To appear in *Mathematics of Operations Research*

CONSTRAINED DISCOUNTED MARKOV DECISION PROCESSES AND HAMILTONIAN CYCLES

Eugene A. Feinberg

303 Harriman Hall
SUNY at Stony Brook
Stony Brook, NY 11794-3775
email: efeinber@fac.har.sunysb.edu

Abstract

This paper establishes new links between stochastic and discrete optimization. We consider the following three problems for discrete time Markov Decision Processes with finite states and action sets: (i) find an optimal deterministic policy for a discounted problem with constraints, (ii) find an optimal stationary policy for a weighted discounted problem with constraints, (iii) find an optimal deterministic policy for a weighted discounted problem with constraints. We formulate mathematical programs for problems (i – iii) and show that the Hamiltonian Cycle Problem is a special case of each of these problems. Therefore problems (i – iii) are NP-hard. We also provide new mathematical programming formulations for the Hamiltonian Cycle and Traveling Salesman Problems.

AMS 1991 subject classification. Primary: 90C40. Secondary: 05C38.

OR/MS Index 1978 subject classification. Primary: 117 Dynamic Programming/ Markov. Secondary: 481 Networks/Graphs.

Key words. Markov decision process, constraint, weighted discounted problem, Hamiltonian cycle.

Running title: Markov Decision Processes and Hamiltonian Cycles

1. Introduction. Several discrete optimization problems could be considered as particular cases of stochastic optimization problems; see for example the shortest path problem in Bertsekas (1987). In other situations, a reduction of a discrete optimization problem to a well-structured stochastic optimization problem may be nontrivial. Such a reduction can potentially lead to three groups of results: (a) complexity estimates for stochastic optimization problems, (b) new approaches to discrete optimization based on the theory developed for stochastic problems, (c) new approaches to stochastic optimization stemming from methods and algorithms for particular discrete optimization problems. In the first paper in this direction, Papadimitrou and Tsitsiklis (1987) estimated the complexity to compute an optimal policy for a partially observable Markov Decision Process (MDP).

Filar and Krass (1994), Chen and Filar (1992), and Filar and Liu (1996) studied the Hamiltonian Cycle Problem (HCP) by reducing it to a constrained singularly perturbed MDP. The Hamiltonian Cycle (HC) in a graph with N vertices is a simple cycle of length N. The HCP is to find an HC or determine that none exist. The HCP is NP-hard.

The approach introduced by J. Filar and his students is directly related to this paper and we briefly describe it here. A starting point of this approach is that there is a natural relation between the HCP and constrained multichain MDPs with average rewards per unit time. For a directed graph, one can consider an MDP with deterministic transition probabilities. The vertices of the graph are the states of this MDP and the outgoing arcs are decisions which also define transitions. One of the vertices (states) is considered to be a starting point of the tour. The one-step rewards do not depend on selected actions and they are equal to 1 for the starting state and to 0 for all other states. Finding a tour is equivalent to finding a nonrandomized stationary policy for this MDP such that the average rewards per unit time are equal to N^{-1}. Finding such a policy is a particular case of finding an optimal policy for a multichain MDP with average rewards and constraints. Average reward multichain MDPs with constraints are intractable. By perturbing the transition probabilities, Filar and Krass (1994) reduced the HCP to a unichain constrained MDP with average rewards per unit time. This is a more tractable problem.

Many real-life problems deal with multiple criteria. A natural approach to problems with multiple criteria is to select one criterion as an objective function and treat other criteria as constraints. Kallenberg (1983) developed a theory of finite state and action MDPs with constraints; see also Heyman and Sobel (1984). For a unichain MDP with average reward criteria and for a discounted MDP, if the problem is feasible then an

optimal randomized stationary policy exists and it can be found via linear programming. This implies that finding optimal randomized policies for constrained discounted MDPs and for constrained unichained MDPs with average rewards per unit time are polynomially solvable problems.

Since it is easier to implement nonrandomized stationary policies than randomized ones, in applications it is natural to try to find the best nonrandomized stationary policy. The results of Filar and Krass (1994) show that computing the best nonrandomized stationary policy for a constrained unichain MDP with average rewards per unit time is an NP-hard problem.

Among various criteria for MDPs, the total discounted criterion is one of the most natural and the easiest one to deal with. This paper investigates the relation between the HCP and discounted MDPs. We also study the relation between the HCP and multiple criteria MDPs in which different criteria have different discount factors. Such problems are particular cases of multiple criteria problems with so-called weighted discounted criteria. Weighted discounted rewards are the sums of several total discounted rewards, each with a different discount factor. The theory of finite state and action MDPs with weighted discounted criteria differs significantly from the theory for standard discounted MDPs; see Feinberg and Shwartz (1994, 1995). For example, stationary policies usually are not optimal for weighted discounted problems. Weighted discounted MDPs model various applications such as project management, management of production, computer, and telecommunication systems, and infrastructure management; see Krass (1989), Krass, Filar and Sinha (1992), Feinberg and Shwartz (1994, 1995, 1999), and Reiman and Shwartz (1997).

In this paper, we show that the HCP can be considered as a particular case of one of the following problems:

(i) find a feasible deterministic policy for a discounted MDP with constraints;

(ii) find a feasible stationary policy for a weighted discounted MDP with constraints.

