Modern Trends in Controlled Stochastic Processes Vol.II (A.Piunovskiy ed.), Luniver Press, Frome, 2015, p.190-212.

Sufficient Classes of Strategies in Continuous-Time Markov Decision Processes with Total Expected Cost

A.B.Piunovskiy
University of Liverpool
Dept. of Mathematical Sciences, Liverpool L69 7ZL, UK
piunov@liv.ac.uk

In this paper we introduce a new wider class of control strategies for continuous-time Markov decision processes (MDP) with the total expected cost in such a way that the exponential semi-Markov decision process (ESMDP) is a special case. After that, we describe the sufficient classes of strategies and justify the transformation to the discrete-time MDP.
Keywords: Markov decision process, continuous-time Markov chain, occupation measure, total expected cost, constrained optimisation, convex analytic approach, linear program.
AMS 2000 subject classification: Primary 90C40, Secondary 90C39

1 Introduction

When studying controlled continuous-time Markov chains, two approaches are known. If the (randomised) action can be changed only at the jump epochs, the model is called ESMDP [5, 8, 15]. Starting from [15], another construction based on the article by Jacod [14] became more popular [6, 9, 10, 16, 19]. The latter model is currently called continuous-time MDP (CTMDP). What is important, CTMDP does not cover ESMDP in the sense that randomised, but constant during the sojourn times, actions cannot be described in the framework of traditional CTMDP. The deep connection between CTMDP and ESMDP for the total discounted cost was studied in [5], but again the author had to introduce those models separately. It looks more appropriate to build one model, but with wider class of strategies in such way that some strategies (called below 'randomised') correspond to ESMDP, and the others (called 'relaxed') correspond to the standard strategies in CTMDP. This plan was realised in [20], and the current paper provides some new insights on that model.

It is worth emphasising that the realisations of a relaxed strategy are usually impossible on practice. For a discussion, see [6, p.78]. Roughly speaking, if the decision maker wants to use two actions with non-zero probabilities at each time moment, then the trajectories of the control process are not measurable. On the opposite, randomised strategies are clearly implementable.

In difference from [5, 6], we consider the total undiscounted cost which transforms to the discounted cost in a special case. What is new and important, the simple randomised strategies (called below 'Markov standard ξ-strategies') are no more sufficient in general even if there are no constraints (see Section 5 'Example'). At the same time, the new class of randomised strategies (called below 'Poisson-related') turns to be sufficient without any restrictions in the framework of constrained optimisation. This makes it possible to prove the equivalence of the continuous-time problem with the corresponding discrete-time MDP, where transitions to the same state (loops) are allowed. Remember that in simple cases that equivalence was known long ago through the so called uniformisation technique [22] which is not directly applicable if the transition rate is unbounded. In the case of discounted cost, the transformation to the discrete-time MDP was justified in [6] without any restrictive conditions.

In Section 2, we introduce the model under study and describe the general and particular classes of strategies. Note that the transition rate may be arbitrarily unbounded and non-conservative. The process is studied up to the accumulation of jumps (if it takes place). In

Sections 3 and 4, the main theoretical results regarding to occupation measures are presented. An example illustrating the theoretical issues is given in Section 5. In Section 6, we show how one can use the modern theory of discrete-time MDP for solving the underlying (constrained) continuous-time problems. The proofs are postponed to Appendix.

2 Model Description

The following notations are frequently used throughout this paper. $\mathbb{N}$ is the set of natural numbers including zero; $\delta_x(\cdot)$ is the Dirac measure concentrated at x, we call such distributions degenerate; $I\{\cdot\}$ is the indicator function. $\mathcal{B}(E)$ is the Borel σ-algebra of the Borel space E, $\mathcal{P}(E)$ is the Borel space of probability measures on E. $\mathcal{F}_1 \bigvee \mathcal{F}_2$ is the smallest σ-algebra containing the two σ-algebras $\mathcal{F}_1$ and $\mathcal{F}_2$. $\mathbb{R}_+ \stackrel{\triangle}{=} (0,\infty)$, $\mathbb{R}_+^0 \stackrel{\triangle}{=} [0,\infty)$, $\bar{\mathbb{R}} = [-\infty,+\infty]$, $\bar{\mathbb{R}}_+ = (0,\infty]$, $\bar{\mathbb{R}}_+^0 = [0,\infty]$. The abbreviation *w.r.t.* (resp. *a.s.*) stands for "with respect to" (resp. "almost surely"); for $b, d \in \bar{\mathbb{R}}$, $b \wedge d = \min\{b,d\}$, $b^+ \stackrel{\triangle}{=} \max\{b,0\}$ and $b^- \stackrel{\triangle}{=} \min\{b,0\}$. Capital letters denote random variables, and little letters are for their values.

