MARKOV DECISION CHAINS IN DISCRETE- AND CONTINUOUS-TIME; A UNIFIED APPROACH

KAREL SLADKÝ

Abstract: In this note we consider Markov decision chains with finite state space in discrete- and continuous-time setting for discounting and averaging optimality criteria. Connections between discounted and averaging optimality along with uniformization methods are employed for producing bounds on optimal discounted and average rewards.

Keywords: discrete-time and continuous-time Markov decision chains, discounted and averaging optimality, connections between discounted and averaging models, uniformization

1 Introduction

In this note, we consider Markov reward processes with finite state and action spaces in discrete- and continuous-time setting. Attention will be primarily focused on connections and similarity between discrete- and continuous-time Markov decision chains useful for finding optimal discounted and averaging control policies. According to the best of our knowledge, in the existing literature generating lower and upper bounds in averaging and discounted optimality was studied only for discrete-time models; in the present note we show how these methods also work in the continuous-time case. Also uniformization methods will be employed for producing bounds on optimal discounted and average rewards.

2 Notations and Preliminaries

We consider Markov decision processes with finite state space $\mathcal{I} = \{1, 2, \ldots, N\}$ both in discrete- and continuous-time. In the discrete-time case, we consider Markov decision chain $X^{\mathrm{d}} = \{X_n, n = 0, 1, \ldots\}$ with finite state space $\mathcal{I} = \{1, 2, \ldots, N\}$, and finite set $\mathcal{A}_i = \{1, 2, \ldots, K_i\}$ of possible decisions (actions) in state $i \in \mathcal{I}$. Supposing that in state $i \in \mathcal{I}$ action $a \in \mathcal{A}_i$ is selected, then state j is reached in the next transition with a given probability $p_{ij}(a)$ and one-stage transition reward r_{ij} will be accrued to such transition.

In the continuous-time setting, the development of the considered Markov decision process $X^{\mathrm{c}} = \{X(t), t \ge 0\}$ (with finite state space $\mathcal{I}$) over time is governed by the transition rates $q(j|i, a)$, for $i, j \in \mathcal{I}$, depending on the selected action $a \in \mathcal{A}_i$. For $j \neq i$ $q(j|i, a)$ is the transition rate from state i into state j, $q(i|i, a) = \sum_{j \in \mathcal{I}, j \neq i} q(j|i, a)$ is the transition rate out of state i. As concerns reward rates, $\tilde{r}(i)$ denotes the rate earned in state $i \in \mathcal{I}$, and $\tilde{r}(i, j)$ is the transition rate accrued to a transition from state i to state j.

A (Markovian) policy controlling the decision process is given either by a sequence of decision at every time point (discrete-time case) or as a piecewise constant right continuous function of time (continuous-time case). In particular, for discrete-time models policy controlling the chain, $\pi = (f^0, f^1, \ldots)$, is identified by a sequence of decision vectors $\{f^n, n = 0, 1, \ldots\}$ where $f^n \in \mathcal{A} \equiv \mathcal{A}_1 \times \ldots \times \mathcal{A}_N$ for every $n = 0, 1, 2, \ldots$, and $f^n_i \in \mathcal{A}_i$ is the decision (or action) taken at the nth transition if the chain X^{d} is in state i. Policy which selects at all times the same decision rule, i.e. $\pi \sim (f)$, is called stationary; $P(f)$ is transition probability matrix with elements $p_{ij}(f_i)$.

Similarly, for the continuous-time case policy controlling the chain, $\pi = f^t$, is a piecewise constant, right continuous vector function where $f^t \in \mathcal{A} \equiv \mathcal{A}_1 \times \ldots \times \mathcal{A}_N$, and $f_i^t \in \mathcal{A}_i$ is the decision (or action) taken at time t if the process $X(t)$ is in state i. Since π is piecewise constant, for each π we can identify time points $0 < t_1 < t_2 \ldots < t_i < \ldots$ at which the policy switches; we denote by $f^{(i)} \in \mathcal{A}$ the decision rule taken in the time interval $(t_{i-1}, t_i]$. Policy which selects at all times the same decision rule, i.e. $\pi \sim (f)$, is called stationary; $Q(f)$ is transition rate matrix with elements $q(j|i, f_i)$.

