Existence of Optimal Policies for Semi-Markov Decision Processes Using Duality for Infinite Linear Programming

Diego Klabjan
Department of Mechanical and Industrial Engineering
University of Illinois at Urbana-Champaign
Urbana, IL
klabjan@uiuc.edu

Daniel Adelman
Graduate School of Business
University of Chicago
Chicago, IL
dan.adelman@gsb.uchicago.edu

July 23, 2004

Abstract

Semi-Markov decision processes on Borel spaces with deterministic kernels have many practical applications, particularly in inventory theory. Most of the results from general semi-Markov decision processes do not carry over to a deterministic kernel since such a kernel does not provide "smoothness." We develop infinite dimensional linear programming theory for a general stochastic semi-Markov decision process. We give conditions, general enough to allow deterministic kernels, for solvability and strong duality of the resulting linear programs. By using the developed linear programming theory we give conditions for the existence of a stationary deterministic policy for deterministic kernels, which is optimal among all possible policies.

1 Introduction

A semi-Markov decision process (SMDP) on Borel state and action spaces is said to have a Dirac's transition law if the state at the next decision epoch is uniquely determined by a given function evaluated at the current state-action pair. Such models are simple to state but turn out to be even more difficult to study and analyze than their true stochastic counterparts. They have many practical applications, for example in inventory routing (Adelman (2003)). In a companion paper, Adelman and Klabjan (2003) provide a new SMDP formulation for a widely studied, classical inventory control problem. This problem generalizes the classical economic order quantity problem to a multi-item setting. Most existing SMDP theory does not apply to Dirac's transition laws.

Nearly all approaches to the question of whether there exists an optimal policy require the transition law to be strongly continuous, but a Dirac's transition law is at best only weakly continuous. Strong continuity ensures that "smoothness" is maintained in the optimality equations. For example, existing approaches are either based on the vanishing discount rate methodology, Hernández-Lerma and Lasserre (1990), Vega-Amaya (1993), or policy iteration, Luque-Vásquez and Hernández-Lerma (1999), Hernández-Lerma and Lasserre (1997) (see also the series of monographs Hernández-Lerma (1989), Hernández-Lerma and Lasserre (1996b, 1999)). An alternative approach presented in Bhattacharya and Majumdar (1989) is to allow weak

continuity of the kernel but to impose equicontinuity of the discounted value functions. Unfortunately, Dirac's transition laws do not provide equicontinuity.

A recent approach in the literature that assumes weak continuity of the transition law is infinite linear programming, developed for the discrete-time case, i.e. Markov decision processes, by Hernández-Lerma and González-Hernández (1998), Hernández-Lerma and Lasserre (1996a), and Hernández-Lerma and Lasserre (1994). However, existing theory also requires that the expected transition times between decision epochs be lower bounded away from zero. The only exception that we are aware of is the work by Vega-Amaya (2003) in the context of zero-sum semi-Markov games, where the author assumes the transition time to be positive and not necessarily bounded away from zero. This condition is trivially satisfied in the case of discrete time periods. When satisfied in the semi-Markov case it is well known that there exists a transformation of the problem into discrete-time, employed for instance by Bhattacharya and Majumdar (1989), Vega-Amaya (1993), Luque-Vásquez and Hernández-Lerma (1999) although not using linear programming. Unfortunately, for the inventory control applications we have in mind, this condition is violated. It is possible to have multiple decision epochs at the same instant of time.

In this paper, we relax both of the above assumptions. We assume instead that the transition law is weakly continuous, and that the expected transition time *plus* current cost, rather than just the former, is lower bounded away from zero. Therefore, all of our results apply to the more restrictive settings in the references above. Instead of seeking transformations to a discrete-time Markov control setting, we work directly with a new infinite linear programming formulation of the SMDP, presented in Section 3.1, which extends the formulation of Fox (1966) in finite spaces to Borel spaces and the infinite linear programming formulation of discrete time MDP by Hernández-Lerma and Lasserre (1994). For a general stochastic SMDP, we establish a set of conditions under which this infinite linear program possesses strong duality, i.e. there is no duality gap and primal/dual optimal solutions are attained. Although the infinite linear programming approach to the Borel setting leads to the existence of an optimal policy that is stationary randomized, to date this approach has not been fruitful in showing the existence of an optimal policy that is stationary deterministic. We provide this result when the transition law is Dirac's under a strong recurrence condition. In a companion paper (Adelman and Klabjan (2003)), we show that all of the conditions in this paper are verifiable in an inventory application.

We owe a debt of gratitude to Hernández-Lerma and González-Hernández (1998), Hernández-Lerma and Lasserre (1996a), and Hernández-Lerma and Lasserre (1994) for their key insight that infinite linear programming can handle weakly continuous transition laws. This was indeed fortuitous, as what originally prompted our interest in it was one of the author's use of it in an approximate dynamic programming framework to generate near optimal control policies in inventory routing, see Adelman (2003). In future work, our duality results will prove useful in devising stronger, and possibly even convergent, approximate dynamic programming methodologies.

In Section 2 we formulate a general semi-Markov decision process. In Section 3 we formulate our primal-dual pair of infinite linear programs and give conditions for strong duality. In Section 4, we provide results specialized to the case of a Dirac's transition law.

2 Semi-Markov Control Model

The semi-Markov control model is defined by $(X, A, \{A(x) : x \in X\}, Q', c')$, where X is the state space and A is the control set. We assume that both X and A are Borel spaces. For each $x \in X$ we are given a non-empty Borel subset $A(x) \subseteq A$, which specifies the set of *admissible controls*, if the state of the system is x. We assume that $K = \{(x, a) : x \in X, a \in A(x)\}$ is a Borel subset of $X \times A$. Let Q' represent the *time-dependant transition law*. If the system is in state $x \in X$ and control action $a \in A(x)$ is taken, then the system's next state is in B after transition time $t \in T = [0, \infty)$ with probability $Q'(t, B|x, a)$, where $B \subseteq X$ is a Borel set. If the system is in state $x \in X$ and control action $a \in A(x)$ is selected leading to a state x' after a transition time t, then the system incurs a cost $c'(t, x', x, a)$. This cost includes the immediate cost of action a as well as any additional cost occurring during the transition to the next state.

For any Borel set $B \subseteq X$, and for any $(x, a) \in K$ the function $Q'(\cdot, B|x, a)$ is a distribution function, i.e.

- $Q'(t, B|x, a) = 0$ for every $t \leq 0$,
- $Q'(t, B|x, a)$ is a monotone lower semi-continuous function in t, and
- $\lim_{t\to\infty} Q'(t, X|x, a) = 1$.

We denote by x_n the state of the system at the nth decision time t_n and by $a_n \in A(x_n)$ the corresponding control action. The *transition time* $\delta_{n+1} = t_{n+1} - t_n$ has distribution $F(\cdot|x_n, a_n) = Q'(\cdot, X|x_n, a_n)$. For every Borel set $B \subseteq X$ and for every $(x, a) \in K$ let $Q(B|x, a) = \lim_{t\to\infty} Q'(t, B|x, a)$ denote the probability that the system is in a state from B in the next decision epoch when action a is chosen in state x. We call Q the *transition law*. Observe that Q is a stochastic kernel on X.

We denote by H_n the state of all *admissible histories* until the nth transition. Formally, $H_0 = X$ and $H_n = (K \times T)^n \times X$, where $h_n = (x_0, a_0, \delta_1, \ldots, x_{n-1}, a_{n-1}, \delta_n, x_n) \in H_n$ encodes the history of the process.

Definition 1. A *policy* π is a sequence $\pi = \{\pi_n\}_{n=0}^{\infty}$ of stochastic kernels π_n on A satisfying $\pi_n(A(x_n)|h_n) = 1$ for every admissible history $h_n \in H_n$ and for every $n \in \mathbb{N}$. A policy π is a *stationary randomized policy* if there exists a stochastic kernel ϕ such that $\pi_n(\cdot|h_n) = \phi(\cdot|x_n)$ for each $h_n \in H_n$ and for each $n \in \mathbb{N}$. A policy $\pi = \{\pi_n\}_{n=0}^{\infty}$ is a *stationary deterministic policy* if there exists a measurable function $f : X \to A$ such that $\pi_n(\cdot|h_n)$ is concentrated at $f(x_n) \in A(x_n)$ for each $n \in \mathbb{N}$. We denote by Π the set of all policies and by Π_{SD} the subset of all stationary deterministic policies.