We provide Mathematical Programs (MPs) that find feasible and optimal solutions for problems (i, ii) and for the following problem which generalizes (i) and is related to (ii):

(iii) find a feasible deterministic policy for a weighted discounted MDP with constraints.

We also provide MPs for the HCP and Traveling Salesman Problem (TSP) that follow from the appropriate MPs for constrained discounted and weighted discounted MDPs. These MPs are based on Linear Programs (LPs) but they have additional constraints that are

either multiplicative, or bilinear, or include absolute values.

According to the results of this paper, problems (i – iii) can be viewed as generalizations of the HCP. Therefore, this problems are NP-hard if we characterize their size as the maximum of the following numbers: (a) the number of states; (b) the maximum of numbers of actions available in each state, (c) the number of constraints. In typical applications, (b) and (c) are significantly smaller than (a).

In Section 2 we introduce constrained discounted MDPs. In Sections 3 we construct MPs to compute the best deterministic policies for constrained discounted MDPs and we construct MPs to compute the best stationary and deterministic policies for weighted discounted MDPs. Section 4 describes the reduction of the HCP to the problems studied in Section 3. Section 5 applies the results of Sections 3 and 4 to the HCP and TSP.

2. Constrained MDPs. We consider a discrete time MDP with a finite state space $I = \{1, \dots, N\}$ and finite action sets $A(i)$, $i \in I$. If in a state i an action $a \in A(i)$ is selected, the process moves to a state $j \in I$ with the probability $p_{i,j}(a)$ where $p_{i,j}(a) \geq 0$ and $\sum_{j=1}^{N} p_{i,j}(a) = 1$, $i, j \in I$ and $a \in A(i)$. Let $A = \bigcup_{i \in I} A(i)$ and $H_n = I \times (A \times I)^n$, $n = 0, 1, \dots$.

A policy $\pi = \{\pi_n : n = 0, 1, \dots\}$ is a sequence of transition probabilities π_n from H_n to A such that $\pi_n(A(i_n)|i_0, a_0, \dots, i_{n-1}, a_{n-1}, i_n) = 1$. Given an initial state i, any policy π defines a stochastic sequence $i_0, a_0, i_1, \dots$. We denote by P_i^π and E_i^π probabilities and expectations for this sequence.

If $\pi_n(a_n|i_0, a_0, \dots, i_n) \in \{0, 1\}$ for all $n = 0, 1, \dots$ and for all $i_0, a_0, \dots, i_n \in H_n$ then the policy is called pure (or nonrandomized). If $\pi_n(a_n|i_0, a_0, \dots, i_n) = \pi_n(a_n|i_n)$ for all $n = 0, 1, \dots$ and for all $i_0, a_0, \dots, i_n \in H_n$ then the policy is called randomized Markov. A randomized Markov policy is called Markov if it is pure. A randomized Markov policy π is called stationary (or randomized stationary) if $\pi_n(a|i) = \pi(a|i)$ for all $n = 0, 1, \dots$, all $i \in I$, and all $a \in A$. A pure stationary policy is called deterministic (or nonrandomized stationary). Let Δ, $\mathbf{S}$ and $\mathbf{D}$ be the set of all, stationary, and deterministic policies respectively.

The triplet $\{I, \{A(i) : i \in I\}, p\}$ defines a probability structure of an MDP. Let r be a reward function, i.e. if an action a is chosen in a state i then an immediate reward is $r(i, a)$. For a discount factor $\beta \in [0; 1[$, we consider the expected total discounted rewards

$$W(i, \pi, \beta, r) = E_i^\pi \sum_{n=0}^{\infty} \beta^n r(i_n, a_n).$$

In many situations there are several criteria. So we have $K+1$ reward functions $r_0, r_1, \ldots r_K$, where K is a nonnegative integer. Let an initial state i be fixed. A standard approach to optimization with several criteria is to maximize one of the criteria subject to constraints on others: for a fixed initial state $i \in I$

$$\max W(i, \pi, \beta, r_0) \tag{2.1}$$

$$\text{subject to} \qquad W(i, \pi, \beta, r_k) \geq R_k, \qquad k = 1, \ldots, K. \tag{2.2}$$

If problem (2.1, 2.2) is feasible then there exists an optimal stationary policy; Kallenberg (1983) and Heyman and Sobel (1984).

In some applications there are several discount factors $\beta_1, \ldots, \beta_K$ and $\beta_k \in [0; 1[$, $k = 0, 1, \ldots, K$. Without loss of generality we assume that $\beta_k \neq \beta_l$ for $k \neq l$. For a nonnegative integer number M we define weighted discounted criteria

$$W_m(i, \pi) = \sum_{k=1}^{K} b_{k,m} W(i, \pi, \beta_k, r_k), \qquad m = 0, \ldots, M,$$

where $b_{k,m}$ are some coefficients.

For a fixed initial state $i \in I$ and for some constants R_m, $m = 0, \ldots, M$, we consider the following problem:

$$\max W_0(i, \pi) \tag{2.3}$$

$$\text{subject to} \qquad W_m(i, \pi) \geq R_m, \qquad m = 1, \ldots, M. \tag{2.4}$$

If problem (2.3, 2.4) is feasible then there exists an optimal policy with the following properties: (i) it is randomized Markov and (ii) it is deterministic from some epoch n onward; Feinberg and Shwartz (1995). Though this policy may be randomized at first n steps, it uses not more than M additional action than a Markov policy would use at all state-time couples; see Feinberg and Shwartz (1995) for details.

