Non-Markovian Policies in Sequential Decision Problems

Csaba Szepesvári *†

Abstract

In this article we prove the validity of the Bellman Optimality Equation and related results for sequential decision problems with a general recursive structure. The characteristic feature of our approach is that also non-Markovian policies are taken into account. The theory is motivated by some experiments with a learning robot.

1 Introduction

The theory of sequential decision problems is an important mathematical tool for studying some problems of cybernetics, e.g. control of robots. Consider for example the robot shown in Figure 1. This robot, called Khepera[1], is equipped with eight infra-red sensors, six in the front and two at the back, the infra- red 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 robot has a vision turret mounted on its top. The vision turret has 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 task of the robot was to find a ball in the arena, bring it to the stick and hit the stick by the ball so as to it jumps out of the gripper. Macro actions such as search, grasp, etc. were defined and the expected number of macro actions taken by the robot until the goal was reached was choosen as the performance measure. Some digital, dynamic filters provide the "state information" necessary for making decisions (for more details concerning these filters see [7]). The robot learnt on-line from the observations (x_t, a_t, c_t), where $x_t \in \mathcal{X}$ is the actual output of the filters ($\mathcal{X}$ is a finite set, called the state space), $a_{t-1} \in \mathcal{A}$ is the previous (macro-)action taken by the robot ($\mathcal{A}$ is also a finite set, called the

*Research Group on Artificial Intelligence, "József Attila" University, Szeged, 6720 Aradi vértanúk tere 1., HUNGARY, e-mail: szepes@math.u-szeged.hu

†Work supported by the Central Research Fund of Hungarian Academy of Sciences (Grant No. T014548)

[1] Khepera is designed and built at Laboratory of Microcomputing, Swiss Federal Institute of Technology, Lausanne, Switzerland and is available commercially.

action set), and c_t is the cost of transition (x_{t-1}, a_{t-1}, x_t) which was 1 until the goal was reached. The task turns out to be well approximated as a Markovian Decision Problem (MDP), i.e. one may assume the existence of transition probabilities of form $p(x,a,y)$, where $p(x,a,y)$ gives the probability of going to state y from state x when action a is used; and the existence of a cost-structure $c(x,a,y)$ s.t. $c_t = c(x_{t-1}, a_{t-1}, x_t)$. The objective is to minimize the total expected discounted cost, $E[\sum_{t=0}^{\infty} \gamma^t c_t]$, $0 < \gamma < 1$, by choosing an appropriate policy, a policy being any function that maps past observations to actions (sometimes to distributions over the action set). Because of the uniform cost structure and the absorbing goal state, the discounted cost criterion can be shown to be equivalent to the undiscounted one, i.e., to minimizing the expected number of steps until the goal is reached. The reason of considering the discounted total cost criterion is that the presence of the discount factor makes the theory of such MDPs quite appealing. In particular, it is well known that policies which, for any given state $x \in \mathcal{X}$, choose the action minimising

$$\sum_{y \in \mathcal{X}} p(x,a,y)(c(x,a,y) + \gamma v^*(y))$$

are optimal. Here v^* is the so-called optimal cost function, defined by

$$v^*(x) = \inf_{\pi \in \Pi} v_\pi(x), \quad x \in \mathcal{X},$$

where Π is the set of policies. More importantly, v^* is known to satisfy the Bellman Optimality Equation

$$v^*(x) = \min_{a \in \mathcal{A}} \sum_{y \in \mathcal{X}} p(x,a,y)(c(x,a,y) + \gamma v^*(y)), \quad x \in \mathcal{X},$$

which is a non-linear equation for v^*. Fortunately, because of the presence of the discount factor, γ, v^* can be found (approximately) in a number of ways. For example, introducing the *optimal cost operator*, $T : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$, defined by

$$(Tv)(x) = \min_{a \in \mathcal{A}} \sum_{y \in \mathcal{X}} p(x,a,y)(c(x,a,y) + \gamma v(y)), \quad x \in \mathcal{X},$$

gives that v^* is the fixed point of T, which can be shown to be a contraction in the sup-norm (in fact, $||Tv - Tu|| \leq \gamma ||v - u||$ holds) and so the Banach fixed-point theorem yields that $v_{n+1} = Tv_n$ converges to v^* in the sup-norm for any choice of v_0. This algorithm, called the *value-iteration algorithm* (or dynamic programming algorithm), served as the basis of the most successful learning algorithm for the above robotic task. The idea of this learning algorithm is to estimate the transition probabilities $p(x,a,y)$ and the costs $c(x,a,y)$ by their respective maximum-likelihood estimates to obtain $p_t(x,a,y)$ and $c_t(x,a,y)$, respectively and then compute v_t, the t^{th} approximation to the optimal cost function v^*, as the fixed point of the approximate optimal cost operator T_t which is defined as T but when p and c are replaced by their respective estimates. After a few hours of learning the performance of the