The more detailed analysis requires to consider the discrete- and continuous-time case separately. In this note we make the following assumption.

Assumption A. There exists state $i_0 \in \mathcal{I}$ that is accessible from any state $i \in \mathcal{I}$ for every $f \in \mathcal{A}$, i.e. for every $f \in \mathcal{A}$ the transition probability matrix $P(f)$ or the transition rate matrix $Q(f)$ is *unichain* (i.e. $P(f)$ or $Q(f)$ have no two disjoint closed sets).

2.1 Discrete-Time Case

We denote by $P(f) = [p_{ij}(f_i)]$ the $N \times N$ transition matrix of the chain X^{d}. Recall that the limiting matrix $P^*(f) = \lim\limits_{m\to\infty} m^{-1} \sum_{n=0}^{m-1} P^n(f)$ exists; in case that the chain is aperiodic even $P^*(f) = \lim\limits_{n\to\infty} (P(f))^n$. In particular, if $P(f)$ is unichain (i.e. $P(f)$ contains a single class of recurrent states) the rows of $P^*(f)$, denoted $p^*(f_i)$, are identical. Obviously, $r_i(f_i) = \sum_{j=1}^{N} p_{ij}(f_i) r_{ij}$ is the expected one-stage reward obtained in state $i \in \mathcal{I}$ and $r(f)$ denotes the corresponding N-dimensional column vector of one-stage rewards. Then $[P(f)]^n \cdot r(f)$ is the (column) vector of rewards accrued after n transitions; its ith entry denotes expectation of the reward obtained at time point n if the process X^{d} starts in state i.

Let $\xi^n_{X_0}(\pi) = \sum_{k=0}^{n-1} r_{X_k}(f^k_{X_k})$ (resp. $\xi^{\beta,n}_{X_0}(\pi) = \sum_{k=0}^{n-1} \beta^k\, r_{X_k}(f^k_{X_k})$) be the (random) total reward (resp. total β-discounted reward) received in the n next transitions of the considered Markov chain X^{d} if policy $\pi = (f^n)$ is followed and the chain starts in state X_0. Then for the total expected reward $v_i^n(\pi)$ and for the total expected discounted reward $v_i^{\beta,n}(\pi)$ we have $v_i^n(\pi) = \mathrm{E}_i^\pi \sum_{k=0}^{n-1} r_{X_k}(f^k_{X_k})$ and $v_i^{\beta,n}(\pi) = \mathrm{E}_i^\pi \sum_{k=0}^{n-1} \beta^k\, r_{X_k}(f^k_{X_k})$ respectively (E_i^π is the expectation if the process starts in state i and policy π is followed). Then for the vectors of total rewards $v^n(\pi)$ and total discounted rewards $v^{\beta,n}(\pi)$ we get

$$v^n(\pi) = \sum_{k=0}^{n-1} \prod_{j=0}^{k-1} P(f^j) r(f^k), \qquad \text{resp.} \qquad v^{\beta,n}(\pi) = \sum_{k=0}^{n-1} \prod_{j=0}^{k-1} \beta^k\; P(f^j) r(f^k). \tag{1}$$

For $n \to \infty$ elements of $v^n(\pi)$ (resp. $v^{\beta,n}(\pi)$) can be typically infinite (resp. bounded by $M/(1-\beta)$ where $M = \max_i \max_k r_i(k)$). Following stationary policy $\pi \sim (f)$ for n tending to infinity there exist vectors of average rewards per transition, denoted $g(f)$ (with elements $g_i(f)$ bounded by M), and vector of total discounted rewards, denoted $v^\beta(f)$ with elements $v_i^\beta(f)$ being the discounted reward if the process starts in state i, where (I denotes the identity matrix)

$$g(f) \;:=\; \lim_{n\to\infty} \frac{1}{n}\, v^n(f) = \; P^*(\pi) r(f) \tag{2}$$

$$v^\beta(f) \;:=\; \sum_{k=0}^{\infty} [\beta\; P(f)]^k r(f) = [\; I - \beta\; P(f)]^{-1}\, r(f) = r(f) + \beta\;\, v^\beta(\pi). \tag{3}$$

