Static and Dynamic Aspects of Optimal Sequential Decision Making

by
Csaba Szepesvári

Thesis

Submitted in partial fulfillment of the requirements
for the
Degree of Doctor of Philosophy in the
Bolyai Institute of Mathematics
at "József Attila" University

1998

Contents

List of Symbols and Abbreviations

$\mathbb{N}, \mathbb{Z}, \mathbb{R}$ – set of natural numbers ($\mathbb{N} = \{0, 1, 2, \ldots\}$), integers, reals

$\|\cdot\|$ – supremum-norm: $f : X \to \mathbb{R}$, $\|f\| = \sup_{x \in X} |f(x)|$.

Part I

$\mathcal{X}, x, y$ – set of states, states

$\mathcal{A}, a, b$ – set of actions, actions

$p(x, a, y)$ – transition probabilities; $p(x, a, y) \geq 0$, $\sum_{y \in \mathcal{X}} p(x, a, y) = 1$

$c(x, a, y)$ – transition costs

$B(\mathcal{X})$ – bounded real-valued functions over $\mathcal{X}$ ($B(\mathcal{X}) = \{f : \mathcal{X} \to \mathbf{R} \mid \|f\| < \infty\}$

$\mathcal{R}(\mathcal{X})$ – extended real-valued functions over $\mathcal{X}$ ($\mathcal{R}(\mathcal{X}) = [-\infty, \infty]^X$)

$\mathcal{Q}$ – cost propagation operator; $\mathcal{Q} : \mathcal{R}(\mathcal{X}) \to \mathcal{R}(\mathcal{X} \times \mathcal{A})$

π, μ, ϕ – policies

v, v_π, v^{*g} – cost-to-go function, cost-to-go function associated to the policy π, optimal cost-to-go function

T – optimal cost-to-go operator; $T : \mathcal{R}(\mathcal{X}) \to \mathcal{R}(\mathcal{X})$, $(Tv)(x) = \inf_{a \in A} (\mathcal{Q}v)(x, a)$

LSC,USC,M,I,D – lower-, upper-semicontinuous, monotone, uniformly increasing, uniformly decreasing

Π, Π^*, Γ – set of policies, set of optimal policies, greedy operator

w.r.t., l.h.s., r.h.s – with respect to, left-hand-side, right-hand-side

$\Pi(X)$ – set of probability distributions over the the finite set X

Part II

γ – $0 < \gamma < 1$, contraction factor

$T_t(\cdot, \cdot)$ – approximating optimal cost-to-go operator; $T_t : \mathcal{B} \times \mathcal{B} \to \mathcal{B}$, where $\mathcal{B}$ is a normed vector space

$P(\cdot), E(\cdot), \mathrm{Var}(\cdot)$ – probability, expectation, variance

$G_t(\cdot), F_t(\cdot)$ – Lipschitz coefficient functions

$\delta_t(\cdot), \Delta_t(\cdot), V_t(\cdot)$ – difference functions, estimated optimal cost-to-go function

w.p.1, a.s., a.e. – with probability one, almost surely, almost everywhere

Foreword

In this thesis the theory of optimal sequential decisions having a general recursive structure is investigated via an operator theoretical approach, the recursive structure (of both of the dynamics and the optimality criterion) being encoded into the so-called cost propagation operator. Decision problems like Markovian Decision Problems with expected or worst-case total discounted/undiscounted cost criterion; repeated zero-sum games such as Markov-games; or alternating Markov-games all admit such a recursive structure. Our setup has the advantage that it emphasizes their common properties as well as it points out some differences.

The thesis consists of two parts, in the first part the model is assumed to be known while in the second one the models are to be explored. The setup of Part I is rather abstract but enables a unified treatment of a large class of sequential decision problems, namely the class when the total cost of decision policies is defined recursively by a so called *cost propagation operator*. Under natural monotonicity and continuity conditions the greedy policies w.r.t. the optimal cost-to-go function turn out to be optimal, due to the recursive structure.

Part II considers the case when the models are unknown, and have to be explored and learnt. The price of considering unknown models is that here we have to restrict ourselves to models with an additive cost structure in order to obtain tractable learning situations. The almost sure convergence of the most frequently used algorithms proposed in the reinforcement learning community is proved. These algorithms are treated as multidimensional asynchronous stochastic approximation schemes and their convergence is deduced from the main theorem of the second part. The key of the method here is the so called rescaling property of certain homogeneous processes. A practical and verifiable sufficient condition for the convergence of on-line learning policies to an optimal policy is formulated and a convergence rate is established.

The algorithms discussed in this thesis has been tried out on a real-robot with some success [36, 37] (the robot is shown in Figure 1). The experiments were analyzed by ANOVA and the results indicated the significant superiority of the model-based learning algorithms over the model-free ones. Although the learnt policy differed from that of a handcrafted policy, the respective performances were indistinguishable.

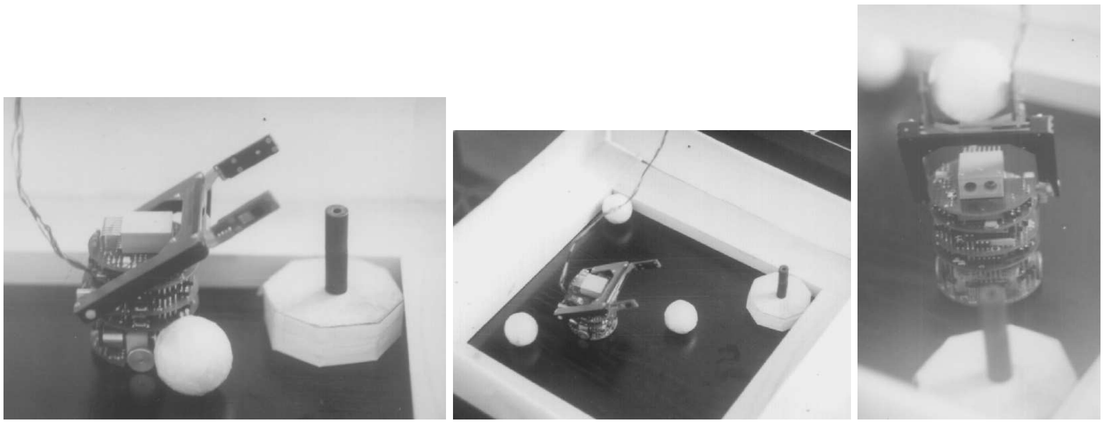

Figure 1: **The Khepera robot**

The figures show a Khepera robot equipped with eight IR *sensors, six in the front and two at the back, the IR sensors measuring the proximity of objects in the range 0-5 cm. The robot has two wheels driven by two independent* DC *motors and a gripper that has two degrees of freedom and is equipped with a resistivity sensor and an object-presence sensor. The vision turret is mounted on the top of the robot as shown. It is an image sensor giving a linear image of the horizontal view of the environment with a resolution of 64 pixels and 256 levels of grey. The horizontal viewing angle is limited to about 36 degrees. This sensor is designed to detect objects in front of the robot situated at a distance spanning 5 to 50 cm. The image sensor has no tilt angle, so the robot observes only those objects whose height is exceeds 5 cm. The task was to find a ball in the arena, bring it to the stick which is in a corner and hit the stick by the ball so as to it jumps out of the gripper. Some macro actions such as search, grasp, etc. were defined and the number of macro actions taken by the robot until the goal was reached were measured. A filtered version of the state space served as the state space and the robot learnt a decision policy by the algorithms investigated in this thesis.*

Acknowledgements

I would like to express my sincere gratitude to András Lőrincz, who introduced me to the field of autonomous robots and reinforcement learning. The present study has grown from our common work during the last few years. My advisor, András Krámli, helped me to formulate and prove my mathematical results. Many thanks to him. Thanks are due to friends and collegues László Balázs, Mark French, Zoltán Gábor, Zsolt Kalmár, Michael L. Littman, Carlos H.C. Ribeiro for many helpful and inspiring discussions with them. Special thanks to Mark French proofread this thesis and Michael L. Littman with whom I coauthored some articles on which this thesis is based despite that we have met personally only after most of our common material was ready. Thanks to my mother & father, as well as to my brother, Szabolcs, who all encoureged me as a child to become a scientist by providing constant positive feedback. Finally, I am grateful to my nucleus family (my wife Beáta, and children Dávid, Réka and Eszter) for their love, encouragement, and support whilst this work was being done.

This work was supported by OTKA (grant Nos. F20132 and T14548), the Hungarian Ministry of Education (grant No. FKFP 1354/1997) and the Soros Foundation.

I dedicate this thesis to the memory of my brother, Gergő.

Part I

Abstract Dynamic Programming

Introduction

Abstract dynamic programming (ADP) analyses *structural questions* associated to sequential decision problems, based only on the recursive structure of such problems. It provides general tools for solving many kinds of sequential decision problems, such as ordinary Markovian decision problems with the total expected discounted cost [13, 65], worst-case cost [4, 28] (see Section 4.5), or exponential utility criterion [32, 6, 16], multi-step games (both alternating and Markov games [60, 17]) (see Example 0.1.4 and Section 4.4) or "mixed" sequential optimization problems where these various criteria and dynamics are combined [31].

The objective of this part of the thesis is to fill in some gaps in the theory of ADP. This part consists of two further chapters. In the next chapter, the evaluation of general (not necessarily Markovian) policies are defined, for the first-time, using cost propagation operators only. It is shown under a positivity or negativity assumption that when the cost propagation operator satisfies certain continuity properties, the optimal cost-to-go function remains the same as the optimal cost-to-go function defined for Markovian policies. Then related problems such as the convergence of the value iteration algorithm and existence of optimal policies are considered.

In the second chapter we consider increasing models under minimal continuity assumptions. It is shown that Howard's policy improvement routine decreases the "long-term cost-to-go" but does not necessarily yield optimal policies even for finite models. We also give a description of the relationships of the key theorems for increasing models under the minimal continuity assumptions (see Figure 2).

0.1 Overview of Problems

0.1.1 Notation

The relation $u \leq v$ will be applied to functions in the usual way: $u \leq v$ means that $u(x) \leq v(x)$ for all x in the domain of u and v. Further, $u < v$ will denote that $u \leq v$ and that there exists an element x of the domain of u and v such that $u(x) < v(x)$. We employ the symbol $\leq$ for operators in the same way, and say that $S_1 \leq S_2$ ($S_1, S_2 : \mathcal{R}(\mathcal{X}) \to \mathcal{R}(\mathcal{X})$) if $S_1 v \leq S_2 v$ for all $v \in \mathcal{R}(\mathcal{X})$. If $S : \mathcal{R}(\mathcal{X}) \to \mathcal{R}(\mathcal{X})$ is an arbitrary operator then S^k ($k = 1, 2, 3, \ldots$) will denote the composition of S with itself k times: $S^0 v = v$, $S^1 v = Sv$, $S^2 v = S(Sv)$, etc.

0.1.2 Definitions

First, we give a definition of abstract sequential decision problems.

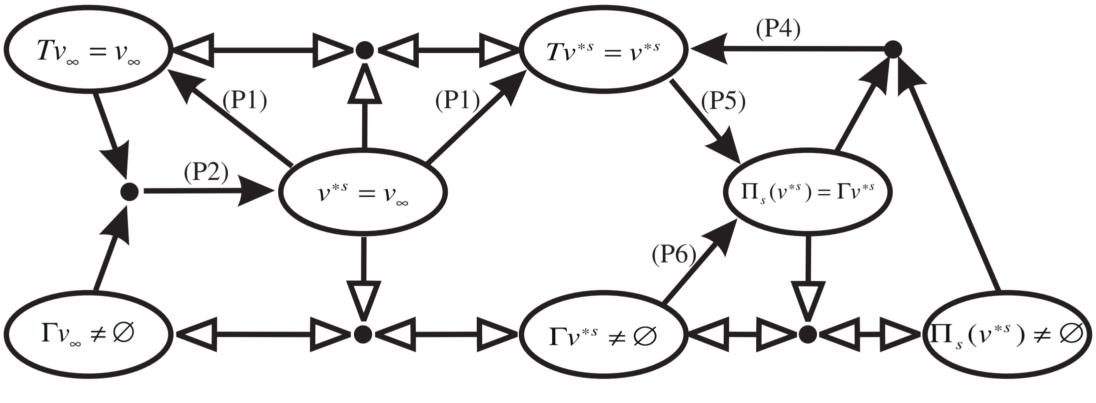

Figure 2: **Relations among the statements concerning best stationary policies in models satisfying M,I and LSC.**
The arrows indicate consequences, i.e., if an arrow points from one node to another then the statement of the goal node is a consequence of the statement of the start node. Black dots denote the "and" operation, so, for example, from $Tv_\infty = v_\infty$ and $\Gamma v_\infty \neq \emptyset$ it follows that $v^{*s} = v_\infty$. The arrows with white heads denote trivial assertions. The non-trivial relations are proved in the text.

Definition 0.1.1 *An* abstract sequential decision problem *(ADP) is a quadruple $(\mathcal{X}, \mathcal{A}, \mathcal{Q}, \ell)$, where $\mathcal{X}$ is the state space of the process, $\mathcal{A}$ is the set of actions, $\mathcal{Q} : [-\infty, \infty]^{\mathcal{X}} \to [-\infty, \infty]^{\mathcal{X}\times\mathcal{A}}$ is the so-called* cost propagation operator *and $\ell \in \mathcal{R}(\mathcal{X})$ is the so-called terminal cost function ($\mathcal{R}(\mathcal{X}) = [-\infty, +\infty]^{\mathcal{X}}$ and $\mathcal{R}(\mathcal{X} \times \mathcal{A}) = [-\infty, +\infty]^{\mathcal{X}\times\mathcal{A}}$).*

The mapping $\mathcal{Q}$ makes it possible to define the cost of a decision (action) sequence in a recursive way: the cost of decision a in state x is given by $(\mathcal{Q}f)(x, a)$ provided the process stops immediately after the first decision and the terminal cost of stopping in state y is given by $f(y)$.

Then action sequences can be evaluated via $\mathcal{Q}$ in the following way. The cost of the finite decision sequence $(a_0, a_1, \ldots, a_t)$ (the first decision is a_0, the second is a_1, etc.) can be built up working backwards from the last decision to the first. If the process starts in state x the cost of $(a_0, a_1, \ldots, a_t)$ is defined as

$$v_{(a_0,a_1,\ldots,a_t)}(x) = (\mathcal{Q}v_{(a_1,a_2,\ldots,a_t)})(x, a_0),$$

where

$$v_{(a_1,\ldots,a_t)}(x) = (\mathcal{Q}v_{(a_2,a_3,\ldots,a_t)})(x, a_1).$$

etc. with the terminal condition $v_{\{a_t\}}(x) = (\mathcal{Q}\ell)(x, a_t)$, i.e., in finite-horizon models the terminal cost function determines the terminal cost: when the final state is y the decision maker incurs a final cost of $\ell(y)$.

Infinite action sequences can be evaluated as the limit of the corresponding finite horizon evaluations provided that the limit exists. If, for example, $(\mathcal{Q}\ell)(\cdot, a) \geq \ell$ holds for all $a \in \mathcal{A}$ then the derived sequence of functions is non-decreasing and thus the limit must exist.

DEFINITION 0.1.2 *Models satisfying*

$$(\mathcal{Q}\ell)(\cdot, a) \geq \ell \quad \text{or} \quad (\mathcal{Q}\ell)(\cdot, a) \leq \ell \tag{1}$$

are called increasing *models (resp. decreasing).*[1] *Accordingly $\mathcal{Q}$ is called increasing (resp. decreasing), and these properties will be denoted by I and D, respectively.*

Markovian decision problems (MDPs) with the expected total cost criterion are the standard example of abstract sequential decision problems.

EXAMPLE 0.1.3 *Finite Markovian decision problems with the expected total cost criterion [7, 55].* $(\mathcal{X}, \mathcal{A}, p, c)$ is called a finite MDP if the following conditions hold:

1. $\mathcal{X}$ and $\mathcal{A}$ are finite sets,
2. $p : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \mathbf{R}$ and for each $a \in \mathcal{A}$, $p(\cdot, a, \cdot)$ is a transition probability matrix, i.e., $0 \leq p(x, a, y) \leq 1$ $\sum_{y \in \mathcal{X}} p(x, a, y) = 1$ for all $x \in \mathcal{X}$,
3. $c : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \mathbf{R}$ is an arbitrary mapping.

For any given action sequence $a_0, a_1, \ldots$ this structure gives rise to a controlled Markov-chain, ξ_t, whose dynamics is given by $P(\xi_{t+1} = y \,|\, \xi_t = x, a_t = a) = p(x, a, y)$, where a_t is the action taken by the decision maker at time t and $\xi_0 = x_0$. Precisely, to any action sequence $a_0, a_1, \ldots$ and initial state x_0 there corresponds a measure $P = P_{x_0, a_0, a_1, \ldots}$ over $\mathcal{X}^{\mathbb{N}}$ which is uniquely defined by the finite dimensional probabilities $P(x_1, \ldots x_n) = p(x_0, a_0, x_1)p(x_1, a_1, x_2) \ldots p(x_{n-1}, a_n, x_n)$. The objective is to select the actions so that the expected total infinite-horizon discounted cost,

$$E_{x_0}[\sum_{t=0}^{\infty} \gamma^t c(\xi_t, a_t, \xi_{t+1})],$$

is minimized for any given initial state x_0, where the expectation is taken w.r.t. $P = P_{x_0, a_0, a_1, \ldots}$. Here $0 < \gamma \leq 1$ is the so-called *discount factor*. If $\gamma = 1$, i.e., when the costs are not discounted then boundedness questions may arise [7]. Let $a_0, a_1, a_2, \ldots$, denote an action sequence and let the evaluation of it, for the initial state $x \in X$, be

[1] The rationale behind this terminology is that with increasing models the costs incurred by the decision maker increase (resp. decrease) with time.

$V_{\{a_0,a_1,\dots\}}(x) = E_x[\sum_{t=0}^{\infty}\gamma^t c(\xi_t, a_t, \xi_{t+1})]$. By the properties of conditional expectation we easily obtain

$$
\begin{aligned}
E_{x_0}[\ \textstyle\sum_{t=0}^{\infty} & \gamma^t c(\xi_t, a_t, \xi_{t+1})] \qquad (2)\\
= \quad & E_{x_0}[c(\xi_0, a_0, \xi_1)] + E_{x_0}[\sum_{t=1}^{\infty}\gamma^t c(\xi_t, a_t, \xi_{t+1})]\\
= \quad & \sum_{y\in\mathcal{X}} p(x_0, a_0, y)\left(c(x_0, a_0, y) + \gamma E_{x_0}[\sum_{t=0}^{\infty}\gamma^t c(\xi_{t+1}, a_{t+1}, \xi_{t+2}) \,|\, \xi_1 = y]\right)\\
= \quad & \sum_{y\in\mathcal{X}} p(x_0, a_0, y)\left(c(x_0, a_0, y) + \gamma V_{\{a_1,a_2,\dots\}}(y)\right)\\
= \quad & \left(\mathcal{Q}V_{\{a_1,a_2,\dots\}}\right)(x_0, a_0) \qquad (3)
\end{aligned}
$$

provided that $\mathcal{Q} : \mathcal{R}(\mathcal{X}) \to \mathcal{R}(\mathcal{X}\times\mathcal{A})$ is defined by

$$(\mathcal{Q}f)(x, a) = c(x, a) + \gamma\sum_{y\in\mathcal{X}} p(x, a, y)V(y).$$

Note that $\mathcal{Q}$ incorporates both the dynamics ($p(x, a, y)$) and the cost-structure ($c(x, a, y)$) of the decision problem and, by Equation 2, action sequences can be evaluated using $\mathcal{Q}$ only without any reference to p or c. If $\ell(x) = 0$ for all $x \in \mathcal{X}$ then (1) is equivalent to $\sum_{y\in\mathcal{X}} p(x, a, y)c(x, a, y) \geq 0$ holding for all (x, a), i.e., that the immediate averaged costs should be non-negative. Thus, inequality (1) can be viewed as the reformulation of the conventional assumption of *negative* dynamic programming [65]. (It is negative since in [65] or [13] the decision problem is given in terms of a reward function which is related to the cost function by $r(x, a) = -c(x, a)$, and so the immediate rewards are negative if the immediate costs are positive.) For a concrete example of a MDP rewritten in terms of the $\mathcal{Q}$ operator see Example 0.1.14.

Another class of decision problems which can be formulated in the framework of abstract dynamic programming are multi-step sequential games. For example two-player, zero-sum alternating Markov games have the following interpretation in our model:

EXAMPLE 0.1.4 *Alternating Two-player Zero-sum Markov Games [60, 17].* Let $\mathcal{X}$ be the state space, $\mathcal{A}$ the action space and assume that $\mathcal{X}$ is divided into two disjoint parts: $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$. When $x \in \mathcal{X}_i$, Player i chooses the action ($i = 1, 2$). The transition probability function, p, is defined and interpreted as in Markovian decision problems. Further, let $c : \mathcal{X}\times\mathcal{A}\times\mathcal{X} \to \mathbf{R}$ be an arbitrary mapping, the one-step cost mapping from the point of view of Player 1. The game is assumed to be zero sum for each step: when Player 1 incurs the cost $c(x, a, y)$, his opponent incurs the cost $-c(x, a, y)$. Player 1 wants to minimize the expected total cost, while Player 2 wants to maximize it. The optimality criterion is minimax optimality: the maximizer (Player 2) should choose actions so as to maximize the cost of Player 1 in the event that it (i.e., Player 1) chooses the best possible counter-policy. An equivalent definition is for the minimizer to choose actions to minimize its cost against the maximizer with the best possible counter-policy. A pair of policies is said to be in *equilibrium* if neither player has any incentive to change policies if the other player's policy remains fixed. From the point of view of Player 1 playing this game is equivalent to the ADP $(\mathcal{X}_1, \mathcal{A}, \mathcal{Q}, 0)$ with $\mathcal{Q} : [-\infty, \infty]^{\mathcal{X}_1} \to [-\infty, \infty]^{\mathcal{X}_1 \times \mathcal{A}}$ given by

$$(\mathcal{Q}f)(x,a) = \sum_{y \in \mathcal{X}} p(x,a,y) \max_{b \in \mathcal{A}} \sum_{z \in \mathcal{X}} p(y,b,z)\Big(c(x,a,y) + c(y,b,z) + f(z)\Big).$$

If Player 1 uses the optimal policy corresponding to $(\mathcal{X}_1, \mathcal{A}, \mathcal{Q}, 0)$ and Player 2 chooses the best possible counter-policy then these policies can be shown to be in equilibrium [60].

The history of a sequential decision process up to the t^{th} stage is a sequence of state-action pairs: $(x_0, a_0, x_1, a_1, \ldots, x_{t-1}, a_{t-1})$. This will be called the history at time t and we will define the tth history space as the possible histories of the process up to the time instant t: $H_t = (\mathcal{A} \times \mathcal{X})^t$ will be called the tth full history-space. For brevity elements of H_t will often be written as the concatenation of their components, i.e., if $h = ((a_t, x_t), \ldots, (a_0, x_0))$ then we will write $h = a_t x_t \ldots a_0 x_0$. Further, for any pair $h_1 = ((a_t, x_t), \ldots, (a_0, x_0))$ and $h_2 = ((a'_s, x'_s), \ldots, (a'_0, x'_0))$ we will denote by $h_1 h_2$ the concatenation of h_1 and h_2: $((a_t, x_t), \ldots, (a_0, x_0), (a'_s, x'_s), \ldots, (a'_0, x'_0))$. Note, that as H_t is a history space the ordering of the components of its elements is important and we admit the assumption that the ordering of the components corresponds to the time order, i.e., (a_t, x_t) is the most recent element of the history.

Definition 0.1.5 *A* policy *is an infinite sequence of mappings:*

$$\pi = (\pi_0, \pi_1, \ldots, \pi_t, \ldots),$$

where $\pi_t : \mathcal{X} \times H_t \to \mathcal{A}$, $t \geq 0$. If π_t depends only on $\mathcal{X}$ then the policy is called Markovian*, otherwise, it is called* non-Markovian*. If a policy is Markovian and $\pi_t = \pi_0$ for all t then the policy is called* stationary*, otherwise, it is called* non-stationary Markovian*. Stationary policies will also be called* selectors*.*

Policies can be evaluated using the $\mathcal{Q}$ operator similarly to the evaluation of action sequences. In order to simplify the notation let us define "policy-evaluation operators".

Definition 0.1.6 *If $\pi \in \mathcal{A}^{\mathcal{X}}$ is an arbitrary selector let the corresponding* policy-evaluation operator $T_\pi : \mathcal{R}(\mathcal{X}) \to \mathcal{R}(\mathcal{X})$ *be defined as*

$$(T_\pi f)(x) = (\mathcal{Q}f)(x, \pi(x)).$$

The evaluation of Markov policies can be defined recursively as follows: If $v \in \mathcal{R}(\mathcal{X})$ is the evaluation of the t-step Markov policy $\pi' = (\pi_1, \ldots, \pi_t)$, i.e., $v(x)$ is the total cost incurred when π' is executed then from the definition of $\mathcal{Q}$ we have that $(\mathcal{Q}v)(x, \pi_0(x)) = (T_{\pi_0}v)(x)$ is the cost of using π_0 in state x and using π' afterwards.

Definition 0.1.7 (Bertsekas, 1977) *The* evaluation function *(or cost-to-go function) of a finite-horizon Markov policy $\pi = (\pi_0, \pi_1, \ldots, \pi_t)$ is defined as $v_\pi = T_{\pi_0}T_{\pi_1}\ldots T_{\pi_t}\ell$, while the evaluation function of an infinite-horizon Markov policy $\pi = (\pi_0, \pi_1, \ldots, \pi_t, \ldots)$ is given by*

$$v_\pi = \lim_{t\to\infty} T_{\pi_0}T_{\pi_1}\ldots T_{\pi_t}\ell. \tag{4}$$

If the policy is stationary ($\pi_t = \pi_0$ for all $t \geq 0$) this reduces to

$$v_\pi = \lim_{n\to\infty} T^n_{\pi_0}\ell. \tag{5}$$

In what follows selectors will be identified with the stationary Markov policy which they define and thus we can speak about the evaluation (the infinite horizon cost associated with the underlying stationary Markov policy) of a selector. The evaluation of arbitrary policies is more complicated and is the subject of the next chapter.

Example 0.1.8 Consider a finite MDP $(\mathcal{X}, \mathcal{A}, p, c)$ (cf. Example 0.1.3). Let π be any policy. Then for all $x_0 \in X$ π generates a unique measure $P = P_{x_0,\pi}$ over $\mathcal{X}^{\mathbb{N}}$ which is defined through

$$P(x_0, x_1, \ldots, x_n) = P(x_0, a_0, x_1)P(x_1, a_1, x_2)\ldots P(x_{n-1}, a_{n-1}, x_n),$$

where $a_0, \ldots, a_{n-1}$ is recursively defined by $a_0 = \pi_0(x_0)$, $a_1 = \pi_1(x_1, a_0x_0)$, $\ldots$, and $a_{n-1} = \pi_{n-1}(x_{n-1}, a_{n-2}x_{n-2}\ldots a_0x_0)$. Clearly, one can construct a random sequence $(\xi_n, \alpha_n) \in \mathcal{X} \times \mathcal{A}$ (the controlled object) s.t. $P(\xi_{n+1}|\alpha_n, \xi_n, \ldots, \alpha_0, \xi_0) = p(\xi_n, \alpha_n, \xi_{n+1})$ where $\alpha_n = \pi_n(\xi_n, \alpha_{n-1}\xi_{n-1}\ldots\alpha_0\xi_0)$.

The definition of policies can be extended to involve randomized policies, which will be needed in Part II. In this case $\pi_n : X \times H_t \to \Pi(A)$ and $P = P_{x_0,\pi}$ becomes a measure over $(\mathcal{X} \times \mathcal{A})^{\mathbb{N}}$ determined by the sample path probabilities

$$\begin{aligned} P(x_0, a_0, \ldots, x_n, a_n) &= \pi_0(x_0)(a_0)P(x_0, a_0, x_1)\pi_1(x_1, a_0x_0)(a_1)\ldots \\ &\quad P(x_{n-1}, a_{n-1}, x_n)\pi_n(x_n, a_{n-1}x_{n-1}\ldots a_0x_0)(a_n), \end{aligned}$$

and α_n of the controlled object (ξ_n, α_n) should now satisfy

$$P(\alpha_n|\xi_n, \alpha_{n-1}, \xi_{n-1}, \ldots \alpha_0, \xi_0) = \pi_n(\xi_n, \alpha_{n-1}\xi_{n-1}\ldots\alpha_0\xi_0)(\alpha_n),$$

while the state-process, ξ_n, is as before.

The evaluation of a policy π in state x_0 is defined to be $E_{x_0,\pi}[\sum_{t=0}^\infty \gamma^t c(\xi_t, \alpha_t, \xi_{t+1})]$, where the expectation is taken w.r.t. $P_{x_0,\pi}$. Later we will see that this definition coincides with Definition 0.1.7 for (non-randomized) Markovian policies.

Notice that in the case of infinite horizon evaluation the intuitive meaning of terminal costs disappears as there is no terminating step. In some cases the evaluation of policies can still explicitly depend on ℓ.

DEFINITION 0.1.9 *If the evaluation of infinite-horizon policies is independent of ℓ, i.e., if any other bounded function ℓ' yields the same cost-to-go functions, then the decision problem is said to be* stable, *otherwise it is said to be* sensitive.[2]

If the decision problem is stable (such as when $\mathcal{Q}$ is a contraction w.r.t. the supremum-norm) then the structure of the decision problem is simple, otherwise it can be quite complicated.

0.1.3 Objectives

The objective of the decision maker is to choose a policy in such a way that the cost incurred during the usage of the policy is minimal. Of course, the best cost that can be achieved depends on the class of policies available for the decision maker.

DEFINITION 0.1.10 *The sets of general, Markov and stationary policies are denoted by Π_g, Π_m and Π_s, respectively. Further, let*

$$v^{*\Delta}(x) = \inf_{\pi \in \Pi_\Delta} v_\pi(x),$$

be the optimal cost-to-go function for the class Π_Δ, where Δ is either g or m or s.

For any $\varepsilon \geq 0$ and fixed $x \in \mathcal{X}$ the decision maker can assure a cost-to-go less than $v^{*\Delta}(x) + \varepsilon$ by the usage of an appropriate policy from Π_Δ. However, this policy may depend on x.

DEFINITION 0.1.11 *Let*

$$\Pi_\Delta(v) = \{\pi \in \Pi_\Delta \mid v_\pi \leq v\},$$

that is $\Pi_\Delta(v)$ contains the policies from Π_Δ whose cost-to-go is uniformly less than or equal to v. A policy is said to be (uniformly) ε-optimal *if it is contained in $\Pi_g(v^{*g} + \varepsilon)$.*[3]

The objective of abstract sequential decision problems is to give conditions under which $\Pi_\Delta(v^{*g} + \varepsilon)$ is guaranteed to be non-empty when $\varepsilon > 0$ or $\varepsilon = 0$. From the algorithmic point of view the question is how to find an element of $\Pi_\Delta(v^{*g} + \varepsilon)$ for a given $\varepsilon > 0$ ($\Delta \in \{a, m, s\}$).

[2]The notion of stability for MDPs has been introduced and discussed in [86].

[3]If v is a real valued function over $\mathcal{X}$ and ε is real then $v + \varepsilon$ stands for the function $v(x) + \varepsilon$.

DEFINITION 0.1.12 *Elements of* $\Pi_g(v^{*g})$, $\Pi_m(v^{*g})$, *and* $\Pi_s(v^{*g})$ *are called optimal, optimal Markovian and optimal stationary policies, respectively.*

Similar questions can be posed when the set available policies is restricted to the set of Markov, or stationary policies.

DEFINITION 0.1.13 *Elements of* $\Pi_m(v^{*m})$ $(\Pi_s(v^{*s}))$ *are called* best Markovian (stationary) *policies.*

It is clear that $v^{*g} \leq v^{*m} \leq v^{*s}$ since $\Pi_s \subseteq \Pi_m \subseteq \Pi_g$, so the existence of best stationary policies (i.e., if $\Pi_s(v^{*s})$ is non-empty) is easier to ensure than that of optimal stationary policies. It may happen that $\Pi_s(v^{*s})$ is non-empty and $\Pi_s(v^{*g})$ is empty but the converse can never hold. A best stationary policy is also optimal if $v^{*s} = v^{*g}$ so the existence of optimal stationary policies can be reduced to the solving of the inequality $\Pi_s(v^{*s}) \neq \emptyset$ and the equality $v^{*s} = v^{*g}$. Thus, it is reasonable to restrict our attention to stationary policies and algorithms that contain elements from $\Pi_s(v^{*s})$ if there are any. This is exactly the approach of Chapter 2.

0.1.4 Algorithms

There are two sorts of algorithms which are best illustrated for Markovian decision problems with the expected total cost criterion, the *value iteration* [4] and the *policy iteration* [32] methods. Both are based on the fact that *greedy* policies w.r.t. the optimal cost-to-go function (v^{*g}) are optimal. A policy π is said to be greedy w.r.t. the function v if it satisfies the equation

$$(\mathcal{Q}v)(x, \pi(x)) = \inf_{a \in \mathcal{A}} (\mathcal{Q}v)(x, a).$$

For convenience, we will write the right-hand side (RHS) of the above equation in the more compact form $(Tv)(x)$, where $T : \mathcal{R}(\mathcal{X}) \to \mathcal{R}(\mathcal{X})$ is given by

$$(Tv)(x) = \inf_{a \in \mathcal{A}} (\mathcal{Q}v)(x, a).$$

Also for notational convenience we will define the greedy policy operator, Γ, as the operator that maps functions of $\mathcal{R}(\mathcal{X})$ to sets of stationary policies with every $\pi \in \Gamma v$ satisfying $T_\pi v = Tv$.

Value iteration is based on the *Bellman optimality equation* which states that

$$v^{*g} = Tv^{*g}.$$

In discounted MDPs greedy policies w.r.t. v^{*g} are optimal, so a knowledge of v^{*g} is sufficient to find an optimal policy. Since v^{*g} is the fixed point of the operator T it is reasonable to seek it by the method of successive approximations that computes successive estimates of the fixed point of T as $v_{n+1} = Tv_n$ with suitable chosen v_0. If $v_0 = \ell$ then $\lim_{n\to\infty} v_n$ is denoted by v_∞. This choice of initial evaluation is reasonable since then v_n is equal to the optimal n-horizon cost-to-go function. (For stable problems v_0 can be any other function.) Further, it can be shown that $v_\infty \le v^{*g}$. In practice for each n a greedy policy π_n w.r.t. v_n is generated. For finite decision problems π_n will be optimal after a finite number of steps [19], meaning that the value-iteration algorithm can be made to terminate after a finite number of steps. The following example shows that without value iteration does not necessarily converge to v^{*g}.

EXAMPLE 0.1.14 [20] Let $\mathcal{A} = \mathbb{N}$ and let $\mathcal{X} = \mathbb{Z}$, $\ell \equiv 0$. Let

$$(\mathcal{Q}f)(x,a) = \begin{cases} f(a+1), & \text{if } x = 0; \\ 1 + f(-1), & \text{if } x = 1; \\ f(x-1), & \text{if } x \ne 0, 1. \end{cases}$$

Then

$$(Tf)(x,a) = \begin{cases} \inf_{a>0} f(a), & \text{if } x = 0; \\ 1 + f(-1), & \text{if } x = 1; \\ f(x-1), & \text{if } x \ne 0, 1; \end{cases}$$

This example corresponds to a deterministic decision problem with additive cost criterion. In state 0 one may choose any state $x > 0$. After this decision the dynamics cannot be controlled anymore: state x is followed by state $x-1$ except when $x = 1$. When $x = 1$ then the next state is -1 and a cost 1 is incurred. The optimal cost-to-go function is given by

$$v^{*g}(x) = v^{*s}(x) = \begin{cases} 1, & \text{if } x \ge 0; \\ 0, & \text{otherwise.} \end{cases}$$

Let $v_n = T^n \ell$:

$$v_n(x) = \begin{cases} 1, & \text{if } 1 \le x \le n; \\ 0, & \text{otherwise.} \end{cases}$$

Then $v_n \nearrow v_\infty$, where

$$v_\infty(x) = \begin{cases} 1, & \text{if } x \ge 1; \\ 0, & \text{otherwise.} \end{cases}$$

On the other hand, $T(\lim_{n\to\infty} v_n) = Tv_\infty = v^{*g} > v_\infty = \lim_{n\to\infty} Tv_n$ meaning that T is not lower-semicontinuous. It is obvious that $\mathcal{Q}$ is monotone, increasing and lower-semicontinuous. Simple case analysis shows that $Tv^{*g} = v^{*g}$. There is no ε-optimal policy when $0 \le \varepsilon < 1$, but for each $n \in \mathbb{N}$ there exists optimal n-horizon policies, even stationary ones: simply let $\pi(0) \ge n$.

Policy iteration (PI) generates a sequence of policies, π_n that satisfy the relations $\pi_{n+1} \in \Gamma v_{\pi_n}$ or $T_{\pi_{n+1}} v_{\pi_n} = Tv_{\pi_n}$. Note that this method involves the calculation of v_{π_n} which can be hard to do in the general (non-linear) case. For Markovian decision problems v_{π_n} drops out as the solution of some linear system of equations. In addition, for discounted Markovian decision problems, the policy iteration method is known to converge in a finite number of steps provided that both $\mathcal{X}$ and $\mathcal{A}$ are finite.

Chapter 1

Non-Markovian Policies

1.1 The Fundamental Equation

In this section we define the evaluation function associated to non-Markovian policies and derive the fundamental equation of dynamic programming.

DEFINITION 1.1.1 *If $\pi = (\pi_0, \pi_1, \ldots, \pi_t, \ldots)$ is an arbitrary policy then π^t denotes the* t-truncation *of π: $\pi^t = (\pi_0, \pi_1, \ldots, \pi_t)$. Further, let $\mathcal{P}_t$ and $\mathcal{P}$ denote the set of t-truncated policies and the set of (infinite horizon) policies, respectively. The s-truncation operator for t-truncated policies is defined similarly if $s \leq t$.*

DEFINITION 1.1.2 *The* shift-operator *$S_{(x,a)}$ for any pair $(x, a) \in \mathcal{X} \times \mathcal{A}$ is defined in the following way:*

$$S_{(x,a)}\pi = (\pi'_0, \pi'_1, \ldots),$$

where π'_t is defined by

$$\pi'_t(y, h) = \pi_{t+1}(y, hax), \quad y \in \mathcal{X}, h \in (\mathcal{A} \times \mathcal{X})^t$$

for all $t \geq 0$. We shall write π^x for $S_{(x,\pi_0(x))}\pi$ and call π^x the derived policy.