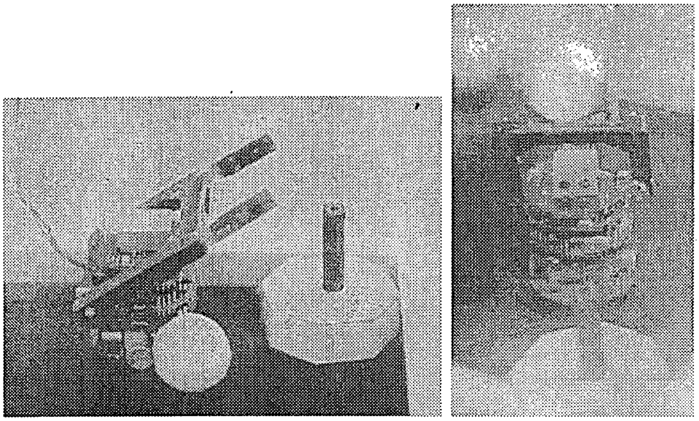

Figure 1: **The Khepera robot**
The figures show a Khepera robot. The description of the sensors and actuators of the robot can be found in the text.

robot using this learning strategy was comparable to that of a well designed control strategy (whose design took a couple of days).

It is important to observe that the *expected* total discounted cost criterion, although suitable in many cases, may yield undesirable behaviour in some cases. For example, in safety-critical applications (like controlling a Mars-rover) the average-case optimal policy may be too bold. Other criteria, such as the minimax criterion which concerns only the long-term worst-case outcomes of the decisions, take safety much better into account. There is a continuum of other criteria which are in between the expected and the minimax criteria. In this article we consider structural questions, such as the validity of the Bellman Optimal Equation, associated with sequential decision problems given by general decision criteria. However, the problem of learning optimal policies will not be considered here. Nevertheless, the theory investigated here is important as the analysis of such learning algorithms should be based on it. For further information on learning issues the reader is referred to the articles [5, 9] and [6]. The main contribution of this article to the "static-theory", which considers structural problems, is that here we do not restrict the analysis to Markovian policies as it is usual in the literature (see, e.g. [1, 10]), but we also consider general policies, which is important since learning policies are non-Markovian by nature.

2 Results

Notation. The set of natural numbers ($\{0,1,2,\ldots\}$), integers and reals will be denoted by $\mathbb{N}, \mathbb{Z}$ and $\mathbb{R}$, respectively. $\mathcal{R}(Z)$ will denote the set of extended real-valued functions over Z: $\mathcal{R}(Z) = [-\infty, \infty]^Z$, and $\mathcal{B}(Z)$ will denote the set of bounded real-valued functions over Z: $\mathcal{B}(Z) \subset \mathbb{R}^Z$, s.t. if $f \in \mathcal{B}(Z)$ then $\|f\| = \sup_{z \in Z} |f(z)| < \infty$. 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 $S_1 \leq S_2$ ($S_1, S_2 : \mathcal{R}(Z) \to \mathcal{R}(Z)$) if $S_1 v \leq S_2 v$ for all $v \in \mathcal{R}(Z)$. If $S : \mathcal{R}(Z) \to \mathcal{R}(Z)$ 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. In the following t, s, n, i, j, k will denote natural numbers.

DEFINITION 2.1 *An operator $S : \mathcal{R}(Z_1) \to \mathcal{R}(Z_2)$ is said to be* Lipschitz *with index $0 < \gamma$ if S maps $\mathcal{B}(Z_1)$ into $\mathcal{B}(Z_2)$ and if for all $f, g \in \mathcal{B}(Z_1)$, $\|Sf - Sg\| \leq \gamma \|f - g\|$. S is said to be a* contraction *with index γ if S is Lipschitz with index $\gamma < 1$.*

DEFINITION 2.2 *An operator $S : \mathcal{R}(Z_1) \to \mathcal{R}(Z_2)$ is said to be (weakly)* continuous *if for all pointwise continuous function sequence $\{f_n\} \subset \mathcal{R}(Z_1)$ with limit function f, also* $\lim_{n \to \infty}(Sf_n)(z) = (Sf)(z)$, $\forall z \in Z_2$.