3. Constrained Discounted MDPs: Optimal Deterministic and Stationary Policies. This section deals with the computation of the best deterministic policies for constrained discounted MDPs and with the computation of the best stationary and deterministic policies for constrained weighted discounted MDPs. These problems can be formulated in the following way.

P1: Solve problem (2.1, 2.2) under the additional constraint $\pi \in \mathbf{D}$;

P2: Solve problem (2.3, 2.4) under the additional constraint $\pi \in \mathbf{S}$.

P3: Solve problem (2.3, 2.4) under the additional constraint $\pi \in \mathbf{D}$.

In order to solve (2.1, 2.2) we consider the following LP

$$\max \sum_{i=1}^{N} \sum_{a \in A(i)} r_0(i,a) x_{i,a} \tag{3.1}$$

$$\text{subject to} \qquad \sum_{a \in A(i)} x_{i,a} - \beta \sum_{j=1}^{N} \sum_{a \in A(j)} p_{j,i}(a) x_{j,a} = \mathbf{1}\{i=1\}, \qquad i = 1, \ldots, N, \tag{3.2}$$

$$\sum_{i=1}^{N} \sum_{a \in A(i)} r_k(i,a) x_{i,a} \geq R_k, \qquad k = 1, \ldots, K, \tag{3.3}$$

$$x_{i,a} \geq 0, \qquad i = 1, \ldots, N, \quad a \in A(i). \tag{3.4}$$

For a policy π we define state-action occupation vectors $x(\beta, \pi)$ with the elements

$$x_{i,a}(\beta, \pi) = \sum_{n=0}^{\infty} \beta^n P_1^{\pi}\{i_n = i, a_n = a\}, \qquad i \in I, \quad a \in A(i). \tag{3.5}$$

For a vector $x = \{x_{i,a} |\ i \in I, a \in A(i)\}$ we denote $x_i = \sum_{a \in A(i)} x_{i,a}$, $i \in I$.

The following statements are known; see section 3.4 in Kallenberg (1983). Problem (2.1, 2.2) has a solution if and only if LP (3.1 – 3.4) is feasible. If LP (3.1 – 3.4) is feasible then it has a solution. If x satisfies (3.2 – 3.4) then $x = x(\beta, \phi)$, where ϕ is a stationary policy defined by

$$\phi(a|i) = \begin{cases} x_{i,a}/x_i, & \text{if } x_i > 0; \\ \text{arbitrary}, & \text{if } x_i = 0. \end{cases} \tag{3.6}$$

If x is a solution of LP (3.1 – 3.4) then a stationary policy defined by (3.6) is optimal among all policies.

For $\Delta' \subseteq \Delta$ we denote by $X(\beta, \Delta') = \{x(\beta, \pi) | \pi \in \Delta'\}$ the set of occupation vectors for the set of policies Δ'. Then $X(\Delta) = X(\mathbf{S})$ and each of these sets of vectors is defined by (3.2 – 3.4); section 3.4 in Kallenberg (1983).

Therefore, the set $X(\mathbf{D})$ is the set of vectors satisfying conditions (3.2 – 3.4) and the additional condition that for each $i = 1, \ldots, N$ at most one element $x_{i,a}$, $a \in A(i)$, is positive. This condition can be written in the following form:

$$|x_{i,a} - x_{i,a^*}| = x_{i,a} + x_{i,a^*}, \qquad i = 1, \ldots, N, \ a, a^* \in A(i), \text{ and } a \neq a^*. \tag{3.7}$$

Let n_i be the number of actions in state $i \in I$. Therefore (3.7) provide us with additional

$$\sum_{i=1}^{N} n_i(n_i - 1)/2 \tag{3.8}$$

constrains to problem (3.1 – 3.4). These constraints make this problem neither linear nor convex. So, the following theorem holds.

Theorem 3.1. **P1** *is feasible if and only if MP (3.1 – 3.4, 3.7) is feasible. If MP (3.1 – 3.4, 3.7) is feasible then this MP and* **P1** *have optimal solutions. Let x be an optimal solution of MP (3.1 – 3.4, 3.7). Then for each $i \in I$ at most one number $\{x_{i,a} | a \in A(i)\}$ is positive. Consider a deterministic policy that selects in each state i an action a with $x_{i,a} > 0$ if $x_i > 0$ and which selects an arbitrary action if $x_i = 0$. This deterministic policy is optimal for* **P1**.

For **P2** we consider the following lemma.

Lemma 3.2. *Let $\alpha, \beta \in]0;1[$ and $x \in X(\alpha, \Delta)$, $y \in X(\beta, \Delta)$. Then*

$$x_{i,a} y_i = y_{i,a} x_i, \qquad i \in I,\ a \in A(i), \tag{3.9}$$

if and only if there exists a stationary policy ϕ such that

$$x = x(\alpha, \phi) \qquad \text{and} \qquad y = x(\beta, \phi). \tag{3.10}$$

Proof. Let (3.10) hold for some stationary policy ϕ. Since all summands in (3.5) are nonnegative, $x_{i,a} > 0$ if and only if $y_{i,a} > 0$. Therefore for each $i \in I$ either $x_i = y_i = 0$ or, in view of (3.6), $x_{i,a}/x_i = y_{i,a}/y_i$ for all $a \in A(i)$. This implies (3.9).