For finite-horizon policies $S_{(x,a)}$ is defined in the same way just now $S_{(x,a)} : \mathcal{P}_t \to \mathcal{P}_{t-1}, t \geq 1$.

The above definition means that $\pi^x \in \mathcal{P}_{t-1}$ holds for any $\pi \in \mathcal{P}_t$ and $x \in \mathcal{X}$. The following proposition follows from the definitions and thus we omit its technical proof.

PROPOSITION 1.1.3 *$\pi^{t,x} = \pi^{x,t-1}$ and thus if $\pi \in \mathcal{P}_t$ then $\pi^{t,x} = \pi^{x,t-1} \in \mathcal{P}_{t-1}$, $t \geq 1$.*

Now we are in the position to give the definition of the evaluation of policies with finite horizon.

DEFINITION 1.1.4 *If* $\pi \in \mathcal{P}_0$, *i.e.*, $\pi = (\pi_0)$ *then* $v_\pi(x) = (\mathcal{Q}\ell)(x, \pi_0(x))$, *where* $\ell \in \mathcal{R}(\mathcal{X})$ *is the terminal cost function. Assume that the evaluation of policies in* $\mathcal{P}_t$ *is already defined. Let* $\pi \in \mathcal{P}_{t+1}$. *Then*

$$v_\pi(x) = (\mathcal{Q}v_{\pi^x})(x, \pi_0(x)). \tag{1.1}$$

Since $\pi^x \in \mathcal{P}_t$, v_{π^x} is already defined and thus (1.1) is well defined. One can interpret this definition as follows: π^x is the policy that is applied after the first decision. The cost of this the derived policy is v_{π^x}. This cost together with the cost of the first decision (the first decision is $\pi_0(x)$ in state x) gives the total cost of the policy.

EXAMPLE 1.1.5 If π is a t-horizon policy in an MDP $(\mathcal{X}, \mathcal{A}, p, c)$ (cf. Example 0.1.3) and we set

$$\hat{v}_\pi^{(t)}(x) = E\left[\sum_{n=0}^{t} \gamma^n c(\xi_n, \alpha_n, \xi_n) \,|\, \xi_0 = x\right],$$

as usual, where $\alpha_n = \pi_n(\xi_n, \alpha_{n-1}\xi_{n-1} \ldots \alpha_0\xi_0)$ and $P(\xi_{n+1} = y \,|\, \alpha_n\xi_n \ldots \alpha_0\xi_0) = p(\xi_n, \alpha_n, y)$, $P(\xi_0 = x) > 0$, $x \in \mathcal{X}$ (i.e., (ξ_n, α_n) is the controlled object corresponding to π but when ξ_0 is random), then one sees easily that $v_\pi^{(t)} = \hat{v}_\pi^{(t)}$, where $v_\pi^{(t)}$ is the evaluation of π in the sense of Definition 1.1.4 in the ADP $(\mathcal{X}, \mathcal{A}, \mathcal{Q}, \ell)$ with

$$(\mathcal{Q}f)(x, a) = \sum_{y \in \mathcal{X}} p(x, a, y)\Big(c(x, a, y) + \gamma f(y)\Big),$$

and $\ell(x) = 0$ for all $x \in \mathcal{X}$.

The evaluation of infinite horizon policies can be defined as the limit of the evaluations of the finite horizon truncations of the policy:

DEFINITION 1.1.6 *Let* $\pi \in \mathcal{P} = \mathcal{P}_\infty$. *Then the total cost of policy for initial state* x *is given by*

$$v_\pi(x) = \liminf_{t \to \infty} v_{\pi^t}(x), \qquad x \in \mathcal{X}.$$

This definition takes an optimistic point of view since it involves the lim inf of the costs.

EXAMPLE 1.1.7 Continuing the above example, if π is an arbitrary policy then (by boundedness)

$$\begin{aligned}
\hat{v}_\pi(x) &\stackrel{\text{def}}{=} E\left[\sum_{n=0}^{\infty} \gamma^n c(\xi_n, \alpha_n, \xi_n) \,|\, \xi_0 = x\right] \\
&= \lim_{t \to \infty} E\left[\sum_{n=0}^{t} \gamma^n c(\xi_n, \alpha_n, \xi_n) \,|\, \xi_0 = x\right],
\end{aligned}$$

and so $v_\pi = \hat{v}_\pi$.

DEFINITION 1.1.8 *$\mathcal{Q}$ is said to be monotone if $\mathcal{Q}v \leq \mathcal{Q}u$ whenever $u \leq v$.*

In what follows we will always assume that $\mathcal{Q}$ is monotone.

DEFINITION 1.1.9 *Operator $\mathcal{Q}$ is called* lower semi-continuous *(LSC) on the set $D(\mathcal{X}) \subseteq \mathcal{R}(\mathcal{X})$ if for every (pointwise) convergent sequence of functions $v_t \in D(\mathcal{X})$ for which $v_t \leq \lim_{t\to\infty} v_t$,*

$$\lim_{t\to\infty} Qv_t = Q(\lim_{t\to\infty} v_t)$$

holds. Similarly $\mathcal{Q}$ is called upper semi-continuous *on the set $D(\mathcal{X})$ if for every (pointwise) convergent sequence of functions $v_t \in D(\mathcal{X})$ for which $v_t \geq \lim_{t\to\infty} v_t$ we have*

$$\lim_{t\to\infty} Qv_t = Q(\lim_{t\to\infty} v_t)$$

If $\mathcal{Q}$ is increasing (decreasing) then the domain $D(\mathcal{X})$ over which the desired property is required is $D(\mathcal{X}) = \{v \in \mathcal{R}(\mathcal{X}) \,|\, \ell \leq v\}$ ($D(\mathcal{X}) = \{v \in \mathcal{R}(\mathcal{X}) \,|\, \ell \geq v\}$). In such cases the domain will not be mentioned.

In his seminal paper Bertsekas investigated operators which are LSC only for *increasing* sequences of functions [6]. Trivially, property LSC implies this property. It is not very hard to show that the reverse implication holds as well, so these too concepts are in fact equivalent. The notation of upper semi-continuity (USC) and the corresponding notion of USC on decreasing function sequences (USCD) are again equivalent.

In harmony with [20] the equation of the next theorem will be called the fundamental equation (FE). Indeed we will find that this equation plays a central role in the development of the theory.

THEOREM 1.1.10 *Assume that at least one of the following conditions hold:*

a) *$\mathcal{Q}$ is LSC and I;*

b) *$\mathcal{Q}$ is USC and D;*

c) *$\mathcal{Q}$ is continuous.*

Then

$$v_\pi(x) = (\mathcal{Q}v_{\pi^x})(x, \pi_0(x)). \tag{1.2}$$

Proof. We prove the equation under Condition *a*. The other cases may be treated similarly. First, we need the following proposition:

PROPOSITION 1.1.11 *Let π be an arbitrary policy. If the ADP is increasing then v_{π^t} is increasing and $\ell \leq v^{*g}$.*

Proof. Let us consider $v_{\pi^{t+1}}$:

$$v_{\pi^{t+1}}(x) = (\mathcal{Q}v_{\pi^{t+1,x}})(x, \pi_0(x)) = (\mathcal{Q}v_{\pi^{x,t}})(x, \pi_0(x)) \tag{1.3}$$

Notice that $v_{\mu^0} \geq \ell$, where μ denotes an arbitrary policy. By definition we have that $v_{\pi^1}(x) = (\mathcal{Q}v_{\pi^{x,0}})(x, \pi_0(x))$. Now, since π^x is just another policy we have that $v_{\pi^{x,0}} \geq \ell$. From the monotonicity of $\mathcal{Q}$ we have that $\mathcal{Q}v_{\pi^{x,0}} \geq \mathcal{Q}\ell$ and consequently that $v_{\pi^1} \geq v_{\pi^0}$.

Now, let us assume that we have already proved up to t that *for all policies* μ there holds an inequality such that $v_{\mu^t} \geq v_{\mu^{t-1}}$. Let π be an arbitrary policy and let us apply the induction hypothesis on $\mu = \pi^x$. This leads to $v_{\pi^{x,t}} \geq v_{\pi^{x,t-1}}$. Again, applying $\mathcal{Q}$ on both sides and using Equation (1.3) we obtain the desired inequality. □

Now let $v_t = v_{\pi^t}$ and let $\mu = \pi^{t+1}$. By definition $v_\mu(x) = (\mathcal{Q}v_{\mu^x})(x, \mu_0(x))$. According to Proposition 1.1.3 $\mu^x = \pi^{t+1,x} = \pi^{x,t}$ and $\mu_0 = \pi_0$ thus

$$v_{\pi^{t+1}}(x) = (\mathcal{Q}v_{\pi^{x,t}})(x, \pi_0(x)). \tag{1.4}$$

Now, let t tend to infinity and consider the lim inf of both sides of the above equation:

$$\begin{aligned}
v_\pi(x) &= \liminf_{t\to\infty} v_{\pi^{t+1}}(x) \\
&= \liminf_{t\to\infty} (\mathcal{Q}v_{\pi^{x,t}})(x, \pi_0(x)) \\
&= (\mathcal{Q}[\lim_{t\to\infty} v_{\pi^{x,t}}])(x, \pi_0(x)) \\
&= (\mathcal{Q}v_{\pi^x})(x, \pi_0(x)),
\end{aligned}$$

where in the first equation the definition $v_{\pi^x} = \liminf_{t\to\infty} v_{\pi^{x,t}}$ and in the second equation Equation (1.4) was exploited, while in the third equation we used that $v_{\pi^{x,t}}$ is an increasing sequence, which was shown above, and that $\mathcal{Q}$ is LSC, and in the last equation $v_{\pi^x} = \lim_{t\to\infty} v_{\pi^{x,t}}$ was utilized which holds since $v_{\pi^{x,t}}$ is an increasing sequence. □

Corollary 1.1.12 *Under the conditions of the above theorem v_{π^t} converges to v_π, i.e., in Definition 1.1.6* lim inf *can be replaced by* lim.

Corollary 1.1.13 *Under the conditions of Theorem 1.1.10 the evaluation function of Markovian policies in the sense of Definition 1.1.6 coincide with their evaluation functions in the sense of Definition 0.1.7. Moreover, if π is a stationary policy and if $\mathcal{Q}$ is I (D) and LSC (USC) then $T_\pi^n \ell$ is increasing (decreasing) and $v_\pi = T_\pi v_\pi$.*

Proof. In Definition 0.1.7 the evaluation of a Markovian policy

$$\pi = (\pi_0, \pi_1, \dots, \pi_t, \dots)$$

was defined as the limit

$$\begin{aligned} \hat{v}_\pi(x) &= \lim_{t\to\infty} \Big(T_{\pi_0} \dots (T_{\pi_{t-1}}(T_{\pi_t}\ell)) \dots \Big) \\ &= \lim_{t\to\infty} T_{\pi_0} \dots T_{\pi_{t-1}} T_{\pi_t}\ell. \end{aligned}$$

However, it is easy to see that $T_{\pi_0} \dots T_{\pi_{t-1}} T_{\pi_t}\ell = v_{\pi^t}$, so by Corollary 1.1.12 the definition of Bertsekas coincides with ours for the case of Markovian policies. The second part comes from the convergence properties of monotone sequences. □

1.2 Uniformly Optimal Policies

DEFINITION 1.2.1 *The* optimal cost-to-go function *is defined by*

$$v^{*g}(x) = \inf_{\pi\in\mathcal{P}} v_\pi(x).$$

A policy is said to be optimal *if* $v_\pi = v^{*g}$.

We will now investigate the properties of v^{*g} and its connection with optimal policies. Firstly, we relax the notion of optimal policies.

DEFINITION 1.2.2 *Policy π is said to be uniformly ε-optimal if*

$$v_\pi(x) \le \begin{cases} v^{*g}(x) + \varepsilon, & \textit{if } v^{*g}(x) > -\infty; \\ -1/\varepsilon, & \textit{otherwise} \end{cases}$$

holds for every $x \in \mathcal{X}$.

THEOREM 1.2.3 *If the FE is satisfied then for all $\varepsilon > 0$ there exists an ε-optimal policy.*

Proof. Pick up an arbitrary $x \in \mathcal{X}$. By the definition of $v^{*g}(x)$ there exists a policy ${}_x\pi$ for which $v_{{}_x\pi}(x) \le v^{*g}(x) + \varepsilon$ when $v^{*g}(x) > -\infty$ and $v_{{}_x\pi}(x) \le -1/\varepsilon$, otherwise. We define a policy which will be ε-optimal by taking the actions prescribed by ${}_x\pi$ when x is the starting state of the decision process. The resulting policy is called the *combination* of the policies ${}_x\pi$. Formally, $\pi_0(x) = {}_x\pi_0(x)$ and

$$\pi_t(x, ha_0x_0) = {}_{x_0}\pi_t(x, ha_0x_0).$$

We claim that $v_\pi(x) = v_{{}_x\pi}(x)$ and thus π is uniformly ε-optimal. Indeed, $\pi^x = ({}_x\pi)^x$ and $\pi_0(x) = {}_x\pi_0(x)$ and so

$$\begin{aligned} v_\pi(x) &= (\mathcal{Q}v_{{}_x\pi^x})(x, \pi_0(x)) \\ &= (\mathcal{Q}v_{{}_x\pi^x})(x, {}_x\pi_0(x)) \\ &= v_{{}_x\pi}(x). \end{aligned}$$

□

1.3 Finite Horizon Problems

DEFINITION 1.3.1 *The optimal cost-to-go function for n-horizon problems is defined by*

$$v_n^{*\Delta}(x) = \inf_{\pi \in \mathcal{P}_n^\Delta} v_\pi,$$

where

$$\mathcal{P}_n^\Delta = \{\, \pi^n \mid \pi \in \Pi_\Delta \,\},$$

$\Delta \in \{g, m, s\}$.

It is clear that $v_n^{*\Delta}(x) = \inf_{\pi \in \Pi_\Delta} v_{\pi^n}$ since since $\mathcal{P}_n^\Delta = \{\, \pi^n \mid \pi \in \Pi_\Delta \,\}$. Further, if $\mathcal{Q}$ is increasing (decreasing) then $v_n^{*\Delta} \le v^{*\Delta}$ ($v_n^{*\Delta} \ge v^{*\Delta}$) and $\{v_n^{*\Delta}\}_n$ is an increasing (decreasing) sequence since for all policy π, $v_{\pi^n} \le v_{\pi^{n+1}} \le v_\pi$ ($v_{\pi^n} \ge v_{\pi^{n+1}} \ge v_\pi$). Moreover, $v_n^{*g} \le v_n^{*m} \le v_n^{*s}$ and $v^{*g} \le v^{*m} \le v^{*s}$.

DEFINITION 1.3.2 *The* optimal cost-to-go operator $T : \mathcal{R}(\mathcal{X}) \to \mathcal{R}(\mathcal{X})$ *associated with the ADP* $(\mathcal{X}, \mathcal{A}, \mathcal{Q}, \ell)$ *is defined by*

$$(Tf)(x) = \inf_{a \in \mathcal{A}(x)} (\mathcal{Q}f)(x, a).$$

THEOREM 1.3.3 (OPTIMALITY EQUATION FOR FINITE HORIZON PROBLEMS) *The optimal cost-to-go functions of the n-stage problem satisfies*

$$v_n^{*g} = v_n^{*m} = T^n \ell \tag{1.5}$$

provided that $\mathcal{Q}$ *is USC and the FE is satisfied.*

Proof. We prove the proposition by induction. One immediately sees that the proposition holds for $n = 1$. Assume that we have already proved the proposition for n. Firstly, we prove that $T^{n+1}\ell \le v_{n+1}^*$. Note that this inequality will follow from the FE and the monotonicity of $\mathcal{Q}$ alone: no continuity assumption is needed here.

Let $\pi \in \mathcal{P}_{n+1}$. We show that $T^{n+1}\ell \le v_\pi$. By the induction hypothesis $(T^{n+1}\ell)(x) = (Tv_n^{*g})(x)$. According to the FE $v_\pi(x) = (\mathcal{Q}v_{\pi^x})(x, \pi_0(x))$. Since $\pi^x \in \mathcal{P}_n$ so $v_{\pi^x} \ge v_n^{*g}$. Since $\mathcal{Q}$ is monotone it follows that

$$\begin{aligned}(Tv_n^{*g})(x) = \inf_{a \in \mathcal{A}(x)} (\mathcal{Q}v_n^{*g})(x, a) &\le \inf_{a \in \mathcal{A}(x)} (\mathcal{Q}v_{\pi^x})(x, a) \\ &\le (\mathcal{Q}v_{\pi^x})(x, \pi_0(x)) = v_\pi(x).\end{aligned}$$

This equation holds for arbitrary $\pi \in \mathcal{P}_{n+1}$ and thus $Tv_n^{*g} \le v_{n+1}^{*g}$. Using the induction hypothesis we find that $T^{n+1}\ell \le v_{n+1}^{*g}$.

Now let us prove the reverse inequality, i.e., that $v_{n+1}^{*g} \le T^{n+1}\ell$ holds. Let us choose a sequence of Markovian policies $\pi_k \in \mathcal{P}_n$ such that v_{π_k} converges to v_n^{*m}. Clearly, $v_{\pi_k} \ge v_n^{*m}$. Now let $\mu_j : \mathcal{X} \to \mathcal{A}$ be a sequence of mappings satisfying

$\lim_{j\to\infty} T_{\mu_j} v_n^{*g} = T v_n^{*g}$. Now consider the policies $\nu_{k,j} = \pi_k \bigoplus \mu_j \in \mathcal{P}_{n+1}$: the first n actions of $\nu_{k,j}$ are the actions prescribed by π_k while the last action is the action prescribed by μ_j. It is clear that $v_{n+1}^{*g} \le v_{n+1}^{*m} \le v_{\nu_{k,j}} = T_{\mu_j} v_{\pi_k}$: the last equality follows from the FE. Taking the limit in k we get that

$$v_{n+1}^{*m} \le \lim_{k\to\infty} T_{\mu_j} v_{\pi_k} = T_{\mu_j}(\lim_{k\to\infty} v_{\pi_k}) = T_{\mu_j} v_n^{*m}$$

holds owing to the choice of the policies π_k and since $\mathcal{Q}$ is USC. Now taking the limit in j the induction hypothesis yields that $v_{n+1}^{*g} \le v_{n+1}^{*m} \le T v_n^{*m} = T^{n+1}\ell$ which finally gives that $v_{n+1}^{*g} = v_{n+1}^{*m} = T^{n+1}\ell$, completing the proof. □

The following examples (taken from [6]) show that the conditions of the previous theorem are indeed essential for the theorem to hold.

Example 1.3.4 Let $\mathcal{X} = \{0\}$ and $\mathcal{A} = (0,1]$, $\ell(0) = 0$, and

$$(\mathcal{Q}f) = \begin{cases} 1, & \text{if } f(0) > 0; \\ a, & \text{otherwise.} \end{cases}$$

Note that $\mathcal{Q}$ is I, LSC, T is LSC but $\mathcal{Q}$ is not USC. It is easy to see that $0 = v_\infty(0) = (T^n\ell)(0) < v_n^{*g}(0) = 1 = v^{*g}(0)$ if $n \ge 2$.

Example 1.3.5 Let $\mathcal{X} = \{0\}$, $\mathcal{A} = (-1, 0]$, $\ell(0) = 0$ and

$$(\mathcal{Q}f) = \begin{cases} a, & \text{if } f(0) > -1; \\ a + f(0), & \text{otherwise.} \end{cases}$$

In this example $\mathcal{Q}$ is D but not USC. $v_n^{*g}(0) = -1$ but $(T^n\ell)(0) = -n$.

1.4 The Optimality Equation and the Convergence of Value iteration

Let us consider the sequence of optimal n-horizon cost-to-go functions, $\{v_n^{*g}\}$. If v_n^{*g} converges to v^{*g} then v^{*g} can be computed as the limit of the function sequence $v_0 = \ell$, $v_{t+1} = T v_t$ provided that $\mathcal{Q}$ is USC and the FE holds. The convergence of v_n^{*g} to v^{*g} expressed in another way means that the inf and lim operations can be interchanged in the definition of v^{*g}:

$$v^{*g} = \inf_{\pi\in\mathcal{P}} \lim_{n\to\infty} v_{\pi^n} = \lim_{n\to\infty} \inf_{\pi\in\mathcal{P}} v_{\pi^n} = \lim_{n\to\infty} v_n^{*g}. \tag{1.6}$$

Definition 1.4.1 *Let $v^\infty \in \mathcal{R}(\mathcal{X})$ denote the function obtained as the limit*

$$v_\infty = \lim_{n\to\infty} T^n \ell.$$

If $\mathcal{Q}$ is increasing and LSC (D and USC) then since $T^n\ell$ is an increasing (decreasing) sequence $T^n\ell$ is convergent, v^∞ is well defined and if T is LSC then $Tv^\infty = v^\infty$.

THEOREM 1.4.2 *The following statements hold:*

1.
$$\limsup_{n\to\infty} v_n^{*g} \le v^{*g}. \tag{1.7}$$

2. *If $\mathcal{Q}$ is D and USC then* $\lim_{n\to\infty} v_n^{*g} = v^{*g} = v^{*m}$ *and*
$$Tv^{*g} = v^{*g}. \tag{1.8}$$

3. *Assume that $\mathcal{Q}$ is I, LSC, T is LSC and there exists a mapping $\pi : \mathcal{X} \to \mathcal{A}$ such that $T_\pi v_\infty = Tv_\infty$. Then $Tv^{*g} = v^{*g}$, $v_\infty = v^{*g} = v^{*s}$ and $\lim_{n\to\infty} v_n^{*g} = v^{*g}$.*

Proof. First, let us prove (1.7). For this choose an arbitrary $x \in \mathcal{X}$ and a number c s.t. $c > v^{*g}(x)$. By the definition of v^{*g} there exists a policy $\pi \in \mathcal{P}$ such that $v_\pi(x) < c$. Furthermore, since $v_\pi(x) = \lim_{n\to\infty} v_{\pi^n}(x)$ there exists a number n_0 such that from $n > n_0$ it follows that $v_{\pi^n}(x) < c$. Thus, if $n > n_0$ then there holds that $v_n^{*g}(x) < c$ and consequently $\limsup_{n\to\infty} v_n^{*g}(x) < c$. Since c and x were arbitrary, we obtain the desired inequality.

Now we show that if $\mathcal{Q}$ is D then $\lim_{n\to\infty} v_n^{*P} = v^{*P}$ where $P \subseteq \mathcal{P}$ is arbitrary. Observe that by Proposition 1.1.11 $v_{\pi^{n+1}} \le v_{\pi^n}$ and thus $v^{*P} \le v_{n+1}^{*P} \le v_n^{*P}$ for all t. Let $\pi \in P$ be arbitrary. Then also $v_n^{*P} \le v_{\pi^n}$. Letting n tend to infinity and combining the result with Inequality 1.7 yields $v^{*P} \le \lim_{n\to\infty} v_n^{*P} \le v_\pi$, and hence
$$v^{*P} = \lim_{n\to\infty} v_n^{*P}. \tag{1.9}$$

Now, we prove that $v^{*g} = Tv^{*g}$. Note that by Theorem 1.1.10 the FE holds so the Finite Horizon Optimality Equations hold (cf. Theorem 1.3.3). By (1.9) $v_n^* \ge v_{n+1}^* \to v^*, n \to \infty$ and $\mathcal{Q}$ is USC so $Tv_n^* \to Tv^*, n \to \infty$. But $Tv_n^{*g} = v_{n+1}^*$ by Theorem 1.3.3 and by Equation (1.9) $v_n^{*g} \to v^{*g}$, so $Tv_n^{*g} \to v^{*g}$ and thus $Tv^{*g} = v^{*g}$. Finally, since by Theorem 1.3.3 $v_n^{*g} = v_n^{*m}$ and for all policy π, $v_\pi \le v_{\pi^n}$ (since $\mathcal{Q}$ is D), it follows that $v^{*g} = v^{*m}$.

Now, let us prove the third part. Since $T^n\ell$ is increasing and converges to v_∞ and $\mathcal{Q}$ is LSC we have that $Tv_\infty = T(\lim_{n\to\infty} T^n\ell) = \lim_{n\to\infty} TT^n\ell = v_\infty$. Since $v_\infty \le v^{*g} \le v^{*s}$ it is sufficient to prove that $v_\infty = v^{*s}$. Let us consider the policy π whose existence is stated in the condition of the proposition: $T_\pi v_\infty = Tv_\infty$. Since $\ell \le v_\infty \le v_\pi$ thus also $T_\pi\ell \le T_\pi v_\infty \le T_\pi v_\pi$. Exploiting the fact that $Tv_\infty = v_\infty$ yields $T_\pi v_\infty = v_\infty$ and also by the LSC of $\mathcal{Q}$ $T_\pi v_\pi = v_\pi$, so $T_\pi \ell \le v_\infty \le v_\pi$. Repeating this argument one gets $T_\pi^n \ell \le v_\infty \le v_\pi$ and letting $n \to \infty$ yields that $v_\pi = v_\infty$, meaning that $v^{*s} \le v_\infty$ and $v^{*s} = v_\infty$. Since $T^n\ell \le v_n^{*g} \le v^{*g}$ we also have that $v_n^{*g} \to v^{*g}$, $n \to \infty$. □

Equation (1.8) is called the Bellman Optimality Equation and plays a fundamental role when solving sequential decision problems. Example 0.1.14 shows that T can be non-LSC even if $\mathcal{Q}$ is increasing and continuous, in which case (1.6) may not hold.

The next example shows that $v^{*g} = v_\infty$ still does not necessarily hold even if *both* $\mathcal{Q}$ and T are I and LSC.

EXAMPLE 1.4.3 [6] Let $\mathcal{X} = \{0, 1\}$, $\mathcal{A} = (-1, 0]$, $\ell \equiv -1$ and

$$(\mathcal{Q}f)(x,a) = \begin{cases} a, & \text{if } f(1) \le -1 \text{ or } x = 1; \\ 0, & \text{otherwise.} \end{cases}$$

Now $\mathcal{Q}$ is I and LSC as is T, but $v_\infty(0) = -1 < 0 = v^{*g}(0)$.

1.5 Existence of Optimal Stationary Policies

DEFINITION 1.5.1 *A stationary policy ϕ is said to be* greedy (myopic) *w.r.t. v if*

$$T_\phi v = Tv,$$

i.e., if for each $x \in \mathcal{X}$ $(\mathcal{Q}v)(x, \phi(x)) = (Tv)(x)$.

In ADPs "greediness" w.r.t. v^{*g} and optimality are intimately related as shown by the next theorem:

THEOREM 1.5.2 *If the FE holds and the stationary policy ϕ is optimal then*

$$T_\phi v^{*g} = v^{*g}. \tag{1.10}$$

If the Bellman Optimality Equation $Tv^{*g} = v^{*g}$ holds then the following statements hold, as well:

1. *If the FE holds then optimal stationary policies are greedy w.r.t. v^{*g};*
2. *If $\mathcal{Q}$ is I (D) and LSC (USC) then if there exists an optimal policy then there is one which is stationary;*
3. *If $\mathcal{Q}$ is I and LSC then ϕ is greedy w.r.t. v^{*g} iff ϕ is optimal.*

Proof. Equation (1.10) follows immediately from the FE (Equation (1.2)) and the equations $v_\phi = v^{*g}$ and $\phi^x = \phi$.

Now, assume that the Bellman Optimality Equation holds. Then immediately $T_\phi v^{*g} = v^{*g} = Tv^{*g}$, showing part 1. Now, let us turn to the proof of part 2. The proof is presented only under the condition that $\mathcal{Q}$ is I and LSC, the proof of the other case follows analogous lines. Let π be an optimal policy with $\pi = (\pi_0, \pi_1, \ldots, \pi_t, \ldots)$. For any selector μ define $P_\mu : \mathcal{R}(\mathcal{X} \times \mathcal{A}) \to \mathcal{R}(\mathcal{X})$ as $(P_\mu v)(x) = v(x, \mu(x))$. It is immediate that P_μ is M. By Theorem 1.1.10 the FE

$$v_\pi(x) = (\mathcal{Q}v_{\pi^x})(x, \pi_0(x))$$

holds, and now it can be written in the form $v_\pi(x) = (P_{\pi_0}\mathcal{Q}v_{\pi^x})(x)$ using the operator just introduced. Now, since $v_{\pi^x} \geq v^{*g}$ and $\mathcal{Q}$ and P_{π_0} are M, we have by the optimality of π that

$$v^{*g}(x) = v_\pi(x) = (P_{\pi_0}\mathcal{Q}v_{\pi^x})(x) \geq (P_{\pi_0}\mathcal{Q}v^{*g})(x) = (T_{\pi_0}v^{*g})(x).$$

Since $T_{\pi_0} \geq T$, so $(T_{\pi_0}v^{*g})(x) \geq (Tv^{*g})(x) = v^{*g}(x)$ and thus

$$v^{*g} \geq T_{\pi_0}v^{*g} \geq v^{*g}.$$

By induction we get that $T^n_{\pi_0}v^{*g} = v^{*g}$ holds for all $n = 1, 2, \ldots$. On the other hand, if $\mathcal{Q}$ is I and LSC then Proposition 1.1.11 yields $\ell \leq v^{*g} \leq v_{\pi_0}$ and thus, by Corollary 1.1.13, $v_{\pi_0} \leftarrow T^n_{\pi_0}\ell \leq T^n_{\pi_0}v^{*g} = v^{*g}$, $n \to \infty$, showing that $v_{\pi_0} = v^{*g}$.

The third part follows easily: Since we know that $Tv^{*g} = v^{*g}$ and by assumption $T_\phi v^{*g} = Tv^{*g}$, thus $T_\phi v^{*g} = v^{*g}$. Consequently for all $n = 1, 2, \ldots$ $T^n_\phi v^{*g} = v^{*g}$. Since $\ell \leq v^{*g} \leq v_\phi$ and since $\mathcal{Q}$ is I and LSC so $T^n_\phi v^{*g}$ converges to v_ϕ and therefore $v_\phi = v^{*g}$. □

Bertsekas proved a somewhat weaker statement, similar to the second part (see Prop. 7 of [6]), namely that if there exists a Markov policy which is Markov-optimal then there exists a stationary policy which is also Markov-optimal.

We have seen that in continuous, increasing models the set of optimal stationary policies coincides with that of the greedy policies w.r.t. v^{*g}. However, we have not obtained any similar results for decreasing models. The following examples show that without additional requirements on $\mathcal{Q}$ we cannot expect to get any such result:

Example 1.5.3 (Idea based on [20]) Let $\mathcal{X} = \{0, 1\}$ and $\mathcal{A} = \{0, 1\}$. Let $\mathcal{Q}$ be defined as follows: $(\mathcal{Q}f)(0, 0) = f(0)$, $(\mathcal{Q}f)(0, 1) = -1 + f(1)$, and $(\mathcal{Q}f)(1, a) = f(1)$, $a \in \mathcal{A}$. Let $\ell \equiv 0$. Then $v^{*g}(0) = -1$ and $v^{*g}(1) = 0$. Clearly, the selector ϕ with $\phi(0) = 0$ is not optimal but $T_\phi v^{*g} = v^{*g}$. In this example $\mathcal{Q}$ is decreasing and continuous and thus by Part 2 of Theorem 1.4.2 also $Tv^{*g} = v^{*g}$.

The second question is whether there exists an optimal solution of Equation (1.10) at all. The following example shows that it is not necessarily the case [20]:

Example 1.5.4 Let $\mathcal{X} = \mathbb{N}$, $\mathcal{A} = \{0, 1\}$ and let $(\mathcal{Q}f)(0, a) = f(0)$, $a \in \mathcal{A}$ and for $x > 0$ let $(\mathcal{Q}f)(x, 0) = -(x-1)/x + f(0)$ and $(\mathcal{Q}f)(x, 1) = f(x+1)$. Let $\ell \equiv 0$. It is easy to see, that $v^{*g}(x) = -1$ if $x > 0$ and $v^{*g}(0) = 0$. The only stationary policy that satisfies Equation (1.10) prescribes action 1 for each non-zero state. However, the evaluation of this policy gives zero everywhere. Here again $\mathcal{Q}$ is decreasing and continuous.

We summarize the results for contraction models, which will be considered in the second part, in the following Corollary:

Corollary 1.5.5 *Assume that $\mathcal{Q}$ is a contraction. Then*

1. *the FE holds;*
2. $v_\pi = \lim_{t\to\infty} v_{\pi^t}$*;*
3. $v_n^{*g} = v_n^{*m} = T^n \ell$*,* $n > 0$*;*
4. $Tv^{*g} = v^{*g}$*;*
5. *Greedy policies w.r.t.* v^{*g} *are optimal and optimal stationary policies are greedy w.r.t.* v^{*g}*; and*
6. *If there exist an optimal policy then there exists one which is stationary.*

Proof. 1 follows from Theorem 1.1.10, 2 from Corollary 1.1.12, 3 from Theorem 1.3.3. We prove 4 in the following way: we know from Theorem 1.4.2/1 that $v_\infty \le v^{*g}$ and since T is a contraction by the definition of v_∞ we get that $v_\infty = Tv_\infty$. It is sufficient to prove that $v^{*g} \le Tv^{*g}$ since then iterating this inequality will yield that $v^{*g} \le v_\infty$. Let π_n be a sequence of $1/n$-uniformly optimal policies. Such policies exist by Theorem 1.2.3. Further, let μ_n be a selector such that $T_{\mu_n} v_{\pi_n} \le Tv_{\pi_n} + 1/n$. Then $v^{*g} \le v_{\mu_n \oplus \pi_n} \le (Tv_{\pi_n}) + 1/n$, and taking the limit of both sides yields the desired inequality. 5 follows since if ϕ is greedy w.r.t. v^{*g} then $T_\phi v^{*g} = Tv^{*g} = v^{*g}$ and if ϕ is an optimal stationary policy then $Tv^{*g} = v^{*g} = v_\phi = T_\phi v_\phi = T_\phi v^*$, showing the greediness of ϕ. Here we exploited that v_ϕ is the fixed point of T_ϕ which follows since $v_\phi = \lim_{n\to\infty} T_\phi^n \ell$ and since T_ϕ is a contraction. 6 follows similarly as Part 2 of Theorem 1.5.2. □

1.6 Discussion

We have defined the evaluation of arbitrary policies based on the notion of the cost propagation operator. The decision problems were investigated under the conditions that the cost propagation operator is increasing or decreasing. It was found that under the decreasing assumption, the upper semi-continuity of the cost propagation operator, $\mathcal{Q}$, was sufficient for the value iteration algorithm to converge to the optimal cost-to-go function, but greedy policies w.r.t. to the optimal cost-to-go function are not necessarily optimal. On the other hand, under the increasing assumption it was much harder to ensure the convergence of value iteration to the optimal cost-to-go function: we had to assume that $\mathcal{Q}$ and T are lower semi-continuous, and that there exists a stationary policy which is greedy w.r.t. the optimal cost-to-go function. However, optimal stationary policies are much easier to find in this case: if $\mathcal{Q}$ is continuous then optimal stationary policies coincide with policies greedy w.r.t. the optimal cost-to-go function. Therefore increasing models are more worthy of study since greediness can be used as the starting point for finding optimal policies. Further properties of these models are considered in the next chapter. One of the reasons for the difference between the increasing and decreasing models is that the corresponding natural concepts of semi-continuity (lower- and upper-semi-continuity in the case of increasing and decreasing models, respectively) carry over differently to the optimal evaluation operator T: in the case of increasing models there is no transfer while in the case of decreasing models there is. For completeness the basic results for contraction models were also derived. The main results of this chapter are published in [72]

To the author's best knowledge there has been no work in ADPs concerning general policies. Some recent related work has been done by Waldmann [83] who developed a highly general model of dynamic-programming problems, with a focus on deriving approximation bounds. Heger [28, 29] extended many of the standard MDP results to cover the risk-sensitive model. Although his work derives many of the important theorems, it does not present these theorems in a generalized way which allow them to be applied to any other models. Verdu and Poor [81] introduced a class of abstract dynamic-programming models that is far more comprehensive than the model discussed here. Their goal, however, was different from ours: they wanted to show that the celebrated "Principle of Optimality" discovered by Bellman relies on the fact that the order of selection of optimal actions and the computation of cumulated costs can be exchanged as desired: in addition to permitting non-additive operators and cost-to-go functions with values from any set (not just the real numbers), they showed how, in the context of finite-horizon models, a weaker "commutativity" condition is sufficient for the principle of optimality to hold. For infinite models they derived only basic results, concerning Markovian policies.[1]

[1] Here is an example of their statements translated into our framework: They first show that from their commutativity condition it follows that $T^n\ell = v_n^{*m}$, where v_n^{*m} is the n-step optimal cost-to-go function for Markovian policies, ℓ is the terminal cost function. Now the statement which concerns infinite horizons goes like this: if v_n^{*m} converges to v^{*m} (Condition 3 in [81]) then $T^n\ell$ converges to v^{*m}. The problem is that in practice it is usually clear that $v_n^{*m} = T^n\ell$, but it is much harder to show that v_n^{*m} converges to v^{*m} (cf. Theorems 1.3.3 and 1.4.2).

Chapter 2

Increasing Models

Throughout this chapter we will assume that $\mathcal{Q}$ is monotone (M), increasing (I) and lower semi-continuous (LSC). This is the minimal set of conditions under which increasing models are worthy of study: without monotonicity the principle of optimality may be violated [64, 48] and without lower semi-continuity even the evaluation of stationary policies may behave strangely.

After reviewing the basic definitions of ADPs in the next section, a classification of increasing ADPs (shown in Figure 2) is given in Section 2.2. Using the classification, the existence of optimal stationary policies can be reduced to more basic problems, such as when the fixed point equation $Tv_\infty = v_\infty$ and the existence of greedy policies w.r.t. v_∞ (i.e., $\Gamma v_\infty \neq \emptyset$) hold. Since previous authors assumed stronger conditions than our minimal set most of our results can be considered as being new. In particular, we can show that Howard's policy improvement routine is valid (Lemma 2.2.11), but may sometimes stop in local optima (Example 2.3.3).

The special properties of policy iteration and value iteration algorithms for finite models are given in Section 2.3. It is shown that policy iteration stops after a finite number of steps (Theorem 2.3.2), but an example is also presented which illustrates that it does not indispensably give the optimal policy (Example 2.3.3). It is proved that value iteration may be stopped after a finite number of steps; the greedy policy w.r.t. the most recent estimate of the optimal cost-to-go function will be optimal if the number of steps is large enough (Theorem 2.3.4).