It is well known that S can be Lipschitz without being continuous and vice versa. continuous in the topology induced by pointwise convergence: Let $S : \mathcal{B}(\mathbb{N}) \to \mathcal{B}(\mathbb{N})$ be defined as $(Sf)(i) = \inf_{j \geq 1} f(j)$ if $i = 0$ and $(Sf)(i) = f(i)$, otherwise. Clearly, S is Lipschitz with index 1. Let $f_n(i) = 1$, if $0 \leq i \leq n$ and $f_n(i) = 0$ otherwise. Now, if we let $f(i) = 1$, $i \in \mathbb{N}$ then $f_n \to f$ pointwise but not in the sup-norm, and $0 = \lim_{n \to \infty}(Sf_n)(0) \neq (Sf)(0) = 1$ showing that S is not continuous in the sense of Definition 2.2.

Sequential Decision Problems.

DEFINITION 2.3 *An* sequential decision problem *(SDP) 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 $\ell \in B(\mathcal{X})$ is the so-called terminal cost function.*

In most of the results we will assume that $\mathcal{Q}$ is a contraction and is continuous in the sense of Definition 2.2.

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

The history of a decision process up to the t^{th} stage is a sequence of state-action pairs: $(a_t, x_t, a_{t-1}, x_{t-1}, \ldots, a_0, x_0)$. Set $H_t = (\mathcal{A} \times \mathcal{X})^t$, $t \geq 0$. For brevity, $h = ((a_t, x_t), \ldots, (a_0, x_0))$ will be written as $h = a_t x_t \ldots a_0 x_0$. 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: $((a_t, x_t), \ldots, (a_0, x_0), (a'_s, x'_s), \ldots, (a'_0, x'_0))$. We admit the assumption that the ordering of the components of $h = a_t x_t \ldots a_0 x_0$ corresponds to the time order, i.e., (a_t, x_t) is the most recent element of the history.

DEFINITION 2.4 *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. *Elements of* $\mathcal{A}^{\mathcal{X}}$ *are called* selectors *and every* $\pi \in \mathcal{A}^{\mathcal{X}}$ *is identified by the associated stationary policy* $(\pi, \pi, \pi, \ldots)$.

DEFINITION 2.5 *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*

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

In the literature the evaluation of Markov policies is defined with the help of the policy-evaluation operators:

DEFINITION 2.6 (BERTSEKAS, 1977) *The* evaluation 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*

$$v_\pi = \lim_{t \to \infty} T_{\pi_0} T_{\pi_1} \ldots T_{\pi_t} \ell, \tag{1}$$

assuming that the limit exists.

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

$$v_\pi = \lim_{n \to \infty} T_{\pi_0}^n \ell. \tag{2}$$

Note that if $\mathcal{Q} : B(\mathcal{X}) \to B(\mathcal{X} \times \mathcal{A})$ *is a contraction then* T_π *is a contraction with the same index* ($\pi \in \mathcal{A}^{\mathcal{X}}$) and so v_π is well defined. The evaluation of arbitrary policies is more complicated and is the subject of the next section, but the following example may shed some light on the forthcoming definitions.

EXAMPLE 2.7 *Finite Markovian decision problems with the expected total cost criterion [2, 8].* $(\mathcal{X}, \mathcal{A}, p, c)$ 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., for all $(x, a, y) \in \mathcal{X} \times \mathcal{A} \times \mathcal{X}$, $0 \leq p(x, a, y) \leq 1$; and for all $(x, a) \in \mathcal{X} \times \mathcal{A}$, $\sum_{y \in \mathcal{X}} p(x, a, y) = 1$;
3. $c : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \mathbf{R}$.

Now let π be any policy. Then, for any X-valued random variable ξ_0, π generates a probability measure $P = P_{\xi_0, \pi}$ over $(\mathcal{X} \times \mathcal{A})^{\mathrm{N}}$ which is uniquely defined by the finite-dimensional probabilities

$$\begin{aligned} P(x_0, a_0, x_1, a_1, \ldots, x_n, a_n) = p(\xi_0 = x_0)\delta(a_0, \pi_0(x_0))p(x_0, a_0, x_1) \ldots \\ \ldots p(x_{n-1}, a_{n-1}, x_n)\delta(a_n, \pi_n(x_n, a_{n-1}x_{n-1} \ldots a_0 x_0), \end{aligned}$$

where $\delta : \mathcal{A}^2 \rightarrow \{0,1\}$ is defined by $\delta(a,b) = 1$ iff $a = b$. Clearly, one can construct a random sequence $(\xi_n, \alpha_n) \in \mathcal{X} \times \mathcal{A}$ (the controlled object) s.t. $P(\xi_{n+1}|\alpha_n, \xi_n, \ldots, \alpha_0, \xi_0) = p(\xi_n, \alpha_n, \xi_{n+1})$ and where $\alpha_n = \pi_n(\xi_n, \alpha_{n-1}\xi_{n-1} \ldots \alpha_0\xi_0)$. If ξ_0 is concentrated on $\{x\}$ for some $x \in X$ then $P_{\xi_0,\pi}$ is denoted by $P_{x,\pi}$.

