|\, S_t, A_t\right],
\end{aligned}$$

where the last equality follows from the definition of $\mathbb{P}_{s,\pi}$. Then using the expression of $(B v_0)(s)$ from Proposition 2.1.3, and applying the above recursively, we get

$$\begin{aligned}
v_0(s) = (B v_0)(s) &= \mathbb{E}_{s,\pi}\left[R_1 + \gamma v_0(S_1)\right] \\
&\geqslant \mathbb{E}_{s,\pi}\left[R_1 + \gamma \mathbb{E}\left[R_2 + \gamma v_0(S_2) \,|\, S_1, A_1\right]\right] \\
&= \mathbb{E}_{s,\pi}\left[R_1 + \gamma\left(R_2 + \gamma v_0(S_2)\right)\right] \\
&\geqslant \cdots \geqslant \mathbb{E}_{s,\pi}\left[\sum_{t=1}^{+\infty} \gamma^{t-1} R_t\right] \\
&= v_\pi(s).
\end{aligned}$$

Therefore, $v_* = \sup_{\pi \in \Pi} v_\pi \leqslant v_0 = v_{\pi_g} \leqslant \sup_{\pi \in \Pi_{0,d}} v_\pi \leqslant \sup_{\pi \in \Pi} v_\pi = v_*$, and the lower and upper bounds being equal, all inequalies are equalities, and the supremums are maximums because they are attained for $\pi_g \in \Pi_{0,d} \subset \Pi$.

Then, we write

$$q_* = \sup_{\pi \in \Pi} q_\pi \geqslant \max_{\pi \in \Pi_{0,d}} q_\pi \geqslant q_{\pi_g} = q_0 = D v_0 = D\left(\max_{\pi \in \Pi} v_\pi\right) \geqslant \sup_{\pi \in \Pi} D v_\pi = \sup_{\pi \in \Pi} q_\pi = q_*$$

where the last inequality holds by monotonicity of D from Proposition 2.1.5 (by writing for $\pi \in \Pi$, $D \max_{\pi \in \Pi} v_\pi \geqslant D v_\pi$ and then taking the supremum over $\pi \in \Pi$) Therefore, all inequalities are equalities are all supremums are maximums. □

Chapter 3

Dynamic programming

3.1 Value iteration for policy evaluation

3.2 Value iteration for control

3.3 Policy iteration for control

3.4 Asynchronous value iteration

Chapter 4

Tabular reinforcement learning

4.1 Asynchronous stochastic approximations

4.2 Stochastic estimators of Bellman equations

4.3 Policy evaluation

4.4 Control

Chapter 5

Value function approximation

Chapter 6

Policy gradient

Chapter 7

Additional methods: actor-critic & model-based