Let (3.9) hold. For an occupational vector z we denote $I^+(z) = \{i \in I |\ z_i > 0\}$. If $I^+(x) = I^+(y)$ then (3.6) implies that (3.10) holds for any ϕ defined by (3.6). So, we shall prove that $I^+(x) = I^+(y)$. Let $I^+ = I^+(x) \cap I^+(y)$. Since $x_1(\beta, \pi) \geq 1$ for any β and π, $1 \in I^+$. Therefore, $I^+ \neq \emptyset$. Consider any stationary policy ϕ^* such that $\phi^*(a|i) = x_{i,a}/x_i$ for $i \in I^+$. Then $\phi^*(a|i) = y_{i,a}/y_i$ for $i \in I^+$. Therefore any two stationary policies, defined by (3.6) for vectors x and y respectively, coincide with ϕ^* on I^+. Let $\tau = \min\{n > 0 | x_n \notin I^+\}$. We have that either $P_1^{\phi^*}\{\tau = \infty\} = 1$ or there exists $j \notin I^+$ such that $P_1^{\phi^*}\{x_\tau = j, \tau < \infty\} > 0$. In the first case, $x_i = y_i = 0$ for all $i \in I \backslash I^+$ and therefore $I^+(x) = I^+(y) = I^+$. In the second case, $x_j > 0$ and $y_j > 0$. These inequalities contradict $j \notin I^+$. ■

We observe that if $x = x(\alpha, \pi)$ and $y = x(\beta, \pi)$ for some $\alpha, \beta \in]0, 1[$ and $\pi \in \Delta$ then for all $i \in I$ and for all $a \in A(i)$ we have that $x_{i,a} > 0$ if and only if $y_{i,a} > 0$. Lemma 3.2 implies the following theorem.

Theorem 3.3. **P2** *is feasible if and only if MP (3.11 – 3.15).*

$$\max \sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{a \in A(i)} b_{k,0} r_{k,0}(i,a) x_{i,a}^{(k)} \tag{3.11}$$

$$\text{subject to } \sum_{a \in A(i)} x_{i,a}^{(k)} - \beta_k \sum_{j=1}^{N} \sum_{a \in A(j)} p_{j,i}(a) x_{j,a}^{(k)} = \mathbf{1}\{i = 1\}, \ k = 1, \ldots, K, i = 1, \ldots, N, \tag{3.12}$$

$$\sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{a \in A(i)} b_{k,m} r_{k,m}(i,a) x_{i,a}^{(k)} \geq R_m, \qquad m = 1, \ldots, M, \tag{3.13}$$

$$x_{i,a}^{(k)} \geq 0, \qquad k = 1, \ldots, K, \quad i = 1, \ldots, N, \quad a \in A(i), \tag{3.14}$$

$$x_{i,a}^{(k)} \sum_{a^* \in A(i)} x_{i,a^*}^{(k+1)} = x_{i,a}^{(k+1)} \sum_{a^* \in A(i)} x_{i,a^*}^{(k)}, \qquad k = 1, \ldots, K-1, i \in I, \ a \in A(i). \tag{3.15}$$

is feasible. If MP (3.11 – 3.15) is feasible then this MP and **P2** *have optimal solutions. Let* $\{x^{(1)}, \ldots, x^{(K)}\}$ *be an optimal solution of MP (3.11 – 3.15). Then a stationary policy* ϕ *defined by (3.6) with* $x = x^{(1)}$ *is optimal for* **P2**.

For **P3** we consider the following lemma.

Lemma 3.4. *Let* $\alpha, \beta \in]0; 1[$ *and* $x \in X(\alpha, \Delta)$, $y \in X(\beta, \Delta)$. *Then*

$$x_{i,a} y_{i,a^*} = 0, i \in I, \ a, a^* \in A(i), \text{ and } a \neq a^*, \tag{3.16}$$

if and only if there exists a deterministic policy ϕ *satisfying (3.10).*

Proof. Let (3.10) hold for a deterministic policy. Then for each i there exists at most one decision $a \in A(i)$ such that $x_{i,a} > 0$. Since $x_{i,a} > 0$ if and only if $y_{i,a} > 0$, (3.16) holds. Now let (3.16) hold. Since (3.16) implies (3.9), we have that (3.10) holds for any stationary policy defined by (3.6). Condition (3.16) implies that this policy can be chosen in a way that it is deterministic. ■

Similarly to Theorem 3.3, Lemma 3.4 implies the following statement.

Theorem 3.5. **P3** *is feasible if and only if MP (3.11 – 3.14, 3.17),*

$$x_{i,a}^{(k)} x_{i,a^*}^{(k+1)} = 0, \qquad k = 1, \ldots, K-1, i \in I,\ a, a^* \in A(i), \text{ and } a \neq a^*, \tag{3.17}$$

is feasible. If MP (3.11 – 3.14, 3.17) is feasible then this MP and **P3** *have optimal solutions. If vector* $\{x^{(1)}, \ldots, x^{(K)}\}$ *satisfies (3.12 – 3.14, 3.17) then: (i)* $x_{i,a}^{(k)} > 0$ *for some* $k = 1, \ldots, K$, *if and only if* $x_{i,a}^{(l)} > 0$ *for any other* $l = 1, \ldots, K$, *where* $i \in I$, $a \in A(i)$; *(ii) for each* $i \in I$ *there exists at most one* $a \in A(i)$ *such that* $x_{i,a}^{(k)} > 0$, $k = 1, \ldots, K$. *Let* $\{x^{(1)}, \ldots, x^{(K)}\}$ *be an optimal solution of MP (3.11 – 3.14, 3.17). Then a deterministic policy* ϕ, *such that* $\phi(i) = a$ *if* $x_{i,a}^{(1)} > 0$ *and* $\phi(i)$ *is arbitrary if* $x_{i,a}^{(1)} = 0$ *for all* $a \in A(i)$, *is optimal for* **P3**.