Several connections are given to the work of other authors, and also connections to different models. We close this chapter with some concluding remarks.

2.1 Notation and Assumptions

As mentioned above, throughout this chapter we make the following assumptions:

Assumption 2.1.1 **(M)** Monotonicity: if $u, v \in \mathcal{R}(\mathcal{X})$ and $u \leq v$ then $\mathcal{Q}u \leq \mathcal{Q}v$.

ASSUMPTION 2.1.2 (**I**) Uniform Increase: $(\mathcal{Q}\ell)(\cdot, a) \geq \ell$ for all $a \in \mathcal{A}$.

ASSUMPTION 2.1.3 (**LSC**) Lower semi-continuity: if $v_n \in \mathcal{R}(\mathcal{X})$ is such that $\lim_{n\to\infty} v_n$ exists and $v_n \leq \lim_{n\to\infty} v_n$ then $\lim_{n\to\infty} \mathcal{Q}v_n = \mathcal{Q}(\lim_{n\to\infty} v_n)$.

The monotonicity assumption implies the monotonicity of T_π and T, where $\pi \in \mathcal{A}^{\mathcal{X}}$ if the $\mathcal{Q}$ operator is monotone. Further, from the definition of T we have that $T \leq T_\pi$ for each selector π. Note that if $\mathcal{Q}$ is LSC then T_π is LSC for each selector π, but as it was already noted and shown in as Example 0.1.14, T is not necessarily LSC even if $\mathcal{Q}$. This causes most of the difficulties with increasing models.

2.2 Relations in Increasing Models

The aim of this section is to prove the relations of Figure 2. The following lemma will be frequently used:

LEMMA 2.2.1 *Assume that $S : \mathcal{R}(\mathcal{X}) \to \mathcal{R}(\mathcal{X})$ is monotone. Let $V \in \mathcal{R}(\mathcal{X})$, $\ell \leq V$. If $SV \leq V$ then $\limsup_{n\to\infty} S^n\ell \leq SV$, so $\limsup_{n\to\infty} S^n\ell \leq V$ and if $SV < V$ then $\limsup_{n\to\infty} S^n\ell < V$*

Proof. This involves a simple induction on n. □

Note that if S is increasing ($S\ell \geq \ell$) then $\limsup_{n\to\infty} S^n\ell = \lim_{n\to\infty} S^n\ell$ holds, as well.

THEOREM 2.2.2 (**P1**) *Assume M, I, LSC and let $v_\infty = \lim_{n\to\infty} T^n\ell$. Then*

$$v_\infty \leq Tv_\infty \leq Tv^{*s} \leq v^{*s}$$

and so if $v_\infty = v^{*s}$ then

$$v_\infty = Tv_\infty = Tv^{*s} = v^{*s} \tag{2.1}$$

Proof. In order to prove this we need two lemmas which we state and prove now.

LEMMA 2.2.3 *Assume M,I. Then v_∞ is well defined and satisfies the following properties:*

a) $\ell \leq v_\infty \leq Tv_\infty$,

b) *if $\ell \leq v$ and $Tv \leq v$ then $v_\infty \leq v$.*

Proof. Since $v_n = T^n\ell$ is non-decreasing by assumptions M and I, v_∞ must be well defined. Since $v_n \leq v_\infty$ for all n and T is increasing then $v_{n+1} = Tv_n \leq Tv_\infty$. Letting $n \to \infty$ yields a). Part b) follows from Lemma 2.2.1 with $S = T$ and $V = v$. □

LEMMA 2.2.4 *Assume M,I,LSC and let μ be an arbitrary selector. Then*

a) $v_\mu = T_\mu v_\mu$, $Tv_\mu \le v_\mu$ *and* $Tv^{*s} \le v^{*s}$;

b) for every $u \in B(\mathcal{X})$ *for which* $\ell \le u \le v_\mu$, $\lim_{n\to\infty} T_\mu^n u = v_\mu$.

Proof. First we prove that if $\mathcal{Q}$ is M and I then (1) $\ell \le v_\mu$, (2) if $v \ge \ell$ then from $T_\mu v \le v$ it follows that $v_\mu \le T_\mu v \le v$, and (3) $v_\mu \le T_\mu v_\mu$.

Indeed, (1) follows from M, (2) follows from Lemma 2.2.1 which can be applied due to (1) with the cast $S = T_\mu$ and $V = v$. (3) follows from (2) by choosing $v = v_\mu$. Now assume that $\mathcal{Q}$ is also LSC. Note that as a consequence T_μ is also LSC. By (3) in order to have $T_\mu v_\mu = v_\mu$ it is sufficient to prove that $T_\mu v_\mu \le v_\mu$. Since $T_\mu^n \ell \nearrow v_\mu$ by LSC of T_μ we have that $v_\mu = \lim_{n\to\infty} T_\mu^{n+1}\ell = T_\mu(\lim_{n\to\infty} T_\mu^n \ell) = T_\mu v_\mu$. Further, since $T \le T_\mu$ then $Tv_\mu \le T_\mu v_\mu = v_\mu$. Finally let μ be an arbitrary selector. Then $Tv^{*s} \le T_\mu v^{*s} \le T_\mu v_\mu = v_\mu$. Taking the infimum of both sides w.r.t. μ we get that $Tv^{*s} \le v^{*s}$ which proves Part a).

Now, let us prove Part b). Since T_μ is increasing and monotone

$$T_\mu^n \ell \le T_\mu^n u \le T_\mu^n v_\mu$$

holds for every natural number n. On the other hand from Part a) we have $T_\mu^n v_\mu = v_\mu$, and finally, by Corollary 1.1.13, $\lim_{n\to\infty} T_\mu^n \ell = v_\mu$. Thus, it must follow that $\lim_{n\to\infty} T_\mu^n u = v_\mu$. □

Continuing the proof of the theorem, we note that the inequality $Tv^{*s} \le v^{*s}$ follows from Lemma 2.2.3, Part a) $v_\infty \le Tv_\infty$ and Lemma 2.2.4, Part a). So it remains to be proved that $v_\infty \le v^{*s}$ so that $Tv_\infty \le Tv^{*s}$ holds because of M. However, this follows immediately from Lemma 2.2.3, Part b) applied for $v = v^{*s}$. □

Note that the corollary (2.1) of $v^{*s} = v_\infty$ is slightly stronger than what was proved in Theorem 1.4.2, Part 3 where the existence of a greedy (stationary) policy w.r.t. v_∞ was also needed. The following example, analogous to that of [20], shows that the converse of this does not hold, i.e., $v_\infty < v^{*s}$ may hold even when $v_\infty = Tv_\infty$ and $Tv^{*s} = v^{*s}$ both hold.

EXAMPLE 2.2.5 In this example $\mathcal{Q}$ is monotone, increasing, Lipschitzian[1] (and thus continuous) and T is also Lipschitzian.
The following relations hold: $Tv_\infty = v_\infty = v^{*g} = v^{*m}$, but $Tv^{*s} = v^{*s} > v_\infty$. $\Gamma v_\infty = \emptyset$, $\Gamma v^{*s} = \emptyset$, $\Pi_s(v^{*g} + \varepsilon) = \emptyset$ if $0 < \varepsilon < 1$, but $\Pi_m(v^{*g} + \varepsilon) \ne \emptyset$.
The model is as follows: $\mathcal{X} = \{0\}$, $\mathcal{A} = \mathbf{Z}^+$. $(\mathcal{Q}f)(0,a) = 1/2^a + (1 - 1/2^a)f(0)$, $\ell \equiv 0$. It is straightforward to see that

$$(Tf)(0) = \begin{cases} f(0), & \text{if } f(0) < 1; \\ 1, & \text{otherwise,} \end{cases}$$

[1] Let B_1 and B_2 be normed vector spaces. Operator $S : B_1 \to B_2$ is called Lipschitzian with index $0 < \alpha$ if $\|Sf - Sg\| \le \alpha \|f - g\|$ holds for all $f, g \in B_1$.

so T is Lipschitz, too.
It is readily seen that $v^{*s}(0) = 1$ (by consideration of the fixed point equation $T_\pi v_\pi = v_\pi$). However, if $\pi_m = (\pi_{m0}, \pi_{m1}, \ldots, \pi_{mt}, \ldots)$ with $\pi_{mt}(0) = m + t$ then $v_{\pi_m}(0) = 1 - \prod_{t=1}^{\infty}(1 - 1/2^{m+t}) \approx 1 - \exp(-1/2^m)$ which goes to zero, as m tends to infinity, showing that $v^{*g}(0) = 0$.

Now we will show that the crucial point of this example was that $\Gamma v_\infty = \emptyset$. First we need some more definitions:

DEFINITION 2.2.6 *Let $\Pi^*_\Delta(v)$ ($\Delta \in \{a, m, s\}$) denote the policies with infinite horizon evaluation exactly equal to v:*

$$\Pi^*_\Delta(v) = \{\pi \in \Pi_\Delta \mid v_\pi = v\}.$$

(The policy sets, Π_Δ, are introduced in Definition 0.1.10.)

Recall that $\Pi_\Delta(v)$ denotes the set of policies with infinite horizon evaluation not greater than v. Of course, $\Pi^*_\Delta(v) \subseteq \Pi_\Delta(v)$ and $\Pi^*_\Delta(v^{*\Delta}) = \Pi_\Delta(v^{*\Delta})$.

THEOREM 2.2.7 *Assume M,I,LSC. Then*

P2 *If $Tv_\infty = v_\infty$ and $\Gamma v_\infty \neq \emptyset$ then $v_\infty = v^{*s}$.*

P5 *If $Tv^{*s} = v^{*s}$ then $\Pi_s(v^{*s}) = \Gamma v^{*s}$.*

P4 *Assume that $\Gamma v^{*s} \cap \Pi_s(v^{*s}) \neq \emptyset$. Then $Tv^{*s} = v^{*s}$. In particular, if $\Gamma v^{*s} = \Pi_s(v^{*s}) \neq \emptyset$ then $Tv^{*s} = v^{*s}$.*

Proof. We need the following lemma:

LEMMA 2.2.8 *Assume M, I, LSC. Then*

a) If $\Gamma v \neq \emptyset$ and $\ell \leq v$ and $Tv \leq v$ then $v^{*s} \leq v$.

b) If $\ell \leq v \leq v^{*s}$ and $Tv = v$ then $\Pi^*_s(v) = \Gamma v$.

c) If $\Gamma v \cap \Pi^*_s(v) \neq \emptyset$ then $Tv = v$.

Proof. Part a) is another application of Lemma 2.2.1. Let $\mu \in \Gamma v$, $S = T_\mu$, $V = v$. Since $T_\mu v = Tv$ & $Tv \leq v$ it follows that $T_\mu v \leq v$ so the lemma assumptions are satisfied. The lemma and Lemma 2.2.4 Part a) yield that $v_\mu = \lim_{n\to\infty} T^n_\mu \ell \leq v$ and consequently that $v^{*s} \leq v$ holds as well. This proves Part a).

Now let us prove Part b). First of all we shall prove that $\Pi^*_s(v) \subseteq \Gamma v$. Let $\pi \in \Pi^*_s(v)$; then $T_\pi v = T_\pi v_\pi = v_\pi = v = Tv$ and thus $\pi \in \Gamma v$. Now assume that $\pi \in \Gamma v$. Then $T_\pi v = Tv = v$. Consequently $T^n_\pi v = v$ for all $n = 1, 2, \ldots$ and by Part b) of Lemma 2.2.4 we get that $v_\pi = v$.

Part c) follows easily since if $\pi \in \Gamma v \cap \Pi^*_s(v)$ then $v = v_\pi = T_\pi v_\pi = T_\pi v = Tv$.
□

Now, let us prove (P2). Part a) of the lemma with the choice $v = v_\infty$ yields the inequality $v^{*s} \le v_\infty$. Since we know from Theorem 2.2.2 that $v_\infty \le v^{*s}$, it follows that $v^{*s} = v_\infty$.

(P5) follows by Part b) of the lemma with the choice $v = v^{*s}$ since we know that $\ell \le v^{*g} \le v^*_m \le v^{*s}$, while (P4) is obtained by Lemma 2.2.8, Part c) with $v = v^{*s}$. □

Corollary 2.2.9 *If $Tv_\infty = v_\infty$ and $\Gamma v_\infty \ne \emptyset$ then $v_\infty = v^{*g} = v^*_m = v^{*s}$.*

Theorem 2.2.10 (**P6**) *Assume M,I, LSC and that $\Gamma v^{*s} \ne \emptyset$ holds. Then the set of best stationary policies and greedy policies w.r.t. v^{*s}, coincide, i.e.,*

$$\Pi_s(v^{*s}) = \Gamma v^{*s}.$$

Proof. Firstly, we prove that greedy policies w.r.t. the optimal evaluation of stationary policies are optimal, i.e.,

$$\Gamma v^{*s} \subseteq \Pi_s(v^{*s}).$$

For this let $\pi \in \Gamma v^{*s}$. We need to show that $v_\pi \le v^{*s}$. Since $T_\pi v^{*s} = Tv^{*s}$ and according to Theorem 2.2.2 $Tv^{*s} \le v^{*s}$, it follows that $T_\pi v^{*s} \le v^{*s}$. Now, applying Lemma 2.2.1 with $S = T_\pi$, $V = v^{*s}$ yields $v_\pi \le v^{*s}$ where we exploited the result of Lemma 2.2.4 Part b), namely that $T_\pi^n v^{*s} \to v_\pi$.

It remains to prove that $\Pi_s(v^{*s}) \subseteq \Gamma v^{*s}$. Let π denote an arbitrary best, stationary policy: $v_\pi = v^{*s}$. Now we would like prove that $Tv_\pi = v_\pi$. Firstly, we know that $Tv_\pi \le v_\pi$. Next, assume that $Tv_\pi < v_\pi$ and let $\hat\pi \in \Gamma v_\pi = \Gamma v^{*s}$. The next lemma, called the Generalized Howard's Policy Improvement Lemma, shows that $v_{\hat\pi}(x_0) < v_\pi(x_0)$ which contradicts the optimality of v_π. So we must have that $Tv_\pi = v_\pi = T_\pi v_\pi$ and $\pi \in \Gamma v_\pi = \Gamma v^{*s}$ thus finishing the proof. □

Lemma 2.2.11 (Generalized Howard's Policy Improvement) *Assume M,I, LSC and let π be an arbitrary selector and $\hat\pi$ and $\hat\pi \in \Gamma v_\pi$. Then $v_{\hat\pi} \le Tv_\pi \le v_\pi$ and if $Tv_\pi < v_\pi$ then $v_{\hat\pi} < v_\pi$.*

The same conclusions hold when $\mathcal{Q}$ is monotone and is a contraction.

Proof. Since $T \le T_\pi$, due to Lemma 2.2.4 Part a) we have $T_{\hat\pi} v_\pi = Tv_\pi \le T_\pi v_\pi = v_\pi$. Now, Lemma 2.2.1 with $S = T_{\hat\pi}$ and $V = v_\pi$ yields $v_{\hat\pi} \le T_{\hat\pi} v_\pi = Tv_\pi$. The second part follows analogously. The proof is identical for monotone, contraction models, by noting that $T_\pi v_\pi = v_\pi$ follows by Corollary 1.5.5. □

An interesting open question is whether from $\Pi_s(v^{*s}) \ne \emptyset$ it follows that $\Pi_s(v^{*s}) = \Gamma v^{*s}$. Unfortunately, we could not prove or disprove this.

2.3 Computational Procedures

If $\mathcal{A}$ is finite then $\Gamma v \neq \emptyset$, for all $v \in \mathcal{R}(\mathcal{X})$. As a consequence we obtain the following corollary:

COROLLARY 2.3.1 *Assume M,I,LSC. Then if $\mathcal{A}$ is finite then all the assertions in Figure 2 are valid. Moreover, $v_\infty = v^{*g} = v^{*m} = v^{*s}$.*

Proof. It is clear that $\Gamma v_\infty \neq \emptyset$. Since T is LSC then $Tv_\infty = v_\infty$ also, and hence all the assertions of Figure 2 are valid. □

In essence, there are two computational approaches that can be utilized to find an optimal policy:[2] policy iteration and value iteration (or successive approximation). First we will consider policy iteration (PI). For MDPs Puterman proved the analogue of the second part of the next theorem [50].

THEOREM 2.3.2 (POLICY ITERATION) *Assume M,I, LSC and that both the action space $\mathcal{A}$ and the state space $\mathcal{X}$ are finite. Let π_0 be an arbitrary policy and let us consider the sequence of policies $\{\pi_t\}$ defined by $\pi_{t+1} \in \Gamma v_{\pi_t}$. Then after a finite number of steps, say τ, the fixed point equation $v_{\pi_{t+1}} = v_{\pi_t}$ $(t \geq \tau)$ and the fixed point equation $v_{\pi_\tau} = Tv_{\pi_\tau}$ is satisfied. Further, $v_{\pi_t} \leq T^t v_{\pi_0}$, i.e., PI converges at least as fast as value iteration when value iteration is started with $v_0 = v_{\pi_0}$.*

Proof. By Lemma 2.2.11 we have that $v_{\pi_{t+1}} \leq v_{\pi_t}$ for all $t \geq 0$. Since there are only a finite number of policies there exists a time τ such that if $t \geq \tau$ then $v_{\pi_t} = v_{\pi_{t+1}}$. However, this means that $Tv_{\pi_\tau} = v_{\pi_\tau}$, otherwise by Lemma 2.2.11 we would have that $v_{\pi_{\tau+1}} < v_{\pi_\tau}$.

The proof of the bound $v_{\pi_t} \leq T^t v_{\pi_0}$ comes from a simple induction on t: Easily the statement holds for $t = 0$. Let us assume that it has been proved for t: $v_{\pi_t} \leq T^t v_{\pi_0}$. Due to M, $Tv_{\pi_t} \leq T^{t+1} v_{\pi_0}$. By Lemma 2.2.11 $v_{\pi_{t+1}} \leq Tv_{\pi_t}$. The combination of the last two inequalities yields $v_{\pi_{t+1}} \leq T^{t+1} v_{\pi_0}$ which completes the induction. □

The following example shows that, in arbitrary increasing models, continuity assumptions alone are insufficient to guarantee the convergence of policy iteration to optimality.

---

[2]In specific problems other procedures may be used, as well. For example, according to Lemma 2.2.4 Part a) and Lemma 2.2.3, Part b) the non-linear variational equation $Tv \leq v$, $||v|| \to \min$ with $\ell \leq v$ can be used to find v^{*s} if $v^{*s} = v_\infty$. Here $||\cdot||$ is an arbitrary norm. This variational equation reduces to linear programming for standard MDPs and the L^1 norm, but in the general case this variational equation may be hard to solve. Another recent efficient method which is available only for deterministic MDPs is based on an observation that such MDPs have a closed semiring formulation [43].

EXAMPLE 2.3.3 In this example $\mathcal{Q}$ is monotone, increasing, continuous and T is continuous. The model is finite. We will show that policy iteration cannot find the optimal policy.
The model is as follows: $\mathcal{X} = \{0\}$, $\mathcal{A} = \{0,1\}$. $(\mathcal{Q}f)(0,0) = 1+(1/2)f(0)$, $(\mathcal{Q}f)(0,1) = 1/4 + 2^{4(f(0)-1)}$.
Now, let π be the policy with $\pi(0) = 0$. Then from $v_\pi = T_\pi v_\pi$ we have that $v_\pi(0) = 2$. $(\mathcal{Q}v_\pi)(0,1) = 1/4 + 4 > 2 = (\mathcal{Q}v_\pi)(0,0)$ thus the PI routine returns π.
On the other hand, the optimal stationary policy is given by $\pi^*(0) = 1$ since from the fixed point equation $T_{\pi^*} v_{\pi^*} = v_{\pi^*}$ we have that $v_{\pi^*}(0) = 1/2$.

Policy iteration always stabilizes after a finite number of steps. Value iteration, on the other hand, generates an infinite sequence of functions $v_n = T^n v_0$. Since greedy policies w.r.t. v^{*s} are optimal, considering v_n as an approximation of v^{*s}, it is natural to ask whether greedy policies w.r.t. v_n are optimal for large enough n. This question has been answered when T is a contraction in [44]. If T is not a contraction then v_n is not guaranteed to converge at all for general v_0. In this case it is convenient to take $v_0 = \ell$. The following theorem shows that $v_n = T^n \ell$ has the desired "absorbtion" property.

THEOREM 2.3.4 (ABSORBTION IN VALUE ITERATION) *Assume M, I, LSC and that v_n converges to v^{*s}, where $v_{n+1} = Tv_n$, $\mathcal{Q}v_n$ converges to $\mathcal{Q}v^{*s}$ and both $\mathcal{A}$ and $\mathcal{X}$ are finite. Then for sufficiently large n we have $\Gamma v_n \subseteq \Pi_s(v^{*s})$.*

Proof. Note that since $\mathcal{A}$ is finite $\Pi_s(v^{*s}) = \Gamma v^{*s}$. So it is sufficient to prove that $\Gamma v_n \subseteq \Gamma v^{*s}$ for large enough n.

Let $x \in \mathcal{X}$ be an arbitrary state and $a \in \mathcal{A}$ be an arbitrary action. Let $\Gamma_{(x,a)} : B(\mathcal{X}) \to 2^{\mathcal{A}^{\mathcal{X}}}$ be defined by

$$\Gamma_{(x,a)} f = \{\, \pi \in \Gamma f \mid \pi(x) = a \,\},$$

that is $\Gamma_{(x,a)} f$ is the set of greedy policies that choose an action a in state x.

Pick up a state x. Since $\mathcal{A}$ is finite, $\Gamma v_n \neq \emptyset$ for all n. So for all n there exists an action $a \in \mathcal{A}$ for which $\Gamma_{(x,a)} v_n \neq \emptyset$. Again, since $\mathcal{A}$ is finite there must be at least one action a for which $\Gamma_{(x,a)} v_n \neq \emptyset$ infinitely many times. Let $\mathcal{A}_x$ be the set of such actions. We claim that

$$G_0 \stackrel{\text{def}}{=} \cap_{x\in\mathcal{X}} \cup_{a\in\mathcal{A}_x} \Gamma_{(x,a)} v^{*s} \subseteq \Gamma v^{*s}, \tag{2.2}$$

and for large enough n: $\Gamma v_n \subseteq G_0$. First of all, let us prove (2.2). Let π be an element of G_0: $\pi(x) \in \mathcal{A}_x$ for each $x \in \mathcal{X}$. Let $n_1(x), n_2(x), \ldots, n_k(x), \ldots$ denote the sequence of indices for which $\Gamma_{(x,\pi(x))} v_{n_k(x)} \neq \emptyset$. Then it follows that $v_{n_k(x)}(x) = (\mathcal{Q}v_{n_k(x)-1})(x, \pi(x)) = (T_\pi v_{n_k(x)-1})(x)$, where $x \in \mathcal{X}$ is arbitrary. Taking the limit $k \to \infty$ we get $v^{*s} = T_\pi v^{*s}$. Since by Corollary 2.3.1 $v^{*s} = Tv^{*s}$, hence $T_\pi v^{*s} = Tv^{*s}$, i.e., $\pi \in \Gamma v^{*s}$ proving (2.2).

It remains to be proved that $\Gamma v_n \subseteq G_0$ if n is large enough. Let $x \in \mathcal{X}$ be an arbitrary state. Since $\mathcal{A}$ is finite there must be a time $n_x < \infty$ after which for an arbitrary policy with $\pi \in \Gamma v_n$, $\pi(x) \in \mathcal{A}_x$. That is

$$\Gamma v_n \subseteq \cup_{a \in \mathcal{A}_x} (\Gamma_{(x,a)} v_n)$$

holds when $n \geq n_x$. Let $N = \max_{x \in \mathcal{X}} n_x$. Since $\mathcal{X}$ is finite $N < \infty$ and if $n \geq N$ and $\pi \in \Gamma v_n$ then $\pi(x) \in \mathcal{A}_x$ for all $x \in \mathcal{X}$, i.e. $\pi \in G_0$. □

Note that since $\mathcal{A}$ is finite, and if $v_0 = \ell$, $\mathcal{Q}$ is increasing, monotone and LSC, and T is also LSC then $v_n = T^n \ell$ converges to v^{*s} and $\mathcal{Q} v_n$ converges to $\mathcal{Q} v^{*s}$ as well and Theorem 2.3.4 applies.

2.4 Discussion

The aim of this chapter was to give a detailed description of the relations between the important sub-problems of ADPs (instead of resolving questions like the existence of optimal stationary policies). In order to answer concrete questions one needs to make further assumptions about $\mathcal{Q}$. For example, if one is interested in the optimality equation $T v_\infty = v_\infty$ then the LSC of T should be ensured. One way of achieving this is to assume that for all $x \in X$ $S_x : \mathcal{R}(\mathcal{X}) \to [-\infty, \infty]^A$ given by $S_x v = (\mathcal{Q} v)(x, \cdot)$ is lower semi-continuous w.r.t. the supremum-norm. Also the moduli of continuity (ω) of $\mathcal{Q}$ could be used for this purpose. Similarly to the method employed in [12] one can show that if $\omega(\delta) \to 0$ as $\delta \to 0$ then T is LSC. Further, if $\omega(\delta) < \delta$ for sufficiently small δ then $T v^{*s} = v^{*s}$ holds as well. The existence of ε-optimal policies can also be studied using the moduli of continuity of $\mathcal{Q}$. Most of the results of this chapter are published in [72].

Bertsekas [6] and Bertsekas & Shreve [9] are the closest to the work in this chapter. Bertsekas assumed that $\mathcal{Q}$ is Lipschitz[3], and if so then it is also continuous at each bounded function u. (However, as Example 0.1.14 shows, $\mathcal{Q}$ can be Lipschitz (even c-Lipschitz) without implying the same property for T.) Of course under the Lipschitz assumption much more can be proved: for example if $\mathcal{Q}$ is monotone, increasing, Lipschitz and lower-semi-continuous (this latter condition could be dropped if we knew there existed a policy with bounded evaluation) then $T v^{*m} = v^{*m}$ [6, Proposition 5], $T v_\infty = v_\infty$ is equivalent to $v_\infty = v^{*m}$,

[3]Under the monotonicity of $\mathcal{Q}$ Condition I.2 of [6] can be shown to be equivalent to $||\mathcal{Q}v - \mathcal{Q}u|| \leq \alpha ||u - v||$, where $|| \cdot ||$ denotes the supremum-norm and $\alpha > 0$. Note that since we consider functions over the extended reals, $\mathcal{Q}$ can be Lipschitz without being continuous. The following Lipschitz-like condition (let us call it the "c-Lipschitz property") implies continuity: if u is in the r ($r > 0$) neighbourhood of v then also $\mathcal{Q}u$ is in the αr neighbourhood of $\mathcal{Q}v$, where the r neighbourhood of a function $V \in \mathcal{R}(\mathcal{X})$ consisting of the functions $U \in \mathcal{R}(\mathcal{X})$ for which $|U(x) - V(x)| < r$ if $|V(x)| < \infty$ and $U(x) < -1/r$, if $V(x) = -\infty$ and $U(x) > 1/r$ if $V(x) = \infty$, $x \in \mathcal{X}$.

[6, Proposition 10] or $\Pi_s(v^{*m}) = \Gamma v^{*m}$ [6, Proposition 7]. Note that under these stronger conditions, $Tv_\infty = v_\infty$ is still not equivalent to $v_\infty = v^{*s}$ (or $v^{*g} = v^{*s}$) as shown by Example 2.2.5. The optimality of stationary policies requires stronger conditions such as a contraction assumption (Assumption C of [6]) or existence-like conditions, such as $\Gamma v_\infty \neq \emptyset$. In Proposition 11 Bertsekas derived a necessary and sufficient condition (based on the epigraph of $\mathcal{Q}$) for $v_\infty = Tv_\infty$ and $\Gamma v_\infty \neq \emptyset$ whose importance is clear from Figure 2.

Another question not fully explored is whether $\Gamma v \neq \emptyset$ for general v. Here, compactness arguments can be put forward. If $(\mathcal{Q}v)(x, \cdot)$ is continuous for all $v \in \mathcal{R}(\mathcal{X})$ and $\mathcal{A}$ is a complete metric space then one can show with Baire's theorem that $\Gamma v \neq \emptyset$.

Increasing models are just one of the usual three models investigated in the abstract setting. The other two assumptions are that $\mathcal{Q}$ is uniformly decreasing on ℓ: $(\mathcal{Q}\ell)(\cdot, a) \leq \ell$ for all $a \in \mathcal{A}$ and that $\mathcal{Q}$ is a contraction for some norm $\|\cdot\|$. The contraction assumption was first considered by Denardo [18] and later revisited by Bertsekas [6] and Bertsekas and Shreve [9]. Contraction models are very well understood, and in [9] it was shown that it is possible to get an approximately optimal policy by an approximate policy iteration routine. Decreasing models were first considered by Strauch [65] for MDPs and later by Bertsekas [6] and Bertsekas and Shreve [9]. Decreasing models are quite different from increasing ones. It is well known, for instance that policy improvement does not work in the decreasing case [65]. Finally we note that without too much effort the present framework could be extended to arbitrary cost (reward) spaces equipped with a partial ordering and thus we could generalize the results that hold for vector-valued DPs [31] to abstract models with optimality criterions differing from the usual total expected discounted cost criterion.

Part II

Reinforcement Learning

Introduction

In this part we describe methods for solving MDPs when the MDP is unknown but the decision maker may "experience" it. Recently many such algorithms has been investigated under the name "reinforcement learning" (RL), but they could also be considered as examples of adaptive controllers. In order to illustrate the idea consider a finite MDP and let the decision criterion be to minimize the total discounted expected cost (cf. Example 0.1.3 of Part 1). Then it follows from Corollary 1.5.5 that the optimal cost-to-go function $v^* = v^{*g}$ is the fixed point of the optimal cost-to-go operator $T : B(\mathcal{X}) \to B(\mathcal{X})$, $(Tv)(x) = \min_{a\in\mathcal{A}} \sum_{y\in\mathcal{X}} p(x,a,y)(c(x,a,y)+\gamma v(y))$, $0 < \gamma < 1$, where $p(x,a,y)$ is the probability of going to state y from state x when action a is used, $c(x,a,y)$ is the cost of this transition and γ is the discount factor. From part I, we also know that greedy policies w.r.t. v^* are optimal. The defining assumption of reinforcement learning is that p and c are unknown, so T is also unknown. Methods of RL can be divided into two categories: optimal cost-to-go function estimation based and policy iteration based methods. Here we will be concerned only with the first class of methods. In this class, an estimate of the optimal cost-to-go function is built gradually from experience (of the decision maker) and sometimes this estimate is simultaneously used for control. Two questions arise then: the convergence of the estimates to the true optimal cost-to-go function and the convergence of the control to optimality. Clearly, the two convergences can affect each other: if the estimates converged to optimality then the control should become asymptotically greedy with respect to the estimates in order to have it converge to optimality, and if the estimates do not converge to the optimal cost-to-go function then neither will the control converge to optimality. A more serious affect is that some control policies prevents the convergence of the cost-to-go function estimates to the optimal cost-to-go function (the decision problem is not "explored" sufficiently) in which case neither the control can converge to optimality. In summary, the control must become asymptotically greedy w.r.t. the cost-to-go function estimate but if the convergence to greediness is too fast then the cost-to-go function estimate may not have enough time to build up, preventing the convergence of control to optimality. The tradeoff between using "exploiting" (greedy) control and the convergence of estimates to optimality is called the *exploration-exploitation tradeoff*, and is well-recognized in the field of adaptive control. The proof that RL algorithms resolve the tradeoff is done in two steps by separating the proof of the convergence of the cost-to-go function estimates from that of the learning policy: first the convergence of the cost-to-go function estimates is shown under quite general conditions which prescribe only the meaning of "sufficient exploration". This is called *off-line* learning since the control policy is not assumed to be coupled to the estimation procedure. Such theorems are useful even from the practical point of view since the off-line algorithms can be viewed as solution methods for large MDPs whose structure is known but for which the explicit solution of the Bellman Optimality Equation would be too laborious to obtain. Then a Monte-Carlo simulation of the system together with a RL algorithm may provide a solution. It is important to note here that off-line RL algorithms differ from dynamic programming in two respects. Firstly, there is an estimation part of the algorithm and second, the algorithm is asynchronous: instead of updating (changing) all components of the estimate to the optimal value-function (as in value-iteration) only some carefully selected components are updated. It is well known that these asynchronous updates can speed up convergence considerably (similarly to the speed-up caused by Jacobi-iteration) [1]. It is exactly the asynchronous nature of these algorithms that makes the analysis of the off-line case non-trivial.

If the system is unknown then on-line learning policies, used during the estimation process to control the system, must be put forth that will satisfy both the "greedy in the limit" and the "sufficient exploration" conditions. This situation is called *on-line* learning since the estimation procedure and the control must be coupled. The technique of providing admissible learning policies is to provide conditions under which the cost-to-go function estimates converge in a restricted sense and finally to show that these conditions can be satisfied by some learning policies.

In the next chapter we derive the main result; in subsequent chapters we give applications of this result to off-line algorithms (Chapter 4) and on-line algorithms (Chapter 5).

Chapter 3

Asynchronous Successive Approximation

Consider an MDP where the decision maker has access to unbiased samples from $p(x,a,\cdot)$ and c; we assume that when the system's state-action transition is (x,a,y) then the decision maker receives a random value c, called the reinforcement signal, whose expectation is $c(x,a,y)$. Assume, moreover, that the decision maker wishes to identify v^*, the optimal cost-to-go function. For example, he might try to approximate p and c using some estimation procedures and use the estimated values, p_t, c_t, to approximate T (the optimal cost-to-go operator) as $T_t = T(p_t, c_t)$ and simultaneously he might try to use the operator sequence T_t to build an estimate of v^* by replacing T in the value iteration procedure by T_t. Or, as in Q learning [85], one might want to directly estimate $\mathcal{Q}v^*$ without ever estimating p or c, where $(\mathcal{Q}f)(x,a) = \sum_{y\in\mathcal{X}} p(x,a,y)(c(x,a,y) + \gamma f(y))$ is the cost propagation operator. The idea of this direct estimation procedure is the following: from the optimality equation $v^* = Tv^*$ it follows that Q^* is the fixed point of the operator $\tilde{T} = \mathcal{Q}\mathcal{N}$, where $\mathcal{N} : B(\mathcal{X}\times\mathcal{A}) \to B(\mathcal{X})$, $(\mathcal{N}Q)(x) = \min_{a\in\mathcal{A}} Q(x,a)$. For any fixed function Q, $\tilde{T}Q$ is easily approximated by averaging, which can be written recursively in the form

$$Q_{t+1}(x,a) = \tag{3.1}$$
$$\begin{cases} \left(1 - \frac{1}{n_t(x,a)}\right) Q_t(x,a) + \frac{1}{n_t(x,a)}\left(c_t + \gamma(\mathcal{N}Q)(x_{t+1})\right), & \text{if } (x,a) = (x_t,a_t); \\ Q_t(x,a), & \text{otherwise,} \end{cases}$$

where $n_t(x,a)$ is the number of times the state-action pair (x,a) was visited by the process (x_t,a_t) before time t plus one and, x_t is a controlled Markov cess with transition laws given by $P(x_{t+1}|x_t,a_t) = p(x_t,a_t,x_{t+1})$, and where $c_t \in \mathbb{R}$ depends stochastically on (x_t,a_t,x_{t+1}) with $E[c_t|x_t,a_t,x_{t+1}] = c(x_t,a_t,x_t)$ and $Var[c_t|x_t,a_t,x_{t+1}] < \infty$. The above iteration can be written in the more compact form

$$Q_{t+1} = T_t(Q_t, Q), \tag{3.2}$$

where T_t is a sequence of appropriately defined random operators. The approximation of Q^* comes then from the "optimistic" replacement of Q in the above iteration by Q_t. The corresponding process, called Q-learning [85], is

$$\hat{Q}_{t+1} = T_t(\hat{Q}_t, \hat{Q}_t). \tag{3.3}$$

Although the convergence of the iteration defined in (3.1) follows trivially from the law of large numbers, since for any fixed pair (x, a), the values $Q_t(x, a)$ as given by (3.1) are simple time averages of $c_t + \gamma(\mathcal{N}Q)(x_{t+1})$ filtered for the time steps when $(x, a) = (x_t, a_t)$, note that the convergence of the iteration given by (3.3) is not so straightforward. Specifically note that the componentwise analysis of the process of (3.3) is no longer possible, i.e., $\hat{Q}_{t+1}(x, a)$ depends on the values of $\hat{Q}_t$ at state-action pairs different from (x, a)— not like the case of Q_{t+1} and Q_t in Equation (3.2).

Interestingly, a large number of algorithms can be put into the form of (3.3) so it is worth choosing this iteration as the basis of our analysis together with the assumption that Q_t as defined in Equation (3.1) converges to $\tilde{T}Q$ for all functions Q. Then our main result is that under certain additional conditions on T_t, the iteration in (3.3) will converge to the fixed point of $\tilde{T}$. In this way, we will be able to prove the convergence of a wide range of algorithms. For example, we will obtain a convergence proof for Q-learning, for the iteration $v_{t+1} = T(p_t, c_t)v_t$ outlined earlier, and similar results for many other related algorithms.

3.1 The Main Result

Let $T : \mathcal{B} \to \mathcal{B}$ be an arbitrary operator, where $\mathcal{B}$ is a normed vector space, and let $\mathcal{T} = (T_0, T_1, \ldots, T_t, \ldots)$ be a sequence of random operators, T_t mapping $\mathcal{B} \times \mathcal{B}$ to $\mathcal{B}$. The following question is investigated here: Under what conditions can the iteration $f_{t+1} = T_t(f_t, f_t)$ be used to find the fixed point of T, provided that $\mathcal{T} = (T_0, T_1, \ldots, T_t, \ldots)$ approximates T in the sense defined next?

Definition 3.1.1 *Let $F \subseteq \mathcal{B}$ be a subset of $\mathcal{B}$ and let $\mathcal{F}_0 : F \to \mathcal{P}(\mathcal{B})$ be a mapping that associates subsets of B_1 with the elements of F. If, for all $f \in F$ and all $m_0 \in \mathcal{F}_0(f)$, the sequence generated by the recursion $m_{t+1} = T_t(m_t, f)$ converges to Tf in the norm of $\mathcal{B}$ with probability 1, then we say that $\mathcal{T}$* approximates *T for initial values from $\mathcal{F}_0(f)$ and on the set $F \subseteq \mathcal{B}$. Further, we say that $\mathcal{T}$* approximates *T at a certain point $f \in \mathcal{B}$ and for initial values from $F_0 \subseteq \mathcal{B}$ if $\mathcal{T}$ approximates T on the singleton set $\{f\}$ and the initial value mapping $\mathcal{F}_0 : F \to \mathcal{B}$ defined by $\mathcal{F}_0(f) = F_0$.*

We will also make use of the following definition.

