Some basic facts concerning minimax sequential decision processes

Csaba Szepesvári

August 1996

TR-96-100

Research **G**roup on **A**rtificial **I**ntelligence
Hungarian Academy of Sciences,
"Attila József" University
Szeged, 6720, Aradi vrt tere 1.
HUNGARY
`http://www.inf.u-szeged.hu/~rgai`

This research was supported by OTKA Grant No. F20132.

Some basic facts concerning minimax sequential decision processes

Csaba Szepesvári

Bolyai Institute of Mathematics,

"József Attila" University,

Szeged 6720, Aradi vrt. tere 1.

Research Group on Artificial Intelligence

Hungarian Academy of Sciences,

"Attila József" University

Szeged, 6720, Aradi vrt tere 1.

HUNGARY

szepes@math.u-szeged.hu

1 Introduction

A finite Markovian decision problem is determined by four objects, (X, A, P, c). X and A are non-empty finite sets, P is a conditional probability on X given $X \times A$, and c is a function on $X \times A \times X$. We interpret X as the set of states of some system, and A as the actions available in each state. (For simplicity, the set of actions available is assumed to be independent of the state.) When the system is in state x and we take action a, we move to a new state y selected with probability

$P_{xy}(a)$ ($P_{x\cdot}(a) = P(\cdot|x,a)$), and we incur a cost $c(x,a,y)$. The process is then repeated from the new state y and we wish to minimize the worst-case total cost over the infinite future.

A *policy* π is a sequence $\pi_0, \pi_1, \ldots$, where π_n maps the history $h = (x_0, a_0, x_1, a_1, \ldots, a_{n-1}, x_n)$ of the process up to the nth stage to an action. Application of a policy means that we always choose the action prescribed by the given policy and thus the application of a policy from a given initial state generates a random element of $(X \times A)^{\mathbf{N}}$. Certain type of policies are of special interest. A *Markov policy* is one in which π_n depends only on x_n. A *stationary policy* is a Markov policy in which $\pi_n = \pi$ for some function $\pi : X \to A$.

Associated with each policy π is a function on X, $v_\pi(x)$, the total worst-case cost starting from state x and using π. Function v_π is called the evaluation of policy π. This total return may well be infinite. Thus we assume that our future cost is discounted with a discount factor γ, $0 < \gamma < 1$, so that the return (or cost) of one unit n stages in the future is worth γ^n now. Then the actual total cost will be bounded.

Our main results are the following: the evaluation function of a stationary policy is the fixed point of the so called policy evaluation operator which is a contraction mapping. From this it follows that Bellman's principle of optimality holds. We also prove that the asynchronous value iteration algorithm converges to the optimal value function.

2 Optimal policies

Let us denote by $c_t^\pi(x)$ the cost incurred at stage t assuming that we follow the policy π and the process was started from state x. By definition

$$v_\pi(x) = \operatorname*{esssup}_\pi \sum_{t=0}^{\infty} \gamma^t c_t^\pi(x),$$

where $\operatorname{esssup}_\pi$ denotes the supremum over the support of the measure defined over the set of all trajectories, the measure being generated by the decision process and π [5].

DEFINITION 2.1 *We say that policy π^* is optimal if*

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

where

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

denotes the optimal value function (the infimum is taken over all possible policies).

3 Iterative evaluation of stationary policies

Let $T(x,a) \subseteq X$ denote the support of $P(\cdot|x,a)$. Further, let us denote by $B(X)$ the set of bounded, real valued functions over X. Let π be a stationary policy and let us define the operator $T_\pi : B(X) \to B(X)$ by

$$(T_\pi v)(x) = \max_{y \in T(x,\pi(x))} \left(c(x,a,y) + \gamma v(y) \right), \tag{1}$$

where $v \in B(X)$ is arbitrary. Let $\|u\|$ denote the supremum of $u \in B(X)$:

$$\|u\| = \sup_{x \in X} |u(x)|. \tag{2}$$

LEMMA 3.1 *Operator T_π is a contraction with index γ. That is for all $u, v \in B(X)$ there holds that $\|T_\pi u - T_\pi v\| \leq \gamma \|u - v\|$.*