If $x^{(k)} = x(\beta_k, \phi)$ for some deterministic policy ϕ and for all $k = 1, \ldots, K$ then

$$|\sum_{k=1}^{K} x_{i,a}^{(k)} - \sum_{k=1}^{K} x_{i,a^*}^{(k)}| = \sum_{k=1}^{K} x_{i,a}^{(k)} + \sum_{k=1}^{K} x_{i,a^*}^{(k)}, \quad i \in I,\ a, a^* \in A(i), \text{ and } a \neq a^*. \tag{3.18}$$

For nonnegative variables $x_{i,a}^{(k)}$, (3.18) implies (3.17). Therefore, MPs (3.11 – 3.14, 3.17) and (3.11 – 3.14, 3.18) are equivalent. We also notice that (3.7) is equivalent to

$$x_{i,a} x_{i,a^*} = 0, \qquad i = 1, \ldots, N,\ a, a^* \in A(i), \text{ and } a \neq a^*. \tag{3.19}$$

4. Constrained Discounted Deterministic MDPs and HCs. We say that an MDP is deterministic if the following two conditions hold:

(a) $p_{i,j}(a) \in \{0, 1\}$ for all $i, j = 1, \ldots, N$ and for all $a \in A(i)$;

(b) if $a, b \in A(i)$ where $i \in I$ and $a \neq b$ then $p_{i,j}(a) \neq p_{i,j}(b)$ for some $j \in I$.

Condition (a) means that all transition probabilities are deterministic and condition (b) means that different actions define different transitions. If we do not consider rewards and discount factors, a Deterministic MDP (DMDP) is a directed graph. Without loss of generality, for a DMDP we denote actions by states to which the corresponding actions move the process. In other words, $A = I$ and if $j \in A(i)$ then $p_{i,j}(j) = 1$ where $i \in I$.

In this section we show the HCP can be viewed as a particular case of each of problems **P1**, **P2**, and **P3**. For a DMDP we always consider that the initial state is 1 and we set

$$r(i, a) = \begin{cases} 1, & \text{if } i = 1; \\ 0, & \text{otherwise.} \end{cases} \tag{4.1}$$

Until the end of these section, we consider DMDPs.

Lemma 4.1. *The following conditions are equivalent:*

(i) $W(1,\pi,\beta,r) = (1-\beta^N)^{-1}$ *for all* $\beta \in [0;1[$;

(ii)

$$P_1^\pi\{i_n = 1\} = \begin{cases} 1, & \text{if } n \in \{0, N, 2N, 3N, \ldots\}; \\ 0, & \text{otherwise.} \end{cases} \tag{4.2}$$

Proof. $W(1,\pi,\beta,r) = \sum_{n=0}^{\infty} \beta^n P_1^\pi\{i_n = 1\}$ and

$$P_1^\pi\{i_n = 1\} = (n!)^{-1} \frac{\partial^n}{\partial \beta^n} W(1,\pi,\beta,r)\Big|_{\beta=0}, \qquad n = 0, 1, \ldots .$$

■

We say that a policy π is Hamiltonian if $i_{nN}, i_{nN+1}, \ldots, i_{(n+1)N}$ is an HC (P_1^π-a.s.) for each $n = 0, 1, \ldots$. A deterministic Hamiltonian policy is an HC in the corresponding directed graph. Given a Hamiltonian policy, it is easy to construct an HC. Obviously, if π is Hamiltonian then (4.2) holds. However, (4.2) may hold for non-Hamiltonian policies. E.g. let $N = 4$, $A(1) = \{2\}$, $A(2) = \{1, 3\}$, $A(3) = \{2, 4\}$, and $A(4) = \{1\}$. Consider a policy π such that: (i) in state 3 it always selects action 2 to go to state 2 and (ii) in state 2 at epoch $n = 0, 1, \ldots$ it selects action 3 to go to state 3 if and only if $n = 4i + 1$, $i = 0, 1, \ldots$. This policy is not Hamiltonian but it satisfies (4.2). Lemmas 4.1, 4.2 and Theorem 4.3 imply that if (4.2) holds for a stationary policy then this policy is Hamiltonian.