Every initial distribution ν (which is a probability measure on X) and every policy π determine a unique probability measure P_ν^π and a stochastic process $\{(x_n, a_n, \delta_n), n = 0, 1, \ldots\}$ on $\Omega = (X \times A \times T)^\infty$ (Theorem of C. Ionescu Tulcea; see e.g. Ash (1972) [pp. 109] for a proof). We denote by E_ν^π the expectation operator with respect to P_ν^π and for $x \in X$ let E_x^π be equal to E_ν^π, where ν is the Dirac measure concentrated on x. The *mean holding time* in state x under a control $a \in A(x)$ is

$$\tau(x, a) = \int_T t\, F(\mathrm{d}t|x, a) = \int_T t\, Q'(\mathrm{d}t, X|x, a)\,.$$

Definition 2. Given an initial distribution ν and a policy π, the *long-run expected average cost* is defined as

$$J(\pi, \nu) = \limsup_{n\to\infty} \frac{E_\nu^\pi(\sum_{k=0}^{n-1} c'(t_k, x_{k+1}, x_k, a_k)}{E_\nu^\pi(t_n)}\,.$$

Let $J^* = \inf_\nu \inf_\pi J(\pi, \nu)$. A pair (ν^*, π^*) is a *minimum pair* if $J(\nu^*, \pi^*) = J^*$. The average cost problem is the problem of finding a minimum pair. For $x \in X$ let $J(x) = \inf_{\pi\in\Pi} J(\pi, x)$ be the *optimal average cost function*. A policy π^* is *average cost optimal* if $J(\pi^*, x) = J(x)$ for every $x \in X$. It is easy to see that

$$J(\pi, \nu) = \limsup_{n\to\infty} \frac{E_\nu^\pi(\sum_{k=0}^{n-1} c(x_k, a_k))}{E_\nu^\pi(\sum_{k=0}^{n-1} \tau(x_k, a_k))}\,,$$

where

$$c(x, a) = \int_X \int_T c'(t, x', x, a) Q'(\mathrm{d}t, \mathrm{d}x'|x, a)\,.$$

Note that a Markov control model is a semi-Markov control model with transition times equal to 1 with probability 1.

A special class of semi-Markov processes includes those with the transition law concentrated at a single state.

Definition 3. The transition law is a *Dirac's transition law* if there exists a measurable function $s : X \times A \to X$ such that

$$Q(B|x, a) = \begin{cases} 1 & s(x, a) \in B \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

for every Borel measurable set $B \subseteq X$.

Definition 4. A transition law Q is *weakly continuous* if $h : X \times A \to \mathbb{R}$ defined by

$$h(x,a) = \int_X u(y)Q(\mathrm{d}y|x,a)$$

is a continuous bounded function on K for every continuous bounded function u on X. Kernel Q is *strongly continuous* if h is a continuous bounded function on K for every measurable bounded function u on X.

If Q is a Dirac's transition law, then $h(x,a) = u(s(x,a))$. In this case Q is weakly continuous if and only if s is a continuous function (consider $u(x) = x$ for all $x \in X$). However, Q is typically not strongly continuous.

For this semi-Markov decision process, the *average cost optimality equation* is

$$u(x) = \inf_{a \in A(x)} \{c(x,a) - g\tau(x,a) + \int_X u(y)Q(\mathrm{d}y|x,a)\} .$$

In the case of a Dirac's transition law, this simplifies to

$$u(x) = \inf_{a \in A(x)} \{c(x,a) - g\tau(x,a) + u(s(x,a))\} . \tag{2}$$

3 Linear Programming and Semi-Markov Control Models with the Average Cost Criterion

In this section we develop a linear programming formulation for the semi-Markov control model. The formulation is based on the prior work on infinite-dimensional linear programming for Markov control models, Hernández-Lerma and Lasserre (1999)[pp. 203-249]. A thorough coverage of infinite-dimensional linear programs is given in Anderson and Nash (1987).

3.1 Linear Programs

Given a Borel space Z and a measurable weight function $f \geq 1$, let $\mathbb{B}_f(Z)$ be the Banach space of measurable functions u with finite *f-norm*

$$\|u\|_f = \sup_Z \frac{|u(s)|}{|f(s)|} .$$

In addition, let $\mathbb{M}_f(Z)$ be the Banach space of signed measures μ on the Borel space on Z with finite f *total variation norm*

$$\|\mu\|_f^{\mathrm{TV}} = \sup_{\|u\|_f \leq 1} |\int_Z u \, \mathrm{d}\mu| .$$

The *total variation norm* of μ is $\|\mu\|_{\mathrm{TV}} = \|\mu\|_1$. It it easy to see that

$$\|\mu\|_{\mathrm{TV}} \leq \|\mu\|_f^{\mathrm{TV}} . \tag{3}$$

See e.g. Hernández-Lerma and Lasserre (1999)[pp. 2-3] for a proof that $\mathbb{B}_f(Z)$ and $\mathbb{M}_f(Z)$ are Banach spaces. Let $\mathcal{B}(Z)$ be the Borel σ-algebra on Z and let $\mathbb{C}_b(Z)$ be the set of all continuous, bounded functions on Z.

Let $w : K \longrightarrow \mathbb{R}, w_0(x) : X \longrightarrow \mathbb{R}$ be defined as

$$w(x,a) = \tau(x,a) + c(x,a) \tag{4}$$
$$w_0(x) = \inf_{a \in A(x)} w(x,a) . \tag{5}$$

In order to define linear programs corresponding to the semi-Markov control process, we need the following assumptions.

Assumption A1. *$w(x,a)$ is lower semi-continuous and $\{a \in A(x) : w(x,a) \leq r\}$ is compact for every $x \in X$ and $r \in \mathbb{R}$.*

Assumption A2. *τ and c are nonnegative measurable functions.*

Assumption A3. *$w(x,a) \geq 1$ for every $(x,a) \in K$.*

Assumption A4. *There exists a finite constant $k \in \mathbb{R}$ such that*

$$\int_X w_0(y) Q(\mathrm{d}y|x,a) \leq k \cdot w(x,a)$$

for every $(x,a) \in K$.

Due to Assumption A1, w_0 is well defined, the infimum can be replaced by the minimum, and, in addition, w_0 is measurable, Rieder (1978). Assumption A3 can be relaxed to $\tau(x,a) + c(x,a) \geq \epsilon$ for every $(x,a) \in K$ and a given $\epsilon > 0$. By Assumption A3, $\mathbb{B}_{w_0}(X)$ is a well defined Banach space, and, by Assumption A1, $\mathbb{M}_{w_0}(X)$ is a well defined Banach space. Since every lower semi-continuous function is measurable, it follows from the same two assumptions that $\mathbb{M}_w(K)$ is a well defined Banach space. Observe also that Assumption A2 implies that $\tau \in \mathbb{B}_w(K)$ and $c \in \mathbb{B}_w(K)$.

Consider the following primal/dual linear programs on the dual pairs $(\mathbb{M}_w(K), \mathbb{B}_w(K))$, $(\mathbb{R} \times \mathbb{M}_{w_0}(X), \mathbb{R} \times \mathbb{B}_{w_0}(X))$. The primal problem is

$$\inf \int_K c(x,a)\mu(\mathrm{d}(x,a)) \tag{6a}$$

$$\int_K \tau(x,a)\mu(\mathrm{d}(x,a)) = 1 \tag{6b}$$

$$\mu((B \times A) \cap K) - \int_K Q(B|x,a)\mu(\mathrm{d}(x,a)) = 0 \qquad \text{for every } B \in \mathcal{B}(X) \tag{6c}$$

$$\mu \geq 0, \mu \in \mathbb{M}_w(K) \tag{6d}$$

and the dual problem reads

$$\sup \rho \tag{7a}$$

$$\tau(x,a)\rho + u(x) - \int_X u(y) Q(\mathrm{d}y|x,a) \leq c(x,a) \qquad \text{for every } (x,a) \in K \tag{7b}$$

$$\rho \in \mathbb{R}, u \in \mathbb{B}_{w_0}(X)\,. \tag{7c}$$

We denote by $\inf(P)$, $\sup(D)$ the optimal value of the primal,dual linear program, respectively.