Proof. Let us fix $u, v \in B(X)$. In order to estimate $||T_\pi u - T_\pi v||$ from above first let us estimate the difference $(T_\pi u)(x) - (T_\pi v)(x)$:

$$\begin{aligned} ((T_\pi u)(x) - (T_\pi v)(x)) &= \max_{y \in T(x,\pi(x))} (c(x,\pi(x),y) + \gamma u(y)) \\ &\quad - \max_{y \in T(x,\pi(x))} (c(x,\pi(x),y) + \gamma v(y)) \\ &= c(x,\pi(x),\bar{y}) + \gamma u(\bar{y}) \\ &\quad - \max_{y \in T(x,\pi(x))} (c(x,\pi(x),y) + \gamma v(y)), \end{aligned}$$

where $\bar{y}$ is such that

$$c(x,\pi(x),\bar{y}) + \gamma u(\bar{y}) = \max_{y \in T(x,\pi(x))} (c(x,\pi(x),y) + \gamma u(y)).$$

From this it follows

$$\begin{aligned}((T_\pi u)(x) - (T_\pi v)(x)) &\le \gamma(u(\bar{y}) - v(\bar{y})) \\ &\le \gamma \sup_{y \in X} (u(y) - v(y)) \\ &\le \gamma \|u - v\|.\end{aligned}$$

Since these estimates hold for all $x \in X$ thus

$$\sup_{x \in X} ((T_\pi u)(x) - (T_\pi v)(x)) \le \gamma \|u - v\|.$$

Exchanging u and v in the above inequality yields

$$\sup_{x \in X} ((T_\pi v)(x) - (T_\pi u)(x)) \le \gamma \|u - v\|.$$

From these follows that

$$\sup_{x \in X} |(T_\pi u)(x) - (T_\pi v)(x)| \le \gamma \|u - v\|,$$

and hence

$$\|T_\pi u - T_\pi v\| \le \gamma \|u - v\|.$$

□

Proposition 3.2 *If π denotes an arbitrary stationary policy then*[1]

$$v_\pi(x) = \lim_{n \to \infty} (T_\pi^n v)(x).$$

Proof. Since we know that T_π is a contraction thus by the Contraction Mapping Theorem $T_\pi^n v$ converges to the fixed point of T_π. Consequently, it is sufficient to show that v_π is the fixed point of T_π, that is

$$v_\pi(x) = \max_{y \in T(x, \pi(x))} \left(c(x, \pi(x), y) + \gamma v_\pi(y) \right).$$

For brevity let us denote $c(x, \pi(x), y)$ by $c_\pi(x, y)$ and further let us denote the infinite sum $\sum_{t=0}^{\infty} \gamma^t c_\pi(x_t, x_{t+1})$ by $c_\pi(x_0, x_1, x_2, \ldots)$. Using this notation $v_\pi(x_0)$ is just equal to

$$\sup_{x_1 \in T_\pi(x_0), x_2 \in T_\pi(x_1), \ldots} c_\pi(x_0, x_1, x_2, \ldots),$$

[1] $T^n v = TT \ldots Tv$, where T is applied n times on v.

where we have used the shorthand notation $T_\pi(x) = T(x, \pi(x))$. Let us now fix an arbitary positive number, ϵ. Then there exists a trajectory with states $x_1^*, x_2^*, \ldots, x_t^*, \ldots$ such that $x_1^* \in T_\pi(x_0)$, $x_2^* \in T_\pi(x_1)$, $\ldots$, $x_{t+1}^* \in T_\pi(x_t^*)$, and