Lemma 4.2. *If* ϕ *is a stationary policy and (4.2) holds for* $\pi = \phi$ *then* ϕ *is deterministic.*

Proof. Every stationary policy ϕ defines a Markov chain on the state space I. We define $m^\phi(i,j) = \min\{k \geq 0 | P_i^\phi(i_k = j) > 0\}$ the minimum number of steps to get from i to j with positive probability. We have that either $m^\phi(i,j) = \infty$ or $m^\phi(i,j) < N$. Let ϕ be a stationary policy which is not deterministic and satisfies (4.2). Since ϕ is not deterministic, there exist $k_1, k_2 \in I$ such that $0 < \phi(k_2|k_1) < 1$. This implies that either $0 < P_1^\phi(i_n = k_2) < 1$ for some $n = 1, \ldots, N-1$ or $P_1^\phi(i_n = k_1) = 0$ for all $n = 0, 1, \ldots, N-1$. Since in view of (4.2) $\sum_{n=1}^{N-1} \sum_{j=2}^{N} P_1^\phi\{i_n = j\} = N - 1$, we have that there exists a state i such that $P_1^\phi(i_n = i) > 0$ and $P_1^\phi(i_l = i) > 0$ for some $l \neq n$ and for $0 < l, n < N$. Let $m = m^\phi(i,1)$. Since $P_1^\phi\{i_l = i) > 0$ and $P_1^\phi(i_N = 1\} = 1$, we have that $m < N$. Then $P_1^\phi(i_{n+m} = 1) > 0$ and $P_1^\phi(i_{l+m} = 1) > 0$ where $0 < n+m, l+m < 2N$ and $n + m \neq l + m$. This contradicts (4.2). ■

Theorem 4.3. *The following statements are equivalent:*

(i) A policy ϕ is deterministic and Hamiltonian.

(ii) A policy ϕ is stationary and Hamiltonian.

(iii) A policy ϕ is deterministic and $W(1,\phi,\beta,r)=(1-\beta^N)^{-1}$ for at least one $\beta\in]0;1[$.

(iv) A policy ϕ is stationary and $W(1,\phi,\beta_k,r)=(1-\beta_k^N)^{-1}$ for $2N-1$ distinct discount factors $\beta_k\in]0;1[$, $k=1,\ldots,2N-1$.

Proof. Obviously, (i)⇒(ii, iii, iv). If ϕ be stationary and Hamiltonian then (4.2) holds. Therefore, Lemma 4.2 implies that (ii)⇒(i). Let (iii) hold. Since ϕ is a deterministic policy, $i_0,i_1,\ldots$ is a finite Markov chain with 0-1 transition probabilities. In particular, $i_0,i_1,\ldots$ is a deterministic sequence and if $i_n=i_m$ then $i_{n+1}=i_{m+1}$, $m,n=0,1,\ldots$. If 1 is a transient state then $W(1,\phi,\beta,r)=1$. If 1 is recurrent then $W(1,\phi,\beta,r)=(1-\beta^n)^{-1}$ where n is a number of states in the ergodic class that contains 1. In this case, $i_0,i_1,\ldots,i_n$ is a cycle and $i_0=i_n=1$. This cycle is Hamiltonian if and only if $n=N$. Thus (iii)⇒(i).

We show that (iv)⇒(i). If ϕ is a stationary policy and $W(1,\phi,\beta,r)=(1-\beta^N)^{-1}$ for all $\beta\in[0;1[$ then Lemmas 4.1 and 4.2 imply that ϕ is deterministic. Therefore, we shall prove that if $W(1,\phi,\beta,r)=(1-\beta^N)^{-1}$ for arbitrary different $2N-1$ points $\beta=\beta_1,\ldots,\beta_{2N-1}$ from $]0;1[$ then this formula takes place for all $\beta\in[0;1[$.

Let R be a vector-column of rewards in states $1,\ldots,N$ and $W(\phi,\beta,R)$ be a vector-column of the expected discounted rewards. We denote by $P(\phi)$ the matrix of transition probabilities of the Markov chain defined by a stationary policy ϕ on the state space I. It is well-known that $W(\phi,\beta,R)=(E-\beta P(\phi))^{-1}R$, where E is the $N\times N$ identity matrix; Blackwell (1962). Therefore

$$W(1,\phi,\beta,r)=\frac{a_{N-1}\beta^{N-1}+\ldots+a_1\beta+1}{b_N\beta^N+\ldots+b_1\beta+1}$$

for some $a_{N-1},\ldots,a_1$ and for some $b_N,\ldots,b_1$. So, $W(1,\phi,\beta,r)=(1-\beta^N)^{-1}$ is equivalent to

$$a_{N-1}\beta^{2N-1}+\ldots+a_1\beta^{N+1}+(b_N+1)\beta^N+(b_{N-1}-a_{N-1})\beta^{N-1}+\ldots+(b_1-a_1)\beta=0. \quad (4.3)$$

Since equation (4.3) has $2N-1$ different non-zero solutions $\beta_1,\ldots,\beta_{2N-1}$, all coefficients in (4.3) are 0. Thus $a_k=b_k=0$ for all $k=1,\ldots,N-1$ and $b_N=-1$. Lemma 4.2 implies that ϕ is deterministic. ■

Theorem 4.3 implies the HCP is equivalent to each of the following two problems for the corresponding DMDP:

P1D: For some $\beta \in]0;1[$ find a deterministic policy ϕ such that

$$W(1,\phi,\beta,r) = (1-\beta^N)^{-1}$$

or determine that none exist.

P2D: For some distinct $\{\beta_k \in]0;1[\}_{k=1}^{2N-1}$ find a stationary policy ϕ such that

$$W(1,\phi,\beta_k,r) = (1-\beta_k^N)^{-1}, \qquad k = 1,2,\ldots,2N-1,$$

or determine that none exist.