To see that (6) and (7) is indeed a primal/dual pair consider the following operators. Let $L_0 : \mathbb{M}_w(K) \longrightarrow \mathbb{M}_{w_0}(X)$ be defined as

$$(L_0\mu)(B) = \mu(B \times A) - \int_K Q(B|x,a)\mu(\mathrm{d}(x,a)) \qquad \text{for every } B \in \mathcal{B}(X)$$

and let $L : \mathbb{M}_w(K) \longrightarrow \mathbb{R} \times \mathbb{M}_{w_0}(X)$ be

$$L\mu = \left(\int_K \tau(x,a)\mu(\mathrm{d}(x,a)), L_0\mu \right)\,.$$

The adjoint operator $L^* : \mathbb{R} \times \mathbb{B}_{w_0}(X) \longrightarrow \mathbb{B}_w(K)$ is given by

$$L^*(\rho,u)(x,a) = \tau(x,a)\rho + u(x) - \int_X u(y) Q(\mathrm{d}y|x,a)$$

for every $(\rho, u) \in \mathbb{R} \times \mathbb{B}_{w_0}(X)$ and $(x, a) \in K$. To see that $L^*(\rho, u) \in \mathbb{B}_w(K)$, let $(\rho, u) \in \mathbb{R} \times \mathbb{B}_{w_0}(X)$. Then

$$\left|\frac{\tau(x,a)\rho}{w_(x,a)}\right| \leq |\rho| \tag{8}$$

$$\left|\frac{u(x)}{w(x,a)}\right| = \left|\frac{u(x)}{w_0(x)}\right| \cdot \frac{w_0(x)}{w(x,a)} \leq \left|\frac{u(x)}{w_0(x)}\right| \leq \|u\|_{w_0} \tag{9}$$

$$\left|\frac{\int_X u(y)Q(\mathrm{d}y|x,a)}{w(x,a)}\right| \leq \|u\|_{w_0} \frac{\int_X w_0(y)Q(\mathrm{d}y|x,a)}{w(x,a)} \leq \|u\|_{w_0} k\,, \tag{10}$$

where (8) follows by Assumption A2 and (4), (9) by definition (5), and (10) by Assumption A4. It follows that the linear operator L is continuous with respect to the weak topology, see e.g. Anderson and Nash (1987)[pp. 35-40], and therefore (7) is a dual linear program to (6). It implies that under Assumptions A1-A4 we can apply results from Anderson and Nash (1987).

3.2 Results

A linear program is *consistent* if it has a feasible solution and it is *solvable* if there is a feasible solution that attains the optimal objective value. If (6), (7) is solvable, then we can replace inf,sup in (6a), (7a) by min, max and we write the corresponding value as $\min(P), \max(D)$, respectively. In this section we discuss the relation between the linear programs and the underlying semi-Markov control model and we give no duality gap and solvability results.

Definition 5. A function g on Z is a *strictly unbounded function* if there is a nondecreasing sequence of compact sets $Z_n \uparrow Z$ such that $\lim_{n\to\infty} \inf\{g(x)|x \notin Z_n\} = \infty$.

If Z is compact, then any function is strictly unbounded by considering $Z_n = Z$ for every n. If Z is open but bounded, then a strictly unbounded function must be discontinuous at the boundary of Z.

We need the following additional assumptions.

Assumption A5. *There is a policy π and an initial distribution ν such that $J(\pi, \nu) < \infty$.*

Assumption A6. *The transition law is weakly continuous.*

Assumption A7. *τ is a nonnegative, continuous, bounded function.*

Assumption A8. *w is strictly unbounded on K.*

Note that Assumptions A1 and A7 imply that c is lower semi-continuous and Assumption A7 yields $\tau \in \mathbb{C}_b(K)$.

Next we give some known results that will be used in subsequent sections. The following theorem is proven in Dynkin and Yushkevich (1979) [pp. 88-89].

Theorem 1. Let μ be a probability measure on $X \times A$ concentrated on K. Then there exists a stochastic kernel π on A such that

$$\mu(B \times C) = \int_B \pi(C|x)\hat{\mu}(\mathrm{d}x) \qquad \text{for every } B \in \mathcal{B}(X), C \in \mathcal{B}(A)\,,$$

where $\hat{\mu}(\cdot) = \mu(\cdot \times A)$ is the marginal of μ on X.

Definition 6. A measure μ on Z is *tight* if for each $\epsilon > 0$ there is a compact set $C \subseteq Z$ such that $\mu(Z \backslash C) < \epsilon$.

The proofs of the following two theorems are given in Billingsley (1968).

Theorem 2. Let Γ be a bounded family of nonnegative measures on Z. Then Γ is tight if and only if there is a strictly unbounded function $g \geq 1$ such that $\sup_{\mu\in\Gamma} \int_Z g\,\mathrm{d}\mu < \infty$. If Γ is a set of probability measures, then the condition $g \geq 1$ can be relaxed to $g \geq 0$.

Theorem 3 (Prohorov). Let Γ be a family of probability measures on a Borel space Z. If Γ is tight, then for each sequence $\{\mu_n\}$ in Γ there is a subsequence $\{\mu_m\}$ and a probability measure μ such that

$$\int_Z u \, \mathrm{d}\mu_m \longrightarrow \int_Z u \, \mathrm{d}\mu \tag{11}$$

for every $u \in \mathbb{C}_b(Z)$.

We say that measures $\{\mu_m\}_m$ converge *weakly* to a measure μ if (11) holds. We will be repeatedly using the following corollary.

Corollary 1. Let Γ be a family of nonnegative measures on a Borel space Z. Assume that there exists a constant $K < \infty$ such that $0 < \|\mu\|_{\mathrm{TV}} < K$. In addition, let there exist a strictly unbounded function $g \geq 1$ such that $\sup_{\mu\in\Gamma} \int_Z g \, \mathrm{d}\mu < \infty$. Then for each sequence $\{\mu_n\}$ in Γ there is a subsequence $\{\mu_m\}$ and a measure μ such that $\{\mu_m\}$ converges weakly to μ.

Proof. Let $\{\mu_n\}$ be a sequence in Γ.

If $\liminf_n \|\mu_n\|_{\mathrm{TV}} = 0$, there there exists a subsequence $\{\mu_m\}$ such that $\lim_m \|\mu_m\|_{\mathrm{TV}} = 0$. But then for any $u \in \mathbb{C}_b(Z)$ and any m we have $|\int_Z u \, \mathrm{d}\mu_m| \leq M\|\mu_m\|_{\mathrm{TV}}$, where $|u(s)| \leq M < \infty$ for any $s \in Z$. Hence $\{\mu_m\}$ converges weakly to the 0 measure.

Let now $\liminf_n \|\mu_n\|_{\mathrm{TV}} > 0$. Without loss of generality we assume that $\|\mu_n\|_{\mathrm{TV}} > m > 0$ for every n. Consider the set $\tilde{\Gamma}$ of probability measures defined as $\{\mu/\|\mu\|_{\mathrm{TV}} : \mu \in \Gamma\}$. We have

$$\sup_{\tilde{\mu}\in\tilde{\Gamma}} \int_Z g \, \mathrm{d}\tilde{\mu} \leq \frac{\sup_{\mu\in\Gamma} \int_Z g \, \mathrm{d}\mu}{m} < \infty$$

by assumption. Therefore by Theorem 2, $\tilde{\Gamma}$ is tight. By Prohorov's theorem we have that there is a weakly convergent subsequence $\{\tilde{\mu}_p\}$ that converges to a measure $\tilde{\mu}$. There is a subsequence $\{\mu_m\}$ of $\{\mu_p\}$ such that $\lim_m \|\mu_m\|_{\mathrm{TV}} = Q$. Clearly $0 < Q < K$. Now for every $u \in \mathbb{C}_b(Z)$ we have

$$\lim_m \int_Z u \, \mathrm{d}\mu_m = \lim_m \left(\int_Z u \, \mathrm{d}\tilde{\mu}_m \cdot \|\mu_m\|_{\mathrm{TV}} \right) = Q \lim_m \int_Z u \, \mathrm{d}\tilde{\mu}_m = Q \int_Z u \, \mathrm{d}\tilde{\mu} \, .$$

Therefore $\{\mu_m\}$ converges weakly to $\mu = Q\tilde{\mu}$. □

3.2.1 Consistency and Solvability

In this section we give results regarding consistency and solvability of (6) and (7). We first address consistency.

Theorem 4. Assume Assumptions A1-A8 hold. (6) and (7) are consistent, and $\inf(P) = J^*$.

The following lemma is proven in Hernández-Lerma and Lasserre (1999) [pp. 225].