Assume that $P(\xi_0 = x) > 0$ for all $x \in \mathcal{X}$. The evaluation of a policy π in state x is defined as

$$\hat{v}_\pi(x) \stackrel{\text{def}}{=} E_{\xi_0,\pi}\left[\sum_{t=0}^{\infty} \gamma^t c(\xi_t, \alpha_t, \xi_{t+1}) \,\middle|\, \xi_0 = x\right] = E_{x,\pi}\left[\sum_{t=0}^{\infty} \gamma^t c(\xi_t^{(x)}, \alpha_t^{(x)}, \xi_{t+1}^{(x)})\right]$$

where $0 < \gamma < 1$ is the discount factor and the expectation is taken w.r.t. $P_{\xi_0,\pi}$ (resp. $P_{x,\pi}$) and $\{(\xi_t, \alpha_t)\}$ (resp. $\{(\xi_t^{(x)}, \alpha_t^{(x)})\}$) is the controlled object corresponding to π and the initial random state ξ_0 (resp. initial state x). The second equality in the above equation comes from the definitions and shows that $\hat{v}_\pi(x)$ is independent of the particular form of ξ_0. Now, if $\pi = (\pi_0, \pi_1, \ldots)$ is any policy then by the law of total probability

$$\hat{v}_\pi(x) = \sum_{y\in\mathcal{X}} P_{x,\pi}\left(\xi_1^{(x)} = y\right) E_{x,\pi}\left[\sum_{t=0}^{\infty} \gamma^t c(\xi_t^{(x)}, \alpha_t^{(x)}, \xi_{t+1}^{(x)}) \,\middle|\, \xi_1^{(x)} = y\right] =$$

$$= \sum_{y\in\mathcal{X}} p(x, \pi_0(x), y)\left(c(x, \pi_0(x), y) + \gamma E_{x,\pi}\left[\sum_{t=0}^{\infty} \gamma^t c(\xi_{t+1}^{(x)}, \alpha_{t+1}^{(x)}, \xi_{t+2}^{(x)}) \,\middle|\, \xi_1^{(x)} = y\right]\right) =$$

$$= \sum_{y\in\mathcal{X}} p(x, \pi_0(x), y)\left(c(x, \pi_0(x), y) + \gamma E_{y,\pi^x}\left[\sum_{t=0}^{\infty} \gamma^t c(\hat{\xi}_t^{(y)}, \hat{\alpha}_t^{(y)}, \hat{\xi}_{t+1}^{(y)})\right]\right) =$$

$$= \sum_{y\in\mathcal{X}} p(x, \pi_0(x), y)\Big(c(x, \pi_0(x), y) + \gamma \hat{v}_{\pi^x}(y)\Big) =$$

$$= \big(\mathcal{Q}\hat{v}_{\pi^x}\big)(x, \pi_0(x)), \tag{3}$$

where π^x denotes the policy executed after the first step, i.e., $\pi^x = (\pi_0^x, \pi_1^x, \ldots)$ with $\pi_t^x(x,h) = \pi_{t+1}(x, h\pi_0(x)x)$, $\{(\hat{\xi}_t^{(y)}, \hat{\alpha}_t^{(y)})\}$ is the corresponding controlled object given that the initial state is y, and $\mathcal{Q} : \mathcal{R}(\mathcal{X}) \rightarrow \mathcal{R}(\mathcal{X} \times \mathcal{A})$ is defined by

$$(\mathcal{Q}f)(x,a) = \sum_{y\in\mathcal{X}} p(x,a,y)(c(x,a,y) + \gamma f(y)).. \tag{4}$$

Equation 3 is called the Fundamental Equation [3] and will be proved to hold for general SDPs in the next section. Note that if π is Markovian then $\pi^x = (\pi_1, \pi_2, \ldots)$ for any $x \in X$, and so Equation 3 yields that $v_{(\pi_0,\pi_1,\ldots)} = T_{\pi_0} \ldots T_{\pi_t} v_{(\pi_{t+1},\pi_{t+2},\ldots)}$. Therefore, for any given $\ell \in B(\mathcal{X})$, $\hat{v}_\pi = \lim_{t\rightarrow\infty} T_{\pi_0} \ldots T_{\pi_t}\ell = v_\pi$ since

$$\|T_{\pi_0} \ldots T_{\pi_t} \hat{v}_{(\pi_{t+1},\pi_{t+2},\ldots)} - T_{\pi_0} \ldots T_{\pi_t}\ell\| \leq \gamma^{t+1}\|\hat{v}_{(\pi_{t+1},\pi_{t+2},\ldots)} - \ell\| \leq \gamma^{t+1} C \overset{t\rightarrow\infty}{\longrightarrow} 0,$$

for some $C > 0$, where we exploited that for any selector π, T_π is a contraction with index γ and that $\sup_{\pi\in\Pi} \|v_\pi\| < \infty$.