Let for arbitrary policy $\pi = (f^n)$ $\hat{v}^\beta := \sup_\pi v^\beta(\pi)$, $\hat{g} := \sup_\pi \liminf_{n\to\infty} \frac{1}{n}\; v^n(\pi)$ where $\hat{v}^\beta_i$, resp. $\hat{g}_i$ (the ith element of $\hat{v}^\beta$, resp. of $\hat{g}$) is the maximal β-discounted reward, resp. maximal average reward, if the process starts in state $i \in \mathcal{I}$. Moreover, under Assumption A for every stationary policy $\pi \sim (f)$ the vector $g(f)$ is a constant vector with elements $\bar{g}(f)$ equal to $p^*(\pi)\, r(f)$.

The following facts are well-known to workers in stochastic dynamic programming (see e.g. [1, 4, 8, 9, 13]).

Fact 1. (i) There exists decision vector $\hat{f}^\beta \in \mathcal{A}$ along with (column) vector $\hat{v}^\beta = v^\beta(\hat{f}^\beta)$, being the unique solution of

$$v^\beta(f) = \max_{f\in\mathcal{A}} \big[r(f) + \beta\; P(f)\;\; v^\beta(f)\big]. \tag{4}$$

In particular, for elements of $\hat{v}^\beta$, denoted $\hat{v}^\beta_i$, we can write

$$\hat{v}^\beta_i = \max_{a\in\mathcal{A}(i)} \big[r_i(a) + \beta \sum_{j\in\mathcal{I}} p_{ij}(a)\hat{v}^\beta_j\big] = r_i(\hat{f}^\beta_i) + \beta \sum_{j\in\mathcal{I}} p_{ij}(\hat{f}^\beta_i)\; \hat{v}^\beta_j. \tag{5}$$

(ii) If Assumption A holds there exists decision vector $\hat{f} \in \mathcal{A}$ along with (column) vectors $\hat{w} = w(\hat{f})$ and $\hat{g} = g(\hat{f})$ (constant vector with elements $\bar{g}(f) = p^*(\hat{f})r(\hat{f})$) being the solution of

$$w(f) +\; g(f) = \max_{f\in\mathcal{A}} \big[r(f) +\; P(f)\cdot\; w(f)\big] \tag{6}$$

where $w(\hat{f})$ is unique up to an additive constant, and unique under the additional normalizing condition $P^*(f)\;\, w(f) =\;\, 0$. In particular, for elements of $\hat{g} = g(\hat{f})$, and $\hat{w} = w(\hat{f})$, denoted $\bar{g}$ and $\hat{w}_i$, we can write

$$\hat{w}_i + \bar{g} = \max_{a\in\mathcal{A}(i)} \big[r_i(a) + \sum_{j\in\mathcal{I}} p_{ij}(a)\hat{w}_j\big] = r_i(\hat{f}_i) + \sum_{j\in\mathcal{I}} p_{ij}(\hat{f}_i)\hat{w}_j. \tag{7}$$

2.2 Continuous-Time Case

Let for $f \in \mathcal{F}$ $Q(f) = [q_{ij}(f_i)]$ be the $N\times N$ matrix whose ijth element $q_{ij}(f_i) = q(j|i, f_i)$ for $i \neq j$ and for the iith element we set $q_{ii}(f_i) = -q(i|i, f_i)$. The sojourn time of the considered process X^{c} in state $i \in \mathcal{I}$ is exponentially distributed with mean value $q(i|i, f_i)$. Hence the expected value of the reward rate obtained in state $i \in \mathcal{I}$ equals $r_i(f_i) = q(i|i, f_i)\, \tilde{r}(i) + \sum_{j\in\mathcal{I}, j\neq i} q(j|i, f_i)\, \tilde{r}(i, j)$ and $r(f)$ is the (column) vector of reward rates at time t.