Lemma 1. Let $\{\mu_n\}$ be a sequence of measures on S and μ a measure on S such that $\{\mu_n\}$ converges weakly to μ. If $c \geq 0$ is a lower semi-continuous function on S, then

$$\liminf_n \int_S c \, \mathrm{d}\mu_n \geq \int_S c \, \mathrm{d}\mu \, .$$

In addition we need the following lemma.

Lemma 2. If $\{\mu_n\}_n$ converges weakly to μ, then for every $v \in \mathbb{C}_b(X)$ we have

$$\lim_{n\to\infty} \int_X v \, \mathrm{d}L_0(\mu_n) = \int_X v \, \mathrm{d}L_0(\mu) \, .$$

Proof. We have

$$\int_X v \,\mathrm{d}L_0(\mu) = \int_K v \cdot \left(1 - \int_X Q(\mathrm{d}y|x,a)\right) \mu(\mathrm{d}(x,a)) = \lim_{n\to\infty} \int_K v \cdot \left(1 - \int_X Q(\mathrm{d}y|x,a)\right) \mu_n(\mathrm{d}(x,a))$$
$$= \lim_{n\to\infty} \int_X v \,\mathrm{d}L_0(\mu_n) = 0\,,$$

where the first equality follows from the definition of the adjoint operator (see Section 3.1), and the second equality follows from Assumption A6 and the definition of weak convergence. □

Proof of Theorem 4. (7) is consistent by taking $\rho = 0, u = 0$.

Next we address consistency of (6). Consider a policy π and an initial distribution ν such that $J(\pi,\nu) < \infty$. For every integer $n \geq 1$ let us define the probability measure on K as

$$\mu_n(\Omega) = \frac{1}{n} \sum_{i=0}^{n-1} P_\nu^\pi \left((x_i, a_i) \in \Omega\right)\,.$$

From Assumption A7 it follows that there exists a constant $M < \infty$ such that $\tau(x,a) \leq M$ for every $(x,a) \in K$. Then

$$\int_K w \,\mathrm{d}\mu_n = \frac{\sum_{k=0}^{n-1} E_\nu^\pi(w(x_k,a_k))}{n} = \frac{\sum_{k=0}^{n-1} E_\nu^\pi(w(x_k,a_k))}{\sum_{k=0}^{n-1} E_\nu^\pi(\tau(x_k,a_k))} \cdot \frac{\sum_{k=0}^{n-1} E_\nu^\pi(\tau(x_k,a_k))}{n} \leq (J(\pi,\nu)+1)\cdot M < \infty\,.$$

This implies that we can use Corollary 1 since by Assumption A8 w is strictly unbounded and $\|\mu_n\|_{\mathrm{TV}} = 1$. Let $\{\mu_m\}_m$ be a subsequence that convergence weakly to μ.

Since every μ_m is a probability measure, so is μ. For a subsequence l of m we have

$$J(\pi,\nu) = \limsup_n \frac{\int_K c \,\mathrm{d}\mu_n}{\int_K \tau \,\mathrm{d}\mu_n} \geq \limsup_m \frac{\int_K c \,\mathrm{d}\mu_m}{\int_K \tau \,\mathrm{d}\mu_m} = \lim_l \frac{\int_K c \,\mathrm{d}\mu_l}{\int_K \tau \,\mathrm{d}\mu_l}\,. \tag{12}$$

In addition, there exists a subsequence k of l such that

$$\liminf_l \int_K c \,\mathrm{d}\mu_l = \lim_k \int_K c \,\mathrm{d}\mu_k\,. \tag{13}$$

It follows from (12) that

$$J(\pi,\nu) \geq \lim_k \frac{\int_K c \,\mathrm{d}\mu_k}{\int_K \tau \,\mathrm{d}\mu_k} \tag{14}$$

By Lemma 1 and (13) we obtain $\lim_k \int_K c \,\mathrm{d}\mu_k \geq \int_K c \,\mathrm{d}\mu$. Since by Assumption A7 $\tau \in \mathbb{C}_b(K)$, we have that $\lim_k \int_K \tau \,\mathrm{d}\mu_k = \int_K \tau \,\mathrm{d}\mu$.

We first show that $\int_K \tau \,\mathrm{d}\mu > 0$. To the contrary, assume that $\int_K \tau \,\mathrm{d}\mu = 0$. Note that $\int_K c \,\mathrm{d}\mu_k = \int_K w \,\mathrm{d}\mu_k \geq \int_K \mathrm{d}\mu_k = 1$. Let us fix an $\epsilon > 0$. There exists an integer k_1 such that for every $k \geq k_1$ we have $\int_K \tau \,\mathrm{d}\mu_k \leq \epsilon$. Then for any $k \geq k_1$ it follows

$$\frac{\int_K c \,\mathrm{d}\mu_k}{\int_K \tau \,\mathrm{d}\mu_k} \geq \frac{1}{\epsilon}\,.$$

Since ϵ is an arbitrarily small number, it follows that $\lim_k \frac{\int_K c \,\mathrm{d}\mu_k}{\int_K \tau \,\mathrm{d}\mu_k} = \infty$, which contradicts (14) and the assumption that $J(\pi,\nu) < \infty$.

We conclude that $0 < \int_K \tau \,\mathrm{d}\mu < M$. This in turn implies that

$$J(\pi,\nu) \geq \lim_k \frac{\int_K c \,\mathrm{d}\mu_k}{\int_K \tau \,\mathrm{d}\mu_k} = \frac{\lim_k \int_K c \,\mathrm{d}\mu_k}{\lim_k \int_K \tau \,\mathrm{d}\mu_k} \geq \frac{\int_K c \,\mathrm{d}\mu}{\int_K \tau \,\mathrm{d}\mu}\,.$$

Next we show that μ satisfies (6c). Let $\mathcal{X}$ denote the characteristic or the indicator function of a set. Since for every $B \in \mathcal{B}(X)$ we have

$$P_\nu^\pi(x_i \in B) = E_\nu^\pi(\mathcal{X}_B(x_i)) = E_\nu^\pi(Q(B|x_{i-1}, a_{i-1}))$$

and for every k

$$\int_K Q(B|x,a)\mathrm{d}\mu_k = \frac{1}{k}\sum_{i=0}^{k-1} E_\nu^\pi(Q(B|x_i,a_i))\,,$$

an easy calculation shows that for every k

$$\mu_k(B\times A) = \int_K Q(B|x,a)\mathrm{d}\mu_k + \frac{P_\nu^\pi(x_{k-1}\in B) - \nu(B)}{k}\,.$$

Note that the last equality can be rewritten as $L_0(\mu_k) = (P_\nu^\pi(x_{k-1}\in B) - \nu(B))/k$. By considering $v = 1$ in Lemma 2 and the above equality, we obtain that $L_0(\mu) = 0$. Hence μ satisfies (6c).

Consider now the measure

$$\tilde{\mu} = \frac{\mu}{\int_K \tau\,\mathrm{d}\mu}\,.$$

Clearly $\tilde{\mu}$ satisfies (6b) and by the above argument it satisfies (6c) as well. We also have

$$\int_K w\,\mathrm{d}\tilde{\mu} = \frac{\int_K w\,\mathrm{d}\mu}{\int_K \tau\,\mathrm{d}\mu} = 1 + \frac{\int_K c\,\mathrm{d}\mu}{\int_K \tau\,\mathrm{d}\mu} \le 1 + J(\pi,\nu) < \infty$$

showing that $\tilde{\mu} \in \mathbb{M}_w(K)$. Therefore $\tilde{\mu}$ is a feasible solution to (6). Note also that $\int_K c\,\mathrm{d}\tilde{\mu} \le J(\pi,\nu)$. Since π is an arbitrary policy and ν an arbitrary initial probability distribution, it follows that $\inf(P) \le J^*$.