The primitives of a continuous-time Markov decision process (CTMDP) are the following elements.

- State space: $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$ (arbitrary Borel).
- Action space: $(\mathbf{A}, \mathcal{B}(\mathbf{A}))$ (arbitrary Borel), $\mathbf{A}(x) \in \mathcal{B}(\mathbf{A})$ is the non-empty space of admissible actions in state $x \in \mathbf{X}$. It is supposed that $\mathbb{K} \stackrel{\triangle}{=} \{(x,a) \in \mathbf{X} \times \mathbf{A} : \ a \in \mathbf{A}(x)\} \in \mathcal{B}(\mathbf{X} \times \mathbf{A})$ and this set contains the graph of a measurable function from $\mathbf{X}$ to $\mathbf{A}$.
- Transition rate: $q(dy|x,a)$, a signed kernel on $\mathcal{B}(\mathbf{X})$ given $(x,a) \in \mathbb{K}$, taking nonnegative values on $\Gamma_{\mathbf{X}} \setminus \{x\}$ with $\Gamma_{\mathbf{X}} \in \mathcal{B}(\mathbf{X})$. We assume that $q(\mathbf{X}|x,a) \le 0$ and $\bar{q}_x = \sup_{a \in \mathbf{A}(x)} q_x(a) < \infty$, where $q_x(a) \stackrel{\triangle}{=} -q(\{x\}|x,a)$.
- Cost rates: measurable $\bar{\mathbb{R}}$-valued functions $c_i(x,a)$ on $\mathbb{K}$, $i = 0, 1, 2, \ldots, N$.
- Initial distribution: $\gamma(\cdot)$, a probability measure on $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$.
- Additional Borel space $(\mathbf{\Xi}, \mathcal{B}(\mathbf{\Xi}))$, the source of the control randomness.

Actually, the space $(\mathbf{\Xi}, \mathcal{B}(\mathbf{\Xi}))$ can be chosen by the decision maker (see Definition 1), but it is convenient to introduce it immediately, in order to describe the sample space. The role of the space $\mathbf{\Xi}$ will become clear after the description of control strategies.

We introduce the artificial isolated point (cemetery) Δ, put $\mathbf{X}_\Delta \stackrel{\triangle}{=} \mathbf{X} \cup \{\Delta\}$, $\mathbf{\Xi}_\Delta = \mathbf{\Xi} \cup \{\Delta\}$, and define $\mathbf{A}(\Delta) \stackrel{\triangle}{=} \mathbf{A}$, $q(\Gamma|\Delta,a) \stackrel{\triangle}{=} 0$ for all $\Gamma \in \mathcal{B}(\mathbf{X}_\Delta)$, $\alpha(x,a) \stackrel{\triangle}{=} q(\{\Delta\}|x,a) \stackrel{\triangle}{=} q_x(a) - q(\mathbf{X} \setminus \{x\}|x,a) \ge 0$ for $(x,a) \in \mathbb{K}$. The state Δ means, the process is over, i.e. escaped from the state space. We also put $c_i(\Delta,a) = 0$.