For any policy $\pi = (f^t)$ the accompanying set of transition rate matrices $\{Q(f^t), t \geq 0\}$ determines a continuous-time (in general, nonstationary) Markov process.

Let $P(\cdot,\cdot,\pi)$ be the $N \times N$ matrix of transition functions associated with Markov chain X^{c}, i.e., for each $0 \leq s \leq t$ the ijth element of $P(s, t, \pi)$, denoted $P_{ij}(s, t, \pi)$, is the probability that the chain is in state j at time t given it was in state i at time s and policy π is followed. Obviously, $P(s, t, \pi) = P(s, u, \pi)\, P(u, t, \pi)$ for each $0 \leq s \leq u \leq t$. The values $P(s, t, \pi)$ are absolutely continuous in t and satisfy the system of differential equations (except possibly where the piecewise constant policy switches)

$$\frac{\partial P(s, t, \pi)}{\partial t} = P(s, t, \pi)\, Q(f^t), \qquad \frac{\partial P(s, t, \pi)}{\partial s} = -Q(f^s)P(s, t, \pi) \tag{8}$$

where $P(s, s, \pi) = I$ (I is an $N \times N$ unit matrix). In what follows it will be often convenient to let $P(t, \pi) = P(0, t, \pi)$. By (8) we then immediately get for any $t \geq 0$

$$\frac{\mathrm{d}P(t, \pi)}{\mathrm{d}t} = P(t, \pi)\, Q(f^t) \Longleftrightarrow P(t, \pi) = I + \int_0^t P(u, \pi)Q(f^u)\mathrm{d}u. \tag{9}$$

In particular, for $\pi \sim (f)$ we have

$$P(t,\pi) = \exp[Q(f)\,t] = \sum_{k=0}^{\infty} \frac{1}{k!}(Q(f)\,t)^k. \tag{10}$$

It is well known that for any stationary policy $\pi \sim (f)$ there exists $\lim_{t\to\infty} P(t,\pi) = P^*(\pi)$ and, moreover, that for any $t \ge 0$ it holds $P(t,\pi)\,P^*(\pi) = P^*(\pi)\,P(t,\pi) = P^*(\pi)\,P^*(\pi) = P^*(\pi)$, $Q(f)\,P^*(\pi) = P^*(\pi)\,Q(f) = 0$.

If policy $\pi = (f^t)$ is followed then for the vector of total rewards and ρ-discounted rewards $V^{t,\rho}(\pi)$ (with discount factor $\rho > 0$) obtained up to time T we get

$$V^T(\pi) = \int_0^T P(t,\pi)\,r(f^t)\mathrm{d}t, \qquad V^{T,\rho}(\pi) = \int_0^T \mathrm{e}^{-\rho t} P(t,\pi)\,r(f^t)\mathrm{d}t \tag{11}$$

(the ith element of $V^{T,\rho}(\pi)$ denoted $V_i^{T,\rho}(\pi)$ is the reward is the process starts in state i).

Following stationary policy $\pi \sim (f)$ for T tending to infinity there exist vectors of average rewards per transition, denoted $G(f)$ (with bounded entries $G_i(f)$) and vector of total discounted rewards, denoted $V^\rho(f)$, such that

$$G(f) \;:=\; \lim_{n\to\infty} \frac{1}{T}\int_0^T P(t,\pi)\,r(f)\mathrm{d}t = \; P^*(\pi)r(f) \tag{12}$$

$$V^\rho(f) \;:=\; \lim_{T\to\infty} \int_0^T \mathrm{e}^{-\rho t}\,P(t,\pi)\,r(f)\mathrm{d}t = \rho^{-1}\,[r(f) + Q(f)V^\rho(f)] \tag{13}$$

The following facts are well-known to workers in stochastic dynamic programming (see e.g. [1, 2, 4, 8, 9, 13]).