It remains to be seen that $J^* \le \inf(P)$. Since (6) is feasible, there exists a feasible solution μ. If $\int_K c\,\mathrm{d}\mu = \infty$, then there is nothing to prove and therefore we assume that $\int_k c\,\mathrm{d}\mu < \infty$. Then by Assumption A3 and feasibility of μ, $0 < \mu(K) \le \int_K w\,\mathrm{d}\mu = 1 + \int_K c\,\mathrm{d}\mu < \infty$ and therefore $\mu(K) < \infty$. By Theorem 1, there exists a policy π such that

$$\frac{\mu(B\times C)}{\mu(X\times A)} = \int_B \pi(C|x)\tilde{\mu}(\mathrm{d}x) \qquad \text{for every } B\in\mathcal{B}(X), C\in\mathcal{B}(A)\,. \tag{15}$$

For any randomized stationary policy π, $n \ge 2$, $x \in X$, $B \in \mathcal{B}(X)$, and a measurable function f on K we denote

$$\begin{aligned}
f(x,\pi) &= \int_A f(x,a)\pi(\mathrm{d}a|x)\\
Q(B|x,\pi) &= \int_A Q(B|x,a)\pi(\mathrm{d}a|x)\\
Q^n(B|x,\pi) &= P_x^\pi(x_n\in B) = \int_X Q^{n-1}(B|y,\pi)Q(\mathrm{d}y|x,\pi)\\
Q^1(B|x,\pi) &= Q(B|x,\pi)\,.
\end{aligned}$$

Then we have

$$\int_K f\,\mathrm{d}\mu/\mu(X\times A) = \int_X f(x,\pi)\tilde{\mu}(\mathrm{d}x) \tag{16}$$

$$E_{\tilde{\mu}}^\pi(f(x_n,a_n)) = \int_X\int_X f(y,\pi)Q^n(\mathrm{d}y|x,\pi)\tilde{\mu}(\mathrm{d}x) \tag{17}$$

$$\tilde{\mu}(B) = \int_X Q^n(B|x,\pi)\tilde{\mu}(\mathrm{d}x)\,, \tag{18}$$

where the first two equalities follow from (15) and aforementioned notation, and the last equality follows by iteratively applying (6c). It follows that

$$J^* \leq J(\pi, \tilde{\mu}) = \frac{\int_K c \,\mathrm{d}\mu / \mu(X \times A)}{\int_K \tau \,\mathrm{d}\mu / \mu(X \times A)} = \int_K c \,\mathrm{d}\mu \,.$$

Since μ is an arbitrary feasible measure to (6), we conclude that $J^* \leq \inf(P)$. □

Next we discuss solvability.

Theorem 5. If Assumptions A1-A8 hold, then (6) is solvable.

Proof. Since (6) is consistent by Theorem 4, for every nonnegative integer n there is a feasible measure μ_n to (6) such that

$$\inf(P) \leq \int_K c(x,a)\mu_n(\mathrm{d}(x,a)) \leq \inf(P) + \frac{1}{n} < \infty \,. \tag{19}$$

Since μ_n is feasible to (6) and from (19) it follows that

$$\|\mu_n\|_w^{\mathrm{TV}} \leq \int_K w \,\mathrm{d}\mu_n = \int_K (\tau + c)\mathrm{d}\mu_n = 1 + \int_K c \,\mathrm{d}\mu_n \leq 2 + \inf(P) \,.$$

If we in addition use (3), we get that $0 < \|\mu_n\|_{\mathrm{TV}} \leq 2 + \inf(P) < \infty$. By Assumption A8 and since $\sup \int_K w \,\mathrm{d}\mu_n$ is bounded, we can use Corollary 1. Let μ_m be a subsequence that converges weakly to a measure μ. We claim that μ is an optimal solution to (6).

From Lemma 1 and (19) it follows that $\int_K c \,\mathrm{d}\mu \leq \inf(P)$. If μ is feasible to (6), then this implies that μ is optimal.

Now we show that μ is feasible to (6). Since by Assumption A7 $\tau \in \mathbb{C}_b(K)$, it follows

$$1 = \int_K \tau \,\mathrm{d}\mu_m \longrightarrow \int_K \tau \,\mathrm{d}\mu$$

and therefore τ satisfies (6b). In turns it implies that

$$\|\mu\|_w^{\mathrm{TV}} \leq \int_K \tau \,\mathrm{d}\mu + \int_K c \,\mathrm{d}\mu \leq 1 + \inf(P)$$

and therefore $\mu \in \mathbb{M}_w(K)$. Since μ_m are feasible, it follows that $L_0(\mu_m) = 0$ for every m and in turn we can apply Lemma 2 with $v = 1$. Therefore μ satisfies (6c). □

Next we address solvability of (7). A sequence $\{(\rho_n, u_n)\}_n$ of feasible solutions to (7) is a *maximizing sequence* if $\lim_{n\to\infty} \rho_n = \sup(D)$.

Theorem 6. Assume that Assumptions A1-A4 hold. If there exists a maximizing sequence $\{(\rho_n, u_n)\}_n$ to (7) such that $\|u_n\|_{w_0} \leq r < \infty$ for a constant r, then (7) is solvable.

Proof. Let $\rho = \sup(D)$ and let us define

$$u(x) = \limsup_{n\to\infty} u_n(x) \,.$$

By assumption $\|u\|_{w_0} \leq r$ and therefore $u \in \mathbb{B}_{w_0}(X)$. For every $y \in X$ we have $|u_n(y)| \leq r w_0(y)$ and by Assumption A4 $\int_X w_0(y) Q(\mathrm{d}y|x,a) \leq k w(x,a) < \infty$, which justifies using Fatou's lemma with respect to $Q(\cdot|x,a)$. Since (ρ_n, u_n) satisfies (7b), we have that for every $(x,a) \in K$ and every n

$$\tau(x,a)\rho_n + u_n(x) \leq \int_X u_n(y) Q(\mathrm{d}y|x,a) + c(x,a) \,.$$

After taking lim sup, using $\lim_n \rho_n = \rho$, and applying Fatou's lemma, we obtain

$$\tau(x,a)\rho + u(x) - \int_X u(y) Q(\mathrm{d}y|x,a) \leq c(x,a) \,.$$

Therefore (ρ, u) is a feasible solution to (7) with value $\sup(D)$ and therefore it is an optimal solution. □

3.2.2 No Duality Gap

In this section we prove that under our assumptions there is no duality gap.

Theorem 7. If Assumptions A1-A8 hold, then $\sup(D) = \inf(P)$.

Proof. Let

$$H = \{(L\mu, \int_K c \,\mathrm{d}\mu + r) : \mu \in \mathbb{M}_w^+(K), r \geq 0\}\,,$$

where $\mathbb{M}_w^+(K)$ is the set of all nonnegative measures in $\mathbb{M}_w(K)$. By a theorem from Anderson and Nash (1987) [pp. 52], if H is closed in the weak topology of $(\mathbb{R} \times \mathbb{M}_{w_0}(X) \times \mathbb{R}, \mathbb{R} \times \mathbb{B}_{w_0}(X) \times \mathbb{R})$, then there is no duality gap.

To this end, let $(D, \geq)$ be a directed set and let $\{\mu_\alpha, r_\alpha\}_{\alpha \in D}$ be a net (see e.g. Ash (1972) for a definition of directed sets and nets) in $\mathbb{M}_w(K) \times \mathbb{R}_+$ such that

$$\int_K \tau \,\mathrm{d}\mu_\alpha \to r_* \tag{20}$$

$$\int_X u \,\mathrm{d}L_0(\mu_\alpha) \to \int_X u \,\mathrm{d}\upsilon_* \qquad \text{for every } u \in \mathbb{C}_b(X)$$

$$\int_K c \,\mathrm{d}\mu_\alpha + r_\alpha \to \rho_*\,. \tag{21}$$

By using Corollary 1 we show that there exists a nonnegative measure $\mu \in \mathbb{M}_w(X)$ and $r \in \mathbb{R}_+$ such that

$$r_* = \int_K \tau \,\mathrm{d}\mu \tag{22}$$

$$\upsilon_* = L_0(\mu) \tag{23}$$

$$\rho_* = \int_K c \,\mathrm{d}\mu + r_*\,. \tag{24}$$

Since $r_\alpha \geq 0, \int_K c \,\mathrm{d}\mu_\alpha \geq 0$ and by (21), it follows that $\int_K c \,\mathrm{d}\mu_\alpha$ are bounded for $\alpha \geq \alpha_0$ for an $\alpha_0 \in D$. Therefore by (20) it follows that there exists $\alpha_1 \in D, \alpha_1 > \alpha_0$ such that $\int_K w \,\mathrm{d}\mu_\alpha$ is bounded and positive for $\alpha \geq \alpha_1$. There exists a constant K such that $\|\mu_\alpha\|_w^{\mathrm{TV}} \leq K$ for every $\alpha \geq \alpha_1$. This in turn implies that $\|\mu_\alpha\|_{\mathrm{TV}} \leq \|\mu_\alpha\|_w^{\mathrm{TV}} \leq K$ for every $\alpha \geq \alpha_1$. We conclude that $\{\mu_\alpha\}_{\alpha \geq \alpha_1}$ is bounded. By Assumption A8 and by using Corollary 1 we obtain that there is a subsequence $\{\mu_m\}_m$ that converges weakly to a measure μ.