Interesting "risk-sensitive" criteria may be obtained if $\mathcal{Q}$ is given by $(\mathcal{Q}f)(x,a) = (\sum_{y\in\mathcal{X}} p(x,a,y)(c(x,a,y) + \gamma f(y))^p)^{1/p}$, $1 \le p < \infty$, where c and f are assumed to be non-negative. This definition can be shown to give the minimax criterion when $p \to \infty$. The results derived below hold for these criteria as well.

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 smallest cost that can be achieved depends on the class of policies available for the decision maker.

DEFINITION 2.8 *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 function for the class Π_Δ, where Δ is either g or m or s.

For any $\varepsilon \ge 0$ and fixed $x \in \mathcal{X}$ the decision maker can assure a cost less than $v^{*\Delta}(x) + \varepsilon$ by the usage of an appropriate policy from Π_Δ but this policy will depend on x. Here we are interested in *uniformly* good policies:

DEFINITION 2.9 *Let $\Pi_\Delta(v) = \{\pi \in \Pi_\Delta \,|\, v_\pi \le v\}$, that is $\Pi_\Delta(v)$ contains the policies from Π_Δ whose cost is uniformly less than or equal to v. A policy is said to be* (uniformly) ε-optimal *if it is contained in $\Pi_g(v^{*g} + \varepsilon)$.*[2]

The objective of sequential decision problems is to give conditions under which $\Pi_\Delta(v^{*g} + \varepsilon)$ is guaranteed to be non-empty when $\varepsilon > 0$ or $\varepsilon = 0$.

DEFINITION 2.10 *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.*

The Fundamental Equation. Now we define the evaluation function associated to non-Markovian policies and derive the fundamental equation.

DEFINITION 2.11 *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 (finite-horizon) policies and the set of (infinite horizon) policies, respectively. The s-truncation operator for t-truncated policies is defined similarly if $s \le t$.*

DEFINITION 2.12 *The* shift-operator, *$S_{(x,a)} : \mathcal{P} \to \mathcal{P}$, for any pair $(x,a) \in \mathcal{X} \times \mathcal{A}$ is defined in the following way:*

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

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

where π'_t is defined by

$$\pi'_t(x,h) = \pi_{t+1}(x, hax)$$

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 t-truncated 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.

PROPOSITION 2.13 $\pi^{t,x} = \pi^{x,t-1}$ *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 2.14 *If $\pi \in \mathcal{P}_0$, i.e., $\pi = (\pi_0)$ then $v_\pi(x) = (Q\ell)(x, \pi_0(x))$, 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*

$$v_\pi(x) = (Qv_{\pi^x})(x, \pi_0(x)). \tag{5}$$

Since $\pi^x \in \mathcal{P}_t$, v_{π^x} is already defined and thus (5) 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 the derived policy is v_{π^x}. This cost together with the cost of the first action (the first being $\pi_0(x)$ in state x) gives the total cost of the policy.

EXAMPLE 2.15 If π is a t-horizon policy in an MDP $(\mathcal{X}, \mathcal{A}, p, c)$ (cf. Example 2.7) and we set

$$\hat{v}^{(t)}_\pi(x) = E\left[\sum_{n=0}^{t-1} \gamma^n c(\xi_n, \alpha_n, \xi_n) \,\middle|\, \xi_0 = x\right],$$

where $\{(\xi_n, \alpha_n)\}$ is the controlled object corresponding to π and the random initial state ξ_0. The argument of Example 2.7 gives that $v^{(t)}_\pi = \hat{v}^{(t)}_\pi$, where $v^{(t)}_\pi$ is the evaluation of π in the sense of Definition 2.14 in the SDP $(\mathcal{X}, \mathcal{A}, Q, \ell)$, with Q given by (4) and where $\ell(x) = 0$ for all $x \in \mathcal{X}$.