Fact 2. (i) There exists decision vector $\hat{f}^{(\rho)} \in \mathcal{A}$ along with (column) vector $\hat{V}^\rho = V^\rho(\hat{f}^{(\rho)})$, being the unique solution of

$$\rho\,V^\rho(f) = \max_{f\in\mathcal{A}} \big[r(f) + Q(f)\,V^\rho(f))\big]. \tag{14}$$

In particular, for elements of $\hat{V}^\rho$, denoted $\hat{V}_i^\rho$, we can write

$$\rho\hat{V}_i^\rho = \max_{a\in\mathcal{A}(i)} \big[r_i(a) + \sum_{j\in\mathcal{I}} q_{ij}(a)\hat{V}_j^\rho\big] = r_i(\hat{f}_i^{(\rho)}) + \sum_{j\in\mathcal{I}} q_{ij}(\hat{f}_i^{(\rho)})\hat{V}_j^\rho. \tag{15}$$

(ii) If Assumption A holds there exists decision vector $\hat{f} \in \mathcal{A}$ along with (column) vectors $\hat{W} = W(\hat{f})$ and $\hat{G} := G(\hat{f}) = P^*(\hat{f})\,r(\hat{f})$ (constant vector with elements $\bar{G}(f) = p^*(\hat{f})r(\hat{f})$) being the solution of

$$G(f) = \max_{f\in\mathcal{A}} \big[r(f) + Q(f)\,W(f)\big] \tag{16}$$

where $W(\hat{f})$ is unique up to an additive constant, and unique under the additional normalizing condition $P^*(f)\,W(f) = \; 0$.

3 Discounted and Averaging Optimality Equations

In this section we discuss connections between optimality equations for discounted and undiscounted models using a simple transformation of discounted model into the undiscounted unichain case. Furthermore, we indicate how continuous-time models can be transformed to discrete state models. The results are adapted from [12] and present a unified approach to various results scattered in the literature (see e.g. [3, 5, 6, 7, 10, 11, 14]).

Theorem 1. The discounted maximal (resp. current) total reward if the process starts in state ℓ equals the maximal (resp. current) average reward of the (not necessarily unichain) Markov reward process if

For the discrete-time case. The transition probability matrix $P(f)$ in (4) is replaced by the transition probability matrix $P^{(\ell)}(f) := \beta P(f) + A^{(\ell)}$ where $A^{(\ell)}$ is a square matrix such that the ℓth column is equal to $(1-\beta)$, and elements of the remaining columns equal zero, and the ℓth element of vector $w(f)$ equals zero. Then $\hat{v}^{\beta}_{\ell}$ equals elements of $(1-\beta)^{-1}\hat{g}$.

For the continuous-time case. The transition rate matrix $Q(f)$ in (14) is replaced by the transition rate matrix $Q^{(\ell)}(f) := \rho^{-1}Q(f) - I + B^{(\ell)}$, where $B^{(\ell)}$ is a square matrix such that only the ℓth column is non-null with elements equal to unity, and the ℓth element of vector $W(f)$ (unique up to additive constant) equals zero. Then $\hat{V}^{\rho}_{\ell}$ equals elements of $\hat{G}$.

Proof. Obviously, results for the current policy follow immediately from results for optimal policy if we shrink the set of feasible policies to a single policy.