Since $\tau \in \mathbb{C}_b(K)$ by Assumption A7, it immediately follows that $r_* = \int_K \tau \,\mathrm{d}\mu$. Hence we have (22). By Lemma 1, we have

$$\int_K w \,\mathrm{d}\mu \leq 1 + \liminf_m \int_K c \,\mathrm{d}\mu_m < \infty$$

and therefore $\mu \in \mathbb{M}_w(K)$. Using again Lemma 1 and taking lim inf in (21) we get

$$\rho_* \geq \liminf_m \int_K c \,\mathrm{d}\mu_n + \liminf_m r_m \geq \int_K c \,\mathrm{d}\mu\,.$$

Thus we can define $r_* = \rho_* - \int_K c \,\mathrm{d}\mu \geq 0$ and we obtain (24). By using Lemma 2 we establish (23) and thus we have shown the theorem. □

3.2.3 Randomized Optimal Policies and Optimality Equation on a Subset of States

In this section we show, under generous assumptions, that there exists a minimum pair and that that the optimality equation has a solution on a subset of states.

Theorem 8. Assume that Assumptions A1-A8 hold and that (7) is solvable. Let μ, (ρ, u) be an optimal solution to (6),(7), respectively, and let $\hat{\mu}$ be the marginal of μ on X. Then

(a) $J^* = \rho$, and there exists a stationary randomized policy π^* and an initial distribution $\hat{\mu}^*$ such that $(\hat{\mu}^*, \pi^*)$ is a minimum pair, and

$$J(x, \pi^*) = \rho \tag{25}$$

holds for $\hat{\mu}^*$-almost all $x \in X$,

(b) [complementary slackness] and for μ-almost all $(x, a) \in K$ we have

$$\tau(x,a)J^* + u(x) = c(x,a) + \int_X u(y)Q(\mathrm{d}y|x,a)\,, \tag{26}$$

(c) if we denote

$$S = \{x \in X : \text{there exists } a \in A(x) \text{ such that (26) holds for } (x,a)\}\,, \tag{27}$$

and

$$S^* = S \cap \{x \in S : u(x) < \infty\}\,,$$

and we assume $S^* \neq \emptyset$, then there exists a stationary policy $f^* \in \Pi_{\mathrm{SD}}$ such that

$$\begin{aligned} u(x) &= \min_{a \in A(x)} \{c(x,a) - \tau(x,a)J^* + \int_X u(y)Q(\mathrm{d}y|x,a)\} \\ &= c(x, f^*(x)) - \tau(x, f^*(x))J^* + \int_X u(y)Q(\mathrm{d}y|x, f^*(x)) \end{aligned} \tag{28}$$

for every $x \in S^*$.

Proof. We first prove (a). Note that by Theorem 7 we have $\rho = J^*$. Since $0 < \mu(X \times A) \leq \int_{X\times A} w\,\mathrm{d}\mu = 1 + J^* < \infty$, we use Theorem 1 for $\mu/\mu(X \times A)$ to decompose this measure into a policy π^* and initial distribution $\hat{\mu}^*$. It follows from the proof of Theorem 4 that $(\hat{\mu}^*, \pi^*)$ is a minimum pair. The individual ergodic theorem, see e.g. Yosida (1978), yields (25).

Next we prove (b). Let q be a measurable function defined by

$$\tau(x,a)J^* + u(x) + q(x,a) = c(x,a) + \int_X u(y)Q(\mathrm{d}y|x,a)\,. \tag{29}$$

Since (ρ, u) is feasible to (7), $q \geq 0$ for every $(x,a) \in K$. After integrating (29) with respect to μ we obtain

$$J^* + \int_K u\,\mathrm{d}\mu + \int_K q\,\mathrm{d}\mu = \int_K c\,\mathrm{d}\mu + \int_K u\,\mathrm{d}\mu\,, \tag{30}$$

where we have used that μ satisfies (6b) and from (6c) it follows

$$\int_K (\int_X u(y)Q(\mathrm{d}y|x,a))\mu(\mathrm{d}(x,a)) = \int_K u\,\mathrm{d}\mu\,.$$

Since $\mu, (\rho, u)$ are optimal for the primal,dual linear programs, respectively, it follows $J^* = \int_K c\,\mathrm{d}\mu$. This together with

$$|\int_K u\,\mathrm{d}\mu| \leq k\|u\|_{w_0} \int_K w\,\mathrm{d}\mu = k\|u\|_{w_0}(1 + J^*) < \infty$$

and (30) yields $\int_K q\,\mathrm{d}\mu = 0$. Since q is nonnegative, we get that $q(x,a) = 0$ for μ-almost all (x,a), which completes the proof of the first statement.

It remains to show the last statement. For every $x \in S$ let $\bar{A}(x)$ be the set of all $a \in A(x)$ such that (x,a) satisfies (26). Note that by definition $\bar{A}(x) \neq \emptyset$. After integrating (26) with respect to $\pi^*(\mathrm{d}a|x)$ we obtain

$$u(x) = \int_{\bar{A}(x)} \left[c(x,a) + \int_X u(y)Q(\mathrm{d}y|x,a) - \tau(x,a)J^*\right] \pi^*(\mathrm{d}a|x)\,.$$

Since $u(x) < \infty$ for $x \in S^*$ it follows from the measurable selection theorem of Blackwell and Ryll-Nardzewski, see e.g. Dynkin and Yushkevich (1979) [pp. 255], that there exists a stationary deterministic policy f^* such that

$$\int_{\bar{A}(x)} \left[c(x,a) + \int_X u(y) Q(\mathrm{d}y|x,a) - \tau(x,a) J^* \right] \pi^*(\mathrm{d}a|x)$$
$$\geq c(x, f^*(x)) + \int_X u(y) Q(\mathrm{d}y|x, f^*(x)) - \tau(x, f^*(x)) J^* \,.$$

The other inequality follows from feasibility of u to (7). This establishes the second part. □

4 Dirac's Transition Laws

Next we study Dirac's kernels. Note that in this case Assumption A4 is equivalent to

$$w_0(s(x,a)) \leq k w(x,a)$$

for every $(x,a) \in K$ and Assumption A6 requires s to be continuous. Under a Dirac's transition kernel the corresponding primal linear program is

$$\inf \int_K c(x,a) \mu(\mathrm{d}(x,a)) \tag{31a}$$
$$\int_K \tau(x,a) \mu(\mathrm{d}(x,a)) = 1 \tag{31b}$$
$$\mu((B \times A) \cap K) - \mu(\{(x,a) \in K : s(x,a) \in B\}) = 0 \qquad \text{for every } B \in \mathcal{B}(X) \tag{31c}$$
$$\mu \geq 0, \mu \in \mathbb{M}_w(K) \tag{31d}$$

and the corresponding dual problem reads

$$\sup \rho \tag{32a}$$
$$\tau(x,a)\rho + u(x) - u(s(x,a)) \leq c(x,a) \qquad \text{for every } (x,a) \in K \tag{32b}$$
$$\rho \in \mathbb{R}, u \in \mathbb{B}_{w_0}(X) \,. \tag{32c}$$

By using a stronger version of Theorem 8 and a more stringent assumption we show the existence of a deterministic stationary optimal policies for all the states.

Assumption A9. *There exist constants $C < \infty$, $\Gamma < \infty$ such that for every measurable subset $S \subseteq X$ there is a measurable function $f : X \setminus S \to A$ with the property that for every $x' \in X \setminus S$ there exists a finite integer N and a set of states $x_0, x_1, \ldots, x_N$ with*

- $x_0 = x'$,
- $a_n = f(x_n) \in A(x_n)$ *for every* $n = 0, \ldots, N-1$,
- $x_{n+1} = s(x_n, a_n)$ *for every* $n = 0, \ldots, N-1$,
- $x_N \in S$,
- $\sum_{n=0}^{N-1} c(x_n, a_n) \leq C$, *and*
- $\sum_{n=0}^{N-1} \tau(x_n, a_n) \leq \Gamma$.

This assumption requires that any two states communicate (select S to be a single state) and the cost and the time of the path between any two states must be uniformly upper bounded.

For Dirac's kernels, we can strengthen Theorem 8 by showing that there exists an optimal policy whose sample path satisfies the average cost optimality equation.