Definition 3.1.2 *The subset $F \subseteq \mathcal{B}$ is* invariant *under $T : \mathcal{B} \times \mathcal{B} \to \mathcal{B}$ if, for all $f, g \in F$ $T(f, g) \in F$. If $\mathcal{T}$ is an operator sequence as above, then F is said to be invariant under $\mathcal{T}$ if for all $i \geq 0$ F is invariant under T_i.*

In many applications it is sufficient to consider the unrestricted case in which $F = \mathcal{B}$ and $\mathcal{F}_0(f) = \mathcal{B}$ for all $f \in \mathcal{B}$. For notational clarity in such cases, the set F and mapping $\mathcal{F}_0$ will not be explicitly mentioned.

The following is our main result.

THEOREM 3.1.3 *Let $\mathcal{X}$ be an arbitrary set and assume that $\mathcal{B}$ is the space of bounded functions over $\mathcal{X}$, $B(\mathcal{X})$, i.e., $T : B(\mathcal{X}) \to B(\mathcal{X})$. Let v^* be a fixed point of T and let $\mathcal{T} = (T_0, T_1, \dots)$ approximate T (w.p.1) at v^* and for initial values from $\mathcal{F}_0 \subseteq B(\mathcal{X})$, and assume that $\mathcal{F}_0$ is invariant under $\mathcal{T}$. Let $V_0 \in \mathcal{F}_0$, and define $V_{t+1} = T_t(V_t, V_t)$. If there exist functions $0 \le F_t(x) \le 1$ and $0 \le G_t(x) \le 1$ satisfying the conditions below w.p.1, then V_t converges to v^* w.p.1 in the norm of $B(\mathcal{X})$:*

1. *for all U_1 and $U_2 \in \mathcal{F}_0$, and all $x \in \mathcal{X}$,*

$$|T_t(U_1, v^*)(x) - T_t(U_2, v^*)(x)| \le G_t(x)|U_1(x) - U_2(x)|;$$

2. *for all U and $V \in \mathcal{F}_0$, and all $x \in \mathcal{X}$,*

$$|T_t(U, v^*)(x) - T_t(U, V)(x)| \le F_t(x)(\|v^* - V\| + \lambda_t),$$

   where $\lambda_t \to 0$ w.p.1. as $t \to \infty$;

3. *for all $k > 0$, $\Pi_{t=k}^{n} G_t(x)$ converges to zero uniformly in x as $n \to \infty$; and,*
4. *there exists $0 \le \gamma < 1$ such that for all $x \in \mathcal{X}$ and large enough t,*

$$F_t(x) \le \gamma(1 - G_t(x)).$$

REMARK 3.1.4 Note that from the conditions of the theorem and the additional conditions that T_t approximates T at every function $V \in B(\mathcal{X})$, it follows that T is a contraction operator at v^* with index of contraction γ (that is, T is a pseudo-contraction at v^* in the sense of [10]).

Proof. [of Remark 3.1.4] Let $V, U_0, V_0 \in B(\mathcal{X})$ be arbitrary and let $U_{t+1} = T_t(U_t, V)$ and $V_{t+1} = T_t(V_t, v^*)$. Let $\delta_t(x) = |U_t(x) - V_t(x)|$. Then, using Condition 1, 2 and 4 of Theorem 3.1.3 we get that $\delta_t(x) \le \hat{\delta}_t(x)$, where $\hat{\delta}_{t+1}(x) = G_t(x)\hat{\delta}_t(x) + \gamma(1 - G_t(x))\|V - v^*\|$. Substracting $\gamma\|V - v^*\|$ from both sides we obtain that $(\hat{\delta}_{t+1}(x) - \gamma\|V - v^*\|) = G_t(x)(\hat{\delta}_t(x) - \gamma\|V - v^*\|)$ and thus, by Condition 3, $\hat{\delta}_t(x)$ converges to $\gamma\|V - v^*\|$, i.e., $\limsup_{t\to\infty} \delta_t(x) \le \gamma\|V - v^*\|$ (see, e.g., the proof of Lemma 3.2.2 of Section 3.2). Since T_t approximates T at v^* and also at V, we have that $U_t \to TV$ and $V_t \to Tv^*$ w.p.1 which shows that δ_t converges to $\|TV - Tv^*\|$ w.p.1 and so $\|TV - Tv^*\| \le \gamma\|V - v^*\|$ also holds. □

One of the most noteworthy aspects of Theorem 3.1.3 is that it shows how to reduce the problem of approximating v^* to the problem of approximating T at a particular point V (in particular, it is enough that T can be approximated at v^*); in many cases, the latter is much easier to prove. For example, the theorem makes the convergence of Q-learning a consequence of the classical Robbins-Monro theory [53].

The most restrictive of the conditions of the theorem is Condition 4, which links the values of $G_t(x)$ and $F_t(x)$ through some quantity $\gamma < 1$. If it were somehow possible to update the values synchronously over the entire state space, i.e., if $V_{t+1}(x)$ depended on $V_t(x)$ only, then the process would converge to v^* even when $\gamma = 1$ provided that it were still the case that $\prod_{t=1}^{\infty}(F_t + G_t) = 0$ uniformly in x. In the more interesting asynchronous case, when $\gamma = 1$, the long-term behavior of V_t is not immediately clear; it may even be that V_t converges to something other than v^* or that it may even diverge, depending on how strict the inequalities of Conditions 4 and (3.4) (below) are. If these are strict, then $\|\delta_t\|$ need not decrease at all. The requirement that $\gamma < 1$ insures that the use of outdated information in the asynchronous updates does not cause a problem in convergence.

Proof. [Theorem 3.1.3] Let $U_0 \in \mathcal{F}_0$ be arbitrary and let $U_{t+1} = T_t(U_t, v^*)$. Since T_t approximates T at v^*, U_t converges to $Tv^* = v^*$ w.p.1 uniformly over $\mathcal{X}$. We will show that $\|U_t - V_t\|$ converges to zero w.p.1, which implies that V_t converges to v^*. Let

$$\delta_t(x) = |U_t(x) - V_t(x)|$$

and let

$$\Delta_t(x) = |U_t(x) - v^*(x)|.$$

We know that $\Delta_t(x)$ converges to zero because U_t converges to v^*.

By the triangle inequality and the conditions on T_t (invariance of $\mathcal{F}_0$ and the Lipschitz conditions), we have

$$\begin{aligned}
\delta_{t+1}(x) &= |U_{t+1}(x) - V_{t+1}(x)| \\
&= |T_t(U_t, v^*)(x) - T_t(V_t, V_t)(x)| \\
&\leq |T_t(U_t, v^*)(x) - T_t(V_t, v^*)(x)| + |T_t(V_t, v^*)(x) - T_t(V_t, V_t)(x)| \\
&\leq G_t(x)|U_t(x) - V_t(x)| + F_t(x)(\|v^* - V_t\| + \lambda_t) \\
&= G_t(x)\delta_t(x) + F_t(x)(\|v^* - V_t\| + \lambda_t) \\
&\leq G_t(x)\delta_t(x) + F_t(x)(\|v^* - U_t\| + \|U_t - V_t\| + \lambda_t) \\
&= G_t(x)\delta_t(x) + F_t(x)(\|\delta_t\| + \|\Delta_t\| + \lambda_t). \qquad (3.4)
\end{aligned}$$

In Lemma 3.4.2 presented below we will show that an inequality similar to Inequality (3.4) holds for a sub-series of $\|\delta_t\|$ and that therefore $\|\delta_t\|$ converges to zero. The perturbation term will be treated by treating δ_t as a homogeneous perturbed processes [33]. $\square$

3.2 Convergence in the Perturbation-free Case

We will need a relaxation of the concept of probability-one-convergence. Recall that by definition a random sequence x_t is said to converge to zero w.p.1 of for all $\eta, \delta > 0$ there exist a finite number $T = T(\eta, \delta)$ such that $P\left(\sup_{t \ge T} |x_t| \ge \delta\right) < \eta$. In this section we address the fact that the bound T might need to be random. Note that usually T is not allowed to be random. However, we show that T can be random and almost sure convergence still holds if T is almost surely bounded.

LEMMA 3.2.1 *Let x_t be a random sequence. Assume that for each $\eta, \delta > 0$ there exist an almost surely finite random index $T = T(\eta, \delta)$ such that*

$$P\left(\sup_{t \ge T} |x_t| \ge \delta\right) < \eta. \tag{3.5}$$

Then x_t converges to zero w.p.1.

Proof. Notice that if $T(\omega) \le k$ then $\sup_{t \ge k} |x_t(\omega)| \le \sup_{t \ge T(\omega)} |x_t(\omega)|$ and thus

$$\left\{\omega \,|\, \sup_{t \ge k} |x_t(\omega)| \ge \delta,\, T(\omega) \le k\right\} \subseteq \left\{\omega \,|\, \sup_{t \ge T(\omega)} |x_t(\omega)| \ge \delta,\, T(\omega) \le k\right\}.$$

Now,

$$\begin{aligned} A &= \left\{\omega \,|\, \sup_{t \ge k} |x_t(\omega)| \ge \delta\right\} \\ &= \Big(A \cap \{\omega \,|T(\omega) \le k\}\Big) \cup \Big(A \cap \{\omega \,|T(\omega) > k\}\Big) \\ &\subseteq \left\{\omega \,|\, \sup_{t \ge T(\omega)} |x_t(\omega)| \ge \delta,\, T(\omega) \le k\right\} \cup \{\omega \,|T(\omega) > k\}. \end{aligned}$$

Thus,

$$P\left(\sup_{t \ge k} |x_t| \ge \delta\right) \le P\left(\sup_{t \ge T} |x_t| \ge \delta\right) + P(T > k).$$

Now, pick up an arbitrary $\delta, \eta > 0$. We want to prove that for large enough $k > 0$ $P(\sup_{t \ge k} |x_t| \ge \delta) < \eta$. Let $T_0 = T(\delta, \eta/2)$ be the random index whose existence is guaranteed by assumption and let $k = k(\varepsilon, \eta)$ be a natural number large enough s.t. $P(T_0 > k) < \eta/2$. Such a number exists since $T_0 < \infty$ w.p.1. Then, $P(\sup_{t \ge k} |x_t| \ge \delta) \le P(\sup_{t \ge T_0} |x_t| \ge \delta) + P(T_0 > k) < \eta$, showing that k is a suitable (non-random) index. $\square$

Now we prove our version of Jaakkola et al.'s Lemma 2 [33] which concerns the convergence of the above process δ_t in the perturbation-free case. Note that both our assumptions and our proof are slightly different from theirs.

Lemma 3.2.2 *Let $\mathcal{Z}$ be an arbitrary set and consider the random sequence*

$$x_{t+1}(z) = g_t(z)x_t(z) + f_t(z)\|x_t\|, z \in \mathcal{Z} \tag{3.6}$$

where $x_0, f_t, g_t \geq 0$, $t \geq 0$, and $\|x_0\| < C < \infty$ w.p.1 for some $C > 0$. Assume that for all $k \geq 0$ $\lim_{n\to\infty} \prod_{t=k}^n g_t(z) = 0$ uniformly in z w.p.1 and $f_t(z) \leq \gamma(1 - g_t(z))$ w.p.1. (g_t and f_t are also random sequences). Then, $\|x_t\|$ converges to 0 w.p.1.

Proof. We will prove that for each $\varepsilon, \delta > 0$ there exist an a.s. bounded index $T = T(\varepsilon, \delta)$ such that

$$P\left(\sup_{t\geq T} \|x_t\| < \delta\right) > 1 - \varepsilon. \tag{3.7}$$

Let $\varepsilon, \delta > 0$ be arbitrary and let $p_0, \ldots, p_t, \ldots$ be a sequence of numbers ($0 < p_t < 1$) to be chosen later.

We have that

$$\begin{aligned} x_{t+1}(z) &= g_t(z)x_t(z) + f_t(z)\|x_t\| \\ &\leq g_t(z)\|x_t\| + f_t(z)\|x_t\| \\ &= (g_t(z) + f_t(z))\|x_t\| \\ &\leq \|x_t\|, \end{aligned}$$

since by assumption $g_t(z) + f_t(z) \leq g_t(z) + \gamma(1 - g_t(z)) \leq 1$. Thus, we have that $\|x_{t+1}\| \leq \|x_t\|$ for all t and, particularly, $\|x_t\| \leq C_0 = \|x_0\|$ holds for all t. Consequently, the process

$$y_{t+1}(z) = g_t(z)y_t(z) + \gamma(1 - g_t(z))C_0, \tag{3.8}$$

with $y_0 = x_0$, estimates the process $\{x_t\}$ from above: $0 \leq x_t \leq y_t$ holds for all $t \geq 0$. The process y_t converges to γC_0 w.p.1 uniformly over $\mathcal{Z}$. (Substract γC_0 from both sides to get $(y_{t+1}(z) - \gamma C_0) = g_t(z)(y_t(z) - \gamma C_0)$. Now convergence of $\|y_t - \gamma C_0\|$ follows since $\lim_{n\to\infty} \prod_{t=k}^n g_t(z) = 0$ uniformly in z). Therefore,

$$\limsup_{t\to\infty} \|x_t\| \leq \gamma C_0$$

w.p.1. Thus, there exists an a.s. bounded index, say M_0, for which if $t > M_0$ then $\|x_t\| \leq (1 + \gamma)/2\, C_0$ with probability p_0. Assume that up to some index $i \geq 0$ we have found numbers M_i such that when $t \geq M_i$ then

$$\|x_t\| \leq \left(\frac{1+\gamma}{2}\right)^i C_0 = C_i \tag{3.9}$$

holds with probability $p_0 p_1 \ldots p_i$. Now, let us restrict our attention to those events for which Inequality (3.9) holds. Then, we see that the process

$$\begin{aligned} y_{M_i} &= x_{M_i} \\ y_{t+1}(z) &= g_t(z)y_t(z) + \gamma(1 - g_t(z))C_i, \ t > M_i \end{aligned}$$

bounds x_t from above from the index M_i. Now, the above argument can be repeated to obtain an index M_{i+1} such that Inequality (3.9) hold for $i+1$ with probability $p_0 p_1 \dots p_i p_{i+1}$.

Since $(1+\gamma)/2 < 1$, there exists an index k for which $((1+\gamma)/2)^k C_0 < \varepsilon$. Then, we get that Inequality (3.7) is satisfied when we choose $p_0, \dots, p_k$ in a way that $p_0 p_1 \dots p_k \geq 1-\varepsilon$ and we set $T = M_k (= M_k(p_0, p_1, \dots, p_k))$. □

When the process of Equation (3.6) is subject to decaying perturbations, say ε_t (see, e.g., the process of Inequality (3.4)), then the proof no longer applies. The problem is that $\|x_t\| \leq \|x_0\|$ (or $\|x_{T+t}\| \leq \|x_T\|$, for large enough T) can no longer be ensured without additional assumptions. For $x_{t+1}(z) \leq \|x_t\|$ to hold, we would need that $\gamma \varepsilon_t \leq (1-\gamma)\|x_t\|$, but if $\liminf_{t\to\infty} \|x_t\| = 0$ (which, in fact, is a consequence of what should be proved), then we could not check this relation *a priori*. Thus, we choose another way to prove Lemma 3.4.2. Notice that the key idea in the above proof is to bound x_t by y_t. This can be done if we assume that x_t is kept bounded artificially, e.g., by scaling. The next subsection shows that such a change of x_t does not effect its convergence properties.

3.3 The Rescaling of Two-variable Homogeneous Processes

The next lemma is about two-variable homogeneous processes, that is, processes of the form

$$x_{t+1} = G_t(x_t, \varepsilon_t), \tag{3.10}$$

where $G_t : \mathcal{B} \times \mathcal{B} \to \mathcal{B}$ is a homogeneous random function ($\mathcal{B}, \mathcal{B}$ denote normed vector spaces as usual), i.e.,

$$G_t(\beta x, \beta \varepsilon) = \beta G_t(x, \varepsilon) \tag{3.11}$$

holds for all $\beta > 0$, x and ε.[1] We are interested in whether x_t converges to zero or not. Note that when the inequality defining δ_t (Inequality (3.4)) is an equality, the process becomes homogeneous. Lemma 3.3.2 below says that, under additional conditions, it is enough to prove the convergence of a modified process *which is kept bounded by rescaling* to zero, namely the process

$$y_{t+1} = \begin{cases} G_t(y_t, \varepsilon_t), & \text{if } \|G_t(y_t, \varepsilon_t)\| \leq C; \\ C\, G_t(y_t, \varepsilon_t)/\|G_t(y_t, \varepsilon_t)\|, & \text{otherwise,} \end{cases} \tag{3.12}$$

[1] In [33] the authors considered a question similar to that which is investigated below in Lemma 3.3.2 for the case of *single-variable* homogeneous processes, which would correspond to the case when $\varepsilon_t = 0$ for all $t \geq 0$ (see Equation (3.10)). The single-variable case follows from our result.

where $C > 0$ is an arbitrary fixed number. The idea of "projecting" on a bounded set to ensure boundedness of stochastic approximation processes has been discussed by Ljung [45] and Kushner and Clark [41].

We denote the solution of Equation (3.10) corresponding to the initial condition $x_0 = w$ and the sequence $\varepsilon = \{\varepsilon_k\}$ by $x_t(w, \varepsilon)$. Similarly, we denote the solution of Equation (3.12) corresponding to the initial condition $y_0 = w$ and the sequence ε by $y_t(w, \varepsilon)$.

DEFINITION 3.3.1 *We say that the process x_t is* insensitive to finite perturbations of ε *if it holds that if $x_t(w, \varepsilon)$ converges to zero then so does $x_t(w, \varepsilon')$, where $\varepsilon'(\omega)$ is an arbitrary sequence that differs only in a finite number of terms from $\varepsilon(\omega)$, where the bound on the number of differences is independent of ω. Further, the process x_t is said to be* insensitive to scaling of ε by numbers smaller than 1, *if for all random $0 < c \leq 1$ it holds that if $x_t(w, \varepsilon)$ converges to zero then so does $x_t(w, c\varepsilon)$.*

LEMMA 3.3.2 (RESCALING LEMMA) *Let $C > 0$, w_0 and the sequence ε be arbitrary. Then, a homogeneous process $x_t(w_0, \varepsilon)$ converges to zero w.p.1 provided that (i) x_t is insensitive to finite perturbations of ε; (ii) x_t is insensitive to scaling of ε by numbers smaller than 1 and (iii) $y_t(w_0, \varepsilon)$ converges to zero.*

Proof. We state that

$$y_t(w, \varepsilon) = x_t(d_t w, c_t \varepsilon) \tag{3.13}$$

for some sequences $\{c_t\}$ and $\{d_t\}$, where $c_t = (c_{t0}, c_{t1}, \ldots, c_{ti}, \ldots)$ and $\{c_t\}$ and $\{d_t\}$ satisfy $0 < d_t, c_{ti} \leq 1$, and $c_{ti} = 1$ if $i \geq t$. Here the product of the sequences c_t and ε should be understood componentwise: $(c_t \epsilon)_i = c_{ti} \epsilon_i$. Note that $y_t(w, \varepsilon)$ and $x_t(w, \varepsilon)$ depend only on $\varepsilon_0, \ldots, \varepsilon_{t-1}$. Thus, it is possible to prove Equation (3.13) by constructing the appropriate sequences c_t and d_t.

Set $c_{0i} = d_i = 1$ for all $i = 0, 1, 2, \ldots$. Then, Equation (3.13) holds for $t = 0$. Let us assume that $\{c_i, d_i\}$ is defined in a way that Equation (3.13) holds for t. Let S_t be the "scaling coefficient" of y_t at step $(n + 1)$ ($S_t = 1$ if there is no scaling, otherwise $0 < S_t < 1$ with $S_t = C/\|G_t(y_t, \varepsilon_t)\|$):

$$\begin{aligned} y_{t+1}(w, \varepsilon) &= S_t G_t(y_t(w, \varepsilon), \varepsilon_t) \\ &= G_t(S_t y_t(w, \varepsilon), S_t \varepsilon_t) \\ &= G_t(S_t x_t(d_t w, c_t \varepsilon), S_t \varepsilon_t). \end{aligned}$$

We claim that

$$S x_t(w, \varepsilon) = x_t(Sw, S\varepsilon) \tag{3.14}$$

holds for all w, ε and $S > 0$.

For $t = 0$, this obviously holds. Assume that it holds for t. Then,

$$\begin{aligned} Sx_{t+1}(w,\varepsilon) &= SG_t(x_t(w,\varepsilon),\varepsilon_t) \\ &= G_t(Sx_t(w,\varepsilon), S\varepsilon_t) \\ &= G_t(x_t(Sw,S\varepsilon), S\varepsilon_t) \\ &= x_{t+1}(Sw,S\varepsilon). \end{aligned}$$

Thus,

$$y_{t+1}(w,\varepsilon) = G_t(x_t(S_t d_t w, S_t c_t \varepsilon), S_t \varepsilon_t),$$

and we see that Equation (3.13) holds if we define $c_{t+1,i}$ as $c_{t+1,i} = S_t c_{ti}$ if $0 \le i \le t$, $c_{t+1,i} = 1$ if $i > t$ and $d_{t+1} = S_t d_t$.

Thus, we get that with the sequences

$$c_{t,i} = \begin{cases} \prod_{j=i}^{t-1} S_j, & \text{if } i < t; \\ 1, & \text{otherwise,} \end{cases}$$

$d_0 = 1$, and

$$d_{t+1} = \prod_{i=0}^{t} S_i,$$

Equation (3.13) is satisfied for all $t \ge 0$.

Now, assume that we want to prove for a particular sequence ε and initial value w that

$$\lim_{t\to\infty} x_t(w,\varepsilon) = 0 \tag{3.15}$$

holds w.p.1. It is enough to prove that Equation (3.15) holds with probability $1-\delta$ when $\delta > 0$ is an arbitrary, sufficiently small number.

We know that $y_t(w,\varepsilon) \to 0$ w.p.1. We may assume that $\delta < C$. Then, there exists an index $M = M(\delta)$ such that if $t > M$ then

$$P(\|y_t(w,\varepsilon)\| < \delta) > 1-\delta. \tag{3.16}$$

Now, let us restrict our attention to those events ω for which $\|y_t(w,\varepsilon(\omega))\| < \delta$ for all $t > M$: $A_\delta = \{\omega : \|y_t(w,\varepsilon)(\omega)\| < \delta\}$. Since $\delta < C$, we get that there is no rescaling after step M: $S_t(\omega) = 1$ if $t > M$. Thus, $c_{t,i} = c_{M+1,i}$ for all $t \ge M+1$ and i, amd specifically $c_{t,i} = 1$ if $i, t \ge M+1$. Similarly, if $t > M$ then $d_{t+1}(\omega) = \prod_{i=0}^{M} S_i(\omega) = d_{M+1}(\omega)$. By Equation (3.13), we have that if $t > M$ then

$$y_t(w,\varepsilon(\omega)) = x_t(d_{M+1}(\omega)w, c_{M+1}(\omega)\varepsilon(\omega)).$$

Thus, it follows from our assumption concerning y_t that $x_t(d_{M+1}(\omega)w, c_{M+1}\varepsilon(\omega))$ converges to zero almost everywhere (a.e.) on A_δ and consequently, by Equation (3.14), $x_t(w, c_{M+1}\varepsilon(\omega)/d_{M+1(\omega)})$ also converges to zero a.e. on A_δ. Since x_t is insensitive to finite perturbations and since in c_{M+1} only a finite number of entries differs from 1, $x_t(w, \varepsilon(\omega)/d_{M+1}(\omega))$ also converges to zero, and, further, since $d_{M+1}(\omega) < 1$, $x_t(w, \varepsilon(\omega)) = x_t(w, d_{M+1}(\omega)(\varepsilon(\omega)/d_{M+1}(\omega)))$ converges to zero, too (x_t is insensitive to scaling of ε by d_{M+1}). All these hold with probability at least $1-\delta$, since, by Equation (3.16), $P(A_\delta) > 1-\delta$. Since δ was arbitrary, the lemma follows. □

3.4 Convergence of Perturbed Processes

We have established that Inequality (3.4) converges if not perturbed. We now extend this to more general perturbed processes so we can complete the proof of Theorem 3.1.3.

The following theorem concerns the stability of certain discrete-time systems:

THEOREM 3.4.1 *Let $\mathcal{X}$ and $\mathcal{Y}$ be normed vector spaces, $U_t : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ ($t = 0, 1, 2, \ldots$) be a sequence of mappings, and $\theta_t \in \mathcal{Y}$ be an arbitrary sequence. Let $\theta_\infty \in \mathcal{Y}$ and $x_\infty \in \mathcal{X}$. Consider the sequences $x_{t+1} = U_t(x_t, \theta_\infty)$, and $y_{t+1} = U_t(y_t, \theta_t)$, and suppose that x_t and θ_t converge to x_∞ and θ_∞, respectively, in the norm of the appropriate spaces.*

Let L_k^θ be the uniform Lipschitz index of $U_k(x, \theta)$ with respect to θ at θ_∞ and, similarly, let L_k^χ be the uniform Lipschitz index of $U_k(x, \theta_\infty)$ with respect to x.[2] *Then, if the Lipschitz constants L_t^χ and L_t^θ satisfy the relations $L_t^\theta \le C(1 - L_t^\chi)$, and $\prod_{m=n}^{\infty} L_m^\chi = 0$, where $C > 0$ is some constant and $n = 0, 1, 2, \ldots$, then $\lim_{t\to\infty} \|y_t - x_\infty\| = 0$.*

Proof. The proof is based on relating the convergence of $\|y_t - x_\infty\|$ and a triangle transformation of $\|\theta_t - \theta_\infty\|$ and can be found in Appendix A.1. □

Observe that if U_t and θ_t are random and all the relations are required to hold a.s., specifically, if "convergence in norm" is replaced by "a.s. convergence in norm", then the above theorem remains valid. We will indeed use this form of the above theorem to show the a.s. convergence of certain random variables holds. Now, we are in the position to prove that Lemma 3.2.2 is immune to decaying perturbations.

LEMMA 3.4.2 *Assume that the conditions of Lemma 3.2.2 are satisfied but Equation (3.6) is replaced by*

$$x_{t+1}(z) = g_t(z)x_t(z) + f_t(z)(\|x_t\| + \varepsilon_t), \tag{3.17}$$

where $\varepsilon_t \ge 0$ and ε_t converges to zero with probability 1. Then, $x_t(z)$ still converges to zero w.p.1 uniformly over $\mathcal{Z}$.

---

[2] That is, for all $x \in \mathcal{X}$ and $\theta \in \mathcal{Y}$ $||U_k(x,\theta) - U_k(x,\theta_\infty)|| \le L_k^\theta ||\theta - \theta_\infty||$ and for all $x, y \in \mathcal{X}$ $||U_k(x,\theta_\infty) - U_k(y,\theta_\infty)|| \le L_k^\chi ||x - y||$.

Proof. First, we show that the process of Equation (3.17) satisfies the assumptions of the Rescaling Lemma (Lemma 3.3.2) and, thus, it is enough to consider the version of Equation (3.17) that is kept bounded by scaling.

First, note that x_t is a homogeneous process of form (3.10) (note that Equation (3.11) is required to hold only for positive scaling numbers). Let us prove that x_t is immune to finite perturbations of ε. To this end, assume that ε'_t differs only in a finite number of terms from ε_t and let

$$y_{t+1}(z) = g_t(z)y_t(z) + f_t(z)(\|y_t\| + \varepsilon'_t).$$

Take

$$k_t(z) = |x_t(z) - y_t(z)|.$$

Then,

$$k_{t+1}(z) \le g_t(z)k_t(z) + f_t(z)(\|k_t(z)\| + |\varepsilon_t - \varepsilon'_t|).$$

For large enough t $\varepsilon_t = \varepsilon'_t$, so

$$k_{t+1}(z) \le g_t(z)k_t(z) + f_t(z)\|k_t(z)\|,$$

which we know to converge to zero by Lemma 3.2.2. Thus, either x_t and y_t converge and converge to the same value, or neither x_t or y_t converges.

The other requirement that we must satisfy to be able to apply the Rescaling Lemma (Lemma 3.3.2) is that x_t is insensitive to scaling of the perturbation by numbers of the interval $[0, 1)$; let us choose a number $0 < c < 1$ and assume that $x_t(w, \varepsilon)$ converges to zero with probability 1. Then, since $0 \le x_t(w, c\varepsilon) \le x_t(w, \varepsilon)$, $x_t(w, c\varepsilon)$ converges to zero w.p.1, too.

Now, let us prove that the process that is obtained from x_t by keeping it bounded converges to zero. The proof is the mere repetition of the proof of Lemma 3.2.2, except a few points that we discuss now. Let us denote by $\hat{x}_t$ the process that is kept bounded and let the bound be C_0. It is enough to prove that $\|\hat{x}_t\|$ converges to zero w.p.1. Now, Equation (3.8) is replaced by

$$y_{t+1}(z) = g_t(z)y_t(z) + \gamma(1 - g_t(z))(C_0 + \varepsilon_t).$$

By Theorem 3.4.1, y_t still converges to γC_0, as the following shows: $\mathcal{X}, \mathcal{Y} := \mathbf{R}$ $\theta_t := \varepsilon_t$, $U_t(x, \theta) := g_t(z)x + \gamma(1 - g_t(z))(C_0 + \theta)$, where $z \in \mathcal{Z}$ is arbitrary. Then, $L_t^{\chi} = g_t(z)$ and $L_t^{\theta} = \gamma(1 - g_t(z))$ satisfying the condition of Theorem 3.4.1.

Since it is also the case that $0 \le \hat{x}_t \le y_t$, the whole argument of Lemma 3.2.2 can be repeated for the process $\hat{x}_t$, yielding that $\|\hat{x}_t\|$ converges to zero w.p.1 and, consequently, so does $\|x_t\|$. This finishes the proof of the lemma. □

We have thus completed the proof of Theorem 3.1.3.

3.5 Relaxation Processes

In this section, we prove a corollary of Theorem 3.1.3 for relaxation processes of the form

$$V_{t+1}(x) = (1 - f_t(x))V_t(x) + f_t(x)[P_t V_t](x), \tag{3.18}$$

where $0 \le f_t(x)$ is a relaxation parameter converging to zero and the sequence $P_t : B(\mathcal{X}) \to B(\mathcal{X})$ is a randomized version of an operator T in the sense that the "averages"

$$U_{t+1}(x) = (1 - f_t(x))U_t(x) + f_t(x)[P_t V](x)$$

converge to TV w.p.1, where $V \in B(\mathcal{X})$. A large number of reinforcement-learning algorithms have this form, which makes these processes of interest. We give some concrete examples in later sections. It is important to note that while $V_{t+1}(x)$ depends on $V_t(y)$ for all $y \in \mathcal{X}$ since $P_t V_t$ depends on all the components of V_t, $U_{t+1}(x)$ depends only on $U_t(x)$, $x \in \mathcal{X}$: the different components are decoupled. This greatly simplifies the proof of convergence of (3.18). Usually, the following, so-called conditional averaging lemma is used to show that the process of (3.18) converges to TV.

Lemma 3.5.1 (Conditional Averaging Lemma) *Let $\mathcal{F}_t$ be an increasing sequence of σ-fields, let $0 \le \alpha_t \le 1$ and w_t be random variables such that α_t and w_{t-1} are $\mathcal{F}_t$ measurable. Assume that the following hold w.p.1: $E[w_t|\mathcal{F}_t, \alpha_t \neq 0] = A$, $E[w_t^2|\mathcal{F}_t] < B < \infty$, $\sum_{t=1}^\infty \alpha_t = \infty$ and $\sum_{t=1}^\infty \alpha_t^2 < C < \infty$ for some $B, C > 0$. Then, the process*

$$Q_{t+1} = (1 - \alpha_t)Q_t + \alpha_t w_t$$

converges to A w.p.1.

Note that this lemma generalizes the Robbins-Monro Theorem in that, here, α_t is allowed to depend on the past of the process, which will prove to be essential in our case, but is less general since $E[w_t|\mathcal{F}_t, \alpha_t \neq 0]$ is not allowed to depend on Q_t. The proof of this Lemma can be found in Appendix A.2 (cf. Lemma A.2.3).

Corollary 3.5.2 *Consider the process generated by iteration (3.18). Assume that the process defined by*

$$U_{t+1}(x) = (1 - f_t(x))U_t(x) + f_t(x)[P_t v^*](x) \tag{3.19}$$

converges to v^* w.p.1. (This condition is called the approximating property *of P_t and f_t.) Assume further that the following conditions hold:*

1. *there exist number $0 < \gamma < 1$ and a sequence $\lambda_t \ge 0$ converging to zero w.p.1 such that $\|P_t V - P_t v^*\| \le \gamma \|V - v^*\| + \lambda_t$ holds for all $V \in B(\mathcal{X})$;*

2. $0 \leq f_t(x) \leq 1$, *and* $\sum_{t=1}^n f_t(x)$ *converges to infinity uniformly in* x *as* $n \to \infty$.

Then, the iteration defined by (3.18) converges to v^* *w.p.1.*

Note that if $\lim_{t\to\infty} \|f_t\| = 0$ w.p.1 then for large enough t $P(\|f_t\| \leq 1) > 1-\epsilon$ for arbitrary $0 < \epsilon < 1$, so if $\|f_t\|$ converges to zero w.p.1 then condition $f_t(x) \leq 1$ can be discarded.

Proof. Let the random operator sequence $T_t : B(\mathcal{X}) \times B(\mathcal{X}) \to B(\mathcal{X})$ be defined by

$$T_t(U,V)(x) = (1 - f_t(x))U(x) + f_t(x)[P_tV](x).$$

T_t approximates T at v^*, since, by assumption, the process defined in (3.19) converges to $v^* = Tv^*$ for all $V \in B(\mathcal{X})$. Moreover, observe that V_t as defined by (3.18) satisfies $V_{t+1} = T_t(V_t, V_t)$. Because of Assumptions 1 and 2, it can be readily verified that the Lipschitz-coefficients $G_t(x) = 1 - f_t(x)$, $F_t(x) = \gamma f_t(x)$ satisfy the rest of the conditions of Theorem 3.1.3, and this yields that the process V_t converges to v^* w.p.1. □

Note that, although a large number of processes of interest admit this relaxation form, there are some important exceptions. In Sections 4.2 and 4.5 we will deal with some processes that are not of the relaxation type and we will show that Theorem 3.1.3 still applies; this shows the broad utility of Theorem 3.1.3. Another class of exceptions are formed by processes when P_t involves some additive, zero-mean, finite conditional variance noise-term which disrupts the pseudo-contraction property (Condition 1) of P_t. With some extra work Corollary 3.5.2 can be extended to work in these cases; in that proof the Rescaling Lemma must be used several times. As a result a proposition almost identical to Theorem 1 of [33] can be deduced. This extension will be presented at the end of the next chapter.

3.6 Convergence Rates

The difficult part of proving the convergence of RL algorithms is to prove that the asynchronous iteration

$$x_{t+1}(z) = g_t(z)x_t(z) + f_t(z)\|x_t\|, \tag{3.20}$$

converges to zero (cf. Equation (3.6) of Lemma 3.2.2). The aim of this section is to strengthen the statement of Lemma 3.2.2 under special assumptions concerning f_t and g_t, and then to give an estimate for the convergence rate of the above process to zero. Here, we assume that the set of possible states Z is finite (we identify this set with $\{1, 2, \ldots, n\}$), and that the process 3.20 takes the form

$$x_{t+1}(i) = \begin{cases} \left(1 - \frac{1}{S_t(i)}\right) x_t(i) + \frac{\gamma}{S_t(i)}\|x_t\|, & \text{if } \eta_t = i; \\ x_t(i), & \text{if } \eta_t \neq i, \end{cases} \tag{3.21}$$

where $t = 1, 2, 3, \ldots$, $i = 1, 2, \ldots, n$, $\eta_t \in \{1, 2, \ldots, n\}$, $S_t(i)$ is the number of times the event $\{\eta_t = i\}$ happened before time $(t+1)$ plus one (i.e., $S_t(i) = 1 + |\{\, s \mid \eta_s = i,\ 0 < s < t+1 \,\}|$) and $\|\cdot\|$ denotes the supremum-norm as before. For example, in the case of Q-learning with learning rates inversely proportional to the visit-times a given state-action pair, the difference process δ_t defined in the proof of Theorem 3.1.3 admits the form of (3.21).

THEOREM 3.6.1 *Assume that $\eta_1, \eta_2, \ldots, \eta_t, \ldots$ is a finite, stationary Markov chain with states $\{1, 2, \ldots, n\}$ and stationary distribution $(p_1, p_2, \ldots, p_n)$, where $p_i > 0$, $1 \le i \le n$. Then the process x_t defined in Equation (3.21) satisfies*

$$\|x_t\| = O\Big(\frac{1}{t^{R(1-\gamma)}}\Big)$$

with probability one (w.p.1)[3], *where $R = \min_i p_i / \max_i p_i$.*

Proof. Let $T_0 = 0$ and

$$T_{k+1} = \min\{\, t \ge T_k \mid \forall i = 1 \ldots n,\ \exists s = s(i) :\ \eta_s = i \,\},$$

i.e. T_{k+1} is the smallest time after time T_k such that during the time interval $[T_k + 1, T_{k+1}]$ all the components of $x_t(\cdot)$ are "updated" in Equation (3.21) at least once. Then

$$x_{T_{k+1}+1}(i) \le \left(1 - \frac{1-\gamma}{S_k}\right) \|x_{T_k+1}\|, \tag{3.22}$$

where $S_k = \max_i S_{T_{k+1}}(i)$. This inequality holds because if $t_k(i)$ is the last time in $[T_k + 1, T_{k+1}]$ when the i^{th} component is updated then

$$\begin{aligned}
x_{T_{k+1}+1}(i) &= x_{t_k(i)+1}(i) = (1 - 1/S_{t_k(i)}(i))x_{t_k(i)}(i) + \gamma/S_{t_k(i)(i)}\|x_{t_k(i)}(\cdot)\| \\
&\le (1 - 1/S_{t_k(i)}(i))\|x_{t_k(i)}(\cdot)\| + \gamma/S_{t_k(i)(i)}\|x_{t_k(i)}(\cdot)\| \\
&= \left(1 - \frac{1-\gamma}{S_{t_k(i)}}(i)\right)\|x_{t_k(i)}(\cdot)\| \\
&\le \left(1 - \frac{1-\gamma}{S_k}\right)\|x_{T_k+1}(\cdot)\|,
\end{aligned}$$

where it was exploited that $\|x_t\|$ is decreasing and that $S_k \ge S_{T_{k+1}}(i) = S_{t_k(i)}(i)$. Now, iterating (3.22) backwards in time yields

$$x_{T_k+1}(\cdot) \le \|x_0\| \prod_{j=0}^{k-1} \left(1 - \frac{1-\gamma}{S_j}\right).$$

[3]In this context $a_t = O(b_t)$ means that $\limsup_{t\to\infty} |b_t|/|a_t| \le C(\omega) < \infty$ for some random variable C and for almost all ω.