For the discrete-time case
(I is an identity matrix, e denotes unit column vector)

$$\begin{aligned}
v^{\beta} &= \max_{f\in\mathcal{A}}\Big[r(f) + \beta P(f) v^{\beta}\Big]\\
&\Updownarrow\\
(1-\beta)\, v^{\beta}_{\ell}\, e &= \max_{f\in\mathcal{A}}\Big[r(f) + (\beta P(f) - I)(v^{\beta} - v^{\beta}_{\ell}\, e)\Big]\\
&\Updownarrow\\
w^{(\ell)} + g\, e &= \max_{f\in\mathcal{A}}\Big[r(f) + P^{(\ell)}(f)\, w^{(\ell)}\Big]\\
\text{where}\qquad & g := (1-\beta)\, v^{\beta}_{\ell},\ \ w^{(\ell)} := v^{\beta} - v^{\beta}_{\ell}\\
& P^{(\ell)}(f) := \beta P(f) + A^{(\ell)};
\end{aligned}$$

(observe that $P^{(\ell)}(f)$ is a stochastic matrix and that $w^{(\ell)}_{\ell} = 0$)

For the continuous-time case

$$\begin{aligned}
\rho V^{\rho} &= \max_{f\in\mathcal{A}}\Big[r(f) + Q(f) V^{\rho}\Big]\\
&\Updownarrow\\
0 &= \max_{f\in\mathcal{A}}\Big[\rho^{-1} r(f) + (\rho^{-1}Q(f) - I) V^{\rho}\Big]\\
&\Updownarrow\\
V^{\rho}_{\ell}\, e &= \max_{f\in\mathcal{A}}\Big[\rho^{-1} r(f) + \rho^{-1}Q(f) V^{\rho} - [V^{\rho} - V^{\rho}_{\ell}\, e]\Big]\\
&\Updownarrow\\
V^{\rho}_{\ell}\, e &= \max_{f\in\mathcal{A}}\Big[\rho^{-1}\, r(f) + \rho^{-1}Q(f)[V^{\rho} - V^{\rho}_{\ell}\, e] - [V^{\rho} - V^{\rho}_{\ell}\, e]\Big]
\end{aligned}$$

Then for $G := V^{\rho}_{\ell}\, e$ and $W^{(\ell)} := V^{\rho} - V^{\rho}_{\ell}\, e$ we can write

$$\begin{aligned}
G &= \max_{f\in\mathcal{A}}\Big[\rho^{-1}\, r(f) + \rho^{-1}Q(f) W^{(\ell)} - W^{(\ell)}\Big]\\
&\Updownarrow\\
G &= \max_{f\in\mathcal{A}}\Big[\rho^{-1}\, r(f) + \rho^{-1}Q^{(\ell)}(f) W^{(\ell)}\Big]\\
\text{for}\qquad & Q^{(\ell)}(f) := \rho^{-1}Q(f) - I + B^{(\ell)}
\end{aligned}$$

(observe that $Q^{(\ell)}(f) = \rho^{-1}Q(f) - I + B^{(\ell)}$ is a transition rate matrix and that $W^{(\ell)}_{\ell} = 0$).

Theorem 2. The continuous-time maximal (resp. current) average reward being the solution of (16) equals the discrete-time maximal (resp. current) average reward if in the optimality equation (6) we set

$$P(f) := B^{-1}Q(f) + I \quad \text{where} \quad B > \max_{f\in\mathcal{A}, i\in\mathcal{I}} \sum_{j\in\mathcal{I}, j\neq i} q_{ij}(f). \tag{17}$$

Then $g(f) = G(f)$ and $w(f) = BW(f)$.

Proof. First observe that elements of $P(f) = B^{-1}Q(f) + I$ are nonnegative, nongreater than unity and all row sums equal unity.

From (16) we get

$$\begin{array}{rcl} G(f) & = & \max\limits_{f\in\mathcal{A}} \left[r(f) + Q(f)\,W(f)\right] \\ & \Updownarrow & \\ W(f) + B^{-1}G(f) & = & \max\limits_{f\in\mathcal{A}} \left[B^{-1}r(f) + [B^{-1}Q(f) + I]\,W(f)\right] \\ & \Updownarrow & \\ BW(f) + G(f) & = & \max\limits_{f\in\mathcal{A}} \left[r(f) + [B^{-1}Q(f) + I]\,BW(f)\right] \\ & \Updownarrow & \\ w(f) + g(f) & = & \max\limits_{f\in\mathcal{A}} \left[r(f) + P(f)\,w(f)\right] \end{array}$$