Theorem 9. Assume that Assumptions A1-A8 hold and that (32) is solvable with (ρ, u) being an optimal solution. Furthermore, assume that there exists a constant N such that $N > u(x) > -N$ for every $x \in X$. Then there exists a stationary deterministic policy $f^* \in \Pi_{\text{SD}}$ and a non empty set $L \subseteq X$ such that the average cost optimality equation (2) holds for every $x \in L$ and

$$J(x) = J(f^*, x) = J^*$$

for every $x \in L$, i.e. f^* is an optimal stationary deterministic policy for all $x \in L$.

The following lemma holds for general kernels.

Lemma 3. If Assumption A3 holds and $J(\pi, \nu) < \infty$ for a policy π and initial distribution ν, then $\lim_{n\to\infty} \sum_{i=0}^{n-1} E_\nu^\pi(\tau(x_i, a_i)) = \infty$.

Proof. Suppose that $0 \leq \lim_{n\to\infty} \sum_{i=0}^{n-1} E_\nu^\pi(\tau(x_i, a_i)) < \infty$. Then there is a constant $K \geq 0$ such that

$$\sum_{i=0}^{n-1} E_\nu^\pi(\tau(x_i, a_i)) \leq K$$

for every n. By assumption we have

$$J(\pi, \nu) = \limsup_{n\to\infty} \frac{E_\nu^\pi(\sum_{k=0}^{n-1} c(x_k, a_k))}{E_\nu^\pi(\sum_{k=0}^{n-1} \tau(x_k, a_k))} = \limsup_{n\to\infty} \frac{E_\nu^\pi(\sum_{k=0}^{n-1} w(x_k, a_k))}{E_\nu^\pi(\sum_{k=0}^{n-1} \tau(x_k, a_k))} - 1 < \infty \,.$$

From Assumption A3 we obtain

$$\infty > 1 + J(\pi, \nu) \geq \limsup_n \frac{n}{K} = \infty \,,$$

which is a contradiction. □

Proof of Theorem 9. We use the same notation as in the proof of Theorem 8. We first show that there exists a trajectory, whose state-action pairs satisfy the optimality equation. For any $\omega \in \Omega$ let us define $r(\omega) = \sum_{i=1}^{\infty} q(x_i, a_i)$. Since u is dual feasible, we clearly have $r \geq 0$. In addition, let $r_n(\omega) = \sum_{i=1}^{n} q(x_i, a_i)$. We note that $r_1 \leq r_2 \leq r_3 \leq \ldots$ and for any $\omega \in \Omega$ we have $\lim_{n\to\infty} r_n(\omega) = r(\omega)$.

Next we show that for every n we have

$$\int_{\omega\in\Omega} r_n(\omega) P_{\hat{\mu}}^\pi(\mathrm{d}\omega) = 0 \,. \tag{33}$$

We show this by induction. We first note that from (16), (18), and complementary slackness for every n it follows

$$0 = \frac{\int_K q \,\mathrm{d}\mu}{\mu(K)} = \int_X q(x, \pi)\hat{\mu}(\mathrm{d}x) = \int_X \int_X q(y, \pi) Q^n(dy|x, \pi)\hat{\mu}(\mathrm{d}x) \,. \tag{34}$$

For $n = 1$ we have

$$\int_{\omega\in\Omega} r_1(\omega) P_{\hat{\mu}}^\pi(\mathrm{d}\omega) = E_{\hat{\mu}}^\pi(q(x_1, a_1)) = \int_X \int_X q(y, \pi) Q(dy|x, \pi)\hat{\mu}(\mathrm{d}x) = 0 \,,$$

where the second equality follows from (17) and the last one from (34). Assume now that (33) holds for $n - 1$. Then

$$\begin{aligned}
\int_{\omega\in\Omega} r_n(\omega) P_{\hat{\mu}}^\pi(\mathrm{d}\omega) &= \int_{\omega\in\Omega} (r_{n-1}(\omega) + q(x_n, a_n)) P_{\hat{\mu}}^\pi(\mathrm{d}\omega) \\
&= \int_{\omega\in\Omega} q(x_n, a_n) P_{\hat{\mu}}^\pi(\mathrm{d}\omega) \qquad (35) \\
&= E_{\hat{\mu}}^\pi(q(x_n, a_n)) = \int_X \int_X q(y, \pi) Q^n(dy|x, \pi)\hat{\mu}(\mathrm{d}x) = 0 \,, \qquad (36)
\end{aligned}$$

where (35) holds by the induction assumption and (36) follows from (17) and (34). Thus we have shown (33) for every n.

By the monotone convergence theorem it follows that

$$\int_{\omega\in\Omega} r(\omega)P^{\pi}_{\hat{\mu}}(\mathrm{d}\omega) = \lim_{n\to\infty}\int_{\omega\in\Omega} r_n(\omega)P^{\pi}_{\hat{\mu}}(\mathrm{d}\omega) = 0\,.$$

Hence there exists ω such that $r(\omega) = 0$, i.e. there is a trajectory that satisfies the optimality equation.

Let L be the set of all $x \in X$ with the property that there exists a trajectory ω with $x_0 = x$ and $r(\omega) = 0$. For every $x \in L$ let $\bar{A}(x) = \{a \in A(x) : q(x,a) = 0, s(x,a) \in L\}$. By definition of L, it follows that $\bar{A}(x) \neq \emptyset$. Now we use the measurable selection theorem of Blackwell and Ryll-Nardzewski as in the proof of Theorem 8. We obtain a stationary deterministic policy f^* satisfying

$$u(x) = c(x, f^*(x)) - \tau(x, f^*(x))J^* + u(s(x, f^*(x))) \tag{37}$$

and such that $q(x, f^*(x)) = 0$ for every $x \in L$. In other words, for every $x \in L$ we have $s(x, f^*(x)) \in L$ and (37) holds.

Let now $x \in L$. Then by iteratively applying (37) for every n it follows that

$$J^* = \frac{\sum_{i=0}^{n-1} c(x_i, f^*(x_i))}{\sum_{i=0}^{n-1} \tau(x_i, f^*(x_i))} + \frac{u(x_n) - u(x)}{\sum_{i=0}^{n-1} \tau(x_i, f^*(x_i))}\,. \tag{38}$$

If $\tau(x_i, f^*(x_i)) = 0$ for every i, then

$$0 = \sum_{i=0}^{n-1} c(x_i, f^*(x_i)) + u(x_n) - u(x)$$

and in turn by Assumption A3

$$0 = \sum_{i=0}^{n-1} w(x_i, f^*(x_i)) + u(x_n) - u(x) \geq n + u(x_n) - u(x)\,.$$

This can be rewritten as $u(x_n) \leq -n + u(x)$. As n tends to infinity, this yields a contradiction since by assumption u is lower bounded.

We conclude that there exists $\bar{i}$ such that $\tau(x_{\bar{i}}, f^*(x_{\bar{i}})) > 0$. For $n \geq \bar{i}$, $\{\sum_{i=0}^{n-1} \tau(x, f^*(x))\}_n$ is a nondecreasing sequence of positive values and it is therefore bounded away from 0. This in turn implies by taking lim sup in (38) and considering u is bounded that $J(f^*, x) < \infty$. As n goes to infinity, the second term goes to 0 since u is bounded in X and Lemma 3. Therefore $J^* = J(f^*, x) = J(x)$. □

Under the conditions stated in Theorem 9, clearly the conclusions of Theorem 8 hold. Before proving the main result, we need two additional statements.

Proposition 1. Let $\bar{x} \in X$ be a fixed state. If Assumption A9 holds and if u is feasible to (32), then there exists a constant M such that $-M \leq u(x) - u(\bar{x}) \leq M$ for every $x \in X$.