The evaluation of an infinite horizon policy is defined as the limit of the evaluations of the finite horizon truncations of that policy:

DEFINITION 2.16 *Let $\pi \in \mathcal{P} = \mathcal{P}_\infty$. Then the total cost of π for initial state x is given by*

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

EXAMPLE 2.17 Continuing the above example, if π is an arbitrary policy then (by the boundedness of c)

$$\hat{v}_\pi(x) \overset{\text{def}}{=} E\left[\sum_{n=0}^{\infty} \gamma^n c(\xi_n, \alpha_n, \xi_n) \mid \xi_0 = x\right] = \lim_{t\to\infty} E\left[\sum_{n=0}^{t-1} \gamma^n c(\xi_n, \alpha_n, \xi_n) \mid \xi_0 = x\right],$$

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

DEFINITION 2.18 *Q is said to be monotone if $Qv \leq Qu$ whenever $u \leq v$.*

In what follows we will always assume that Q is monotone.

Equation 6 below, which in harmony with [3] we call the fundamental equation (FE), has already been derived for MDPs in Example 2.7. Here we show that it holds in general SDPs when Q is continuous.

THEOREM 2.19 *If Q is continuous then*

$$v_\pi(x) = (Qv_{\pi^x})(x, \pi_0(x)). \tag{6}$$

Proof. Let $v_t = v_{\pi^t}$ and let $\mu = \pi^{t+1}$. By definition $v_\mu(x) = (Qv_{\mu^x})(x, \mu_0(x))$. According to Proposition 2.13 $\mu^x = \pi^{t+1,x} = \pi^{x,t}$ and $\mu_0 = \pi_0$, therefore

$$v_{\pi^{t+1}}(x) = (Qv_{\pi^{x,t}})(x; \pi_0(x)). \tag{7}$$

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

$$v_\pi(x) = \liminf_{t\to\infty}(Qv_{\pi^{x,t}})(x, \pi_0(x)) = \left(Q[\liminf_{t\to\infty} v_{\pi^{x,t}}]\right)(x, \pi_0(x)) = (Qv_{\pi^x})\,(x, \pi_0(x)),$$

where in the first equation we exploited the definition v_π, in the second equation we used that Q is monotone and is continuous, and in the third equation the definition of v_{π^x} was utilised. □

COROLLARY 2.20 *Assume that Q is a contraction and is continuous. Then v_{π^t} converges to v_π, i.e., in Definition 2.16* lim inf *can be replaced by* lim, *and for any Markovian policy π, the evaluation function associated to π in the sense of Definition 2.16 coincides with the evaluation function in the sense of Definition 2.6. Moreover, if π is stationary then $T_\pi^n v_0$ converges to v_π, where $v_0 \in B(\mathcal{X})$ is arbitrary, and $v_\pi = T_\pi v_\pi$.*

Proof. Recall that in Definition 2.6 the evaluation of a Markovian policy $\pi = (\pi_0, \pi_1, \ldots, \pi_t, \ldots)$ was defined as the limit $\hat{v}_\pi(x) = \lim_{t\to\infty}\left(T_{\pi_0} \cdots (T_{\pi_{t-1}}(T_{\pi_t}\ell)) \ldots\right)$. Easily, $T_{\pi_0} \ldots T_{\pi_{t-1}} T_{\pi_t}\ell = v_{\pi^t}$, so the definition of Bertsekas coincides with that of ours. The rest of the statement follows from the Banach fixed-point theorem. □

Uniformly Optimal Policies.

DEFINITION 2.21 *Policy π is said to be uniformly ε-optimal if, for all $x \in \mathcal{X}$:*

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

THEOREM 2.22 *If the FE is satisfied then for all $\varepsilon > 0$ there exists an ε-optimal policy.*

Proof. Fix an arbitrary $x \in \mathcal{X}$. By the definition of $v^{*g}(x)$ there exists a policy ${}_x\pi = ({}_x\pi_0, {}_x\pi_1, \ldots)$ 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, called the *combination* of the policies ${}_x\pi$, is given formally by $\pi_0(x) = {}_x\pi_0(x)$ and $\pi_t(y, hax) = {}_x\pi_t(y, hax)$. 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 $v_\pi(x) = (Qv_{{}_x\pi^x})(x, \pi_0(x)) = (Qv_{{}_x\pi^x})(x, {}_x\pi_0(x)) = v_{{}_x\pi}(x)$. □

Finite Horizon Problems.

DEFINITION 2.23 *The optimal cost function for n-horizon problems is defined by*

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

where $\mathcal{P}_n^\Delta = \{\, \pi^n \mid \pi \in \Pi_\Delta \,\}$, *and* $\Delta \in \{g, m, s\}$.