Theorem 3. The vector of continuous-time maximal (resp. current) ρ-discounted reward being the solution of (14) equals the discrete-time maximal (resp. current) β-discounted reward if the optimality equation (4) takes on the following form

$$v(f) = \max_{f\in\mathcal{A}} \left[B^{-1}r(f) + [B^{-1}(Q(f) - \rho I) + I]v(f)\right] \tag{18}$$

where $B > \max\limits_{f\in\mathcal{A}, i\in\mathcal{I}} \sum\limits_{j\in\mathcal{I}, j\neq i} q_{ij}(f)$.

Proof. First observe that elements of the matrix $\tilde{P}(f) = [B^{-1}(Q(f) - \rho I) + I]B^{-1}$ are nonnegative, nongreater than unity and all row sums equal $(1 - \rho)$.

From (14) we get

$$\begin{array}{rcl} \rho V(f) & = & \max\limits_{f\in\mathcal{A}} \left[r(f) + Q(f)\,V(f)\right] \\ & \Updownarrow & \\ V(f) & = & \max\limits_{f\in\mathcal{A}} \left[B^{-1}r(f) + [B^{-1}(Q(f) - \rho I) + I]\,V(f)\right] \\ \Updownarrow & & \\ v(f) & = & \max\limits_{f\in\mathcal{A}} \left[B^{-1}r(f) + [B^{-1}(Q(f) - \rho I) + I]v(f)\right] \end{array}$$

4 Conclusions

In this note we focus attention on optimality equations for discrete- and continuous time Markov decision chains if discounted and averaging optimality criteria are considered. Using a suitable data transformation we shown connections between discounted and averaging optimality equations, and using the uniformization technique also connections between discrete- and continuous-time models.

Acknowledgement. This research was supported by the Czech Science Foundation under Grants 402/08/0107, 402/10/0956 and 402/10/1610.

References

[1] D. P. Bertsekas: Dynamic Programming and Stochastic Control. Academic Press, New York 1976.

[2] Xi Guo and O. Hernandez-Lerma: Continuous-Time Markov Decision Processes; Theory and Applications. Springer, Heidelberg 2009.

[3] N. A. J. Hastings: Bounds on the gain of a Markov decision processes. Oper. Res. *19* (1971), 240–243.

[4] R. A. Howard: Dynamic Programming and Markov Processes. MIT Press, Cambridge, Mass. 1960.

[5] J. MacQueen: A modified dynamic programming method for Markov decision problems. J. Math. Anal. Appl. *14* (1966), 38–43.

[6] A. R. Odoni: On finding the maximal gain for Markov decision processes. Oper. Res. *17* (1969), 857–860.

[7] M. L. Puterman and M. C. Shin: Modified policy iteration algorithms for discounted Markov decision problems. Manag. Sci. *24* (1978), 1127–1137.

[8] M. L. Puterman: Markov Decision Processes – Discrete Stochastic Dynamic Programming. Wiley, New York 1994.

[9] S. M. Ross: Applied Probability Models with Optimization Application. Holden Day, San Francisco 1970.

[10] R. F. Serfozo: An equivalence between continuous time and discrete time Markov decision processes. Oper. Res. *27* (1979), 3, 616–620.

[11] K. Sladký: O metodě postupných aproximací pro nalezení optimálního řízení markovského řetězce (On successive approximation method for finding optimal control of a Markov chain). Kybernetika *4* (1969), 2, 167–176.

[12] K. Sladký: Identification of optimal policies in Markov decision processes. Kybernetika *46* (2010), 3, 561–573.

[13] H. C. Tijms: A First Course in Stochastic Models. Wiley, Chichister 2003.

[14] D. J. White: Dynamic programming, Markov chains and the method of successive approximation. J. Math. Anal. Appl. *6* (1963), 296–306.

KAREL SLADKÝ

Department of Econometrics
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 4, 182 08 Praha 8, Czech Republic
`e-mail: sladky@utia.cas.cz`