Proof. Consider $x \in X$ and let (u, ρ) be a feasible solution to (32). Then by Assumption A9 with $x' = x$ and $S = \{\bar{x}\}$ there is a sequence of state-action pairs $(x_i, a_i), a_i \in A(x_i)$ for $i = 0, 1, \ldots, N-1$ such that $x_0 = x, x_N = \bar{x}$. By iteratively using (32b) for $x_i, i = 0, 1, \ldots, N-1$ and then summing up the inequalities we obtain that

$$u(x) \leq \sum_{i=0}^{N-1} c(x_i, a_i) - \rho \sum_{i=0}^{N-1} \tau(x_i, a_i) + u(\bar{x}) \leq C + |\rho| \cdot \Gamma + u(\bar{x}) \leq C + J^* \cdot \Gamma + u(\bar{x})\,.$$

On the other hand, again by Assumption A9 there exists a sequence of state-action pairs $(x_i, a_i), a_i \in A(x_i)$ for $i = 0, 1, \ldots, N$ with $x_0 = \bar{x}$ and $x_N = x$. Similarly as above we obtain

$$u(\bar{x}) \leq \sum_{i=0}^{N-1} c(x_i, a_i) - \rho \sum_{i=0}^{N-1} \tau(x_i, a_i) + u(x) \leq C + |\rho| \cdot \Gamma + u(x) \leq C + J^* \cdot \Gamma + u(x)\,.$$

This completes the proof by taking $M = C + J^* \cdot \Gamma$. □

We are now ready to prove solvability of (32).

Corollary 2. Under Assumptions A1-A4 and Assumption A9, (32) is solvable.

Proof. Let $\{\rho_n, u_n\}_n$ be a maximizing sequence. Note that if (ρ, u) is feasible to (32), then for every $r \in \mathbb{R}$ the pair $(\rho, u - r)$ is feasible as well. Therefore $\{\rho_n, \hat{u}_n\}_n$ is a maximizing sequence as well, where $\hat{u}_n = u_n - u_n(\bar{x})$ and $\bar{x} \in X$ is a fixed state. By Proposition 1 $\hat{u}_n$ are bounded since $\hat{u}_n(\bar{x}) = 0$. By Theorem 6, we get that (32) is solvable. □

We summarize the linear programming results in the following proposition.

Theorem 10. Assume that Assumptions A1-A9 hold. The problems (31) and (32) are consistent, solvable, and there is no duality gap. There exists a nonempty set $L \subseteq X$, a deterministic stationary policy f^*, and a function $u \in \mathbb{B}_{w_0}(X)$ such that the average cost optimality equation

$$\begin{aligned} u(x) &= \min_{a \in A(x)} \{c(x, a) - J^* \tau(x, a) + u(s(x, a))\} \\ &= c(x, f^*(x)) - J^* \tau(x, f^*(x)) + u(s(x, f^*(x)))\lambda) \end{aligned} \tag{39}$$

holds for every $x \in L$ and $s(x, f^*(x)) \in L$ for every $x \in L$. In addition, for every $x \in L$, f^* is the optimal policy and

$$J^* = J(f^*, x) = J(x)$$

for every $x \in L$.

Proof. The first statement has already been proven. The last statement follows from Theorem 9 and Corollary 2. □

We are now ready to state the main result in the Dirac's case.

Theorem 11. Under Assumptions A1-A9, for every $x_0 = x \in X$ there exists an optimal deterministic stationary policy f^*. For every $x \in X$ we have $J(x) = J(f^*, x) = J^*$.

Proof. Let L and f^* be as in Theorem 10 and let f be as in Assumption A9 with respect to this particular L. Consider the deterministic stationary policy $\hat{f}$ defined for any $x \in X$ as

$$\hat{f}(x) = \begin{cases} f(x) & x \in X \setminus L \\ f^*(x) & x \in L\,. \end{cases}$$

We claim that the value of this policy is J^* for any $x_0 = x \in X$, which shows the statement.

Let $x_0 = x \in X$ be an initial state. By Assumption A9 policy $\hat{f}$ leads in at most N steps to a state in L and then the policy follows f^*. It is clear that $\sum_{k=0}^{\infty} \tau(x_k, \hat{f}(x_k)) > 0$ and therefore $J(\hat{f}, x) < \infty$. By Lemma 3 it follows that

$$\sum_{k=0}^{\infty} \tau(x_k, \hat{f}(x_k)) = \infty\,. \tag{40}$$

For any $n \geq N$ we have

$$\begin{aligned}
\frac{\sum_{k=0}^{n-1} c(x_k, \hat{f}(x_k))}{\sum_{k=0}^{n-1} \tau(x_k, \hat{f}(x_k))} &= J^* + \frac{\sum_{k=0}^{n-1} \Big(c(x_k, \hat{f}(x_k)) - J^* \tau(x_k, \hat{f}(x_k)) \Big)}{\sum_{k=0}^{n-1} \tau(x_k, \hat{f}(x_k))} \\
&= J^* + \frac{\sum_{k=0}^{N-1} \Big(c(x_k, \hat{f}(x_k)) - J^* \tau(x_k, \hat{f}(x_k)) \Big) + u(\hat{x}) - u(x_n)}{\sum_{k=0}^{n-1} \tau(x_k, \hat{f}(x_k))} \qquad (41) \\
&\leq J^* + \frac{2M}{\sum_{k=0}^{n-1} \tau(x_k, \hat{f}(x_k))} \,,
\end{aligned}$$

where M is as in Proposition 1. (41) follows since for $x_k, k \geq N$ we have $c(x_k, \hat{f}(x_k)) - J^* \tau(x_k, \hat{f}(x_k)) = c(x_k, f^*(x_k)) - J^* \tau(x_k, f^*(x_k)) = u(x_k) - u(x_{k+1})$ by using (39).

Taking the lim sup over n on both sides and considering (40) we obtain $J(\hat{f}, x) \leq J^*$, which completes the proof. □

Acknowledgements

The authors thank two anonymous referees for many helpful comments. In particular, we acknowledge an anonymous referee for pointing out a subtle error in an early version of the manuscript.

References

Adelman, D. (2003). Price-directed replenishment of subsets: Methodology and its application to inventory routing. *Manufacturing & Service Operations Management*, **5**, 348–371.

Adelman, D. and Klabjan, D. (2003). Duality and existence of optimal policies in generalized joint replenishment. *Mathematics of Operations Reserach*. To appear.

Anderson, E. and Nash, P. (1987). *Linear Programming in Infinite-dimensional Spaces*. John Wiley & Sons.

Ash, R. (1972). *Real Analysis and Probability*. Academic Press.

Bhattacharya, R. and Majumdar, M. (1989). Controlled semi-Markov models under long-run average rewards. *Journal of Statistical Planning and Inference*, **22**, 223–242.

Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley and Sons.

Dynkin, E. and Yushkevich, A. (1979). *Controlled Markov Processes*. Springer-Verlag.

Fox, B. (1966). Markov renewal programming by linear fractional programming. *SIAM Journal on Applied Mathematics*, **14**, 1418–1432.

Hernández-Lerma, O. (1989). *Adaptive Markov Control Processes*. Springer-Verlag.

Hernández-Lerma, O. and González-Hernández, J. (1998). Infinite linear programming and multichain Markov control processes in uncountable spaces. *SIAM Journal on Control and Optimization*, **36**, 313–335.

Hernández-Lerma, O. and Lasserre, J. (1990). Average cost optimal policies for Markov control processes with Borel state space and unbounded costs. *Systems and Control Letters*, **15**, 349–356.

Hernández-Lerma, O. and Lasserre, J. (1994). Linear programming and average optimality of Markov control processes on Borel spaces-unbounded costs. *SIAM Journal on Control and Optimization*, **32**, 480–500.

Hernández-Lerma, O. and Lasserre, J. (1996a). Average optimality in Markov control processes via discounted cost problems and linear programming. *SIAM Journal on Control and Optimization*, **34**, 295–310.

Hernández-Lerma, O. and Lasserre, J. (1996b). *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer-Verlag.

Hernández-Lerma, O. and Lasserre, J. (1997). Policy iteration for average cost Markov control processes on Borel spaces. *Acta Applicandae Mathematicae*, **47**, 125–154.

Hernández-Lerma, O. and Lasserre, J. (1999). *Further Topics on Discrete-Time Markov Control Processes*. Springer-Verlag.

Luque-Vásquez, F. and Hernández-Lerma, O. (1999). Semi-Markov control models with average costs. *Applicationes Mathematicae*, **26**, 315–331.

Rieder, U. (1978). Measurable selection theorems for optimization problems. *Manuscripta Mathematica*, **24**, 115–131.

Vega-Amaya, O. (1993). Average optimality in semi-Markov control models on Borel spaces: unbounded cost and controls. *Boletín de la Sociedad Matemática Mexicana*, **38**, 47–60.

Vega-Amaya, O. (2003). Zero-sum average semi-Markov games: Fixed-point solutions of the Shapley equation. *SIAM Journal on Control and Optimization*, **42**, 1876–1894.

Yosida, K. (1978). *Functional Analysis*. Springer-Verlag.