5. HC and TSP. Consider a directed graph $\{I, A\}$ where $I = \{1,\ldots,N\}$ is the set of vertices and A is the set of directed arcs. For each vertex i we denote $A(i) = \{j \in I |\ (i,j) \in A\}$ the set of following vertices and $B(i) = \{j \in I |\ (j,i) \in A\}$ the set of preceding vertices. If for some $i \in I$ we have that either $A(i) = \emptyset$ or $B(i) = \emptyset$ then there is no HC. Each graph defines a DMDP with the state space I, action sets $A(i)$, the reward function r defined by (4.1). Let $\beta \in]0;1[$ be a discount factor. Consider variables $x_{i,j}$, $(i,j) \in A$. Then constraints (3.2 – 3.4, 3.7) can be re-written for **P1D** in the following form:

$$\sum_{j\in A(i)} x_{i,j} - \beta \sum_{j\in B(i)} x_{j,i} = \mathbf{1}\{i=1\}, \qquad i = 1,\ldots,N, \tag{5.1}$$

$$\sum_{j\in A(1)} x_{1,j} = (1-\beta^N)^{-1}, \tag{5.2}$$

$$|x_{i,j} - x_{i,l}| = x_{i,j} + x_{i,l}, \qquad i = 1,\ldots,N,\ j,l \in A(i),\ \text{and } j \neq l, \tag{5.3}$$

$$x_{i,j} \geq 0, \qquad (i,j) \in A. \tag{5.4}$$

Theorems 3.1 and 4.3 imply that the existence of an HC is equivalent to the existence of a vector x that satisfies (5.1 – 5.4). Furthermore, any vector x, feasible for (5.1 – 5.4), has exactly N nonnegative coordinates and the set of arcs (i,j) with $x_{i,j} > 0$ forms an HC.

Because of (3.19), nonlinear constrains (5.3) can be replaced with the equivalent multiplicative constraints

$$x_{i,j}x_{i,l} = 0, \qquad i = 1,\ldots,N,\ j,l \in A(i),\ \text{and } j \neq l. \tag{5.5}$$

In order to apply Theorems 3.3 and 3.5 to HCs, consider $2N-1$ different numbers $\beta_1,\ldots,\beta_{2N-1}$ from the interval $]0,1[$ and consider vectors $x = \{x_{i,j}^{(k)} |\ (i,j) \in A, k =$

$1, \ldots, 2N-1\}$. For **P2D** constraints (3.12 –3.15) can be re-written in the following form

$$\sum_{j \in A(i)} x_{i,j}^{(k)} - \beta_k \sum_{j \in B(i)} x_{j,i}^{(k)} = \mathbf{1}\{i = 1\}, \ k = 1, \ldots, 2N-1, i = 1, \ldots, N, \tag{5.6}$$

$$\sum_{j \in A(1)} x_{1,j}^{(k)} = (1 - \beta_k^N)^{-1}, \qquad k = 1, \ldots, 2N-1, \tag{5.7}$$

$$x_{i,j}^{(k)} \geq 0, \qquad k = 1, \ldots, 2N-1, \quad (i,j) \in A, \tag{5.8}$$

$$x_{i,j}^{(k)} \sum_{l \in A(i)} x_{i,l}^{(k+1)} = x_{i,j}^{(k+1)} \sum_{l \in A(i)} x_{i,l}^{(k)}, \qquad k = 1, \ldots, 2N-2, i \in I, \ j \in A(i). \tag{5.9}$$

Theorem 3.3 implies that the existence of an HC is equivalent to the existence of a vector x which is feasible for (5.6 – 5.9). If a feasible vector x exists then there are exactly N positive numbers $x_{i_1,j_1}^{(1)}, \ldots, x_{i_N,j_N}^{(1)}$ and the set of arcs $\{(i_k, j_k) |\ k = 1, \ldots, N\}$ forms an HC. In view of Theorem 3.5, constraints (5.9) in MP (5.6 – 5.9) are equivalent to

$$x_{i,j}^{(k)} x_{i,l}^{(k+1)} = 0, \qquad k = 1, \ldots, 2N-2, i = 1, \ldots, N, j, l \in A(i), \text{ and } j \neq l. \tag{5.10}$$

Since all variables are nonnegative, (5.10) is equivalent to

$$|x_{i,j}^{(k)} - x_{i,l}^{(k+1)}| = x_{i,j}^{(k)} + x_{i,l}^{(k+1)}, \qquad k = 1, \ldots, 2N-2, i = 1, \ldots, N, j, l \in A(i), \text{ and } j \neq l. \tag{5.11}$$

Let each arc (i,j) carry cost $c(i,j)$. The TSP is to find an HC with the minimal total cost. When we consider multiple discount factors, we denote $\beta = \beta_1$ and $x_{i,j} = x_{i,j}^{(1)}$. We observe that if $(i_1, j_1), (i_2, j_2), \ldots, (i_N, j_N)$ is an HC and its arcs are written in a natural order, i.e $i_1 = i_N = 1$ and $j_n = i_{n+1}$, $n = 1, \ldots, N-1$ then for any vector x that satisfies either (5.1 – 5.4) or (5.6 – 5.9) we have that $x_{i_n,j_n} = \beta^{n-1}/(1-\beta^N)$, $n = 1, \ldots, N$. Therefore if we add the objective function

$$\min \sum_{(i,j) \in A} c(i,j) x_{i,j} \tag{5.12}$$

to either (5.1 – 5.4) or (5.6 – 5.9) and consider an HC defined by an optimal solution of one of these MP, we get an HC which minimizes the objective function

$$\sum_{n=1}^{N} \beta^{n-1} c(i_n, j_n). \tag{5.13}$$

We recall that for MP (5.12, 5.6 – 5.9) we denoted $x_{i,j} = x_{i,j}^{(1)}$ and variables $x_{i,j}^{(2)}, \ldots, x_{i,j}^{(2N-1)}$ do not participate in the objective function.