Given the above primitives, let us construct the underlying (measurable) sample space $(\Omega, \mathcal{F})$. Having firstly defined the measurable space $(\Omega^0, \mathcal{F}^0) \stackrel{\triangle}{=} (\mathbf{\Xi} \times (\mathbf{X} \times \mathbf{\Xi} \times \mathbb{R}_+)^\infty, \mathcal{B}(\mathbf{\Xi} \times (\mathbf{X} \times \mathbf{\Xi} \times \mathbb{R}_+)^\infty))$, let us adjoin all the sequences of the form

$$(\xi_0, x_0, \ \xi_1, \ \theta_1, \ x_1, \ \xi_2, \ \ldots, \ \theta_{m-1}, \ x_{m-1}, \ \xi_m, \ \theta_m, \ \Delta, \ \Delta, \ \infty, \ \Delta, \ \Delta, \ \ldots)$$

to Ω^0, where $m \ge 1$ is some integer, $\xi_m \in \mathbf{\Xi}$, $\theta_m \in \bar{\mathbb{R}}_+$, $\theta_l \in \mathbb{R}_+$, $x_l \in \mathbf{X}$, $\xi_l \in \mathbf{\Xi}$ for all nonnegative integers $l \le m-1$. After the corresponding modification of the σ-algebra $\mathcal{F}^0$, we obtain the basic sample space $(\Omega, \mathcal{F})$.

Below,

$$\omega = (\xi_0,\ x_0,\ \xi_1,\ \theta_1,\ x_1,\ \xi_2,\ \theta_2,\ x_2, \ldots).$$

For $n \in \mathbb{N} \setminus \{0\}$, introduce the mapping $\Theta_n :\ \Omega \to \bar{\mathbb{R}}_+$ by $\Theta_n(\omega) = \theta_n$; for $n \in \mathbb{N}$, the mappings $X_n :\ \Omega \to \mathbf{X}_\Delta$ and $\Xi_n :\ \Omega \to \mathbf{\Xi}_\Delta$ are defined by $X_n(\omega) = x_n$ and $\Xi_n(\omega) = \xi_n$. As usual, the argument ω will be often omitted. The increasing sequence of random variables T_n, $n \in \mathbb{N}$ is defined by $T_n = \sum_{i=1}^n \Theta_i$; $T_\infty = \lim_{n\to\infty} T_n$. Here, Θ_n (resp. T_n, X_n) can be understood as the sojourn times (resp. the jump moments, the states of the process on the intervals $[T_n, T_{n+1})$). We do not intend to consider the process after T_∞; the isolated point Δ will be regarded as absorbing; it appears when $\theta_m = \infty$ or when $\theta_m < \infty$ and the jump $x_{m-1} \to \Delta$ is realized with intensity $\alpha(x, a)$. The meaning of the ξ_n components will be described later. Finally, for $n \in \mathbb{N}$,

$$H_n = (\Xi_0, X_0, \Xi_1, \Theta_1, X_1, \ldots, \Xi_n, \Theta_n, X_n)$$

is the n-term (random) history.

The random measure μ is a measure on $\mathbb{R}_+ \times \mathbf{\Xi} \times \mathbf{X}_\Delta$ with values in $\mathbb{N} \cup \{\infty\}$, defined by

$$\mu(\omega;\ \Gamma_{\mathbb{R}} \times \Gamma_{\mathbf{\Xi}} \times \Gamma_{\mathbf{X}}) = \sum_{n\geq 1} I\{T_n(\omega) < \infty\} \delta_{(T_n(\omega), \Xi_n(\omega), X_n(\omega))}(\Gamma_{\mathbb{R}} \times \Gamma_{\mathbf{\Xi}} \times \Gamma_{\mathbf{X}});$$

the right continuous filtration $(\mathcal{F}_t)_{t\in\mathbb{R}^0_+}$ on $(\Omega, \mathcal{F})$ is given by

$$\mathcal{F}_t = \sigma\{H_0\} \vee \sigma\{\mu(]0, s] \times B) :\ s \leq t,\ B \in \mathcal{B}(\mathbf{\Xi} \times \mathbf{X}_\Delta)\}.$$

The controlled process of our interest

$$X(\omega, t) \stackrel{\triangle}{=} \sum_{n\geq 0} I\{T_n \leq t < T_{n+1}\} X_n + I\{T_\infty \leq t\}\Delta$$

takes values in $\mathbf{X}_\Delta$ and is right continuous and adapted. The filtration $\{\mathcal{F}_t\}_{t\geq 0}$ gives rise to the predictable σ-algebra on $\Omega \times \mathbb{R}^0_+$ defined by $\mathcal{P} \stackrel{\triangle}{=} \sigma\{\Gamma \times \{0\}\ (\Gamma \in \mathcal{F}_0), \Gamma \times (s, \infty)\ (\Gamma \in \mathcal{F}_{s-}, s > 0)\}$, where $\mathcal{F}_{s-} \stackrel{\triangle}{=} \bigvee_{t<s} \mathcal{F}_t$.