DEFINITION 2.24 *The* optimal cost operator $T : \mathcal{R}(\mathcal{X}) \to \mathcal{R}(\mathcal{X})$ *associated with the SDP* $(\mathcal{X}, \mathcal{A}, \mathcal{Q}, \ell)$ *is defined by*

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

It is immediate from the definition and by the triangle inequality that if $\mathcal{Q}$ is a contraction with index γ then the optimal cost operator is a contraction with the same index.

DEFINITION 2.25 $\mathcal{Q}$ *is called* upper semi-continuous *if for every (pointwise) convergent sequence of functions* $v_t \in \mathcal{R}(\mathcal{X})$ *for which* $v_t \ge \lim_{t\to\infty} v_t$ *we have*

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

THEOREM 2.26 (OPTIMALITY EQUATION FOR FINITE HORIZON PROBLEMS) *The optimal cost functions of the n-horizon problem satisfies*

$$v_n^{*g} = v_n^{*m} = T^n \ell \tag{8}$$

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 proven 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, $v_\pi(x) = (Qv_{\pi^x})(x, \pi_0(x))$. Since $\pi^x \in \mathcal{P}_n$ so $v_{\pi^x} \ge v_n^{*g}$. Since Q is monotone it follows that

$$(Tv_n^{*g})(x) = \inf_{a \in \mathcal{A}(x)} (Qv_n^{*g})(x,a) \le \inf_{a \in \mathcal{A}(x)} (Qv_{\pi^x})(x,a) \le (Qv_{\pi^x})(x, \pi_0(x) = v_\pi(x).$$

This 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 $\lim_{j\to\infty} T_{\mu_j} v_n^{*g} = Tv_n^{*g}$. 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 Q is USC. Now taking the limit in j the induction hypothesis yields that $v_{n+1}^{*g} \le v_{n+1}^{*m} \le Tv_n^{*m} = T^{n+1}\ell$ which finally gives that $v_{n+1}^{*g} = v_{n+1}^{*m} = T^{n+1}\ell$, completing the proof. □

The following example shows that the conditions of the previous theorem are indeed essential.

EXAMPLE 2.27 [1] Let $\mathcal{X} = \{0\}$ and $\mathcal{A} = (0,1]$, $\ell(0) = 0$, and $(Qf)(0,a) = 1$, if $f(0) > 0$; and $(Qf)(0,a) = a$, otherwise. Note that 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 $n \ge 2$.

The Bellman Optimality Equation. According to Theorem 2.26, if v_n^{*g} converges to v^{*g} then v^{*g} can be computed as the limit of the function sequence $v_0 = \ell$, $v_{t+1} = Tv_t$ provided that 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}:

$$v^{*g} = \inf_{\pi\in\mathcal{P}} \lim_{n\to\infty} v_{\pi^n} = \lim_{n\to\infty} \inf_{\pi\in\mathcal{P}} v_{\pi^n} = \lim_{n\to\infty} v_n^{*g}. \tag{9}$$

THEOREM 2.28 *(i) Assume that Q is continuous and set $v_\infty = \limsup_{n\to\infty} T^n\ell$. Then*

$$v_\infty \le v^{*g}. \tag{10}$$

(ii) If we further assume that Q is a contraction then $\lim_{n\to\infty} v_n^{*g} = \lim_{n\to\infty} T^n v_0 = v^{*g} = v^{*m}$, where $v_0 \in B(\mathcal{X})$ is arbitrary, and

$$Tv^{*g} = v^{*g}. \tag{11}$$

Proof. (i) Note that by Theorems 2.19 & 2.26 $v_\infty = \limsup_{n\to\infty} v_n^{*g}$. Let $x \in \mathcal{X}$ and let c be a number 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 $v_n^{*g}(x) < c$ and consequently $\limsup_{n\to\infty} v_n^{*g}(x) < c$. Since c and x were arbitrary, we obtain the desired inequality.

(ii) By the Banach fixed-point theorem $v_\infty = \lim_{n\to\infty} T^n \ell = \lim_{n\to\infty} T^n v_0$ and $Tv_\infty = v_\infty$. It is sufficient to prove that $v^{*g} \le Tv^{*g}$ since then iterating this inequality will yield $v^{*g} \le T^n v^{*g} \to v_\infty, n \to \infty$, which together with Part (i) shows (11). Let π_n be a sequence of $1/n$-uniformly optimal policies. Such policies exist by Theorem 2.22. 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 in n yields the desired inequality. □

Existence of Optimal Stationary Policies.

DEFINITION 2.29 *A stationary policy ϕ is said to be* greedy *w.r.t. $v \in \mathcal{R}(\mathcal{X})$ if*

$$T_\phi v = Tv,$$

i.e., if for each $x \in \mathcal{X}$, $(Qv)(x, \phi(x)) = (Tv)(x) = \inf_{a\in\mathcal{A}}(Qv)(x,a)$.