$$v_\pi(x_0) - \epsilon < c(x_0, x_1^*, x_2^*, \ldots, x_t^*, \ldots).$$

Clearly,

$$\begin{aligned}
c(x_0, x_1^*, x_2^*, \ldots, x_t^*, \ldots) &\leq \sup_{x_2 \in T_\pi(x_1^*), x_3 \in T_\pi(x_2), \ldots} c(x_0, x_1^*, x_2, \ldots, x_t, \ldots) \\
&\leq \max_{x_1 \in T_\pi(x_0)} \Big(\sup_{x_2 \in T_\pi(x_1), x_3 \in T_\pi(x_2), \ldots} c(x_0, x_1, x_2, \ldots, x_t, \ldots) \Big) \\
&= \max_{x_1 \in T_\pi(x_0)} \Big(\sup_{x_2 \in T_\pi(x_1), x_3 \in T_\pi(x_2), \ldots} c_\pi(x_0, x_1) + \gamma \sum_{t=0}^{\infty} \gamma^t c_\pi(x_{t+1}, x_{t+2}) \Big) \\
&= \max_{x_1 \in T_\pi(x_0)} \Big(c_\pi(x_0, x_1) + \gamma \sup_{x_2 \in T_\pi(x_1), x_3 \in T_\pi(x_2), \ldots} \sum_{t=0}^{\infty} \gamma^t c_\pi(x_{t+1}, x_{t+2}) \Big).
\end{aligned}$$

Observe that

$$\sup_{x_2 \in T_\pi(x_1), x_3 \in T_\pi(x_2), \ldots} \sum_{t=0}^{\infty} \gamma^t c_\pi(x_{t+1}, x_{t+2})$$

is equal to $v_\pi(x_1)$ and thus yields

$$v_\pi(x_0) - \epsilon < \max_{x_1 \in T_\pi(x_0)} (c_\pi(x_0, x_1) + \gamma v_\pi(x_1)) = (T_\pi v_\pi)(x_0).$$

Since ϵ was arbitrary, thus $v_\pi \leq T_\pi v_\pi$.

On the other hand,

$$\sup_{x_1 \in T_\pi(x_0), x_2 \in T_\pi(x_1), \ldots} c_\pi(x_0, x_1, x_2, \ldots) \geq \max_{x_1 \in T_\pi(x_0)} \Big(\sup_{x_2 \in T_\pi(x_1), \ldots} c_\pi(x_0, x_1, x_2, \ldots) \Big) = (T_\pi v_\pi)(x_0),$$

and thus $T_\pi v_\pi = v_\pi$. □

4 Myopic and optimal policies

Definition 4.1 *We say that the policy π is* myopic *w.r.t. the function v if*

$$\pi(x) \in \operatorname*{Argmin}_{a \in A} \max_{y \in T(x,a)} (c(x, a, y) + \gamma v(y)).^2$$

[2] By definition $\operatorname{Argmin}_{a \in A} v(a)$ denotes the subset of A on which v attaches its minimum.

In this section we show that optimal policies exists and are myopic w.r.t. the optimal value function $v^*(x)$. First we need the following technical lemma:

Lemma 4.1 *There exist numbers λ_{min} and λ_{max} such that for all (λ_1, λ_2) pair satisfying $\lambda_1 \le \lambda_{min}$ and $\lambda_2 \ge \lambda_{max}$ there holds that if $v_{min}(x) = \lambda_1, v_{max}(x) = \lambda_2$ $(x \in X)$, then*

$$v_{min} \le Gv_{min}, \qquad Gv_{max} \le v_{max}.$$

Proof. Let $\lambda \in \mathbf{R}$ and $v(x) = \lambda, (x \in X)$. Then

$$(Gv)(x) = \min_{a\in A} \max_{y\in T(x,a)} (c(x,a,y) + \gamma v(y)) = \min_{a\in A} \max_{y\in T(x,a)} c(x,a,y) + \gamma\lambda.$$