Now fix an $\epsilon > 0$. Then there exists an integer $N = N(\epsilon) > 0$ and an event set A_ϵ such that $P(A_\epsilon) \geq 1 - \epsilon$ and

$$|S_j - (j+1)R_0^{-1}| \leq (j+1)\epsilon \tag{3.23}$$

holds when $\omega \in A_\epsilon$ and $j \geq N$ ($S_j = S_j(\omega)$). Now, if ϵ is sufficiently small, $k \geq N(\epsilon)$ and $\omega \in A_\epsilon$ then:

$$\begin{aligned}
\prod_{j=0}^{k-1}\left(1 - \frac{1-\gamma}{S_j}\right) &= \prod_{j=0}^{N-1}\left(1 - \frac{1-\gamma}{S_j}\right)\prod_{j=N}^{k-1}\left(1 - \frac{1-\gamma}{S_j}\right) \\
&\leq \prod_{j=N}^{k-1}\left(1 - \frac{(1-\gamma)/(j+1)}{R_0^{-1}+\epsilon}\right) \\
&= \exp\left(\frac{1-\gamma}{R_0^{-1}+\epsilon}\sum_{j=N}^{k-1}\frac{1}{j+1}\right) \\
&\leq \exp\left(\frac{1-\gamma}{R_0^{-1}+\epsilon}\log(k/(N+1))\right) \\
&= \left(\frac{N(\epsilon)+1}{k}\right)^{\frac{1-\gamma}{R_0^{-1}+\epsilon}} \\
&\leq \left(\frac{N(\epsilon)+1}{k}\right)^{\frac{1-\gamma}{R}}.
\end{aligned}$$

In the last inequality we have used $R_0 > p_{\min}/p_{\max} = R$ and that ϵ was assumed to be sufficiently small.

Now, by defining $s = T_k + 1$ so that $s/C \approx k$ we get

$$\|x_s\| = \|x_{T_k+1}\| \leq \|x_0\|\left(\frac{1}{k}\right)^{R(1-\gamma)} \approx \|x_0\|\left(\frac{C}{s}\right)^{R(1-\gamma)}.$$

Therefore, also $\|x_t\| = O(1/t^{R(1-\gamma)})$ holds due to the monotonicity of x_t and the monotonicity of $\{1/k^{R_0(1-\gamma)}\}$ in k. All these hold on A_ϵ, i.e., with probability $1-\epsilon$, thus, finishing the proof. □

Now, assume $\gamma > 1/2$. Then the same convergence rate holds for the perturbed process

$$x_{t+1}(i) = \left(1 - \frac{1}{S_t(i)}\right)x_t(i) + \frac{\gamma}{S_t(i)}\Big(\|x_t\| + \varepsilon_t\Big), \tag{3.24}$$

(cf. also (3.17)) where $\varepsilon_t = O(\sqrt{\log\log t/t})$ is a decreasing sequence. We state that the convergence rate of ϵ_t is faster than that of x_t. Define the process

$$z_{t+1}(i) = \begin{cases} \left(1 - \frac{1-\gamma}{S_t(i)}\right)z_t(i), & \text{if } \eta_t = i; \\ z_t(i), & \text{if } \eta_t \neq i. \end{cases} \tag{3.25}$$

This process clearly lower bounds the perturbed process, x_t. Obviously, the convergence rate of x_t is slower than that of z_t, whose convergence rate is $o(1/t^{1-\gamma})$, which is slower than the convergence rate of ϵ_t provided that $\gamma > 1/2$, proving that ϵ_t must be faster than x_t. Thus, asymptotically $\epsilon_t \le (1/\gamma - 1)x_t$, and so $\|x_t\|$ is decreasing for large enough t. Then, by an argument similar to that of used in the derivation of (3.22), we get

$$x_{T_{k+1}+1}(i) \le \left(1 - \frac{1-\gamma}{S_k}\right) \|x_{T_k+1}\| + \frac{\gamma}{s_k}\epsilon_{T_k}, \tag{3.26}$$

where $s_k = \min_i S_{T_{k+1}}(i)$. Finally, by some approximation arguments similar to that of Theorem 3.6.1, together with the bound $(1/n^\eta)\sum_s^n s^{\eta-3/2}\sqrt{\log\log s} \le s^{-1/2}\sqrt{\log\log s}$, $1 > \eta > 0$, which follows from the mean-value theorem for integrals and the law of integration by parts, we get that $\|x_t\| \approx O(1/t^{R(1-\gamma)})$.

Now, consider a relaxation type-of approximation process

$$V_{t+1}(i) = \begin{cases} (1 - 1/S_t(i))V_t(i) + 1/S_t(i)[P_tV_t](i), & \text{if } \eta_t = i; \\ V_t(i), & \text{otherwise,} \end{cases} \tag{3.27}$$

(which can be e.g. Q-learning) and assume that $\lambda_t = 0$ in Condition 1 of Corollary 3.5.2. Then ϵ_t corresponds to $\|V^* - U_t\|$, where U_t is an n-dimensional stochastic approximation process, where there is no coupling among different components:

$$U_{t+1}(i) = \begin{cases} (1 - 1/S_t(i))U_t(i) + 1/S_t(i)[P_tv^*](i), & \text{if } \eta_t = i; \\ U_t(i), & \text{otherwise,} \end{cases} \tag{3.28}$$

(cf. the proof of Theorem 3.1.3). Therefore, the Law of Iterated Logarithm applies to U_t if P_tv^* has a bounded range, showing that

$$\epsilon_t = O(\sqrt{\log(\log(t/p_{\min}))/(t/p_{\min})}),$$

where $p_{\min} = \min_i p_i$ [46]. Thus, $\|V_t - v^*\| = \|x_t\| = O(1/t^{R(1-\gamma)})$.

3.7 Discussion

Most of the results of this chapter were presented in the paper of the author and Littman [74] and the paper of the author [71]. Some ideas in the proof of the main convergence result (Section 3.1) came from the paper of Jaakkola, Jordan and Singh who were the first together with and independently of Tsitsiklis [79] who developed the connection between stochastic-approximation theory and reinforcement learning in MDPs. Our work is more similar in spirit to that of Jaakkola, et al. We believe the form of Theorem 3.1.3 makes it particularly convenient for proving the convergence of reinforcement-learning algorithms; our theorem reduces the proof of the convergence of an asynchronous process to a simpler proof of convergence of a corresponding “synchronized” one. This idea enables us to prove the convergence of asynchronous stochastic processes whose underlying synchronous process is not of the Robbins-Monro type (e.g., risk-sensitive MDPs, model-based algorithms, etc.) in unified way (see e.g., Section 4.5).

The Main Convergence Lemma, presented in Section 3.2 is central to our results. A similar, but different lemma was proposed by Jaakkola et al. [33]. Similarly, the result of Section 3.3 comes from an idea after the same paper of Jaakkola et al. However, Lemma 3.4.2 is original.

The relaxation process formalism, which will prove to be very useful in the subsequent sections, and the corresponding results of Section 3.5 are new, as are the convergence rate results. This latter result is published in [69].

Chapter 4

Off-line Learning

This section makes use of Theorem 3.1.3 to prove the convergence of various off-line learning algorithms. First we give the proofs for the basic reinforcement learning algorithms, particularly for Q-learning [85] and for the adaptive real-time dynamic programming (value iteration) algorithms. Then the extension of Q-learning to an algorithm that may use a function-approximator to store the Q values $Q_t(x, a)$, and finally the convergence of the equivalent of Q-learning for risk-sensitive models are considered. The defining characteristics of these proofs are that they make use of the *sufficient-exploration* (SE) condition which requires that in the MDP every state-action pair is visited infinitely often.

4.1 Q-learning

In Chapter 3 we presented the Q-learning algorithm, but we repeat this definition here for the convenience of the reader. Consider an MDP with the expected total-discounted cost criterion and with discount factor $0 < \gamma < 1$. Assume that at time t we are given a 4-tuple $\langle x_t, a_t, y_t, c_t \rangle$, where $x_t, y_t \in \mathcal{X}$, $a_t \in \mathcal{A}$ and $c_t \in \mathbb{R}$ are the decision maker's actual and next states, the decision maker's action, and a randomised cost received at step t, respectively. We assume that the following holds on $\langle x_t, a_t, y_t, c_t \rangle$:

ASSUMPTION 4.1.1 (SAMPLING ASSUMPTIONS) Consider a finite MDP , $(\mathcal{X}, \mathcal{A}, p, c)$. Let $\{(x_t, a_t, y_t, c_t)\}$ be a fixed stochastic process, and let $\mathcal{F}_t$ be an increasing sequence of σ-fields (the history spaces) for which

$$\{x_t, a_t, y_{t-1}, c_{t-1}, \ldots, x_0\}$$

are measurable (x_0 can be random). Assume that the followings hold:

1. $P(y_t = y | x = x_t, a = a_t, \mathcal{F}_t) = p(x, a, y)$
2. $E[c_t | x = x_t, a = a_t, y = y_t, \mathcal{F}_t] = c(x, a, y)$ and $Var[c_t | \mathcal{F}_t]$ is bounded independently of t, and

3. y_t and c_t are independent given the history $\mathcal{F}_t$: $P(y_t \in \mathcal{X}_0, c_t \in U \,|\, \mathcal{F}_t) = P(y_t \in \mathcal{X}_0 \,|\, \mathcal{F}_t) P(c_t \in U \,|\, \mathcal{F}_t)$, where $\mathcal{X}_0 \subseteq \mathcal{X}$ is arbitrary and $U \subseteq \mathbb{R}$ is measurable.

Note that one may set $x_{t+1} = y_t$, which corresponds to the situation in which the decision maker gains its experience in a real-system; this is in contrast to Monte-Carlo simulations, in which $x_{t+1} = y_t$ does not necessarily hold. The Q-learning algorithm is given by

$$Q_{t+1}(x,a) = (1 - \alpha_t(x,a)) Q_t(x,a) + \alpha_t(x,a) \left(c_t + \gamma \min_b Q_t(y_t, b) \right), \qquad (4.1)$$

where $\alpha_t(x,a) = 0$ unless $(x,a) = (x_t, a_t)$, which is intended to approximate the optimal Q function, Q^*, of the MDP.

DEFINITION 4.1.1 *We say that the process* $\{(x_t, a_t)\}$ *satisfies the* sufficient exploration (SE) condition *if* $(x = x_t, a = a_t)$ *i.o. holds w.p.1, for all* $(x,a) \in \mathcal{X} \times \mathcal{A}$.

We have the following theorem:

THEOREM 4.1.2 *Assume 4.1.1 and that the following holds w.p.1:*

1. $0 \le \alpha_t(x,a) \le 1$, $\sum_{t=0}^{\infty} \alpha_t(x,a) = \infty$, $\sum_{t=0}^{\infty} \alpha_t^2(x,a) < \infty$, *and*

2. $\alpha_t(x,a) = 0$ *if* $(x,a) \ne (x_t, a_t)$.

Then, the values defined by (4.1) converge to the optimal Q function Q^* *w.p.1.*

REMARK 4.1.3 Note that usually one defines

$$\alpha_t(x,a) = \begin{cases} 1/(n_t(x,a)+1); & \text{if } (x,a) = (x_t,a_t), \\ 0; & \text{otherwise,} \end{cases} \qquad (4.2)$$

where $n_t(x,a)$ is the number of steps the pair (x,a) was visited by the process (x_t, a_t) before time t. Then the *Robbins-Monro condition* (Condition 1 above, see [53]) on the learning rates requires every state-action be visited infinitely often, i.e., the SE condition. The theorem leaves open the question whether this can be fulfilled by any particular learning-policy and also the form of such learning policies. We shall return to this question in Chapter 5.

Proof. The proof relies on the observation that Q-learning is a relaxation process, so we may apply Corollary 3.5.2. We identify the state set of Corollary 3.5.2, $\mathcal{X}$, by the set of possible state-action pairs $\mathcal{X} \times \mathcal{A}$. If we let

$$f_t(x,a) = \begin{cases} \alpha_t(x,a), & \text{if } (x,a) = (x_t,a_t); \\ 0, & \text{otherwise,} \end{cases}$$

and

$$(P_tQ)(x,a) = c_t + \gamma \max_{b \in \mathcal{A}} Q(y_t, b)$$

(P_t does not depend on a), then we see that Conditions 1 and 2 of Corollary 3.5.2 on f_t and P_t are satisfied because of our Condition 1. So it remains to show that for a fixed function $Q \in B(\mathcal{X} \times \mathcal{A})$ the process

$$\hat{Q}_{t+1}(x,a) = (1 - \alpha_t(x,a))\hat{Q}_t(x,a) + \alpha_t(x,a)\left(c_t + \gamma \min_b Q(y_t, b)\right) \tag{4.3}$$

converges to TQ, where T is defined by

$$(TQ)(x,a) = \sum_{y \in \mathcal{X}} p(x,a,y)\left(c(x,a,y) + \gamma \min_b Q(y,b)\right). \tag{4.4}$$

Using the conditional averaging lemma (Lemma 3.5.1), this should be a routine: First, observe that the different components of $\hat{Q}_t$ are decoupled, i.e., $\hat{Q}_{t+1}(x,a)$ does not depend on $\hat{Q}_t(x',a')$ and vice versa whenever $(x,a) \neq (x',a')$. Thus, it is sufficient to prove the convergence of the one-dimensional process $\hat{Q}_t(x,a)$ to $(TQ)(x,a)$ for any fixed pair (x,a). So pick up any such pair (x,a) and identify Q_t of the lemma with $\hat{Q}_t(x,a)$ defined by (4.3). Let $\mathcal{F}_t$ be the σ-field that is adapted to

$$(x_t, a_t, \alpha_t(x,a), y_{t-1}, c_{t-1}, x_{t-1}, a_{t-1}, \alpha_{t-1}(x,a), y_{t-2}, c_{t-2}, \ldots, x_0, a_0)$$

if $t \geq 1$ and let $\mathcal{F}_0$ be adapted to (x_0, a_0), $\alpha_t = \alpha_t(x,a)$, $w_t = c_t + \gamma \min_b Q(y_t, b)$. The conditions of Lemma 3.5.1 are satisfied, namely,

1. $\mathcal{F}_t$ is an increasing sequence of σ-fields by its definition;
2. $0 \leq \alpha_t$ by the same property of $\alpha_t(x,a)$ (Condition 1);
3. α_t and w_{t-1} are $\mathcal{F}_t$ measurable because of the definition of $\mathcal{F}_t$;
4. $E[w_t | \mathcal{F}_t, \alpha_t \neq 0] = E[c_t + \gamma \min_b Q(y_t, b) | \mathcal{F}_t] = \sum_{y \in \mathcal{X}} p(x,a,y)(c(x,a,y) + \gamma \min_b Q(y,b)) = (TQ)(x,a)$ because of the first part of Condition 2;
5. $E[w_t^2 | \mathcal{F}_t]$ is uniformly bounded because y_t can take on finite values ($\mathcal{X}$ is finite), the bounded variance of c_t given the past (cf. second part of Condition 2) and the independence of c_t and y_t (Condition 3);
6. $\sum_{t=1}^{\infty} \alpha_t = \infty$ and $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$ (Condition 1).

Thus, $\hat{Q}_{t+1}(x,a)$ converges to $E[w_t | \mathcal{F}_t, \alpha_t \neq 0] = (TQ)(x,a)$, which proves the theorem. □

4.2 Model-based Learning Methods

Q-learning shows that optimal cost-to-go functions can be estimated without ever explicitly learning c and p; however, building estimates of c and p can make more efficient use of experience at the expense of additional storage and computation [47]. The parameters of c and p can be learned from experience by keeping statistics for each state-action pair on the expected cost and the proportion of transitions to each next state. In model-based reinforcement learning, c and p are estimated on-line, and the estimate of the optimal cost-to-go function is updated according to the approximate dynamic-programming operator derived from these estimates. Interestingly, although this process is not of the relaxation form, still Theorem 3.1.3 implies their convergence for a wide variety of models and methods. In order to capture this generality, let us introduce the class of MDPs in which the cost-propagation operator $\mathcal{Q}$ takes the special form

$$(\mathcal{Q}V)(x,a) = \bigoplus_{y\in\mathcal{X}}^{(x,a)} \left(c(x,a,y) + \gamma V(y)\right).$$

Here, $\bigoplus^{(x,a)} f(\cdot)$ might take the form $\sum_{y\in\mathcal{X}} p(x,a,y) f(y)$, which corresponds to the case of expected total-discounted cost criterion, or it may take the form

$$\max_{y:p(x,a,y)>0} f(y),$$

which corresponds to the case of the risk-averse worst-case total discounted cost criterion. One may easily imagine a heterogeneous criterion, when $\bigoplus^{(x,a)}$ would be of the expected-value form for some (x,a) pairs, while it would be of the worst-case criterion form for other pairs expressing a state-action dependent risk attitude of the decision maker. In general, we require only that the operation $\bigoplus^{(x,a)} : B(\mathcal{X}) \to \mathbb{R}$ should be a non-expansion with respect to the supremum-norm, i.e., that

$$\left| \bigoplus^{(x,a)} f(\cdot) - \bigoplus^{(x,a)} g(\cdot) \right| \le \|f - g\|$$

for all $f, g \in B(\mathcal{X})$. See our earlier work [44, 75], for a detailed discussion of non-expansion operators.

As was noted above, in model-based reinforcement learning, c and p are estimated by some quantities c_t and p_t. As long as every state-action pair is visited infinitely often, there are a number of simple methods for computing c_t and p_t that converge to c and p. Model-based reinforcement-learning algorithms use the latest estimates of the model-parameters (e.g. c and p) to approximate operator $\mathcal{Q}$, and in particular operator $\bigoplus$. In some cases, a bit of care is needed to insure that $\bigoplus_t$, the latest estimate of $\bigoplus$, converges to $\bigoplus$, however (here, convergence should be understood in the sense that $\| \bigoplus_t f \to \bigoplus f \| \to 0, t \to \infty$ holds for all $f \in B(\mathcal{X})$). There is no problem with expected-cost models; here the convergence of p_t to p guarantees the convergence of $\bigoplus_t^{(x,a)} f = \sum_{y \in \mathcal{X}} p_t(x,a,y) f(y)$ to $\bigoplus$. On the other hand, for worst-case-cost models, it is necessary to approximate p in a way that insures that the set of y such that $p_t(x,a,y) > 0$ converges to the set of y such that $p(x,a,y) > 0$. This can be accomplished easily, however, by setting $p_t(x,a,y) = 0$ if no transition from x to y under a has been observed.

In this framework, the adaptive real-time dynamic-programming algorithm [2] takes the form

$$V_{t+1}(x) = \begin{cases} \min_{a \in \mathcal{A}} \bigoplus_t^{(x,a)} \left(c_t(x,a,\cdot) + \gamma V_t(\cdot)\right), & \text{if } x \in \tau_t \\ V_t(x), & \text{otherwise,} \end{cases} \tag{4.5}$$

where $c_t(x,a,y)$ is the estimated cost-function and τ_t is the set of states updated at time step t. This algorithm is called "real time" if the decision maker encounters its experiences in the real system and $x_t \in \tau_t$, where x_t denotes the actual state of the decision maker at time step t, i.e., the value of the actual state is always updated.

Theorem 4.2.1 *Consider a finite* MDP *and for any pair $(x,a) \in \mathcal{X} \times \mathcal{A}$ let $\bigoplus_t^{(x,a)}, \bigoplus : B(\mathcal{X}) \to \mathbb{R}$. Assume that the following hold w.p.1:*

1. *$\bigoplus_t \to \bigoplus$ in the sense that*

$$\lim_{t\to\infty} \max_{(x,a)\in\mathcal{X}\times\mathcal{A}} \left| \bigoplus_t^{(x,a)} f(\cdot) - \bigoplus^{(x,a)} f(\cdot) \right| = 0 \quad a.s.$$

2. *$\bigoplus_t^{(x,a)}$ is a non-expansion for all $(x,a) \in \mathcal{X} \times \mathcal{A}$ and t.*

3. *$c_t(x,a,y)$ converges to $c(x,a,y)$ for all (x,a,y).*

4. *$0 \le \gamma < 1$.*

5. *Every state x is updated infinitely often (i.o.), that is $x \in \tau_t$ i.o. for all $x \in \mathcal{X}$.*

Then, V_t defined in Equation (4.5) converges to the fixed point of the operator $T : B(\mathcal{X}) \to B(\mathcal{X})$, where

$$(TV)(x) = \min_{a \in \mathcal{A}} \bigoplus_{y \in \mathcal{X}}^{(x,a)} \left(c(x,a,y) + \gamma V(y)\right).$$

Proof. We apply Theorem 3.1.3. Let the appropriate approximate dynamic-programming operator sequence $\{T_t\}$ be defined by

$$T_t(U,V)(x) = \begin{cases} \min_{a \in \mathcal{A}} \bigoplus_t^{(x,a)} \left(c_t(x,a,\cdot) + \gamma V(\cdot)\right), & \text{if } x \in \tau_t \\ U(x), & \text{otherwise.} \end{cases}$$

Using Theorem 3.4.1 it is a routine to prove that T_t approximates T. However, the following simple direct argument can also be used: Let $x \in \mathcal{X}$ and let $U_{t+1} = T_t(U_t, V)$. Then $U_{t+1}(x) = U_t(x)$ if $x \notin \tau_t$. Since in the other case, when $x \in \tau_t$, $U_{t+1}(x)$ does not depend on U_t and since $x \in \tau_t$ i.o., it is sufficient to show that $D_t = |\min_{a\in\mathcal{A}} \bigoplus_t^{(x,a)} (c_t(x,a,\cdot) + \gamma V(\cdot)) - (TV)(x)|$ converges to zero as $t \to \infty$. Now,

$$
\begin{aligned}
D_t &\le \max_{a\in\mathcal{A}} \left| \bigoplus\nolimits_t^{(x,a)} (c_t(x,a,\cdot) + \gamma V(\cdot)) - \bigoplus\nolimits^{(x,a)} (c(x,a,\cdot) + \gamma V(\cdot)) \right| \\
&\le \max_{a\in\mathcal{A}} \left| \bigoplus\nolimits_t^{(x,a)} (c_t(x,a,\cdot) + \gamma V(\cdot)) - \bigoplus\nolimits_t^{(x,a)} (c(x,a,\cdot) + \gamma V(\cdot)) \right| \\
&\quad + \max_{a\in\mathcal{A}} \left| \bigoplus\nolimits_t^{(x,a)} (c(x,a,\cdot) + \gamma V(\cdot)) - \bigoplus\nolimits^{(x,a)} (c(x,a,\cdot) + \gamma V(\cdot)) \right| \\
&\le \max_{a\in\mathcal{A}} \max_{y\in\mathcal{X}} |c_t(x,a,y) - c(x,a,y)| \\
&\quad + \max_{a\in\mathcal{A}} \left| \bigoplus\nolimits_t^{(x,a)} (c(x,a,\cdot) + \gamma V(\cdot)) - \bigoplus\nolimits^{(x,a)} (c(x,a,\cdot) + \gamma V(\cdot)) \right|,
\end{aligned}
$$

where we made use of the triangle inequality and Condition 2. The first term on the r.h.s. converges to zero because of our Condition 3, while the second term converges to zero because of our Condition 1. This, together with Condition 5 implies that $D_t \to 0$, which, since $x \in \mathcal{X}$ was arbitrary, shows that T_t indeed approximates T.

Returning to checking the conditions of Theorem 3.1.3, we find that the functions

$$G_t(x) = \begin{cases} 0, & \text{if } x \in \tau_t; \\ 1, & \text{otherwise,} \end{cases}$$

and

$$F_t(x) = \begin{cases} \gamma, & \text{if } x \in \tau_t; \\ 0, & \text{otherwise,} \end{cases}$$

satisfy the remaining conditions of Theorem 3.1.3, as long as $\bigoplus_t$ is a non-expansion for all t (which holds by Condition 2) and each x is included in the τ_t sets infinitely often (this is required by Condition 3 of Theorem 3.1.3) and the discount factor γ is less than 1 (cf. Condition 4 of Theorem 3.1.3). But these hold by our Conditions 5 and 4, respectively, and therefore the proof is complete. □

This theorem generalizes the results of [26], which deal only with the expected total-discounted cost criterion.

Note that in the above argument, $\min_{a\in\mathcal{A}}$ could have been replaced by any other non-expansion operation (this holds also for the other algorithms presented in this article). As a consequence of this, model-based methods can be used to find optimal policies in MDPs, alternating Markov games, Markov games, risk-sensitive models, and exploration-sensitive (i.e., SARSA) models [35, 57]. Also, if $c_t = c$ and $p_t = p$ for all t, this result implies that asynchronous dynamic programming converges to the optimal cost-to-go function [3, 10, 2].

4.3 Q-learning with Multi-State Updates

Now let us return to direct (or model-free) methods. Ribeiro argued that the use of available information in Q-learning is inefficient: in each step it is only the actual state and action whose Q value is reestimated [51] (i.e., only $Q_t(x, a)$ is changed). The training process is local both in space and time. If some *a priori* knowledge of the "smoothness" of the optimal Q value is available, then one can make the updates of Q-learning more efficient by introducing a so-called "spreading mechanism," which updates the Q values of state-action pairs in the "vicinity" of the actual state-action pair also.

The rule studied by Ribeiro is as follows: let Q_0 be arbitrary and

$$\begin{aligned} Q_{t+1}(z,a) &= (1-\alpha_t(z,a)s(z,a,x_t))Q_t(z,a) + \\ &\quad \alpha_t(z,a)s(z,a,x_t)\left(c_t + \gamma \min_b Q_t(y_t,b)\right), \end{aligned} \tag{4.6}$$

where $\alpha_t(z, a) \geq 0$ is the learning rate associated to the state-action pair (z, a), which is 0 of $a \neq a_t$; $s(z, a, x)$ is a fixed "similarity" function satisfying $0 \leq s(z, a, x)$; and $\langle x_t, a_t, y_t, c_t \rangle$ is the experience of the decision maker at time t. The difference between the above and the standard Q-learning rule is that here we may allow $\alpha_t(z, a) \neq 0$ even if $x_t \neq z$, i.e., the values of states different from the state actually experienced may be updated, too. The similarity function $s(z, a, x)$ weighs the relative strength at which these updates occur. (One could use a similarity which extends spreading over actions. For simplicity, we do not consider that case here.)

Our aim here is to show that, under the appropriate conditions, this learning rule converges; also, we will be able to derive a bound on how far the limit values of this rule are from the optimal Q function of the underlying MDP.

THEOREM 4.3.1 *Consider the learning rule (4.6) and assume that the sampling conditions 4.1.1 are satisfied and further*

1. *the states, x_t, are sampled from a probability distribution $p^\infty \in \Pi(\mathcal{X})$*
2. *$0 \leq s(z, a, \cdot)$ and $s(z, a, z) \not\equiv 0$,*
3. *$\alpha_t(z, a) = 0$ if $a \neq a_t$, and*

$$0 \leq \alpha_t(z,a) \leq 1, \qquad \sum_{t=0}^{\infty} \alpha_t(z,a) = \infty, \qquad \sum_{t=0}^{\infty} \alpha_t^2(z,a) < \infty.$$

Then Q_t, as given by Equation (4.6), converges to the fixed point of the operator $\hat{T} : B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X} \times \mathcal{A})$,

$$(\hat{T}Q)(z,a) = \sum_{x \in \mathcal{X}} \hat{s}(z,a,x) \sum_{y \in \mathcal{X}} p(x,a,y) \left(c(x,a,y) + \gamma \min_b Q(y,b) \right), \tag{4.7}$$

where

$$\hat{s}(z,a,x) = \frac{s(z,a,x)p^\infty(x)}{\sum_y s(z,a,y)p^\infty(y)}.$$

Note that $\hat{T}$ is a contraction with index γ since $\sum_x \hat{s}(z,a,x) = 1$ for all (z,a).
Proof. Since the process (4.6) is of the relaxation type, we apply Corollary 3.5.2. As in the proof of the convergence of Q-learning in Theorem 4.1.2, we identify the state set $\mathcal{X}$ of Corollary 3.5.2 by the set of possible state-action pairs $\mathcal{X} \times \mathcal{A}$. We let

$$(P_t Q)(x,a) = c_t + \gamma \max_{b \in \mathcal{A}} Q(y_t, b),$$

but now we set $f_t(z,a) = s(z,a,x_t)\alpha_t(z,a)$. For large enough t the conditions on f_t and P_t are satisfied by Conditions 2 and the conditions on the learning rates $\alpha_t(x,a)$, so it remains to prove that for a fixed function $Q \in B(\mathcal{X} \times \mathcal{A})$, the process

$$\begin{aligned} Q_{t+1}(z,a) &= (1 - \alpha_t(z,a)s(z,a,x_t))Q_t(z,a) + \\ &\quad \alpha_t(z,a)s(z,a,x_t) \left(c_t + \gamma \min_b Q(y_t,b) \right), \end{aligned} \tag{4.8}$$

converges to $\hat{T}Q$. We apply a modified form of the Conditional Averaging Lemma, which concerns processes of form $Q_{t+1} = (1 - \alpha_t s_t)Q_t + \alpha_t s_t w_t$ and which is presented and proved in Appendix A.2 (cf. Lemma A.2.3). This lemma states that under some bounded-variance conditions Q_t converges to $E[s_t w_t|\mathcal{F}_t]/E[s_t|\mathcal{F}_t]$, where $\mathcal{F}_t$ is an increasing sequence of σ-fields which is adapted to $\{s_{t-1}, w_{t-1}, \alpha_t\}$. In this case let $\mathcal{F}_t$ of Lemma A.2.3 be the σ-field generated by

$$(a_t, \alpha_t(x,a), y_{t-1}, c_{t-1}, x_{t-1}, \ldots, a_1, \alpha_1(x,a), y_0, c_0, x_0, a_0, \alpha_0(x,a))$$

if $t \geq 1$ and let $\mathcal{F}_0$ be adapted to $(a_0, \alpha_0(x,a))$. Easily,

$$(\hat{T}Q)(z,a) = \frac{E[s(z,a,x_t)(c_t + \gamma \min_{b \in \mathcal{A}} Q(y_t,b))|\mathcal{F}_t, \alpha_t(z,a) \neq 0]}{E[s(z,a,x_t)|\mathcal{F}_t, \alpha_t(z,a) \neq 0]}.$$

By Conditions 2 & 3 $E[s^2(z,a,x_t)(c_t+\gamma \min_a Q(y_t,a))^2|x_t,\mathcal{F}_t] < B < \infty$ for some $B > 0$. Moreover $E[s(z,a,x_t)|\mathcal{F}_t] = \sum_{x \in \mathcal{X}} p^\infty(x)s(z,a,x) > 0$ by Conditions 1 & 2, and $E[s^2(z,a,x_t)|\mathcal{F}_t] = \sum_{x \in \mathcal{X}} p^\infty(x)s^2(z,a,x) < \hat{B} < \infty$ for some $\hat{B} > 0$, by the finiteness of $\mathcal{X}$. Finally, $\alpha_t(z,a)$ obviously satisfies the assumptions of

Lemma A.2.3 and, therefore, all the conditions of the quoted lemma are satisfied. So, $Q_t(z,a)$, defined by (4.8), converges to $(\hat{T}Q)(z,a)$. □

Note that if we set $s(z,a,x) = 1$ if and only if $z = x$ and $s(z,a,x) = 0$ otherwise, then (4.6) becomes the same as the Q-learning update rule (see Equation (4.1)). However, the condition on the sampling of x_t is quite strict, so Theorem 4.3.1 is less general than Theorem 4.1.2.

It is interesting and important to ask how close $\hat{Q}^*$, the fixed point of $\hat{T}$ is to the "true" optimal Q^*, which is the fixed point of T (defined by 4.4), if s is different from the above "no-spreading" version. The answer can be derived as a corollary of the following proposition:

PROPOSITION 4.3.2 *Let $\mathcal{B}$ be a normed vector space, $T : \mathcal{B} \to \mathcal{B}$ be a contraction and $F : \mathcal{B} \to \mathcal{B}$ be a non-expansion. Further, let $\hat{T} : \mathcal{B} \to \mathcal{B}$ be defined by $\hat{T}Q = F(TQ)$, $Q \in \mathcal{B}$. Let Q^* be the fixed point of T and $\hat{Q}^*$ be the fixed point of $\hat{T}$. Then*

$$\|\hat{Q}^* - Q^*\| \leq \frac{2\inf_Q\{\,\|Q - Q^*\| \,:\, FQ = Q\,\}}{1-\gamma}. \tag{4.9}$$

Proof. Let Q denote an arbitrary fixed point of F. Then, since $\|T\hat{Q}^* - Q^*\| = \|T\hat{Q}^* - TQ^*\| \leq \gamma\|\hat{Q}^* - Q^*\|$, $\|\hat{Q}^* - Q^*\| = \|FT\hat{Q}^* - Q^*\| \leq \|FT\hat{Q}^* - Q\| + \|Q - Q^*\| = \|FT\hat{Q}^* - FQ\| + \|Q - Q^*\| \leq \|T\hat{Q}^* - Q\| + \|Q - Q^*\| \leq \|T\hat{Q}^* - Q^*\| + 2\|Q - Q^*\| \leq \gamma\|\hat{Q}^* - Q^*\| + 2\|Q - Q^*\|$. Rearranging the terms and taking the infimum over all possible Q yields (4.9). □

Inequality (4.9) helps us to define the spreading coefficients $s(z,a,x)$. Specifically, let $n > 0$ be fixed and let

$$s(z,a,x) = \begin{cases} 1, & \text{if } i/n \leq Q^*(z,a), Q^*(x,a) < (i+1)/n \text{ for some } i; \\ 0, & \text{otherwise,} \end{cases} \tag{4.10}$$

then we have $(1-\gamma)\|\hat{Q}^* - Q^*\| \leq 1/n$. Of course, the problem with this definition is that we do not know in advance the optimal Q values. However, the above definition gives us a guideline for how to define a "good" spreading function: $s(z,a,x)$ should be small (zero) for states z and x for which $Q^*(z,a)$ and $Q^*(x,a)$ differ substantially, otherwise $s(z,a,x)$ should take on larger values. In other words, it is a good idea to define $s(z,a,x)$ as a degree of the expected difference between $Q^*(z,a)$ and $Q^*(x,a)$.

Note that the above learning process is closely related to learning on aggregated states. An aggregated state is simply a subset $\mathcal{X}_i$ of $\mathcal{X}$. The idea is that the size of the Q table (which stores the $Q_t(x,a)$ values) could be reduced if we assigned a common value to all of the states in the same aggregated state $\mathcal{X}_i$. By defining the aggregated states $\{\mathcal{X}_i\}_{i=1,2,\ldots,n}$ in a clever way, one may achieve that the common value assigned to the states in $\mathcal{X}_i$ are close to the actual values of the states. In order to avoid ambiguity, the aggregated states should be disjoint, i.e., $\{\mathcal{X}_i\}$ should form a partitioning of $\mathcal{X}$. For convenience, let us introduce the equivalence relation $\approx$ among states with the definition that $x \approx y$ if and only if x and y are elements of the same aggregated state. We say that the function $Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is compatible with $\approx$ if $Q(x,a) = Q(z,a)$ holds for all $a \in \mathcal{A}$ if $x \approx z$.

Now, assume that Q_0 is initialized so that it is compatible with the "$\approx$" relation, set $s(z,a,x) = 1$ if and only if $z \approx x$ and $s(z,a,x) = 0$ otherwise and consider the function Q_t obtained by Iteration (4.6). It is easy to see by induction that for all $t \geq 0$ Q_t will be compatible with $\approx$. This enables us to rewrite Equation (4.6) in terms of the indices of the aggregated states:

$$Q_{t+1}(i,a) = \begin{cases} (1-\alpha_t(i,a))Q_t(i,a) & \\ \quad +\alpha_t(i,a)\left\{c_t + \gamma \min_a Q_t(i(y_t),a)\right\}, & \text{if } i(x_t) = i, a_t = a; \\ Q_t(i,a), & \text{otherwise.} \end{cases} \tag{4.11}$$

Here, $i(z)$ stands for the index of the aggregated state to which z belongs. Then we have the following:

Proposition 4.3.3 *Let $\underline{n} = \{1,2,\ldots,n\}$ and let $\tilde{T} : B(\underline{n} \times \mathcal{A}) \to B(\underline{n} \times \mathcal{A})$ be given by*

$$(\tilde{T}\tilde{Q})(i,a) = \sum_{x \in \mathcal{X}_i, y \in \mathcal{X}} P(\mathcal{X}_i, x) p(x,a,y) \left\{ c(x,a,y) + \gamma \min_b \tilde{Q}(i(y),b) \right\},$$

where $P(\mathcal{X}_i, x) = p^\infty(x) / \sum_{y \in \mathcal{X}_i} p^\infty(y)$. Then, under the conditions of Theorem 4.3.1, $Q_t(i,a)$ converges to the fixed point of $\tilde{T}$, where $Q_t(i,a)$ is defined by (4.11).