Note that the finiteness of $\mathcal{A}$ assures the existence of greedy policies w.r.t. any function $v \in \mathcal{R}(\mathcal{X})$. If $\mathcal{A}$ is infinite special continuity assumptions are needed on Q for the existence of greedy policies (see [1] for further information on this). The next theorem shows that greediness is a useful concept under the appropriate conditions since the knowledge of the optimal cost function can be sufficient to find optimal stationary policies.

THEOREM 2.30 *If Q is a contraction and is continuous then optimal stationary policies are greedy w.r.t. v^{*g}; and vice versa.*

Proof. If ϕ is greedy w.r.t. v^{*g} then $T_\phi v^{*g} = Tv^{*g} = v^{*g}$ and by induction we get that $T_\phi^n v^{*g} = v^{*g}$ holds for all n. Since by Corollary 2.20 the l.h.s. converges to v_ϕ as $n \to \infty$, we get that $v_\phi = v^{*g}$, i.e., ϕ is optimal. Now, 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. □

Theorems 2.28 & 2.30 are at the very core of the learning algorithms used in the robotic experiments. In particular, a theorem was proven in [5] which shows that in contractive models (i.e., when Q is a contraction) value iteration can be combined with learning processes without effecting the convergence. In [9] and [6] examples are shown for asymptotically optimal learning policies which use the adaptive value iteration scheme.

Final Remarks. Similar statements hold for models when $(Q\ell)(\cdot, a) \geq \ell$ or when $(Q\ell)(\cdot, a) \leq \ell$ $(a \in \mathcal{A})$ in which cases we require Q to be lower (resp. upper) semicontinuous on the set of functions $\{v \in \mathcal{R}(\mathcal{X}) \,|\, v \geq \ell\}$ (resp. $\{v \in \mathcal{R}(\mathcal{X}) \,|\, v \leq \ell\}$). In such cases the analysis should be based on the monotonicity of the various function sequences involved. However, problems related to the existence of stationary optimal policies become more complicated: in fact for models satisfying $(Q\ell)(\cdot, a) \geq \ell$ (these are called increasing models) value iteration does not necessarily converge to v^{*g}, but greedy policies w.r.t. v^{*g} are optimal; whilst in models satisfying the opposite inequality, $(Q\ell)(\cdot, a) \leq \ell$ (these are called decreasing models) value iteration does always converge to v^{*g} but greedy policies w.r.t. v^{*g} are not necessarily optimal. It is also worth noting that Howard's policy improvement theorem [4] is valid in increasing or contractive models, and when iterated converges to optimum in contractive models [5] but does not necessarily converge to optimum in increasing ones. In certain contractive models one can estimate the speed of convergence of both the value and the policy iteration methods which turns out to be pseudo-polynomial [5].

Acknowledgements. This work was partially supported by OTKA Grants. F20132 and T014548 and by a grant provided by the Hungarian Educational Ministry under contract no. FKFP 1354/1997. I would like to express my sincere gratitude to András Krámli for the many discussions on the topics of this article.

References

[1] D. P. Bertsekas. Monotone mappings with application in dynamic programming. *SIAM J. Control and Optimization*, 15(3):438–464, 1977.

[2] D. P. Bertsekas. *Dynamic Programming: Deterministic and Stochastic Models.* Prentice-Hall, Englewood Cliffs, NJ, USA, 1989.

[3] E. Dynkin and A. Yushkevich. *Controlled Markov Processes.* Springer-Verlag, Berlin, 1979.

[4] R. A. Howard. *Dynamic Programming and Markov Processes.* The MIT Press, Cambridge, MA, 1960.

[5] M. L. Littman and Cs. Szepesvári. A Generalized Reinforcement Learning Model: Convergence and applications. In *Int. Conf. on Machine Learning*, pages 310–318, 1996.

[6] Cs. 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.

[7] Zs. Kalmár, Cs. Szepesvári, and A. Lőrincz. Module based reinforcement learning for a real robot. In *Proc. of the 6th European Workshop on Learning Robots*, pages 22–32, 1997.

[8] S. Ross. *Applied Probability Models with Optimization Applications.* Holden Day, San Francisco, California, 1970.

[9] S. Singh, T. Jaakkola, M. L. Littman, and Cs. Szepesvári. On the convergence of single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 1997. submitted.

[10] S. Verdu and H. Poor. Abstract dynamic programming models under commutativity conditions. *SIAM J. Control and Optimization*, 25(4):990–1006, 1987.

Received October, 1997