Let

$$\hat{c}(x) = \min_{a\in A} \max_{y\in T(x,a)} c(x,a,y),$$

then $(Gv)(x) = \gamma\lambda + \hat{c}(x)$. Now, let

$$\lambda_{min} = \frac{1}{1-\gamma} \min_{x\in X} \hat{c}(x)$$

and

$$\lambda_{max} = \frac{1}{1-\gamma} \max_{x\in X} \hat{c}(x).$$

We state that these λ_{min}, λ_{max} satisfy the conditions of the lemma. First, let $\lambda_1 \le \lambda_{min}$ be arbitrary and let $v(x) \equiv \lambda_1$. We show that for all $x \in X$ there holds $(Gv)(x) \ge v(x)$. Indeed,

$$(Gv)(x) = \gamma\lambda_1 + \hat{c}(x) \ge \gamma\lambda_1 + \min_{x\in X} \hat{c}(x) = \gamma\lambda_1 + (1-\gamma)\lambda_{min} \ge \lambda_1.$$

On the other hand, let $\lambda_2 \ge \lambda_{max}$ be arbitary and let $v(x) = \lambda_2, x \in X$. Then $(Gv)(x) \le v(x)$ holds for all $x \in X$ since

$$(Gv)(x) = \gamma\lambda_2 + \hat{c}(x) \le \gamma\lambda_2 + \max_{x\in X} \hat{c}(x) = \gamma\lambda_2 + (1-\gamma)\lambda_{max} \le \lambda_2.$$

□

Note that G is a contraction and thus its fixed point is well defined. Let us denote it by v_∞. We need the following lemma:

Lemma 4.2

$$v_\infty = v^*.$$

Proof. Let

$$v_s^*(x) = \inf_{\pi \text{ stationary}} v_\pi(x)$$

be the optimal value function corresponding to stationary policies. It is immediate that $v^* \le v_s^*$ and further that $v_\infty \le v^*$.[3] Thus in order to prove the lemma it is sufficient to prove that $v_s^* \le v_\infty$.

Let μ be an arbitrary myopic policy w.r.t. v_∞. Then $T_\mu \ell \le T_\mu v_\infty = v_\infty$ and thus $T_\mu^{n+1}\ell \le T_\mu^{n+1} v_\infty = T_\mu(T_\mu^n v_\infty) \le T_\mu v_\infty \le G v_\infty = v_\infty$ by induction. Let n tend to infinity yields by Proposition 3.2 that $v_\mu \le v_\infty$ and thus even more that $v_s^* \le v_\infty$. □

Now we can prove the main result of this section:

PROPOSITION 4.2 *If π is myopic w.r.t. v^* then π is optimal.*

Proof. Remember, that by definition

$$v^*(x) = \inf_\pi v_\pi(x), \qquad x \in X.$$

For brevity, let us define the operator $G : B(X) \to B(X)$ as

$$(Gv)(x) = \min_{a\in A} \max_{y \in T(x,a)} (c(x,a,y) + \gamma v(y)) \tag{3}$$

Let π by a stationary policy which is myopic w.r.t. v^*. Clearly, $T_\pi v^* = G v^*$ by the definition of π, T_π and G.

Since by the above lemma also $Gv^* = v^*$ thus $T_\pi v^* = v^*$, i.e., v^* is the fixed point of T_π. But T_π is a contraction and the evaluation of π, v_π is the fixed of T_π, too so we must have that $v_\pi = v^*$. □

5 The Bellman Optimality Equation

[3]This last inequality follows from the relations $G^n \ell \le G^n 0 \le v_\pi^{[n]}$, where π denotes an arbitrary policy, n is a natural number and $\ell \in B(X)$ is defined by $\ell(x) = \lambda_{min}$, $x \in X$ and $v_\pi^{[n]}$ denotes the n-horizon evaluation of π: $v_\pi^{[n]}(x) = \sum_{t=0}^n \gamma^t c_t^\pi(x)$.

COROLLARY 5.1 *There exists an optimal policy and the optimal value function satisfies the following, so called* Bellman (type) Optimality Equation *[1]:*

$$v^*(x) = \min_{a \in A} \max_{y \in T(x,a)} \left(c(x,a,y) + \gamma v^*(y) \right).$$

Morover, the optimal value function satisfies $v^* = \lim_{n\to\infty} G^n v$, where $v \in B(X)$ is arbitrary.