Proof. Since $\tilde{T}$ is a contraction its fixed point is uniquely defined. The proposition follows from Theorem 4.3.1:[1] Indeed, let $Q_0(x,a) = Q_0(i(x),a)$ for all (x,a) pairs. Then Theorem 4.3.1 yields that $Q_t(x,a)$ converges to $\hat{Q}^*(x,a)$, where $\hat{Q}^*$ is the fixed point of operator $\hat{T}$. Observe that $\hat{s}(z,a,x) = 0$ if $z \not\approx x$ and $\hat{s}(z,a,x) = P(\mathcal{X}_i(z), x)$ if $z \approx x$. The properties of $\hat{s}$ yield that if Q is compatible with the partitioning (i.e., if $Q(x,a) = Q(z,a)$ if $x \approx z$), then $\hat{T}Q$ will also be compatible with the partitioning since the right-hand side of the following equation depends only on the index of z and $\tilde{Q}(i,b)$, which is the common Q

[1] Note that Corollary 3.5.2 could also be applied directly to this rule. Another alternate way to deduce the above convergence result is to consider the learning rule over the aggregated states as a standard Q-learning rule for an induced MDP whose state space is $\{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$, whose transition probabilities are $p(\mathcal{X}_i, a, \mathcal{X}_j) = \sum_{x \in \mathcal{X}_i, y \in \mathcal{X}_j} p^\infty(x) p(x,a,y)$ and whose cost-structure is $c(\mathcal{X}_i, a, \mathcal{X}_j) = \sum_{x \in \mathcal{X}_i, y \in \mathcal{X}_j} p^\infty(x) p(x,a,y) c(x,a,y)$.

value of state-action pairs for which the state is the element of $\mathcal{X}_i$:

$$\begin{aligned}(\hat{T}Q)(z,a) &= \sum_{x\in\mathcal{X}_{i(z)},y\in\mathcal{X}} P(\mathcal{X}_{i(z)},x)p(x,a,y)\left\{c(x,a,y)+\gamma\min_b Q(y,b)\right\}\\ &= \sum_{x\in\mathcal{X}_{i(z)},y\in\mathcal{X}} P(\mathcal{X}_{i(z)},x)p(x,a,y)\left\{c(x,a,y)+\gamma\min_b \tilde{Q}(i(y),b)\right\}.\end{aligned}$$

Since $\hat{T}$ is compatible with the partitioning, its fixed point must be compatible with the partitioning, and, further the fixed point of $\tilde{T}$ and that of $\hat{T}$ are equal when we identify functions of $B(\mathcal{X}\times\mathcal{A})$ that are compatible with $\approx$ with the corresponding functions of $B(\underline{n}\times\mathcal{A})$ in the natural way. Putting the above pieces together yields that Q_t as defined in Equation (4.11) converges to the fixed point of $\tilde{T}$. $\square$

Note that Inequality (4.9) still gives an upper bound for the largest difference between $\hat{Q}^*$ and Q^*, and Equation (4.10) defines how a $1/n$ partitioning should look.

The above results can be extended to the case when the decision maker follows a fixed stationary policy that guarantees that every state-action pair is visited infinitely often and that there exists a non-vanishing limit probability distribution over the states $\mathcal{X}$. However, if the actions that are tried depend on the estimated Q_t values then there does not seem to be any simple way to ensure the convergence of Q_t unless randomized policies are used during learning whose rate of change is slower than that of the estimation process [39].

4.4 Q-learning for Markov Games

In an MDP, a single agent selects actions to minimize its expected discounted cost in a stochastic environment. A generalization of this model is the *alternating Markov game*, in which two players, the maximizer and the minimizer, take turns selecting actions—the minimizer tries to minimize its expected discounted cost, while the maximizer tries to maximize the cost to the other player. The update rule for alternating Markov games is a simple variation of Equation 4.5 in which a max replaces a min in those states in which the maximizer gets to choose the action; this makes the optimality criterion discounted minimax optimality. Theorem 4.2.1 implies the convergence of Q-learning for alternating Markov games because both min and max are both non-expansions.

Markov games are a generalization of both MDPs and alternating Markov games in which the two players simultaneously choose actions at each step in the process [42]. The basic model is defined by the tuple $\langle\mathcal{X},\mathcal{A},\mathcal{B},p,c\rangle$ and discount factor γ. As in alternating Markov games, the optimality criterion is one of discounted minimax optimality, but because the players move simultaneously, the Bellman equations take on a more complex form:

$$v^*(x) = \min_{\rho\in\Pi(\mathcal{A})} \max_{b\in\mathcal{B}} \sum_{a\in\mathcal{A}} \rho(a) \left\{ c(x, \langle a, b\rangle) + \gamma \sum_{y\in\mathcal{X}} p(x, \langle a, b\rangle, y) v^*(y) \right\}. \quad (4.12)$$

In these equations, $c(x, \langle a, b\rangle)$ is the immediate cost for the minimizer for taking action a in state x at the same time the maximizer takes action b, $p(x, \langle a, b\rangle, y)$ is the probability that state y is reached from state x when the minimizer takes action a and the maximizer takes action b, and $\Pi(\mathcal{A})$ represents the set of discrete probability distributions over the set $\mathcal{A}$. The sets $\mathcal{X}$, $\mathcal{A}$, and $\mathcal{B}$ are finite.

Optimal policies are in equilibrium, meaning that neither player has any incentive to deviate from its policy as long as its opponent adopts its policy. There is always a pair of optimal policies that are stationary [49]. Unlike MDPs and alternating Markov games, the optimal policies are sometimes stochastic; there are Markov games in which no deterministic policy is optimal (the classic playground game of "rock, paper, scissors" is of this type). The stochastic nature of optimal policies explains the need for the optimization over probability distributions in the Bellman equations, and stems from the fact that players must avoid being "second guessed" during action selection. An equivalent set of equations can be written with a stochastic choice for the maximizer, and also with the roles of the minimizer and maximizer reversed.

The obvious way to extend Q-learning to Markov games is to define the cost propagation operator $\mathcal{Q}$ analogously to the case of MDPs from the fixed point Equation (4.12). This yields the definition $\mathcal{Q} : B(\mathcal{X}) \to B(\mathcal{X} \times \Pi(\mathcal{A}))$ as

$$(\mathcal{Q}V)(x, \rho) = \max_{b\in\mathcal{B}} \sum_{a\in\mathcal{A}} \rho(a) \left\{ c(x, \langle a, b\rangle) + \gamma \sum_{y\in\mathcal{X}} p(x, \langle a, b\rangle, y) V(y) \right\}.$$

Note that $\mathcal{Q}$ is a contraction with index γ.

However, because $Q^* = \mathcal{Q}v^*$ would be a function of an infinite space (all probability distributions over the action space), we have to choose another representation. If we redefine $\mathcal{Q}$ to map functions over $\mathcal{X}$ to functions over the finite space $\mathcal{X} \times (\mathcal{A} \times \mathcal{B})$:

$$[\mathcal{Q}V](x, \langle a, b\rangle) = \left\{ c(x, \langle a, b\rangle) + \gamma \sum_{y\in\mathcal{X}} p(x, \langle a, b\rangle, y) V(y) \right\}$$

then, for $Q^* = \mathcal{Q}v^*$, the fixed-point equation (4.12) takes the form

$$v^*(y) = \min_{\rho\in\Pi(\mathcal{A})} \max_{b\in\mathcal{B}} \sum_{a\in\mathcal{A}} \rho(a) Q^*(y, \langle a, b\rangle).$$

Applying $\mathcal{Q}$ on both sides yields

$$Q^*(x, \langle a_0, b_0\rangle) = c(x, \langle a_0, b_0\rangle) + \gamma \sum_{y \in \mathcal{X}} \left\{ p(x, \langle a_0, b_0\rangle, y) \min_{\rho \in \Pi(\mathcal{A})} \max_{b \in \mathcal{B}} \sum_{a \in \mathcal{A}} \rho(a) Q^*(y, \langle a, b\rangle) \right\}. \quad (4.13)$$

The corresponding Q-learning update rule given the step t observation

$$\langle x_t, a_t, b_t, y_t, c_t\rangle$$

has the form

$$\begin{aligned} Q_{t+1}(x_t, \langle a_t, b_t\rangle) &= (1 - \alpha_t(x_t, \langle a_t, b_t\rangle))\, Q_t(x_t, \langle a_t, b_t\rangle) \quad (4.14) \\ &\quad + \alpha_t(x_t, \langle a_t, b_t\rangle) \left\{ c_t + \gamma (\textstyle\bigotimes Q_t)(y_t) \right\}, \end{aligned}$$

where

$$(\textstyle\bigotimes Q)(y) = \min_{\rho \in \Pi(\mathcal{A})} \max_{b \in \mathcal{B}} \sum_{a \in \mathcal{A}} \rho(a) Q\,(y, \langle a, b\rangle)$$

and the values of Q_t not shown in (4.14) are left unchanged. In general, it is necessary to solve a linear program to compute $(\bigotimes Q)(y)$.

THEOREM 4.4.1 *Learning rule (4.14) converges to Q^* defined by (4.13) provided that the conditions 4.1.1 hold, the learning rates satisfy* $0 \le \alpha_t(x, \langle a, b\rangle)$, $\sum_{t=0}^{\infty} \alpha_t(x, \langle a, b\rangle) = \infty$, $\sum_{t=0}^{\infty} \alpha_t^2(x, \langle a, b\rangle) < \infty$ *and* $\alpha_t(x, \langle a, b\rangle) = 0$ *if* $(x, a, b) \neq (x_t, a_t, b_t)$.

Proof. The above update rule is identical to Equation (4.1), except that actions are taken to be simultaneous pairs for both players. So a proof identical to that of Theorem 4.1.2 proves the theorem. □

It is possible that Theorem 3.1.3 can be combined with the results of [82] on solving Markov games by "fictitious play" to prove the convergence of a linear-programming-free version of Q-learning for Markov games.

4.5 Risk-sensitive Models

The optimality criterion for MDPs in which only the *worst* possible value of the next state makes a contribution to the value of a state is called the *worst-case total cost criterion*. An optimal policy under this criterion is one that avoids states for which a bad outcome is possible, even if it is not probable; for this reason, the criterion has a risk-averse quality to it. Following [28], this can be expressed by changing the expectation operator of MDPs used in the definition of the cost propagation operator $\mathcal{Q}$ to

$$(\mathcal{Q}V)(x, a) = \max_{y: p(a,x,y) > 0} \left(c(x, a, y) + \gamma V(y) \right).$$

The argument in Section 4.2 shows that model-based reinforcement learning can be used to find optimal policies in risk-sensitive models, as long as p is estimated in a way that preserves its zero vs. non-zero nature in the limit. Analogously, a Q-learning-like algorithm, called $\hat{Q}$-learning (Q-hat learning) can be shown and will be shown here to converge to optimal policies. In essence, the learning algorithm uses an update rule that is quite similar to the rule in Q-learning, but has the additional requirement that the initial Q function be set optimistically; that is, $Q_0(x,a) \leq Q^*(x,a)$ for all x and a.[2] Like Q-learning, this learning algorithm is a generalization of the LRTA* algorithm of [40] to stochastic environments.

THEOREM 4.5.1 *Assume that both $\mathcal{X}$ and $\mathcal{A}$ are finite. Let*

$$Q_{t+1}(x,a) = \begin{cases} \max\{c_t + \gamma \min_{b\in\mathcal{A}} Q_t(y_t,b), Q_t(x,a)\}; & \text{if } (x,a)=(x_t,a_t); \\ Q_t(x,a); & \text{otherwise,} \end{cases} \tag{4.15}$$

where $\langle x_t, a_t, y_t, c_t\rangle$ is the data observed by the decision maker at time t, y_t is selected at random according to $p(x,a,\cdot)$, and c_t is a random variable satisfying the following condition: If $t_n(x,a,y)$ is the subsequence of ts for which $(x,a,y) = (x_t,a_t,y_t)$, then $c_{t_n(x,a,y)} \leq c(x,a,y)$ and $\limsup_{n\to\infty} c_{t_n(x,a,y)} = c(x,a,y)$ w.p.1. Then, Q_t converges to $Q^* = \mathcal{Q}v^*$ provided that $Q_0 \leq Q^*$ and every state action pair is updated infinitely often.

Proof. The proof is an application of Theorem 3.1.3 but here the definition of the appropriate operator sequence T_t needs some more care. Let the set of "critical states" for a given (x,a) pair be given by

$$\mathcal{M}(x,a) = \{y \in \mathcal{X} \mid p(x,a,y) > 0,\ Q^*(x,a) = c(x,a,y) + \gamma \min_{b\in\mathcal{A}} Q^*(y,b)\}. \tag{4.16}$$

$\mathcal{M}(x,a)$ is non-empty, since $\mathcal{X}$ is finite. Since the costs c_t satisfy $c_{t_n(x,a,y)} \leq c(x,a,y)$, $n \geq 0$, and

$$\limsup_{n\to\infty} c_{t_n(x,a,y)} = c(x,a,y)$$

we may also assume (by possibly redefining $t_n(x,a,y)$ to become a subsequence of itself) that

$$\lim_{n\to\infty} c_{t_n(x,a,y)} = c(x,a,y). \tag{4.17}$$

[2] The necessity of this condition is clear since in the $\hat{Q}$-learning algorithm we need to estimate the operator $\max_{y:p(x,a,y)>0}$ from the observed transitions, and the underlying iterative method is consistent with $\max_{y:p(x,a,y)>0}$ only if the initial estimate is overestimating. Since we require only that T_t approximates T at Q^*, it is sufficient for the initial value of the process to satisfy $Q_0 \leq Q^*$. Note that $Q_0 = -M/(1-\gamma)$ satisfies this condition, where $M = \max_{(x,a,y)} c(x,a,y)$.

Now, let $T(x,a,y) = \{t_k(x,a,y) \mid k \geq 0\}$ and $T(x,a) = \cup_{y\in\mathcal{M}(x,a)} T(x,a,y)$. Consider the following sequence of random operators:

$$T_t(Q',Q)(x,a) = \begin{cases} \max\left(c_t + \gamma \min_{b\in\mathcal{A}} Q(y_t,b), Q'(x,a)\right); & \text{if } t \in T(x,a), \\ Q'(x,a); & \text{otherwise,} \end{cases}$$

and the sequence $Q'_0 = Q_0$ and $Q'_{t+1} = T_t(Q'_t, Q'_t)$ with the set of possible initial values taken from

$$\mathcal{F}_0 = \{Q \in B(\mathcal{X}\times\mathcal{A}) \mid Q(x,a) \leq Q^*(x,a) \text{ for all } (x,a) \in \mathcal{X}\times\mathcal{A}\}.$$

Clearly, $\mathcal{F}_0$ is invariant under T_t. We claim that it is sufficient to consider the convergence of Q'_t. Since there are no more updates (increases of value) in the sequence Q'_t than in Q_t, we have that $Q^* \geq Q_t \geq Q'_t$ and, thus, if Q'_t converges to Q^*, then so does Q_t. It is immediate that T_t approximates T at Q^* (since w.p.1 there exist an infinite number of times t such that $t \in T(x,a)$), and also that with the choice of functions

$$G_t(x,a) = \begin{cases} 0; & \text{if } (x,a) = (x_t,a_t) \text{ and } y_t \in \mathcal{M}(x,a), \\ 1; & \text{otherwise,} \end{cases}$$

Condition 1 of Theorem 3.1.3 is satisfied since $T_t(Q,Q^*)(x,a) = Q^*(x,a)$ if $(x,a) = (x_t,a_t)$ and $y_t \in \mathcal{M}(x,a)$.

Now, let us bound $|T_t(Q',Q)(x,a) - T_t(Q',Q^*)(x,a)|$. For this assume first that $t \in T(x,a)$. This means that $(x,a) = (x_t,a_t)$ and $y_t \in \mathcal{M}(x,a)$. Assume that $Q, Q' \in \mathcal{F}_0$, i.e., $Q, Q' \leq Q^*$. Then

$$\begin{aligned} & |T_t(Q',Q)(x,a) - T_t(Q',Q^*)(x,a)| \leq \\ & \qquad \left(c(x,a,y_t) + \gamma \min_{b\in\mathcal{A}} Q^*(y_t,b)\right) \\ & \qquad - \max\left(c_t + \gamma \min_{b\in\mathcal{A}} Q(y_t,b), Q'(x,a)\right) \\ & \leq \left(c(x,a,y_t) + \gamma \min_{b\in\mathcal{A}} Q^*(y_t,b)\right) - \left(c_t + \gamma \min_{b\in\mathcal{A}} Q(y_t,b)\right) \\ & \leq \gamma\|Q^* - Q\| + |c(x,a,y_t) - c_t|. \end{aligned} \tag{4.18}$$

where we have used that $T_t(Q',Q^*)(x,a) \geq T_t(Q',Q)(x,a)$ (since T_t is monotone in its second variable) and that

$$\begin{aligned} T_t(Q',Q^*)(x,a) &\leq \max(c(x,a,y_t) + \gamma \min_{b\in\mathcal{A}} Q^*(y_t,b), Q'(x,a)) \\ &\leq \max(c(x,a,y_t) + \gamma \min_{b\in\mathcal{A}} Q^*(y_t,b), Q^*(x,a)) \\ &= c(x,a,y_t) + \gamma \min_{b\in\mathcal{A}} Q^*(y_t,b) \end{aligned}$$

which holds since $Q' \leq Q^*$ and $y_t \in \mathcal{M}(x,a)$.

Let $\sigma_t(x,a) = |c(x,a,y_t) - c_t|$. Note that by (4.17),

$$\lim_{t\to\infty, t\in T(x,a)} \sigma_t(x,a) = 0$$

w.p.1. In the other case (when $t \notin T(x,a)$), $|T_t(Q',Q)(x,a) - T_t(Q',Q^*)(x,a)| = 0$. Therefore

$$|T_t(Q',Q)(x,a) - T_t(Q',Q^*)(x,a)| \leq F_t(x,a)(\|Q - Q^*\| + \lambda_t),$$

where

$$F_t(x,a) = \begin{cases} \gamma; & \text{if } t \in T(x,a), \\ 0; & \text{otherwise,} \end{cases}$$

and $\lambda_t = \sigma_t(x_t,a_t)/\gamma$ if $t \in T(x,a)$, and $\lambda_t = 0$, otherwise. Thus, we satisfy Condition 2 of Theorem 3.1.3 since λ_t converges to zero w.p.1.

Condition 3 of the same theorem is satisfied if and only if $t \in T(x,a)$ i.o. But this must hold due to the assumptions on the sampling of (x_t,a_t) and y_t, and since $p(x,a,y) > 0$ for all $y \in \mathcal{M}(x,a)$. Finally, Condition 4 is satisfied, since for all t, $F_t(x) = \gamma(1 - G_t(x))$, and so Theorem 3.1.3 yields that "$\hat{Q}$-learning" converges to Q^* w.p.1. $\square$

Note that this process is not of the relaxation-type (cf. Equation (3.18)) but Theorem 3.1.3 still applies to.

4.6 Discussion

The Q-learning algorithm analyzed in Section 4.1 was described first by Watkins in his thesis [84]. He presented a simulation-type of proof. The connection to stochastic approximation was first observed by Jaakkola, Jordan, and Singh [33] and Tsitsiklis [79]. The proof presented here was first presented in the paper of Michael Littman and the author [74].

Model-based learning methods for expected value models (Section 4.2) were proposed by Barto et al. [1]. The first theoretical analysis of this algorithm is due to Gullapalli and Barto [27]. The proof presented here allows a more general treatment than that earlier and can be found in the paper of Michael Littman and the author [74].

In Section 4.3 several generalization of Q-learning are treated at once. The algorithm presented here was proposed by Ribeiro [51] and a variant was analyzed theoretically by Ribeiro and the author in [52].[3] The proof presented here is

[3] The variant considered the case when the time-independent spreading coefficients of (4.6) are replaced by time-dependent coefficients and also a function of the actual action, that is the spreading coefficient of (z,a) at time t is given by $s_t(z,a,x_t,a_t)$. By using Theorem 3.1.3 Ribeiro and the author have shown that if the convergence rate of $s_t(z,a,x_t,a_t) - \delta(z = x_t, a = a_t)$ to zero is not slower than that of $\alpha_t(z,a)$, and the expected time between two successive visits is bounded for any state-action pair then Q_t, as defined by the appropriately modified version of Equation (4.6), converges to the true optimal Q-function, Q^* [52].

new, although Gordon [24] and Tsitsiklis and Van Roy [80] commented on these questions in the context of using function-approximators together with value iteration. Proposition 4.3.2 is due to Gordon [24, Theorem 6.2]. The idea of state-aggregation is discussed in the context of MDPs in [59] & [8] and the context of RL in [61].

The basic model of Markov-games presented in Section 4.4 was developed by Shapley [60] (see also [17]). The corresponding Q-learning update rule was proposed by Littman [42]. The theorem and the proof were first presented in the paper of Littman and the author [74].

It is hard to trace back the origin of risk-sensitive models (Section 4.5). Bellman already described a multistep minimax model in his book [4]. Bertsekas and Shreve [9] also analyzed this criterion as a nice example for their abstract dynamic programming framework. Recently, Heger [28, 29] considered the learning aspects of this criterion. He proposed the $\hat{Q}$-learning algorithm whose proof is presented here which appeared first in the paper of Littman and the author [74]. Heger also proved the convergence of this algorithm in his thesis [30]. The author also proved rigorously that the decomposition of this risk-sensitive criterion through $\mathcal{Q}$ is valid [68] and that estimating a model does not conflict with acting myopically under this criterion [67, 70].

Chapter 5

On-line Learning

All the algorithms of the previous chapter required that *every state-action pair is visited infinitely often* – which is called the *"sufficient exploration" (SE)* assumption. In this chapter we assume that the decision maker interacts with the system through an appropriate learning policy, i.e. learning is *on-line*. The learning policy determines then the actual visits to the state-action pairs and thus should be chosen in a way that the SE condition is met. For example, when the environment is a communicating MDP (i.e., each state can be reached from every other state with positive probability) then a random-walk learning-policy (i.e., when the actions are chosen randomly in each step) satisfies this condition. However, the random-walk learning policy is not very efficient in that the decision maker will never behave optimally. Think about choosing in each step the action that seems to be the best according to the momentary knowledge of the decision maker. For certain environments such a purely exploiting learning-policy does not satisfy the SE condition and neither the estimate of the cost-to-go function nor the learning-policy converges to optimality. Three questions arise.

1. What are the environments when pure exploitation yields convergence?

2. What is a universal form of learning-policies which results in the convergence of both the cost-to-go function estimates to the optimal cost-to-go function and the learning-policy to the optimal policy?

3. What happens if the SE condition is not met: do we still have convergence (to somewhere)?

In this chapter we try to answer the last two question and some related question. The first question was partially answered by the author in [67, 70] where it was shown that under minimax optimality pure exploitation does not conflict with convergence to optimality, so we considere here the remaining two questions.

5.1 Convergence of Q-learning without the SE Condition

In the subsequent results the following sampling assumptions will be standard:

Assumption 5.1.1 (On-line Sampling Assumptions) Let us consider a finite MDP , $(\mathcal{X}, \mathcal{A}, p, c)$, and let $\pi = (\pi_0, \pi_1, \ldots, \pi_t, \ldots)$ be a fixed randomized policy, where $\pi_t : \mathcal{X} \times (\mathbb{R} \times \mathcal{A} \times \mathcal{X})^t \to \Pi(A)$, and let (x_t, a_t, c_t) be the stochastic process over $\mathcal{X} \times \mathcal{A}$, induced by π and the MDP, which is constructed recursively in a way that is satisfies the followings: If $\{\mathcal{F}_t\}$ denotes an increasing sequence of σ-fields adapted to $(x_t, c_{t-1}, a_{t-1}, x_{t-1}, \ldots, c_0, a_0, x_0)$ then

1. $P(x_{t+1} = y | x = x_t, a = a_t, \mathcal{F}_t) = p(x, a, y)$;
2. $P(a_t = a | \mathcal{F}_t) = \pi(x_t, c_{t-1} a_{t-1} x_{t-1} \ldots c_0 a_0 x_0)(a)$;
3. $E[c_t | x = x_t, a = a_t, y = x_{t+1}, \mathcal{F}_t] = c(x, a, y)$ and $Var[c_t | \mathcal{F}_t]$ is bounded;
4. x_{t+1} and c_t are independent given $\sigma(\mathcal{F}_t, a_t)$.

We further assume that the learning-rates $\alpha_t(x, a)$ in the stochastic approximation processes considered in the following are of the form

$$\alpha_t(x,a) = \begin{cases} 1/(n_t(x,a)+1); & \text{if } (x,a) = (x_t, a_t), \\ 0; & \text{otherwise}, \end{cases} \tag{5.1}$$

where $n_t(x, a)$ denotes the number of times the pair (x, a) was visited before time t by the process $\{(x_t, a_t)\}$.

Definition 5.1.1 *We say that the policy π satisfies the* sufficient exploration (SE) condition *(or is sufficiently exploring) if, for the process $\{(x_t, a_t)\}$ induced by π, $(x = x_t, a = a_t)$ i.o. holds w.p.1, for all $(x, a) \in \mathcal{X} \times \mathcal{A}$.*

A reasonable weakening of the SE condition is when in states visited infinitely often every action is tried infinitely often – this will be called the *weak SE condition (WSE).*

Definition 5.1.2 *We say that the policy π satisfies the* weak sufficient exploration (WSE) condition *(or is weakly sufficiently exploring) if for all $(x, a) \in \mathcal{X} \times \mathcal{A}$, and for the process $\{(x_t, a_t)\}$ induced by π, $(x = x_t, a = a_t)$ i.o. holds on $\{\omega : x = x_t(\omega)\text{i.o.}\}$ a.s. We will denote by X_∞ the set of states visited i.o. by x_t, i.e., $X_\infty(\omega) = \{\, x \in \mathcal{X} \mid x = x_t(\omega)\,\text{i.o.}\,\}$.*

When designing a given learning policy we shall write $\pi(a_t = a | \mathcal{F}_t)$ to denote that given the information/history $\mathcal{F}_t$, action a is chosen with a given (a-priori) probability. We begin by a simple observation:

PROPOSITION 5.1.3 *Consider Q-learning given by (4.1) and assume 5.1.1. If the policy π is sufficiently exploring then Q_t converges to Q^* w.p.1.*

Proof. The proposition follows immediately from Theorem 4.1.2 since

$$\sum_{t=0}^{\infty} \alpha_t(x,a) = \sum_{n=0}^{\infty} 1/(n+1) = \infty$$

w.p.1 and

$$\sum_{t=0}^{\infty} \alpha_t^2(x,a) = \sum_{n=0}^{\infty} 1/(n+1)^2 < \infty$$

w.p.1. □

The following theorem shows that the WSE condition is sufficient for convergence to optimality to hold in a restricted sense.

THEOREM 5.1.4 *Consider Q-learning given by (4.1) and assume 5.1.1 and that the learning policy meets the WSE condition. Then for all $(x,a) \in \mathcal{X} \times \mathcal{A}$ and almost all ω s.t. $x \in \mathcal{X}_\infty(\omega)$, $\lim_{t\to\infty} Q_t(x,a)(\omega) = Q^*(x,a)$, i.e., $Q_t|_{\mathcal{X}_\infty \times \mathcal{A}} \to Q^*|_{\mathcal{X}_\infty \times \mathcal{A}}$, as $t \to \infty$ w.p.1.*

Proof. Let $\mathcal{X}_1, \ldots, \mathcal{X}_k \subset \mathcal{X}$ denote those subsets of $\mathcal{X}$ for which $P(\mathcal{X}_\infty = \mathcal{X}_i) > 0$ and pick up some i, $1 \le i \le k$. Then for all $(x,y) \in \mathcal{X}_i \times (\mathcal{X} \setminus \mathcal{X}_i)$ and for all $a \in \mathcal{A}$, $p(x,a,y) = 0$ since otherwise by the WSE property of the control-policy we would have $P(\mathcal{X}_\infty = X_i) = P(\mathcal{X}_\infty = X_i, y \in X_\infty) + P(\mathcal{X}_\infty = X_i, y \notin X_\infty) = 0$. Now, let us restrict our attention to the event $\{X_\infty = X_i\}$. By the above property the restriction of the original MDP to $\mathcal{X}_i$ is well-defined. Let Q_i^* denote the corresponding optimal Q-function. The Bellman Optimality Equation for Q_i^* gives

$$Q_i^*(x,a) = \sum_{y \in \mathcal{X}_i} p(x,a,y)\Big(c(x,a,y) + \gamma \min_{b \in A} Q_i^*(y,b)\Big), \quad \forall (x,a) \in \mathcal{X}_i \times \mathcal{A}$$

and for Q^* gives

$$\begin{aligned} Q^*(x,a) &= \sum_{y \in \mathcal{X}} p(x,a,y)\Big(c(x,a,y) + \gamma \min_{b \in A} Q^*(y,b)\Big) \\ &= \sum_{y \in \mathcal{X}_i} p(x,a,y)\Big(c(x,a,y) + \gamma \min_{b \in A} Q^*(y,b)\Big), \quad \forall (x,a) \in \mathcal{X}_i \times \mathcal{A}. \end{aligned}$$

So both Q_i^* and Q^* satisfy the same fixed point equation. Since $\gamma < 1$ this fixed point equation has a unique solution and so we must have

$$Q_i^* = Q^*|_{\mathcal{X}_i \times \mathcal{A}}. \tag{5.2}$$

Let $\tau \in \mathbb{N}$ and set $\Omega_{i,\tau} = \{\omega : x_t \in \mathcal{X}_i, t \geq \tau\}$. Then Q_t converges to Q_i^* a.s. on $\Omega_{i,\tau}$ by Proposition 5.1.3 since on $\Omega_{i,\tau}$ and after time τ the Q-learning algorithm works as a Q-learning algorithm for the MDP whose state space is restricted to $\mathcal{X}_i$, and the SE condition for this restricted state space is satisfied on $\Omega_{i,\tau}$. Since $\cup_{\tau\in\mathbb{N}}\Omega_{i,\tau} = \{\omega : \mathcal{X}_\infty(\omega) = \mathcal{X}_i\}$ a.s., so $Q_t(\omega)|_{\mathcal{X}_i\times\mathcal{A}} \to Q_i^*$, $t \to \infty$ holds a.e. on $\mathcal{X}_i = \mathcal{X}_\infty$, and the theorem follows by (5.2). □

Note that without any condition put on the learning policy convergence of Q-learning could be still proved (to some random function), but in general there is no simple relation between Q^* and the limit values.

5.2 Asymptotically Optimal Learning Strategies

In order to present sufficient conditions for asymptotic optimality of learning strategies we need to specialize the form of on-line learning policies:

Assumption 5.2.1 (Statistics Based Learning Policies) Let us assume that Assumptions 5.2.1 hold and that also the following hold: Let s_t be some $\mathcal{F}_t$-measurable statistics which is computed recursively on the basis of (x_t, a_t, c_t): $s_t = s_t(x_t, c_{t-1}a_{t-1}x_{t-1}\dots c_0a_0x_0)$.[1] Then, assume that the learning policy π depends on the history of the decision process only through s_t:

$$\begin{aligned} P(a_t = a|\mathcal{F}_t) &= \pi(x_t, c_{t-1}a_{t-1}x_{t-1}\dots c_0a_0x_0)(a) &\quad (5.3)\\ &= \pi'(a|x_t, s_t) &\quad (5.4)\end{aligned}$$

In the following we will identify π with π'.

Note that it is sufficient to specify π at times when $x_t = x$, i.e., as a function $\pi(a|k, s, x) = \pi(a|x, s_{t_k(x)})$, where $t_k(x)$ denotes the time when x_t visits x the kth time. If s'_t is any other statistics of $x_t, c_{t-1}, a_{t-1}, x_{t-1}, \dots, c_0, a_0, x_0$ then

$$P(a_{t_k(x)} = a|x, s'_{t_0(x)}, s'_{t_1(x)}, \dots, s_{t_k(x)}) = P(a_{t_k(x)} = a|x, s_{t_k(x)}) \qquad (5.5)$$

since, by assumption, the probability of choosing action a in state x at any time given the history of process depends only on the statistics s at that time and the learning policy $\pi(\cdot)$.

The following lemma gives a sufficient condition for a learning policy to satisfy the WSE condition.

[1] In the examples below this statistics will be composed of the "table" $Q_t(\cdot,\cdot)$, $\{n_t(x)\}_{x\in\mathcal{X}}$, and $\{n_t(x,a)\}_{(x,a)\in\mathcal{X}\times\mathcal{A}}$. Here $n_t(x)$ is the number of times state x was visited by $\{x_j\}_{j\leq t}$.

LEMMA 5.2.1 *Assume 5.2.1 and that the policy* π *is given in terms of the statistics* s_t *and the probabilities* $\pi(a|x_t, s_t)$ *of choosing action* a*. Let* $t_k(x)$ *denote the time when* x_t *visits* x *the* k*th time. If for all* $x \in \mathcal{X}$

$$\sum_{k=1}^{\infty} \pi(a|x, s_{t_k(x)}) = \infty \quad \text{a.s.} \tag{5.6}$$

then π *satisfies the WSE condition.*

Below we will give two examples when $\pi(a|x_t, s_t)$ is given explicitly and condition (5.6) is verifiable.
Proof. Note that the WSE condition is equivalent to that for all $x \in \mathcal{X}$ and every action $a \in \mathcal{A}$, and for almost every ω s.t. $x \in \mathcal{X}_\infty(\omega)$ there holds that $\{a = a_{t_k(x)}\}$ i.o.

For the proof we need the following lemma whose proof follows the same lines as the proof of Corollary 5.29 in [14, pp.96].

LEMMA 5.2.2 (EXTENDED BOREL-CANTELLI LEMMA) *Let* $\hat{\mathcal{F}}_k$ *be an increasing sequence of* σ*-fields and let* A_k *be* $\hat{\mathcal{F}}_k$*-measurable. Then*

$$\{\, \omega \,:\, \sum_{k=1}^{\infty} P(A_k|\hat{\mathcal{F}}_{k-1}) = \infty \,\} = \{\, \omega \,:\, \omega \in A_k \text{ i.o.} \,\}.$$

holds w.p.1.

Assume that the given policy is used for control and let $x \in \mathcal{X}$ be fixed. Then for all ω s.t. $x \in \mathcal{X}_\infty(\omega)$ the sequence $t_k(x)$ is well-defined and can be continued up to infinity. In Lemma 5.2.2 choose $A_k = \{a_{t_k(x)} = a\}$ and $\hat{\mathcal{F}}_k$ be the σ-algebra generated by $s_{t_0(x)}, s_{t_1(x)}, \ldots, s_{t_{k+1}(x)}$ and $A_0, A_1, \ldots, A_k$. Then A_k is $\hat{\mathcal{F}}_k$-measurable and thus by Equation (5.5):

$$\begin{aligned} P(A_k|\hat{\mathcal{F}}_{k-1}) &= P(a_{t_k(x)} = a | s_{t_0(x)}, s_{t_1(x)}, \ldots, s_{t_k(x)}, A_0, A_1, \ldots, A_{k-1}) \\ &= \pi(a|x, s_{t_k(x)}). \end{aligned}$$

So a.e. on $\{x \in \mathcal{X}^\infty\}$, $\sum_{k=1}^{\infty} P(A_k|\hat{\mathcal{F}}_{k-1}) = \infty$ and thus by Lemma 5.2.2 also $\{a = a_{t_k(x)}\}$ i.o. holds a.e. on $\{x \in \mathcal{X}^\infty\}$. □

DEFINITION 5.2.3 *A policy* π *is said to be* asymptotically optimal *if for all* $x \in X$*,* $\lim_{t\to\infty} P(a_t \in \operatorname{Argmin} Q^*(x,a)|\mathcal{F}_t, x = x_t) = 1$ *holds w.p.1.*

The following theorem concerns policies which achieve asymptotic optimality by becoming greedy w.r.t. Q_t sufficiently slowly, where $Q_t \in B(\mathcal{X} \times \mathcal{A})$ is the t^{th} estimate of the optimal action-value function computed (e.g.) by Q-learning. In order to state the theorem we need $\pi(a_t \in \operatorname{Argmin}_{a\in\mathcal{A}} Q_t(x,a)|x = x_t, \mathcal{F}_t) \stackrel{\text{def}}{=}$

$\sum_{a\in \mathrm{Argmin}_{b\in\mathcal{A}} Q_t(x,b)} \pi(a|x=x_t,\mathcal{F}_t)$, so $\pi(a_t \in \mathrm{Argmin}_{a\in\mathcal{A}} Q_t(x,a)|x=x_t,\mathcal{F}_t)$ can be interpreted as the conditional probability of choosing a myopic action given the history. This is well defined because of the measurability conditions on π and since Q_t is $\mathcal{F}_t$-measurable.

THEOREM 5.2.4 *Consider Q-learning given by (4.1) and assume 5.2.1. Let π be a policy such that for all $x \in \mathcal{X}$:*

$$\sum_{k=1}^{\infty} \pi(a|x, s_{t_k(x)}) = \infty \quad \text{and} \tag{5.7}$$

$$\lim_{t\to\infty} \pi\left(a_t \in \operatorname*{Argmin}_{a\in\mathcal{A}} Q_t(x,a) \middle| x = x_t, \mathcal{F}_t\right) = 1. \tag{5.8}$$

Then π is asymptotically optimal.

Proof. Condition 5.7 ensures that the condition of Lemma 5.2.1 is satisfied which shows that the conditions of Theorem 5.1.4 are satisfied. This latter Theorem yields that $Q_t|_{\mathcal{X}_\infty\times\mathcal{A}} \to Q^*|_{\mathcal{X}_\infty\times\mathcal{A}}$ as t tends to infinity w.p.1 and so by the finiteness of $\mathcal{A}$ and (5.8) we get the asymptotic optimality of π. $\square$

It is immediate that the learning policy that, on the kth visit to state x chooses a myopic action with a probability proportional to $1-1/k$, and chooses non-myopic actions randomly, i.e., for which

$$\pi(a|x, s_{t_k(x)}) = \begin{cases} \frac{1}{C_{t_k}}\left(1-\frac{1}{k+1}\right); & \text{if } a \in \mathrm{Argmin}_{a\in\mathcal{A}} Q_{t_k(x)}(x,a), \\ 1-\frac{1}{C_{t_k}}\left(1-\frac{1}{k+1}\right); & \text{otherwise,} \end{cases}$$

where $C_t(x)$ is the cardinality of $\mathrm{Argmin}_{a\in\mathcal{A}} Q_t(x,a)$, satisfies the conditions of Theorem 5.2.4. Here $s_t = (Q_t, \{n_t(x)\}_{x\in\mathcal{X}})$. Note that here the whole history of the process up to time t is projected into the function Q_t.

The following corollary gives conditions under which the Boltzmann-exploration policy becomes asymptotically optimal.