Since the number of all HCs is finite we have that if β is large enough, the HC obtained from a solution of one of these MPs is a solution of the TSP. Let all $c(i,j)$ be integer. Then it is possible to write an explicit lower bound for β such that the appropriate solution of either MP (5.12, 5.1-5.4) or MP (5.12, 5.6 – 5.9) defines a solution of the TSP.

Let $(i_1, j_1), \ldots, (i_N, j_N)$ be an HC defined by appropriate positive solutions of MP (5.12, 5.1-5.4) or MP (5.12, 5.6 – 5.9). We have that

$$\sum_{n=1}^{N} \beta^{n-1} c(i_n, j_n) \leq \sum_{n=1}^{N} \beta^{n-1} c(i_n^*, j_n^*), \tag{5.14}$$

where $(i_1^*, j_1^*), \ldots, (i_N^*, j_N^*)$ is another HC, is equivalent to

$$\sum_{n=1}^{N} c(i_n, j_n) \leq \sum_{n=1}^{N} c(i_n^*, j_n^*) + \epsilon \tag{5.15}$$

with $\epsilon = \sum_{n=2}^{N} (1 - \beta^{n-1})[c(i_n, j_n) - c(i_n^*, j_n^*)]$. If $\epsilon < 1$ for any HC $(i_1^*, j_1^*), \ldots, (i_N^*, j_N^*)$ then the HC $(i_1, j_1), \ldots, (i_N, j_N)$ is an optimal solution of the TSP. Let $C = \max\{c(i,j)|\ (i,j) \in A\} - \min\{c(i,j)|\ (i,j) \in A\}$. In a non-trivial case $C > 0$. Then $\epsilon \leq [(N-1) - \sum_{n=2}^{N} \beta^{n-1}]C$. And β can be selected in a way that

$$[(N-1) - \sum_{n=2}^{N} \beta^{n-1}]C < 1. \tag{5.16}$$

We observe that (5.16) holds for $\beta \geq \beta^* = 1 - [(N-1)C]^{-1}$. For $\beta \in [\beta^*, 1[$ the HCs, defined by solutions of either MP (5.12, 5.1-5.4) or MP (5.12, 5.6 – 5.9), are also solutions of the TSP. The case of rational $c(i,j)$ could be reduced to the case of integer $c(i,j)$ by multiplying all $c(i,j)$ by some number.

Acknowledgement. I would like to thank Jerzy A. Filar and John N. Tsitsiklis for interesting discussions related to this paper. This research was partially supported by NSF Grant DMI-9500746.

References

Bertsekas, D.P. (1987). *Dynamic Programming: Deterministic and Stochastic Models.* Prentice-Hall, NJ.

Blackwell, D. (1962). Discrete Dynamic Programming. *Ann. Math. Statist.* **33** 719–726.

Chen, M. and Filar, J.A. (1992). Hamiltonian Cycles, Quadratic Programming, and Ranking of Extreme Points. In *Global Optimization* (C. Floudas and P. Pardalos, Eds.), Princeton University Press, 32–39.

Feinberg, E.A. and Shwartz, A. (1994). Markov Decision Processes with Weighted Discounted Criteria. *Math. Oper. Res.* **19** 152–168.

Feinberg, E.A. and Shwartz, A. (1995). Constrained Markov Decision Processes with Weighted Discounted Criteria. *Math. Oper. Res.* **20** 302–320.

Feinberg, E.A. and Shwartz, A. (1999). Constrained Dynamic Programming with Two Discount Factors: Applications and an Algorithm. *IEEE Trans. Automatic Control.* **44** 628-631.

Filar, J.A. and Krass, D. (1994). Hamiltonian Cycles and Markov Chains. *Math. Oper. Res.* **19** 223–237.

Filar, J.A. and Liu, K. (1996). Hamiltonian Cycle Problem and Singularly Perturbed Markov Decision Process. In *Statistics, Probability and Game Theory. Papers in Honor of David Blackwell.* (T.S. Ferguson, L.S. Shapley and J.B. MacQueen, Eds.), Institute of Mathematical Statistics. Lecture Notes – Monograph Series **30**, 45–63.

Heyman, D.P. and Sobel, M. J. (1984). *Stochastic Models in Operations Research, Volume II: Stochastic Optimization.* McGraw-Hill, NJ.

Kallenberg, L.C.M. (1983). *Linear Programming and Finite Markovian Control Problems.* Math Centre Tracts 148, Mathematisch Centrum, Amsterdam.

Krass, D., Filar, J.A. and Sinha, S.S. (1992). A Weighted Markov Decision Process. *Oper. Res.* **40** 1180–1187.

Krass, D. (1989). *Contributions to the Theory and Applications of Markov Decision Processes.* Ph. D. Thesis, Johns Hopkins University, Baltimore, MD.

Papadimitriou, C.H. and Tsitsiklis, J.N. (1987). Complexity of Markov Decision Processes. *Math. Oper. Res.* **12** 441–450.

Reiman, M.I. and Shwartz, A. (1997) Call Admission: a New Approach to Quality of Service, CC Pub. 216, Technion, and Bell Labs Manuscript.