Proof. Since A is finite, for all $v \in B(X)$ there exist a myopic policy w.r.t v. Thus there exists a myopic policy w.r.t. v^*, which is optimal by Proposition 4.2.

Let π denote any myopic policy w.r.t. v^*. Then $T_\pi v^* = v^*$ (since v_π is the fixed point of T_π and $v_\pi = v^*$) and since π is myopic, we have that $T_\pi v^* = Gv^*$ and thus $Gv^* = v^*$. The last part of the assertion follows from the fact that G is a contraction. □

6 Asynchronous Dynamic Programming

6.1 A general asynchronous convergence theorem

In this section we review parts of the work of Bertsekas and Tsitsiklis concerning the convergence of asynchronous iterations [3]. Our aim here is to introduce the concepts. Although the proofs of the theorems presented here are not too complicated they are outside of the scope of this report. The interested reader may found the proofs (in a more general form) in [3].

DEFINITION 6.1 *Let $T : \mathbf{R}^X \to \mathbf{R}^X$ be an arbitary mapping, $v_0 \in \mathbf{R}^X$, and let $U_t \subseteq \{1, \ldots, n\}$ denote a sequence of non-empty sets ($t \geq 0$). Then the iteration*

$$v_{t+1}(x) = \begin{cases} v_t(x), & x \notin U_t; \\ (Tv_t)(x), & x \in U_t. \end{cases}$$

is called a weakly asynchronous *iteration of T. We say that $x \in X$ is "updated" at the tth stage of the iteration if $x \in U_t$.*

THEOREM 6.2 (BERTSEKAS AND TSITSIKLIS, 1989) *Let* $\{V(k)\}$ *be a sequence of non-empty sets satisfying*

$$\ldots \subseteq V(k+1) \subseteq V(k) \subseteq \ldots \subseteq \mathbf{R}^X,$$

and also the following two conditions:

1. Synchronous Convergence Condition:

$$T(V(k)) \subseteq V(k+1)$$

and if v_k *is a sequence such that for all* k*:* $v_k \in V(k)$*, then* v_k *converges and* $\lim_{k\to\infty} v_k = v^*$*, where* $Tv^* = v^*$*.*

2. Box Condition: *For all* k *there holds*

$$V(k) = V_1(k) \times \ldots \times V_n(k),$$

where $|X| = n$*, i.e., the set* $V(k)$ *is closed under the exchange of components of any two vectors from* $V(k)$*.*

Then for all $v_0 \in V(0)$ *the weakly asycnhronous iteration converges to some limit point of* T provided *that all* $x \in X$ *is updated infinitely many times, i.e.,* x *occurs infinitely often in the sets* $\{U_t\}_{t\in\mathbf{N}}$*.*

6.2 Convergence of Asynchronous Dynamic Programming

In this section we prove that the asynchronous iteration of operator G yields convergent sequences which converge to v^*.

THEOREM 6.3 *Let* $v_0 \in B(X)$ *be arbitrary and consider the following weakly asynchronous iteration:*

$$v_{t+1}(x) = \begin{cases} v_t(x), & x \notin U_t; \\ (Gv_t)(x), & x \in U_t; \end{cases},$$

where $U_t \subseteq X$ *are arbitrary non-empty sets. Then* $\lim_{t\to\infty} v_t = v^*$ *provided that every* $x \in X$ *is updated infinitely often.*

Proof. We apply Theorem 6.2. First we shall construct the sets $V(k)$, and then we have to show that the Synchronous and the Box Conditions are satisfied.