COROLLARY 5.2.5 *Consider Q-learning given by (4.1) and assume 5.2.1 and that the random immediate costs, c_t, are uniformly bounded by some $R > 0$ ($c_t \in [-R, R]$ for all $t \geq 0$) and the learning policy has the form of the so-called Boltzmann-exploration policy, i.e.,*

$$\pi(a|x=x_t,\mathcal{F}_t) = \frac{1}{N_t(x)} \exp\left(\frac{-Q_t(x,a)}{T_t(x)}\right),$$

where

$$N_t(x) = \sum_{b\in\mathcal{A}} \exp\left(\frac{-Q_t(x,b)}{T_t(x)}\right)$$

is the partition function. Assume that the "temperature" parameter $T_t(x)$ $(0 < T_t(x))$ *is a function of the number of visits of state* x, $n_t(x)$, *i.e.,*

$$T_t(x) = T(n_t(x)).$$

If

$$T(v) = \frac{2R}{(1-\gamma)\log(v)}, \tag{5.9}$$

where $0 < \gamma < 1$ *is the discount factor of the* MDP *then the Boltzmann-exploration policy is asymptotically optimal.*

Proof. Since $T_t(x) \to 0$ a.e. on $x \in \mathcal{X}^\infty$, (5.8) is satisfied. First note that since the random immediate costs are uniformly bounded Q_t remains bounded by $R/(1-\gamma)$ w.p.1. (this comes directly from the form of the Q-learning equation). $R/(1-\gamma)$ is therefore an upper bound on $\max_b Q_t(x,b)$ and similarly $-R/(1-\gamma)$ is a lower bound on $\min_b Q_t(x,b)$:

$$\pi(a|x, s_{t_k(x)}) \geq \frac{1}{1 + (m-1)\exp(\frac{2R}{(1-\gamma)T(k)})}, \tag{5.10}$$

where m is the number of actions (we used the identities $n_{t_k(x)} = k$ and $T_{t_k(x)}(x) = T(n_{t_k(x)}) = T(k)$.) Substituting (5.9) yields

$$\pi(a|x, s_{t_k(x)}) \geq \frac{1}{(m-1)k+1}$$

and thus (5.7) is also satisfied. Therefore, Theorem 5.2.4 yields that π is asymptotically optimal. $\square$

Since the WSE condition is only an asymptotic requirement and the bound $R/(1-\gamma)$ may be asymptotically strengthened to $(|\max_{a\in\mathcal{A}} Q^*(x,a)| + \varepsilon)$, it is plausible that Boltzmann-learning with $T_x(v) = \frac{2(|\max_{a\in\mathcal{A}} Q^*(x,a)|+a_k)}{\log v}$ satisfies the WSE condition with some $a_v \to 0$, $v \to \infty$. However, a lower bound on the rate of convergence of a_v towards zero is not trivial to obtain. Another approach is to make the temperature T_t depend explicitly on the Q_t values. This way one can give asymptotically optimally policies with the assumption on the bounded range of the immediate costs removed [62].

Note that although these theorems enable us to construct asymptotically optimal learning policies, the convergence rate of such general learning policies to optimality could still be optimized. Policies which optimize this convergence rate are called *optimal learning policies*. Some types of optimal learning policies, which minimize the total expected loss caused by adaption as compared to the performance of an optimal policy, has recently been determined for the *long-run average cost* criterion [15, 25]. We would like to note that this optimization property is an asymptotic requirement which fits the long-run average cost criterion, but does not fit the discounted cost criterion where the transient performance plays an important role. A suitable concept of optimality for the discounted case is discussed in [58]. Note that once we know that the algorithms are asymptotically optimal the ODE method to stochastic approximations (see, e.g. [45, 41, 5]) could be used for further analysis, e.g. to derive central limit type of theorems.

REMARK 5.2.6 It is important to note that in the proofs of this and the previous sections we have not exploited the actual form of Q-learning. The only property that was used is that the for any process $\{(x_t, a_t)\}$ which satisfies the SE condition the Q_t values should converge to Q^* w.p.1. This means that we may use any algorithm that satisfies this property to generate the Q_t values an the results of these two sections will still hold. Moreover, the same holds if Q^* is replaced by any other function (corresponding to a different optimization criterion) which is the fixed point of a contraction of form $(TQ)(x,a) = \sum_{y\in\mathcal{X}} p(x,a,y)(T'Q)(x,a,y)$, where $T' : B(\mathcal{X}\times\mathcal{A}) \to B(\mathcal{X}\times\mathcal{A}\times\mathcal{X})$ is another contraction. This shows that the results of this chapter remain true for RL algorithms with multi-state updates (see Section 4.3).

5.3 Exploration-Sensitive Learning

In this section we investigate in some sense an opposite situation to that of the previous section, in that here the probability of the selection of exploring actions (non-greedy actions w.r.t. Q_t) will be kept constant. Such learning policies could respond to the changes in a non-stationary environment much faster than the ones with decaying exploration. Nevertheless, we investigate their properties in stationary environments so as to capture the price of their faster response times.

We will consider semi-uniform learning policies:

DEFINITION 5.3.1 *Given a Q function and a small value $\varepsilon > 0$, when in state x, take the action* $\operatorname{Argmin} Q(x,a)$ *with probability* $1-\varepsilon$ *and a random action from* $\mathcal{A}$ *with probability* ε *(i.e., the probability of choosing an action from* $\operatorname{Argmin} Q(x,a)$ *is* $(1-\varepsilon) + \varepsilon/|\operatorname{Argmin} Q(x,a)|$*). Such a learning-policy is referred to as* $(1-\varepsilon)$-greedy *w.r.t. Q, and is denoted by $\pi(Q,\varepsilon)$ ($\pi(Q,\varepsilon) : \mathcal{X} \to \Pi(\mathcal{A})$).*

The following theorem gives the form of the optimal semi-uniform policy.

THEOREM 5.3.2 *Let*

$$Q^*(x,a) = \sum_{y\in\mathcal{X}} p(x,a,y)\Big(c(x,a,y) + \gamma\varepsilon\,\frac{\sum_b Q^*(y,b)}{|A|} + (1-\varepsilon)\min_b Q^*(y,b)\Big). \quad (5.11)$$

Then $v_{\pi(Q^*,\varepsilon)} \le v_{\pi(Q,\varepsilon)}$ for all $Q \in B(\mathcal{X}\times\mathcal{A})$.

Proof. We prove a slightly more general version form of this theorem. For this let us define the operator $P : B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X})$ by

$$(Pf)(x) = \inf_{\pi \in P_0} (P_\pi f)(x),$$

where P_0 is a fixed set of (randomized) stationary policies, π denotes a random stationary policy ($\pi : \mathcal{X} \to \Pi(\mathcal{A})$, or in an equivalent representation $\pi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, where $\pi(x, \cdot)$ is a probability distribution) and $(P_\pi h)(x) = \sum_a \pi(x,a)h(x,a)$. Note that P is a non-expansion [74]. Let $T : B(\mathcal{X}) \to B(\mathcal{X})$ be given by $T = P\,\mathcal{Q}$, where $\mathcal{Q} : B(\mathcal{X}) \to B(\mathcal{X} \times \mathcal{A})$ is defined in the usual way:

$$(\mathcal{Q}V)(x,a) = \sum_{y \in \mathcal{X}} \Big(c(x,a,y) + \gamma p(x,a,y) V(y) \Big).$$

Since $\mathcal{Q}$ is a contraction with index γ and P is a non-expansion we see that T is also a contraction with index γ. More importantly, $TV \le T_\pi V = P_\pi \mathcal{Q} V$ for all $\pi \in P$ and T is monotone. From these it follows by the reasoning of Corollary 1.5.5 that the optimal cost-to-go function given by

$$v^*(x) = \inf_{\pi \in P_0} v_\pi(x)$$

is the fixed point of T and the T-greedy policy w.r.t. the optimal cost-to-go function is optimal, i.e., it evaluates to v^*.

For the class of greedy policies considered in the previous paragraph

$$(P_{\pi_Q} Q)(x) = \varepsilon \frac{\sum_{a \in \mathcal{A}} Q(x,a)}{|A|} + (1-\varepsilon) \min_{a \in \mathcal{A}} Q(x,a),$$

so if we set $Q^* = \mathcal{Q} v^*$ then $v^* = PQ^* = P_{\pi_Q} Q^*$ which yields that Q^* indeed satisfies (5.11). $\square$

Corollary 5.3.3 *Assume 5.2.1, let $\varepsilon > 0$ be fixed and consider the learning equation*

$$\begin{aligned} Q_{t+1}(x_t,a_t) &= (1-\alpha_t(x_t,a_t))Q_t(x_t,a_t) + \\ &\quad \alpha_t(x_t,a_t) \left\{ c_t + \gamma(\varepsilon \frac{\sum_b Q_t(y_t,b)}{|A|} + (1-\varepsilon) \min_b Q_t(y_t,b)) \right\}, \end{aligned} \tag{5.12}$$

where a_t is selected by the ε-greedy policy w.r.t. Q_t, π. (Values not shown in (5.12) are left unchanged.) Then, for any fixed $x \in \mathcal{X}$, a.s. on $\{x \in \mathcal{X}_\infty\}$ and for all $a \in \mathcal{A}$, $Q_t(x,a) \to Q^*(x,a)$, where Q^* is the solution of (5.11) and $P(a_t = a | x = x_t, \mathcal{F}_t) \to \pi(Q^*, \varepsilon)(x,a)$, i.e., π is asymptotically optimal in the class of ε-greedy policies.

Proof. Note that the "natural" learning rule corresponding to the fixed point equation (5.11) (as follows from the constructions of Chapter 4) is

$$Q_{t+1}(x_t,a_t) = (1-\alpha_t(x_t,a_t))Q_t(x_t,a_t) + \alpha_t(x_t,a_t)\left\{c_t + \gamma \min_{\pi\in P_0}(T_\pi Q_t)(y_t)\right\}, \tag{5.13}$$

where $P_0 = \{\pi(Q,\varepsilon)\,|Q \in B(\mathcal{X}\times\mathcal{A})\}$ and the required convergence of Q_t defined by this learning rule to the fixed point of (5.11) follows by Remark 5.2.6. Since the Lemma of Three-Years-Old yields that

$$\min_{\pi\in P_0}(T_\pi Q)(x) = \varepsilon\,\frac{\sum_a Q(x,a)}{|A|} + (1-\varepsilon)\min_a Q(x,a),$$

the Q-learning update rule (5.13) corresponds to (5.12). Now let $x \in X$ be arbitrary. Since by construction $P(a_t = a|\mathcal{F}_t, x = x_t) = \pi(Q_t,\varepsilon)(x,a)$, a.s. on $\{x \in X_\infty\}$ where a.s. $Q_t(x,a) \to Q^*(x,a)$, also $P(a_t = a|\mathcal{F}_t, x = x_t) \to \pi(Q_t,\varepsilon)(x,a)$ a.s. on $\{x \in X_\infty\}$. □

An alternative to the rule (5.12) is given by

$$Q_{t+1}(x_t,a_t) = (1-\alpha_t(x_t,a_t))Q_t(x_t,a_t) + \alpha_t(x_t,a_t)\left(c_t + \gamma Q_t(y_t,b_t)\right), \tag{5.14}$$

where b_t is chosen stochastically as the action in state y_t according to the learning policy, i.e., in the case of $(1-\varepsilon)$-greedy learning policies

$$P(b_t \in \operatorname*{Argmin}_b Q_t(x_t,b)|\mathcal{F}_t) = (1-\varepsilon) + \varepsilon\frac{|\operatorname{Argmin}_b Q_t(x_t,b)|}{|\mathcal{A}|}$$

. The above rule can be viewed as an action-sampled version of the update rule (5.12) and is called SARSA since the update rule is based on the actual state, action, the immediate reward (cost) and the next state and action. We will investigate the behaviour of this learning rule for the class of *rank-based* learning policies.

DEFINITION 5.3.4 *Assume that $Q_t \in B(\mathcal{X}\times\mathcal{A})$ is a sequence of $\mathcal{F}_t$-measureable functions. A stationary* rank-based *learning policy $\mathcal{P}$ is given in terms of a probability distribution $(P_1, P_2, \ldots, P_m)$ over the action space $\mathcal{A}$, where $P_i > 0$ for $i = 1,2,\ldots,m$, and $\pi(a|\mathcal{F}_t, x = x_t) = P_{\rho(Q_t,x,a)}$, where $\rho(Q,x,a)$ denotes the rank of action a in the table $Q(x,\cdot)$ ($\rho(Q,x,a) = 1$ if $a =$ $\operatorname{argmin}_b Q(x,b)$ and ties are assumed to be broken in an arbitrarily fixed manner).*

Assuming that $\operatorname{Argmin} Q_t(x,a)$ is a singleton the above $(1-\varepsilon)$-greedy learning policy is given by $P_1 = (1-\varepsilon) + \varepsilon/m$ and $P_i = \varepsilon/m$, where $|\mathcal{A}| = m \geq i > 1$.

THEOREM 5.3.5 *Assume 5.2.1 and that the SARSA learning equation (Equation (5.14)) is used to compute Q_t. Further, assume that a rank-based learning* *policy $\pi = \pi_P$, where $P = (P_1, \ldots, P_m)$ is a distribution, is used for control. Then, for all $(x, a) \in \mathcal{X} \times \mathcal{A}$, $Q_t(x, a)(\omega) \to Q^*(x, a)$, $t \to \infty$ holds a.e. on $\{x \in X_\infty\}$, where Q^* is the function defined by the fixed-point equation*

$$Q^*(x,a) = \sum_{y \in \mathcal{X}} p(x,a,y)\Big(c(x,a,y) + \gamma \sum_{b \in \mathcal{A}} P_{\rho(Q^*,y,b)} Q^*(y,b)\Big). \tag{5.15}$$

Moreover, π is asymptotically optimal w.p.1 in the class of rank-based policies determined by P.

Proof. By Remark 5.2.6 it is sufficient to prove that (5.14) converges when the SE condition is satisfied, so let us assume this. Since the learning rule is of the relaxation form, we can apply Corollary 3.5.2. Let $a(Q, x, k)$ denote the action whose rank in table $Q(x, \cdot)$ is k (i.e., $\rho(Q, a(Q, x, k)) = k$). Let us write the SARSA update rule in the form

$$Q_{t+1}(x_t, a_t) = (1 - \alpha_t(x_t, a_t))Q_t(x_t, a_t) + \alpha_t(x_t, a_t)\left(c_t + \gamma Q_t(y_t, a(Q, x_{t+1}, k_t))\right), \tag{5.16}$$

where the k_ts are appropriately defined independent, identically distributed random variables with $P(k_t = k) = P_k$. These random variables should be defined in such a way that $a_{t+1} = a(Q_t, x_{t+1}, k_t)$ w.p.1. Notice the condition on action-selection guarantees the existence of these random variables. Let $T : B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X} \times \mathcal{A})$ denote the operator whose fixed point is Q^* as given by (5.15):

$$(TQ)(x,a) = \sum_{y \in \mathcal{X}} p(x,a,y)\Big(c(x,a,y) + \gamma \sum_{b \in \mathcal{A}} P_{\rho(Q,y,b)} Q(y,b)\Big).$$

Note that rank-based averaging is a non-expansion (see Lemma 7 and Theorem 9 of [74]) so T is a contraction. Once again, we identify the state set $\mathcal{X}$ Corollary 3.5.2 by the set of possible state-action pairs $\mathcal{X} \times \mathcal{A}$, we further let

$$f_t(x,a) = \begin{cases} \alpha_t(x,a), & \text{if } (x,a) = (x_t, a_t); \\ 0, & \text{otherwise,} \end{cases}$$

and

$$(P_t Q)(x,a) = c_t + \gamma Q(x_{t+1}, a(Q_t, x_{t+1}, k_t)).$$

The conditions on f_t and the approximation property of (f_t, P_t) are readily verifiable, but the condition for P_t is not immediate in this case. We show:

$$\begin{aligned} |(P_t Q_1)(x,a) - (P_t Q_2)(x,a)| &\le \\ &\gamma |Q_1(x_{t+1}, a(Q_1, x_{t+1}, k_t)) - Q_2(x_{t+1}, a(Q_2, x_{t+1}, k_t))| \\ \le\ & \gamma \max_{a \in \mathcal{A}} |Q_1(x_{t+1}, a) - Q_2(x_{t+1}, a)| \\ \le\ & \gamma \|Q_1 - Q_2\|. \end{aligned}$$

Here the second inequality comes from Lemma 7 of [74] which states that the absolute difference between the ith largest values of two (discrete-valued) functions is smaller than or equal that the maximum-norm difference of the functions. Since all the conditions of Corollary 3.5.2 are satisfied, the convergence of Q_t to Q^* follows. □

We conjecture that if the action a_{t+1} would be allowed to depend on the actual values of Q_t and not only the ranks of actions in $Q_t(x_{t+1}, \cdot)$ (consider e.g. the Boltzmann-learning policy) then the SARSA rule would not necessarily converge. In fact, for a non-rank based learning policy the Lipschitz property of the appropriate T_t operators w.r.t. their second argument may not hold.

5.4 SARSA with Asymptotically Greedy Learning-Policies

In this section we still consider the SARSA learning-rule (5.14), but here we assume that the greedy actions (i.e., elements of $\mathrm{Argmin}_{b \in \mathcal{A}} Q_t(x_{t+1}, b)$) are selected with a probability going to one. If this convergence is slow enough then, as suggested by Theorem 5.2.4, this learning policy might be asymptotically optimal. Note that the SARSA learning rule is cheaper to compute at the onset of learning since initially there is no need to compute $\min_{a \in \mathcal{A}} Q_t(x_t, a)$. This may save essential computational resources if $\mathcal{A}$ is large, such as when $\mathcal{A}$ is a discretized version of some continuous action space (initially a rough discretization may be used).

If in this learning rule actions are still selected based on their ranks in $Q_t(x_{t+1}, \cdot)$ then an argument similar to the proof of Theorem 5.3.5 can be used to show the convergence of Q_t to Q^*. However, if we do not make any such restrictions on the action-selection procedure apart from that the probability of the choice of greedy actions goes to one, then this type of proof cannot be used any more. The reason for this is that

$$(P_t Q)(x, a) = c_t + \gamma Q(x_{t+1}, a_{t+1}(Q, x_{t+1}))$$

is not necessarily a contraction, but can be viewed as the sum of a contraction ($E[P_t Q|\mathcal{F}_t]$, where $\mathcal{F}_t$ represents the history of the process up to time t) plus a zero-mean term ($P_t Q - E[P_t Q|\mathcal{F}_t]$) whose variance depends on Q. In order to analyze this new situation we need an extension of Lemma 3.2.2 and Theorem 3.1.3 to conditional probability spaces.

LEMMA 5.4.1 *Let $\mathcal{Z}$ be an arbitrary set and consider the sequence*

$$x_{t+1}(z) = g_t(z)x_t(z) + f_t(z)N_t(z), \tag{5.17}$$

where $z \in \mathcal{Z}$ and $\|x_1\| < C < \infty$ w.p.1 for some $C > 0$. Assume that for all $k \geq 0$ $\lim_{n\to\infty} \prod_{t=k}^{n} g_t(z) = 0$ uniformly in z w.p.1, $f_t(z) \leq \gamma(1 - g_t(z))$ w.p.1,

$\sum_{t=0}^{\infty} f_t^2(z) < D < \infty$ *w.p.1,* $\|E[N_t(\cdot)|\mathcal{F}_t]\| \leq \|x_t\| + \lambda_t$*, where* $0 \geq \lambda_t \to 0$*, and* $\mathrm{Var}[N_t(z)|\mathcal{F}_t] \leq C(z)$ *for some non-negative constants* $C(z)$*. Here,* $\mathcal{F}_t$ *is an increasing sequence of* σ*-fields adapted to the process* $\{N_{t-1}, f_t, g_t\}$ *and* $\mathcal{F}_0$ *is adapted to* x_0*. Then,* $\|x_t\|$ *converges to* 0 *with probability 1.*

Proof. By defining $r_t(z) = N_t(z) - E[N_t(z)\,|\,\mathcal{F}_t]$, we can decompose x_t to the sum of δ_t and w_t, where

$$\begin{aligned} \delta_{t+1}(z) &= g_t(z)\delta_t(z) + f_t(z)E[N_t(z)\,|\,\mathcal{F}_t] \\ w_{t+1}(z) &= g_t(z)w_t(z) + f_t(z)r_t(z), \end{aligned}$$

with $\delta_0 = x_0$ and $w_0 = 0$. Since, by assumption, $Var(r_t(z)\,|\,\mathcal{F}_t)$ is bounded by $C(z)$, we get by the Conditional Averaging Lemma (Lemma 3.5.1) that w_t converges to $E[r_t\,|\,\mathcal{F}_t] = 0$ w.p.1. Moreover,

$$\begin{aligned} \delta_{t+1}(z) &= g_t(z)\delta_t(z) + f_t(z)||\delta_t + w_t|| \\ &\leq g_t(z)\delta_t(z) + f_t(z)(\|\delta_t\| + \|w_t\| + \lambda_t) \end{aligned}$$

and thus, since $\|w_t\| + \lambda_t \to 0$, we get by Lemma 3.4.2 that also δ_t converges to zero w.p.1, thus finishing the proof. □

The next lemma is a non-trivial extension of the above lemma to the case when the conditional variance may grow with $\|x_t\|$:

LEMMA 5.4.2 *Let us consider the process* x_t *defined in Lemma 5.4.1 and assume that the conditions of that lemma are satisfied, except that here we assume only that* $\mathrm{Var}[N_t(z)|\mathcal{F}_t] \leq C(z)(1 + \|x_t\|)^2$*. Then,* $\|x_t\|$ *converges to* 0 *with probability 1.*

Proof. Define $s_t(z) = r_t(z)/(1 + \|x_t\|)$, where $r_t(z) = N_t(z) - E[N_t(z)\,|\,\mathcal{F}_t]$ as above. Then $E[s_t(z)\,|\,\mathcal{F}_t] = 0$ and $E[s_t^2(z)\,|\,\mathcal{F}_t] < C(z) < \infty$, since x_t is $\mathcal{F}_t$-measurable. This yields the decomposition $r_t(z) = s_t(z) + s_t(z)\|x_t\|$. Now let us define the sequences δ_t, u_t and v_t by the recursions

$$\begin{aligned} \delta_{t+1}(z) &= g_t(z)\delta_t(z) + f_t(z)E[N_t(z)\,|\,\mathcal{F}_t] \\ &= g_t(z)\delta_t(z) + f_t(z)\left(\|u_t + \delta_t\| + a_t(z)\right) \\ u_{t+1}(z) &= g_t(z)u_t(z) + f_t(z)s_t(z)\|x_t\| \\ &= g_t(z)u_t(z) + f_t(z)s_t(z)\left(\|u_t + \delta_t\| + b_t\right) \\ v_{t+1}(z) &= g_t(z)v_t(z) + f_t(z)s_t(z), \end{aligned}$$

with $\delta_0 = x_0$, $u_0 = v_0 = 0$, and

$$\begin{aligned} a_t(z) &= E[N_t(z)\,|\,\mathcal{F}_t] - \|u_t + \delta_t\|, \\ b_t &= \|x_t\| - \|u_t + \delta_t\|. \end{aligned}$$

Then $x_t(z) = \delta_t(z) + u_t(z) + v_t(z)$. By the Conditional Averaging Lemma (Lemma 3.5.1) $v_t \to 0$ w.p.1, that is we can think of v_t as a perturbation converging to zero. Now let

$$\begin{aligned} G_t([\delta, u], \varepsilon_t)(z) = \; [\; & g_t(z)\delta(z) + f_t(z)\left(\|u+\delta\| + a_t(z)\right), \\ & g_t(z)u(z) + f_t(z)s_t(z)\left(\|u+\delta\| + b_t\right)], \end{aligned}$$

where $\varepsilon_t(z) = [a_t(z), b_t]$. Then $[\delta_{t+1}, u_{t+1}] = G_t([\delta_t, u_t], \varepsilon_t)$, G_t is homogeneous, and by linearity $X_{t+1} = G_t(X_t, \varepsilon_t)$ is insensitive to finite perturbations of $\varepsilon = \{\varepsilon_t\}_{t\geq 0}$ and also scaling ε by positive numbers smaller than one. Therefore the Rescaling Lemma (Lemma 3.3.2) applies to X_t, i.e., in order to show that X_t converges to zero w.p.1 it is sufficient to show this under the assumption that X_t is kept bounded by rescaling it. Let $[\delta^{\text{sc.}}, u_t^{\text{sc.}}]$ denote the rescaled version and $0 < S_t \leq 1$ the scaling coefficient at time t. Since

$$u_{t+1}^{\text{sc.}}(z) = S_t g_t(z) u_t^{\text{sc.}} + S_t f_t(z) s_t(z)\left(\|u_t^{\text{sc.}} + \delta_t^{\text{sc.}}\| + b_t\right),$$

by the Conditional Averaging Lemma (Lemma A.2.3, Part 2) $u_t(z)$ converges to zero. In order to show this apply the identification $w_t = s_t(z)\left(\|u_t^{\text{sc.}} + \delta_t^{\text{sc.}}\| + b_t\right)$, and the rest of the identifications should go without saying. In fact, the conditions on g_t and f_t imply the conditions on α_t in the Lemma, and the conditions on w_t are implied by

$$E\left[s_t(z)\left(\|u_t^{\text{sc.}} + \delta_t^{\text{sc.}}\| + b_t\right) \mid \mathcal{F}_t\right] = \left(\|u_t^{\text{sc.}} + \delta_t^{\text{sc.}}\| + b_t\right) E\left[s_t(z) \mid \mathcal{F}_t\right] = 0$$

and

$$\begin{aligned} Var\left[s_t(z)\left(\|u_t^{\text{sc.}} + \delta_t^{\text{sc.}}\| + b_t\right) \mid \mathcal{F}_t\right] &= E\left[s_t^2(z)\left(\|u_t^{\text{sc.}} + \delta_t^{\text{sc.}}\| + b_t\right)^2 \mid \mathcal{F}_t\right] \\ &\leq 3KE[s_t^2(z) \mid \mathcal{F}_t] \qquad (5.18) \\ &\leq 3KC(z), \end{aligned}$$

where K is a bound on $[\delta^{\text{sc.}}, u_t^{\text{sc.}}]$ and in (5.18) $b_t \to 0$ w.p.1 was used. Also $\delta_t^{\text{sc.}}$ converges to zero since by the triangle inequality $|a_t(z)| \leq \|a_t\| \leq b_t + \lambda_t$ and

$$\begin{aligned} |\delta_{t+1}^{\text{sc.}}(z)| &\leq S_t\Big(g_t(z)|\delta_t^{\text{sc.}}(z)| + f_t(z)\left(\|u_t^{\text{sc.}} + \delta_t^{\text{sc.}}\| + |a_t(z)|\right)\Big) \\ &\leq g_t(z)|\delta_t^{\text{sc.}}(z)| + f_t(z)\left(\|u_t^{\text{sc.}}\| + \|\delta_t^{\text{sc.}}\| + b_t + \lambda_t\right). \end{aligned}$$

If we let $\delta_0^+ = \delta_0^{\text{sc.}}$ and $\delta_{t+1}^+(z) = g_t(z)|\delta_t^+(z)| + f_t(z)\left(\|\delta_t^+\| + \|u_t^{\text{sc.}}\| + b_t + \lambda_t\right)$ then (by induction on t) we get $|\delta_t^{\text{sc.}}(z)| \leq \delta_t^+(z)$, which holds for all $t \geq 0$. Since we know that u_t^{sc} converges to zero, thus by Lemma 3.4.2 also δ_t^+ and therefore $\delta_t^{\text{sc.}}$ converge to zero w.p.1. This proves that the rescaled version of X_t converges to zero, and therefore by the Rescaling Lemma so must do the unscaled version. But this means that $x_t = \delta_t + u_t + v_t$ converges to zero w.p.1, which is just what we wanted to prove. □

The next theorem concerns relaxation processes of form

$$V_{t+1}(x) = (1 - f_t(x))V_t(x) + f_t(x)([P_t V_t](x) + Z_t(V_t)(x)), \tag{5.19}$$

where f_t satisfies the usual properties, P_t is typically a pseudo-contraction and Z_t represents some kind of zero-mean "noise." Note that the relaxation processes considered earlier could also be put in this form, but, there, the noise term still remains a contraction! Our aim here is to consider the case when Z_t is no longer a contraction but is a "real" noise.

THEOREM 5.4.3 *Let $\mathcal{X}$ be an arbitrary set, $v^* \in B(\mathcal{X})$ and consider the stochastic process given by Equation 5.19. Assume that the following conditions are satisfied:*

1. *The process defined by*
$$U_{t+1}(x) = (1 - f_t(x))U_t(x) + f_t(x)[P_t v^*](x) \tag{5.20}$$
converges to v^* w.p.1;
2. $0 \le f_t(x) \le 1$*;*
3. $\sum_{t=1}^n f_t(x)$ *converges to infinity uniformly in x as $n \to \infty$ and $\sum_{t=1}^\infty f_t^2(x) < D < \infty$ for some $D > 0$;*
4. *there exist number $0 < \gamma < 1$ and a sequence $\lambda_t \ge 0$ converging to zero w.p.1 such that $\|P_t U - P_t V\| \le \gamma \|U - V\| + \lambda_t$ holds for all $U, V \in B(\mathcal{X})$;*
5. $E[Z_t(V)(x) \,|\, \mathcal{F}_t] = 0$, $\mathrm{Var}[Z_t(V)(x) \,|\, \mathcal{F}_t] < C(x)(1 + \|V\|^2) < \infty$ *for some $C > 0$, for all $V \in B(\mathcal{X})$, where $\mathcal{F}_t$ is an increasing sequence of σ-fields adapted to the process (G_t, F_t, P_t, Z_{t-1}) and V_0 is $\mathcal{F}_0$-measurable.*

Then, V_t converges to v^* w.p.1 uniformly over $\mathcal{X}$.

Proof. Let $T_t(U,V)(x) = (1 - f_t(x))U(x) + f_t(x)([P_t V](x) + Z_t(V)(x))$ and let $\hat{T}_t(U,V) = E[T_t(U,V) \,|\, \mathcal{F}_t]$, so $T_t(U,V) = \hat{T}_t(U,V) + f_t Z_t(V)$. If we let $\hat{U}_{t+1} = \hat{T}_t(\hat{U}_t, V)$ and $U_{t+1} = T_t(U_t, V)$ then

$$\hat{U}_{t+1}(x) - U_{t+1}(x) = (1 - f_t(x))\Big(\hat{U}_t(x) - U_t(x)\Big) + f_t(x) Z_t(V)(x)$$

converges to zero w.p.1 by the Conditional Averaging Lemma (Lemma 3.5.1). Therefore we get that T_t approximates T if and only if $\hat{T}_t$ approximates T. Next, observe that $\hat{T}_t$ gives rise to a relaxation process of the type considered in Corollary 3.5.2, so $\hat{T}_t$ approximates T at v^* and $\hat{V}_{t+1} = \hat{T}_t(\hat{V}_t, \hat{V}_t)$, $\hat{V}_0 = V_0$ converges to v^* w.p.1. Since

$$\begin{aligned} V_{t+1}(x) - \hat{V}_{t+1}(x) &= (1 - f_t(x))(V_t(x) - \hat{V}_t(x)) \\ &\quad + f_t(x)\Big((P_t V_t)(x) - (P_t \hat{V}_t)(x) + Z_t(V_t)(x)\Big), \end{aligned}$$

by identifying $\mathcal{Z}$, x_t, f_t, g_t and N_t of Lemma 5.4.2 with $\mathcal{X}$, $V_t - \hat{V}_t$, γf_t, $1 - f_t$ and $(P_t V_t) - (P_t \hat{V}_t) + Z_t(V_t)$, respectively, we have that $V_t - \hat{V}_t$ converges to zero w.p.1 if

$$E_t = \|E[(P_t V_t) - (P_t \hat{V}_t) + Z_t(V_t) \,|\, \mathcal{F}_t]\| \leq \gamma \|V_t - \hat{V}_t\| + \lambda_t \tag{5.21}$$

and

$$\mathrm{Var}[(P_t V_t) - (P_t \hat{V}_t) + Z_t(V_t) \,|\, \mathcal{F}_t] \leq \hat{C}(x)(1 + \|V_t - \hat{V}_t\|)^2 < \infty \tag{5.22}$$

for some $\hat{C}(x) > 0$. However, since P_t is a non-expansion, we have that

$$\begin{aligned} E_t &= \|E[(P_t V_t)(\cdot) - (P_t \hat{V}_t)(\cdot) \,|\, \mathcal{F}_t] + E[Z_t(V_t)(\cdot) \,|\, \mathcal{F}_t]\| \\ &= \|E[(P_t V_t)(\cdot) - (P_t \hat{V}_t)(\cdot) \,|\, \mathcal{F}_t]\| \\ &= \|(P_t V_t)(\cdot) - (P_t \hat{V}_t)(\cdot)\| \\ &\leq \gamma \|V_t - \hat{V}_t\| + \lambda_t \end{aligned}$$

where in the the $\mathcal{F}_t$ measurability of $V_t, \hat{V}_t$ and P_t was used. Finally, since $\hat{V}_t$ is bounded,

$$E[\,|Z_t(V_t)(x)|^2 \,|\, \mathcal{F}_t] \leq C(x)(1 + \|V_t\|^2) \leq D(x)(1 + \|V_t - \hat{V}_t\|^2)$$

for some $D(x) > 0$, thus proving (5.22). □

Using this theorem, it is easy to show the convergence of SARSA with decaying exploration.

THEOREM 5.4.4 *Consider the learning rule 5.14, assume 5.2.1 and that the learning policy satisfies the WSE condition. Further, assume that the learning policy π is constructed so that*

$$\lim_{t \to \infty} \pi(a_{t+1} \in \operatorname*{Argmin}_a Q_t(x_{t+1}, a) \,|\, \mathcal{F}_t) = 1 \tag{5.23}$$

where $\mathcal{F}_t$ is an increasing sequence of σ-fields adapted to $\{x_{t+1}, Q_t, a_t, c_{t-1}, \alpha_t\}$.

Then, for all $(x, a) \in \mathcal{X} \times \mathcal{A}$, $Q_t(x, a)(\omega) \to Q^*(x, a)$, $t \to \infty$ holds a.e. on $\{x \in X_\infty\}$. Moreover, the learning policy is asymptotically optimal w.p.1.

Proof. By Remark 5.2.6 it is sufficient to prove that (5.14) converges when the SE condition is satisfied, so let us assume this.

The process (5.14) is a relaxation process of form (5.19), where V_t corresponds to Q_t, $\mathcal{X}$ to $\mathcal{X} \times \mathcal{A}$, $f_t(x)$ to $\alpha_t(x, a)$ if $(x, a) = (x_t, a_t)$ and to zero, otherwise, and P_t is given by

$$(P_t Q)(x, a) = \begin{cases} c(x, a, x_{t+1}) + \\ \qquad \gamma \sum_{b \in \mathcal{A}} P(a_{t+1} = b \,|\, \mathcal{F}_t) Q(x_{t+1}, b), & \text{if } (x, a) = (x_t, a_t); \\ 0, & \text{otherwise,} \end{cases}$$

and

$$Z_t(Q)(x,a) = \begin{cases} c_t + \gamma Q(x_{t+1}, a_{t+1}) - (P_t Q)(x,a), & \text{if } (x,a) = (x_t, a_t); \\ 0, & \text{otherwise.} \end{cases}$$

Since $\mathcal{F}_t$ is adapted to $(x_{t+1}, a_t, Q_t, c_{t-1}, \alpha_t)$, indeed $E[c_t + \gamma Q_t(x_{t+1}, a_{t+1}) \,|\, \mathcal{F}_t] = (P_t Q_t)(x_t, a_t)$. By assumption the learning rates satisfy the conditions of Theorem 5.4.3. So in order to apply Theorem 5.4.3 we need to check only if

$$\begin{aligned} \hat{Q}_{t+1}(x_t, a_t) &= (1 - \alpha_t(x_t, a_t))\hat{Q}_t(x_t, a_t) + \alpha_t(x_t, a_t)(c_t + \gamma \hat{Q}(x_{t+1}, a_{t+1})) \\ &= (1 - \alpha_t(x_t, a_t))\hat{Q}_t(x_t, a_t) + \alpha_t(x_t, a_t)\left(c_t + \gamma \min_{b \in \mathcal{A}} \hat{Q}(x_{t+1}, b)\right) + \\ &\qquad \gamma \alpha_t(x_t, a_t)\left(\hat{Q}(x_{t+1}, a_{t+1}) - \min_{b \in \mathcal{A}} \hat{Q}(x_{t+1}, b)\right), \end{aligned}$$

converges to TQ (T being the ordinary Q value operator of the MDP : $(TQ)(x) = \sum_{y \in \mathcal{X}} p(x,a,y)(c(x,a,y) + \gamma \min_{b \in \mathcal{A}} Q(y,b)))$ and the properties of P_t and Z_t. The Conditional Averaging Lemma (Lemma 3.5.1) together with Theorem 3.4.1 show that (5.24) indeed converges to TQ. The properties of P_t and Z_t are easy to verify. By definition, $E[Z_t(Q)(x,a) \,|\, \mathcal{F}_t] = 0$. Since $Var[c_t \,|\, \mathcal{F}_t] < C < \infty$, and c_t & a_{t+1}, and c_t & x_{t+1} are independent given $\mathcal{F}_t$, and there are a finite number of actions, so there exists a function $C(x,a) > 0$ such that $Var[Z_t(Q)(x,a) \,|\, \mathcal{F}_t] < C(x,a)(1 + \|Q\|^2)$. $\square$

Note that the learning policies suggested after Theorem 5.2.4 are concrete examples when all the conditions of Theorem 5.4.4 are met.

5.5 Discussion

Most of the material presented in this chapter is original. The results on the convergence of Q-learning algorithms *without* the SE (sufficient exploration) condition, presented in Section 5.1, are new (Theorem 5.1.4).

The main results of Sections 5.2,5.3 and 5.4 are based on the joint work of the author with Jaakkola, Singh and Littman [62]. Nevertheless, the proofs presented here differ from those of [62] in order the connection with the previous chapters are clear. Strictly speaking, Lemma 5.2.1, Theorem 5.2.4, Lemma 5.4.1, Lemma 5.4.2, Theorem 5.4.3 are entirely new. A similar result to that of Theorem 5.4.3 appeared in [11] (Proposition 4.5, pp.157) with an informal proof.

Notice that the result on how to decay the temperature in Boltzmann-exploration (Corollary 5.2.5) so as to keep the SE condition satisfied is the outcome of a worst-case analysis. Indeed, recently Kalmár, Lőrincz and the author found that in a real-world domain (controlling an autonomous robot) Boltzmann exploration is not needed for convergence [37]. Such situation may arise as a consequence of the strong ergodicity of the underlying Markov-chain.