We choose

$$V(k) = \{v \in \mathbf{R}^X | G^k v_{min} \leq v \leq G^k v_{max}\},$$

where

$$v_{min}(x) = \min\Big(\lambda_{min}, \min(v_0(x), v^*(x)\Big), \qquad x \in X$$

and

$$v_{max}(x) = \max\Big(\lambda_{max}, \max(v_0(x), v^*(x)\Big), \qquad x \in X,$$

where λ_{min} and λ_{max} are defined in Lemma 4.1. Let us first show that $V(k+1) \subseteq V(k)$ holds for all k. Indeed, since $v_{min} \equiv \lambda \leq \lambda_{min}$, thus by the Lemma —reflem:tl $v_{min} \leq Gv_{min}$. Since G is monotonic the application of G on both sides preserves the relation and thus $Gv_{min} \leq G^2 v_{min}$. Continuing in this way it follows by induction that

$$G^k v_{min} \leq G^{k+1} v_{min}.$$

Similarly, one can prove that

$$G^k v_{max} \geq G^{k+1} v_{max}.$$

Then by the definition of $V(k+1)$ we get that $G^k v_{min} \leq v \leq G^k v_{max}$, i.e., $v \in V(k)$.

Now let us prove that

$$G(V(k)) \subseteq V(k+1).$$

For this let $v \in V(k)$ be arbitrary. Then

$$G^k v_{min} \leq v \leq G^k v_{max}.$$

By the application of G on both sides we get

$$G^{k+1} v_{min} \leq Gv \leq G^{k+1} v_{max},$$

i.e, $Gv \in V(k+1)$. Now, let v_k be a sequence of functions satisfying $v_k \in V(k)$, $k \in \mathbf{N}$. We show that v_k converges to the fixed point of G. Since $G^k v_{min} \leq v_k \leq G^k v_{max}$ holds for all k and $\lim_{k\to\infty} G^k v_{min} = v^* = \lim_{k\to\infty} G^k v_{max}$ (since G is a contraction) thus we have that $\lim_{k\to\infty} v_k = v^*$.

The Box Condition is satisfied since if we let

$$V_x(k) = \{r \in \mathbf{R} | (G^k v_{min})(x) \leq r \leq (G^k v_{max})(x)\}, \qquad x \in X$$

then $V(k) = \times_{x \in X} V_x(k)$. Thus all the conditions of Theorem 6.2 are satisfied and this proves our theorem. □

7 Related work

After proving Proposition 3.2 and Lemma 4.1 there are several options to continue the proofs. For example, one might use the results of any of the "abstract dynamic programming" approaches when it assumed that the evaluation of policies can be defined recursively by a generating operator or local income function [8,4,7,2,9,10, 11]. The approach we have taken here is most closely related to the approach of [10]. Heger [6] followed a different line in proving the above statements. We would like to note that the above results and their proofs holds if the set of admissible actions changes with the states.

References

[1] R. Bellman. *Dynamic Programming.* Princeton University Press, Princeton, New Jersey, 1957.

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

[3] D.P. Bertsekas and J.N. Tsitsiklis. *Parallel and Distributed Computation (Numerical Methods).* Prentice-Hall, Englewood Cliffs, NJ, USA, 1989.

[4] E.V. Denardo. Contraction mappings in the theory underlying dynamic programming. *SIAM Rev.*, 9:165–177, 1967.

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

[6] M. Heger. *Risk-sensitive decision making.* PhD thesis, Zentrum für Kognitionwissenschaften, Universität Bremen, FB3 Informatik, Postfach 330 440, 28334 Bremen, Germany, 1996.

[7] R.M. Karp and M. Held. Finite state processes and dynamic programming. *SIAM J. Appl. Math.*, 15:693–718, 1967.

[8] L.G. Mitten. Composition principles for synthesis of optimum multi-stage processes. *Operations Research*, 12:610–619, 1964.

[9] E. Porteus. Conditions for characterizing the structure of optimal strategies in infinite horizon dynamic programs. *J. Optim. Theory Appl.*, 36:419–432, 1982.

[10] Cs. Szepesvári. A general framework for reinforcement learning. In *Proc. of ICANN'95*, pages 165–170, 1995.

[11] Cs. Szepesvári and M. Littman. Generalized Markov Decision Processes: Dynamic programming and reinforcement learning algorithms. *Operations Research*, 1996. in preparation.