John [35, 34] devised the algorithm presented in Section 5.3 which is intended to work with persistent exploration. He found that better learning performance can be achieved if the Q-learning rule is changed to incorporate the condition of persistent exploration. More precisely, in some domains John's learning rule performs better than standard Q-learning when exploration is retained, i.e., the discounted cumulated cost during learning was higher for his learning rule. The convergence of this algorithm was first investigated in the paper of Littman and the author [74]. This update rule was also described by Rummery [56] in the context of variations of the TD(λ) rule. In addition, Rummery explored SARSA, the action-sampled variant of the exploration-sensitive rule, too. This rule has also been studied by John [34], and by Singh and Sutton [63, 66]. The first proof published for the convergence of SARSA appeared in [62].

The convergence rate of on-line RL algorithms is can be investigated from different perspectives. An asymptotic convergence rate can be obtained directly from the results of Section 3.6. Non-asymptotic results are also of interest. Such results has been obtained for particular learning policies by [21] and more recently by Singh and Jaakkola [38]. These results show that the problem of learning optimal policies is PAC-learnable, i.e., almost optimal policies (δ-optimal policies) can be found with high probability $(1-\varepsilon)$ within polynomial time in the size of the MDP, $1/(1-\gamma)$, $1/\varepsilon$ and $1/\delta$. The proofs rely on the idea that the structure of MDPs can be estimated quickly by appropriate learning policies and they use Hoeffding (or Large Deviations) type of inequalities to provide upper-bounds on the precision of estimates.

Chapter 6

Summary

In summary, we list the contributions of the author:

1. For the first time the operator-theoretical treatment of sequential decision problems is extended to *non-stationary* policies in a consistent way (Corollary 1.1.13).

2. The usual theorems concerning the fundamental equation of Dynamic Programming (Theorem 1.1.10), the convergence of the finite-horizon optimal cost-to-go functions to the infinite-horizon optimal cost-to-go function (Theorem 1.3.3), the validity of Bellman's fixed point equation (Theorem 1.4.2), existence and characterisation of optimal stationary policies (Theorem 1.5.2), etc. are transferred to this case.

3. It is shown that in increasing models (i.e. when the cost of policies increases by time) if v_∞ denotes the function obtained by value iteration (dynamic programming) then

   (a) if the optimal cost-to-go function for stationary policies is equal to v_∞ then both functions satisfy the Bellman optimality equation (Theorem 2.2.2)

   (b) if v_∞ satisfies the Bellman optimality equation and there exists a greedy policy w.r.t. v_∞ then v_∞ equals to the optimal cost-to-go function and the myopoic policies w.r.t. v_∞ are optimal (Theorem 2.2.7).

4. Also for increasing models it is shown that Howard's policy improvement routine is valid under mild continuity assumptions (Lemma 2.2.11) but does not necessarily give rise to optimal policies when iterated (Example 2.3.3).

5. For finite and increasing models it is shown that the greedy policies w.r.t. the function given by the t^{th} step of the value iteration algorithm are optimal after a finite number of steps (Theorem 2.3.4).

6. Convergence of the so called reinforcement learning algorithms which are asynchronous, estimation-based variations of the value-iteration algorithm is proved based on a new operator-theoretical treatment (Theorem 3.1.3).

7. The asymptotic rate of convergence of value-iteration based reinforcement learning algorithms is given (Theorem 3.6.1).

8. As an application a rigorous proof of the convergence of the so-called $\hat{Q}$-learning algorithm which approximates the optimal action-value function in minimax problems is put forth (Theorem 4.5.1).

9. For the first time, conditions for the convergence of the learning policy to an optimal policy are given (Theorem 5.2.4).

10. It is proved that the on-line algorithm called SARSA(0) converges to optimality (Theorem 5.4.4).

11. Some experiments have been performed using these algorithms on a real-robot [36, 37] (the robot is shown in Figure 1). The experiments were analyzed by ANOVA and the results indicated the significant superiority of the model-based learning algorithms over the model-free ones. Although the learnt policy differed from that of a handcrafted policy, the respective performances were indistinguishable.

12. In the case of the worst-case total discounted cost criterion the learning policy which always selects the actions that seem to be the best converges to optimality [70, 67].

13. Stability results based on the Liapunov method were derived for a class of continuous-time adaptive control for plants whose control gain varies by the state of the plant [76, 77, 73, 78].

14. Worst-case upper performance bounds for a class of model-based adaptive control systems are given in [22, 23] under the assumption that the controlled system is known up to a L^2 measure of uncertainty.

Results 1–5 can be found in [44, 72, 74], results 6–9 were published in [44, 74, 71, 69], while result 10 was published in [62].

Appendix A

A.1 Convergence of Composite Processes

In this appendix, we prove Theorem 3.4.1, which we restate below for the convenience of the reader.

THEOREM 3.4.1 *Let $\mathcal{X}$ and $\mathcal{Y}$ be normed vector spaces, $U_t : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ $(t = 0, 1, 2, \ldots)$ be a sequence of mappings, and $\theta_t \in \mathcal{Y}$ be an arbitrary sequence. Let $\theta_\infty \in \mathcal{Y}$ and $x_\infty \in \mathcal{X}$. Consider the sequences*

$$x_{t+1} = U_t(x_t, \theta_\infty), \tag{A.1}$$

and

$$y_{t+1} = U_t(y_t, \theta_t), \tag{A.2}$$

and suppose that x_t and θ_t converge to x_∞ and θ_∞, respectively, in the norm of the corresponding spaces.

Let L_k^θ be the uniform Lipschitz index of $U_k(x, \theta)$ with respect to θ at θ_∞ and, similarly, let L_k^X be the uniform Lipschitz index of $U_k(x, \theta_\infty)$ with respect to x. Then, if the Lipschitz constants L_t^X and L_t^θ satisfy the relations $L_t^\theta \le C(1 - L_t^X)$, and $\prod_{m=n}^{\infty} L_m^X = 0$, where $C > 0$ is some constant and $n = 0, 1, 2, \ldots$, then $\lim_{t\to\infty} \|y_t - x_\infty\| = 0$. Without the loss of generality we will assume that $x_0 = y_0$. The theorem is proved through a series of simple lemmas. The assumption $x_0 = y_0$ will be used to simplify the expressions in Lemmas A.1.2 and A.1.5 below. *In this part we admit the convention that $\prod_{t=a}^{b} A_t = 1$ if $a > b$* which also makes some expressions shorter.

We start with an elementary, though useful lemma concerning iterated linear inequalities.

LEMMA A.1.1 *Let $\{\delta_t\}, \{a_t\}, \{b_t\}$ be sequences of real numbers which satisfy the inequality*

$$\delta_{t+1} \le a_t \delta_t + b_t, \quad n = 0, 1, 2, \ldots.$$

Then,

$$\delta_{t+1} \le \delta_0 \prod_{i=0}^{t} a_i + \sum_{i=0}^{t} b_i \prod_{j=i+1}^{t} a_j. \tag{A.3}$$

Proof. By assumption, $\delta_1 \leq a_0\delta_0 + b_0$ and $\delta_2 \leq a_1\delta_1 + b_1$. Substituting the estimate of δ_1 into the latter inequality yields $\delta_2 \leq a_1a_0\delta_0 + a_1b_0 + b_1$. Continuing in this way we get (A.3). □

Since in Theorem 3.4.1 $||y_t - x_\infty|| \leq ||y_t - x_t|| + ||x_t - x_\infty||$, it is sufficient to show that $\|y_t - x_t\|$ converges to zero. First we bound this difference:

LEMMA A.1.2 *Let $x_t, y_t, L_t^\theta, L_t^\chi, \theta_t$ be the series as defined in Theorem 3.4.1. Then,*

$$||y_{t+1} - x_{t+1}|| \leq \sum_{s=0}^{t} ||\theta_s - \theta_\infty|| L_s^\theta \prod_{p=s+1}^{t} L_p^\chi. \tag{A.4}$$

Proof. The proof is the application of Lemma A.1.1 to the sequence $\delta_t = ||x_t - y_t||$. Indeed, by the triangle inequality we have that

$$\begin{aligned} \delta_{t+1} &= ||U_t(x_t, \theta_\infty) - U_t(y_t, \theta_t)|| \\ &\leq ||U_t(x_t, \theta_\infty) - U_t(y_t, \theta_\infty)|| + ||U_t(y_t, \theta_\infty) - U_t(y_t, \theta_t)|| \\ &\leq L_t^\chi \delta_t + L_t^\theta ||\theta_t - \theta_\infty||. \end{aligned}$$

Thus, δ_t satisfies the conditions of Lemma A.1.1 with $a_t = L_t^\chi$ and $b_t = L_t^\theta||\theta_t - \theta_\infty||$. □

Observe that the right-hand side of (A.4) can be taken as the "triangle" transformation of the sequence $||\theta_t - \theta_\infty||$ in the following sense. If Δ_t is a sequence of numbers, and $a_{t,m}$, $0 \leq m \leq t$ is a triangular matrix, then the transformed sequence

$$b_t = \sum_{m=0}^{t} a_{t,m}\Delta_m$$

is called the *triangle transform* of Δ_t with transformation sequence $a_{t,m}$. The transformation itself is a $\mathbb{R}^\mathbb{N} \to \mathbb{R}^\mathbb{N}$ linear operator.

DEFINITION A.1.3 *A triangle transformation defined by the sequence $\{a_{t,m}\}$, $0 \leq m \leq t, t, m \in \mathcal{N}$, is called* regular *if the following two conditions hold:*

1. $\lim_{t\to\infty} a_{t,m} = 0$ *for all* m, *and*
2. $\lim_{t\to\infty} \left(\sum_{m=0}^{t} a_{t,m}\right) = c$.

In this case the sequence $\{a_{t,m}\}$ is called regular, too. The constant c is called the multiplier of the transformation.

Regular triangle transformations actually map $C_0(\mathbb{N})$ to $C_0(\mathbb{N})$, as it is shown in the following lemma.

LEMMA A.1.4 *Let $a = \{a_{t,m}\}$ ($t, m \in \mathbb{N}, 0 \le m \le t$) be an arbitrary regular sequence with multiplier c. Let Δ_t be arbitrary with $\lim_{t\to\infty} \Delta_t = \Delta$ and let b_t be the "triangle transform" of Δ_t with transformation sequence $a_{t,m}$:*

$$b_t = \sum_{m=0}^{t} a_{t,m} \Delta_m.$$

Then, $b_t \to c\Delta$, $t \to \infty$ and:

$$|b_t - c\Delta| \le |\varepsilon_t \Delta| + 2|c| H_{K+1} + H_0 \sum_{m=0}^{K} |a_{t,m}|, \tag{A.5}$$

where $0 \le K \le t$ is arbitrary (but it may depend on t), $\varepsilon_t = c - \sum_{m=0}^{t} a_{t,m}$, and

$$H_K = \sup_{t \ge K} |\Delta_t - \Delta|.$$

Proof. Let $0 \le K \le t$. Then

$$\begin{aligned} |b_t - c\Delta| &= \left| \sum_{m=0}^{t} a_{t,m}\Delta_m - (\varepsilon_t + \sum_{m=0}^{t} a_{t,m})\Delta \right| \qquad (A.6) \\ &\le \left| \sum_{m=0}^{t} a_{t,m}(\Delta_m - \Delta) \right| + |\varepsilon_t \Delta| \\ &\le \left| \sum_{m=0}^{K} a_{t,m}(\Delta_m - \Delta) \right| + \sup_{i \ge K+1} |\Delta_i - \Delta| \left| \sum_{m=K+1}^{t} a_{t,m} \right| + |\varepsilon_t \Delta| \\ &\le H_0 \sum_{m=0}^{K} |a_{t,m}| + 2|c| H_{K+1} + |\varepsilon_t \Delta|. \end{aligned}$$

Hence, (A.5) follows.

Now, let us consider the convergence of b_t. Pick up an arbitrary $\varepsilon > 0$. Then, let us choose the number K so that $2|c|H_{K+1} < \varepsilon/2$, and choose $N \ge K$ such that $|\varepsilon_t \Delta| + H_0 \sum_{m=0}^{K} |a_{t,m}| < \varepsilon/2$ is satisfied whenever $t \ge N$. Such numbers exist because H_K and ε_t converge to zero and also $a_{t,m}$ converges to zero as t tends to infinity for any fixed m. Then, by (A.7), if $t \ge N$, then $|b_t - c\Delta| < \varepsilon$. Since ε was arbitrary, b_t converges to $c\Delta$. □

The following lemma completes the proof of Theorem 3.4.1.

LEMMA A.1.5 *Let $x_t, y_t, L_t^\theta, L_t^\chi, \theta_t$ be the sequences as defined in Theorem 3.4.1. Assume that there exists a constant $C > 0$ such that*

$$L_t^\theta \le C(1 - L_t^\chi) \tag{A.7}$$

holds for all $t \geq 0$ *and also for all* $k \geq 0$

$$\prod_{t=k}^{\infty} L_t^{\chi} = 0. \tag{A.8}$$

Then, $\lim_{t\to\infty} ||x_t - y_t|| = 0$. *(Note that Condition (A.7) requires that* $L_t^{\chi} \leq 1$ *holds for all* t *since* $L_t^{\theta} \geq 0$.*)*

Proof. Let $\Delta_t = ||\theta_t - \theta_\infty||$ and $a_{t,m} = L_m^{\theta} \prod_{j=m+1}^{t} L_j^{\chi}$. Observe that by Lemma A.1.2, $||x_{t+1} - y_{t+1}||$ can be estimated from above by the triangle transform of Δ_t with coefficients $a_{t,m}$. Since $L_m^{\theta} \leq C(1 - L_m^{\chi})$, then $a_{t,m} \leq b_{t,m}$, where $b_{t,m} = C(1 - L_m^{\chi}) \prod_{j=m+1}^{t} L_j^{\chi}$. Therefore, since $0 \leq \sum_{m=0}^{t} a_{t,m}\Delta_t \leq \sum_{m=0}^{t} b_{t,m}\Delta_t$, it is sufficient to estimate the sum $\sum_{m=0}^{t} b_{t,m}\Delta_t$ from the above to obtain an upper bound on $\|x_t - y_t\|$. Now, observe that $(b_{t,m})$ satisfies the assumptions of Lemma A.1.4: $b_{t,m} \geq 0$,

$$\lim_{t\to\infty} b_{t,m} = C(1 - L_m^{\chi}) \prod_{j=m+1}^{\infty} L_j^{\chi} = 0$$

and

$$\sum_{m=0}^{t} b_{t,m} = C \sum_{m=0}^{t} (1 - L_m^{\chi}) \prod_{j=m+1}^{t} L_j^{\chi},$$

which is a telescopic sum that reduces to $C(1 - \prod_{j=0}^{t} L_j^{\chi})$ after expansion. Therefore,

$$\lim_{t\to\infty} \sum_{m=0}^{t} b_{t,m} = C(1 - \prod_{j=0}^{\infty} L_j^{\chi}) = C > 0.$$

Thus, Lemma A.1.2 can be applied, and we see that $\|x_t - y_t\|$ converges to zero. □

Note that this proved Theorem 3.4.1.

A.2 Convergence of Certain Stochastic Approximation Processes

Theorem A.2.1 *Suppose the following assumptions are satisfied. Let* $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_t \subseteq \mathcal{F}_{t+1} \subseteq \ldots$ *be an increasing sequence of* σ*-fields and consider the process*

$$x_{t+1} = x_t + H_t(x_t), \quad t = 0, 1, 2, \ldots \tag{A.9}$$

where $H_t(\cdot)$ *is real-valued and a.s. bounded function. Assume that* x_t *is* $\mathcal{F}_t$*-measurable and let* $h_t(x_t) = E[H_t(x_t)\,|\mathcal{F}_t]$. *Assume that the following assumptions are satisfied:*

1. *A number x^* exists s.t.*

   (a) $(x - x^*)h_t(x) \leq 0$ *for all* $t \geq 0$,
   and if for any fixed $\varepsilon > 0$ *we let*

   $$\overline{h}_t(\varepsilon) = \sup_{\varepsilon \leq |x - x^*| \leq 1/\varepsilon} \frac{h_t(x)}{x - x^*},$$

   then w.p.1;

   (b) $\sum_{t=0}^{\infty} \overline{h}_t(\varepsilon) = -\infty$;

   (c) $\sum_{t=0}^{\infty} \overline{h}_t^{+}(\varepsilon) < \infty$, *where* $r^+ = (r + |r|)/2$; *and*

2. $E[H_t^2(x_t)\,|\mathcal{F}_t] \leq C_t(1 + (x_t - x^*)^2)$, *for some non-negative random sequence* C_t *which satisfies* $\sum_{t=1}^{\infty} C_t < \infty$ *w.p.1.*

Then x_t *converges to* x^* *w.p.1.*

Proof. The proof is based on the following "super-martingale" type lemma due to [54]:

Lemma A.2.2 *Suppose that* Z_t, B_t, C_t, D_t *are finite, non-negative random variables, adapted to the* σ*-field* $\mathcal{F}_t$*, which satisfy*

$$E[Z_{t+1}\,|\mathcal{F}_t] \leq (1 + B_t)Z_t + C_t - D_t. \tag{A.10}$$

Then on the set $\{\sum_{t=0}^{\infty} B_t < \infty, \sum_{t=0}^{\infty} C_t < \infty\}$, *we have* $\sum_{t=0}^{\infty} D_t < \infty$ *and* $Z_t \to Z < \infty$ *a.s.*

In our case let $Z_t = (x_t - x^*)^2$. Then

$$\begin{aligned} E[Z_{t+1}\,|\mathcal{F}_t] &\leq Z_t + C_t(1 + Z_t) + 2(x_t - x^*)h_t(x_t) \\ &\leq (1 + C_t)Z_t + C_t + 2(x_t - x^*)h_t(x_t) \end{aligned}$$

and therefore by the above lemma (since by assumption $C_t \geq 0$, $\sum_{t=0}^{\infty} C_t < \infty$ and $(x_t - x^*)h_t(x_t) \leq 0$) $Z_t \to Z < \infty$ w.p.1 for some random variable Z and $\sum_{t=0}^{\infty}(x_t - x^*)h_t(x_t) > -\infty$. If $\infty > Z(\omega) \neq 0$ for some ω then there exist an $\varepsilon > 0$ and $N > 0$ (which may depend on ω) s.t. if $t \geq N$ then $\varepsilon \leq |x_t(\omega) - x^*| \leq \frac{1}{\varepsilon}$. Consequently

$$\begin{aligned} -\infty &< \sum_{s=0}^{\infty}(x_s(\omega) - x^*)h_s(x_s(\omega)) \\ &\leq \sum_{s=0}^{\infty}(x_s(\omega) - x^*)^2\,\overline{h}_s(\varepsilon;\omega) \end{aligned}$$

$$\begin{aligned}
&\le \sum_{s=0}^{N-1}(x_s(\omega)-x^*)^2\,\overline{h}_s(\varepsilon;\omega)+ \\
&\qquad \varepsilon^2 \sum_{s\ge N,\overline{h}_s(\varepsilon;\omega)\le 0} \overline{h}_s(\varepsilon;\omega)+\frac{1}{\varepsilon^2}\sum_{s\ge N,\overline{h}_s(\varepsilon;\omega)>0}\overline{h}_s(\varepsilon;\omega) \\
&= -\infty
\end{aligned}$$

by Condition 1b. This means that $\{\omega \,|\, Z(\omega) \neq 0\}$ must be a null-set which finishes the proof of the theorem. □

The theorem could easily be extended to vector-valued processes. Then the definition of $\overline{h}_t(\varepsilon)$ would become $\overline{h}_t(\varepsilon) = \sup_{\varepsilon\le\|x-x^*\|_2\le 1/\varepsilon}(x-x^*)^T h_t(x)$ and Condition 1a should become $(x-x^*)^T h(x) \le 0$. Otherwise, the proof is identical if one defines $Z_t = \|x_t - x^*\|_2^2$. Note that Theorem A.2.1 includes as a special case *i)* the standard Robbins-Monro process of form $x_{t+1} = x_t + \gamma_t H(x_t,\eta_t)$, where η_t are random variables whose distribution depend only on x_t; $\gamma_t \ge 0$, $\sum_t \gamma_t = \infty$ and $\sum_t \gamma_t^2 < \infty$, and *ii)* one form of the Dvoretzky process $x_{t+1} = T_t + \eta_t$, where $T_t = G_t(x_t - x^*) + x^*$, $E[\eta_t|G_t, \eta_{t-1}, G_{t-1}, \ldots, \eta_0, G_0] = 0$, $\sum_t E[\eta_t^2] < \infty$, $G_t \le 1$, and $\sum_t (G_t - 1) = -\infty$.

Next an extension of Lemma 3.5.1 is proved.

LEMMA A.2.3 (CONDITIONAL AVERAGING LEMMA) *Let $\mathcal{F}_t$ be an increasing sequence of σ-fields, let $0 \le \alpha_t$, s_t and w_t be random variables such that α_t, w_{t-1} and s_{t-1} are $\mathcal{F}_t$ measurable. Assume that the following hold w.p.1: $E[s_t|\mathcal{F}_t, \alpha_t \neq 0] = \hat{A} > 0$, $E[s_t^2|\mathcal{F}_t] < \hat{B} < \infty$, $E[s_t w_t|\mathcal{F}_t, \alpha_t \neq 0] = A$, $E[s_t^2 w_t^2|\mathcal{F}_t] < B < \infty$, $\sum_{t=1}^{\infty}\alpha_t = \infty$, and $\sum_{t=1}^{\infty}\alpha_t^2 < C < \infty$ for some $B, C > 0$. Then, the process*

$$Q_{t+1} = (1 - s_t\alpha_t)Q_t + \alpha_t s_t w_t \tag{A.11}$$

converges to $A/\hat{A}$ w.p.1.
Further,

$$Q_{t+1} = S_t\Big((1-\alpha_t)Q_t + \beta_t w_t\Big) \tag{A.12}$$

converges to zero, if α_t, β_t, S_t and w_{t-1} are $\mathcal{F}_t$-measurable, $E[w_t|\mathcal{F}_t] = 0$, $\mathrm{Var}[w_t|\mathcal{F}_t] < B$, $0 \le S_t \le 1$ and the relations $\alpha_t \ge 0$, $\sum_{t=0}^{\infty}\alpha_t = \infty$ and $\sum_{t=0}^{\infty}\max(\alpha_t,\beta_t)^2 < \infty$ hold w.p.1.

Proof. Without the loss of generality we may assume that $E[s_t|\mathcal{F}_t] = \hat{A}$ and $E[s_t w_t|\mathcal{F}_t] = A$. Rewriting the process (A.11) in the form of (A.9) we get $Q_{t+1} = Q_t + \alpha_t s_t (w_t - Q_t)$ and thus $h_t(Q) = E[\alpha_t s_t(w_t - Q)|\mathcal{F}_t] = \alpha_t(E[s_t w_t|\mathcal{F}_t] - QE[s_t|\mathcal{F}_t]) = \alpha_t\hat{A}(A/\hat{A} - Q)$ and $\overline{h}_t(\varepsilon) = -\alpha_t\hat{A}$ independently of ε. Due to the identity $|x| \le 1 + x^2$, $|E[s_t^2 w_t|\mathcal{F}_t]| \le E[s_t^2|w_t|\,|\mathcal{F}_t] \le E[s_t^2(1+w_t^2)\,|\mathcal{F}_t] \le \hat{B} + B$ and making use of $|x| \le 1 + x^2$ again, we have $E[H_t^2(Q_t)|\mathcal{F}_t] = \alpha_t^2 E[s_t^2(w_t - Q_t)^2|\mathcal{F}_t] \le \alpha_t^2(B + 2(\hat{B}+B)(1+Q_t^2) + \hat{B}Q_t^2) \le \alpha_t^2 C'(1 + (Q_t - A/\hat{A})^2)$ for some $C' > 0$. Thus, the lemma follows from Theorem A.2.1.

The second part is proved from Lemma A.2.2 directly. Let $Z_t = |Q_t|^2$. Note that by the non-negativity of S_t

$$|Q_{t+1}| \leq S_t((1-\alpha_t)|Q_t| + \beta_t w_t),$$

so

$$E[Z_{t+1}|\mathcal{F}_t] \leq S_t^2(Z_t - 2\alpha_t|Q_t| + \beta_t^2 B + \alpha_t^2 Z_t).$$

If we let $C_t^2 = \max(B,1)\max(\beta_t,\alpha_t)^2$ then we obtain

$$E[Z_{t+1}|\mathcal{F}_t] \leq Z_t - (1-S_t^2)Z_t - 2\alpha_t S_t^2|Q_t| + S_t^2 C_t^2(1+Z_t)$$

Since $0 \leq S_t \leq 1$ and since for large enough t, $1 \geq \alpha_t \geq 0$ w.p.1., we have $((1-S_t^2)Z_t + 2\alpha_t S_t^2|Q_t|) \geq ((1-S_t^2) + S_t^2(2\alpha_t))\min(|Q_t|, Z_t) \geq 2\alpha_t \min(|Q_t|, Z_t)$ and so

$$E[Z_{t+1}|\mathcal{F}_t] \quad \leq \quad (1+C_t^2)Z_t + C_t^2 - 2\alpha_t \min(\sqrt{Z_t}, Z_t).$$

Now the same argument that lead to the proof of Theorem A.2.1 yields that here also $Z_t \to 0$ w.p.1, thus proving the Lemma. $\square$

Bibliography

[1] A. Barto, S. Bradtke, and S. Singh. Real-time learning and control using asynchronous dynamic programming. Technical report 91-57, Computer Science Department, University of Massachusetts, 1991.

[2] A. Barto, S. J. Bradtke, and S. Singh. Learning to act using real-time dynamic programming. *Artificial Intelligence*, 1(72):81–138, 1995.

[3] A. Barto, R. Sutton, and C. J. C. H. Watkins. Learning and sequential decision making. Technical Report 89-95, Department of Computer and Information Science, University of Massachusetts, Amherst, Massachusetts, 1989. Also published in *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, M. Gabriel and John Moore, editors. The MIT Press, Cambridge, Massachusetts, 1991.

[4] R. Bellman. *Dynamic Programming.* Princeton University Press, Princeton, New Jersey, 1957.

[5] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive Algorithms and Stochastic Approximations.* Springer Verlag, New York, 1990.

[6] D. P. Bertsekas. Monotone mappings with application in dynamic programming. *SIAM J. Control and Optimization*, 15(3):438–464, 1977.

[7] D. P. Bertsekas. *Dynamic Programming: Deterministic and Stochastic Models.* Prentice-Hall, Englewood Cliffs, NJ, USA, 1989.

[8] D. P. Bertsekas and D. A. Castañon. Adaptive aggregation for infinite horizon dynamic programming. *IEEE Transactions on Automatic Control*, 34(6):589–598, 1989.

[9] D. P. Bertsekas and S. Shreve. *Stochastic Optimal Control (The Discrete Time Case).* Academic Press, New York, 1978.

[10] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods.* Prentice-Hall, Englewood Cliffs, NJ, 1989.

[11] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming.* Athena Scientific, Belmont, MA, 1996.

[12] P. Bhakta and S. Choudhury. Some existence theorems for functional equations arising in dynamic programming, II. *J. of Math. Anal. and Appl.*, 131:217–231, 1988.

[13] D. Blackwell. Discounted dynamic programming. *Annals of Math. Statistics*, 36:226–235, 1965.

[14] L. Breiman. *Probability.* Addison-Wesley Publishing Company, Reading, 1968.

[15] A. Burnetas and M. Katehakis. Optimal adaptive policies for Markov Decision Processes. *Mathematics of Operations Research*, 22(1), 1997.

[16] K. Chung and M. Sobel. Discounted MDPs: Distribution functions and exponential utility maximization. *SIAM J. Control and Optimization*, 25(1):49–62, 1987.

[17] A. Condon. The complexity of stochastic games. *Information and Computation*, 96(2):203–224, February 1992.

[18] E. Denardo. Contraction mappings in the theory underlying dynamic programming. *SIAM Rev.*, 9:165–177, 1967.

[19] S. Dreyfus and A. Law. *The Art and Theory of Dynamic Programming*, volume 130 of *Mathematics in Science and Engineering*. Academic Press, New York, San Francisco, London, 1977.

[20] E. Dynkin and A. Yushkevich. *Controlled Markov Processes*. Springer-Verlag, Berlin, 1979.

[21] C. Fiechter. Efficient reinforcement learning. In *Proceedings of the Seventh Annual ACM Conference on Computational Learning Theory*, pages 88–97. Association of Computing Machinery, 1994.

[22] M. French, C. Szepesvári, and E. Rogers. Uncertainty, performance, and model dependency in approximate adaptive nonlinear control. In *Proc. of 1997 IEEE Conference on Decision and Decision*, San Diego, California, December 1997. IEEE. in press.

[23] M. French, C. Szepesvári, and E. Rogers. Uncertainty, performance, and model dependency in approximate adaptive nonlinear control. *IEEE Transactions on Automatic Control*, 1997. (submitted).

[24] G. J. Gordon. Stable function approximation in dynamic programming. In A. Prieditis and S. Russell, editors, *Proceedings of the Twelfth International Conference on Machine Learning*, pages 261–268, San Francisco, CA, 1995. Morgan Kaufmann.

[25] T. Graves and T. Lai. Asymptotically efficient adaptive choice of control laws in controlled Markov chains. *SIAM J. Contr. and Opt.*, 35(3):715–743, 1997.

[26] V. Gullapalli and A. Barto. Convergence of indirect adaptive asynchronous value iteration algorithms. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems 6*, pages 695–702. Morgan Kaufmann, April 1994.

[27] V. Gullapalli and A. Barto. Convergence of indirect adaptive asynchronous value iteration algorithms. In J. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems 6*, pages 695–702. San Francisco: Morgan Kauffmann, 1994.

[28] M. Heger. Consideration of risk in reinforcement learning. In *Proc. of the Eleventh International Machine Learning Conference*, pages 105–111, San Francisco, CA, 1994. Morgan Kaufmann.

[29] M. Heger. The loss from imperfect value functions in expectation-based and minimax-based tasks. *Machine Learning*, 22:197–225, 1996.

[30] M. Heger. *Risk-sensitive decision making*. PhD thesis, Zentrum für Kognitionwissenschaften, Universität Bremen, FB3 Informatik, Postfach 330 440, 28334 Bremen, Germany, 1996.

[31] M. Henig. Vector-valued dynamic programming. *SIAM J. Control and Optimization*, 21(3):490–499, 1983.

[32] R. Howard. *Dynamic Probabilistic Systems*. John Wiley, New York, 1970.

[33] T. Jaakkola, M. Jordan, and S. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6(6):1185–1201, November 1994.

[34] G. John. When the best move isn't optimal: Q-learning with exploration. Technical report, Stanford University, 1995. Available on the web.

[35] G. H. John. When the best move isn't optimal: Q-learning with exploration. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, page 1464, Seattle, WA, 1994.

[36] Z. Kalmár, C. Szepesvári, and A. Lőrincz. Module based reinforcement learning for a real robot. *Machine Learning*, 1997. joint special issue on "Learning Robots" with the J. of Autonomous Robots; in press.

[37] Z. Kalmár, C. Szepesvári, and A. Lőrincz. Module based reinforcement learning for a real robot. In *Proc. of the 6th European Workshop on Learning Robots*, Lecture Notes in Artificial Intelligence. Springer, 1998. in press.

[38] M. Kearns and S. Singh. Near-optimal performance for reinforcement learning in polynomial time. Personal Communication, 1998.

[39] V. Konda and V. Borkar. Learning algorithms for Markov decision processes. *SIAM Journal on Control and Optimization*, 1997. submitted.

[40] R. Korf. Real-time heuristic search. *Artificial Intelligence*, 42:189–211, 1990.

[41] H. Kushner and D. Clark. *Stochastic approximation methods for constrained and unconstrained systems.* Springer-Verlag, Berlin, Heidelberg, New York, 1978.

[42] M. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, pages 157–163, San Francisco, CA, 1994. Morgan Kaufmann.

[43] M. Littman. *Algorithms for Sequential Decision Making.* PhD thesis, Brown University, Computer Science Department, 1996.

[44] M. Littman and C. Szepesvári. A Generalized Reinforcement Learning Model: Convergence and applications. In *Int. Conf. on Machine Learning*, pages 310–318, 1996.

[45] L. Ljung. Analysis of recursive stochastic algorithms. *IEEE Tran. Automat. Control*, 22:551–575, 1977.

[46] P. Major. A law of the iterated logarithm for the Robbins-Monro method. *Studia Scientiarum Mathematicarum Hungarica*, 8:95–102, 1973.

[47] A. W. Moore and C. G. Atkeson. Prioritized sweeping: Reinforcement learning with less data and less real time. *Machine Learning*, 13, 1993.

[48] T. Morin. Monotonicity and the principle of optimality. *J. of Math. Anal. and Appl.*, 86:665–674, 1982.

[49] G. Owen. *Game Theory: Second edition.* Academic Press, Orlando, Florida, 1982.

[50] M. Puterman. *Markov Decision Processes — Discrete Stochastic Dynamic Programming.* John Wiley & Sons, Inc., New York, NY, 1994.

[51] C. Ribeiro. Attentional mechanisms as a strategy for generalisation in the Q-learning algorithm. In *Proc. of ICANN'95*, volume 1, pages 455–460, 1995.

[52] C. Ribeiro and C. Szepesvári. Convergence of Q-learning combined with spreading. In *Proc. of ICNAI'96*, 1996. submitted.

[53] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[54] H. Robbins and D. Siegmund. A convergence theorem for non-negative almost supermartingales and some applications. In J. Rustagi, editor, *Optimizing Methods in Statistics*, pages 235–257. Academic Press, New York, 1971.

[55] S. Ross. *Applied Probability Models with Optimization Applications.* Holden Day, San Francisco, California, 1970.

[56] G. Rummery. *Problem solving with reinforcement learning.* PhD thesis, Cambridge University Engineering Department, 1994.

[57] G. A. Rummery and M. Niranjan. On-line Q-learning using connectionist systems. Technical Report CUED/F-INFENG/TR 166, Cambridge University Engineering Department, 1994.

[58] M. Schäl. Estimation and control in discounted dynamic programming. *Stochastics*, 20:51–71, 1987.

[59] P. J. Schweitzer. Aggregation methods for large Markov chains. In G. Iazola, P. J. Coutois, and A. Hordijk, editors, *Mathematical Computer Performance and Reliability*, pages 275–302. Elsevier, Amsterdam, Holland, 1984.

[60] L. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences of the United States of America*, 39:1095–1100, 1953.

[61] S. Singh, T. Jaakkola, and M. Jordan. Reinforcement learning with soft state aggregation. In *Proceedings of Neural Information Processing Systems*, 1995.

[62] S. Singh, T. Jaakkola, M. Littman, and C. Szepesvári. On the convergence of single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 1997. accepted.

[63] S. Singh and R. Sutton. Reinforcement learning with replacing eligibility traces. *Machine Learning*, To appear.

[64] M. Sniedovich. Dynamic programming and principles of optimality. *J. of Math. Anal. and Appl.*, 65:586–606, 1978.

[65] R. Strauch. Negative dynamic programming. *Annals of Math. Statistics*, 37:871–890, 1966.

[66] R. S. Sutton. Generalization in reinforcement learning: Successful examples using sparse coarse coding. *Advances in Neural Information Processing Systems*, 8, 1996.

[67] C. Szepesvári. Certainty equivalence policies are self-optimizing under minimax optimality. Technical Report 96-101, Research Group on Artificial Intelligence, JATE-MTA, Szeged 6720, Aradi vrt tere 1., HUNGARY, August 1996. e-mail: szepes@math.u-szeged.hu, http://www.inf.u-szeged.hu/ rgai.

[68] C. Szepesvári. Some basic facts concerning minimax sequential decision problems. Technical Report 96-100, Research Group on Artificial Intelligence, JATE-MTA, Szeged 6720, Aradi vrt tere 1., HUNGARY, August 1996. e-mail: szepes@math.u-szeged.hu, http://www.inf.u-szeged.hu/ rgai.

[69] C. Szepesvári. The asymptotic convergence-rate of Q-learning. In *Neural Information Processing Systems*, 1997. in press.

[70] C. Szepesvári. Learning and exploitation do not conflict under minimax optimality. In M. Someren and G. Widmer, editors, *Machine Learning: ECML'97 (9th European Conf. on Machine Learning, Proceedings)*, volume 1224 of *Lecture Notes in Artificial Intelligence*, pages 242–249. Springer, Berlin, 1997.

[71] C. Szepesvári. An asynchronous stochastic approximation theorem and its applications. *Alkalmazott Matematikai Lapok*, 1998. accepted (in Hungarian).

[72] C. Szepesvári. Non-markovian policies in sequential decision problems. *Acta Cybernetica*, 1998. accepted.

[73] C. Szepesvári, S. Czimmer, and A. Lőrincz. Neurocontroller using dynamic state feedback for compensatory control. *Neural Networks*, pages 1691–1708, 1997.

[74] C. Szepesvári and M. Littman. A unified analysis of value-function-based reinforcement-learning algorithms. *Neural Computation*, 1997. submitted.

[75] C. Szepesvári and M. L. Littman. Generalized Markov decision processes: Dynamic-programming and reinforcement-learning algorithms. Technical Report CS-96-11, Brown University, Providence, RI, 1996.

[76] C. Szepesvári and A. Lőrincz. Inverse dynamics controllers for robust control: Consequences for neurocontrollers. In *Proc. of ICANN'96*, pages 697–702, 1996.

[77] C. Szepesvári and A. Lőrincz. *Approximate inverse-dynamics based robust control using static and dynamic state feedback*, pages 151–197. World Scientific, Singapore, 1997.

[78] C. Szepesvári and A. Lőrincz. High precision neurocontrol of a chaotic bioreactor. *Nonlinear Analysis*, 30(3):1669–1676, 1997. Proc. of Second World Cong. of Nonlinear Analysts (WCNA96).

[79] J. N. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16(3), September 1994.

[80] J. N. Tsitsiklis and B. Van Roy. Feature-based methods for large scale dynamic programming. *Machine Learning*, 22:59–94, 1996.

[81] S. Verdu and H. Poor. Abstract dynamic programming models under commutativity conditions. *SIAM J. Control and Optimization*, 25(4):990–1006, 1987.

[82] O. J. Vrieze and S. H. Tijs. Fictitious play applied to sequences of games and discounted stochastic games. *International Journal of Game Theory*, 11(2):71–85, 1982.

[83] K.-H. Waldmann. On bounds for dynamic programs. *Mathematics of Operations Research*, 10(2):220–232, May 1985.

[84] C. Watkins. *Learning from Delayed Rewards.* PhD thesis, King's College, Cambridge, 1990.

[85] C. Watkins and P. Dayan. Q-learning. *Machine Learning*, 3(8):279–292, 1992.

[86] P. Whittle. Stability and characterisation conditions in negative programming. *J. Appl. Prob.*, 17:635–645, 1980.