Definition 1 *A <u>control strategy</u> is defined as follows*

$$S = \{\mathbf{\Xi}, p_0, \langle p_n, \pi_n\rangle,\ n = 1, 2, \ldots\},$$

where $p_0(d\xi_0)$ is a probability distribution on $\mathbf{\Xi}$; for $x_{n-1} \in \mathbf{X}$, $p_n(d\xi_n|h_{n-1})$ is a stochastic kernel on $\mathbf{\Xi}$ given $\mathbf{H}_{n-1}$ (the space of $(n-1)$-component histories); $\pi_n(da|h_{n-1}, \xi_n, s)$ is a stochastic kernel on $\mathbf{A}(x_{n-1})$ given $\mathbf{H}_{n-1} \times \mathbf{\Xi} \times \mathbb{R}_+$. If $x_{n-1} = \Delta$, then we assume that $p_n(d\xi_n|h_{n-1}) = \delta_\Delta(d\xi_n)$ and $\pi_n(da|h_{n-1}, \Delta, s) = \delta_\Delta(da)$.

The p_n components mean the randomizations of controls; the π_n components mean relaxations.

If the randomizations are absent, that is, the kernels π_n do not depend on the ξ-components, then we deal with a <u>relaxed</u> strategy. One can omit the ξ_n components; as a result we obtain the standard control strategy $\{\pi_n,\ n = 1, 2, \ldots\}$. Such models were built and investigated by many authors [5, 6, 16, 19].

On the other hand, if the relaxations are absent, that is, all kernels π_n are degenerate and concentrated at singletons

$$\varphi_n(\xi_0, x_0, \theta_1, \ldots, x_{n-1}, \xi_n, s) \in \mathbf{A}(x_{n-1}), \tag{1}$$

then the control process $A(t)$ can be defined like follows

$$\begin{aligned} A(\omega,t) &= \sum_{n\geq 1} I\{T_{n-1}<t\leq T_n\}\varphi_n(\Xi_0,X_0,\Xi_1,\Theta_1,\ldots,X_{n-1},\Xi_n,t-T_{n-1}) \\ &\quad +I\{T_\infty\leq t\}\Delta. \end{aligned} \tag{2}$$

Below, we call such (purely randomized) strategies as ξ-strategies; they are defined by sequences $\{\boldsymbol{\Xi}, p_0, \langle p_n,\varphi_n\rangle,\ n=1,2,\ldots\}$. According to (2), after the history H_{n-1} is realized, the decision maker flips a coin resulting in the value of Ξ_n having the distribution p_n. Afterwards, up to the next jump epoch T_n, the control $A(t)$ is just a (deterministic measurable) function φ_n.

Definition 2 *ξ-strategies were defined just above. Purely relaxed strategies introduced earlier will be called π-strategies. General strategies S can be called π-ξ-strategies. If $\pi_n(da|x_0,\theta_1,x_1,\theta_2,\ldots,x_{n-1},s) = \pi_n^M(da|x_{n-1},s)$ for all $n=1,2,\ldots$ then the π-strategy is called Markov.*

Suppose a π-ξ-strategy S is fixed. The dynamics of the controlled process can be described like follows. First of all, $\Xi_0=\xi_0$ is realized based on the chosen distribution $p_0(d\xi_0)$. If p_0 is a combination of two Dirac measures, then in the future this or that control will be applied: p_0 is responsible for the mixtures of simpler control strategies. After that, the initial state X_0, having the distribution $\gamma(dx)$, is realized. Later, when the realized state $x_{n-1}\in\mathbf{X}$ becomes known at the realized jump epoch t_{n-1} $(n=1,2,\ldots)$, the dynamics is controlled in the following way. The decision maker flips a coin resulting in the $\Xi_n=\xi_n$ component having distribution $p_n(d\xi_n|h_{n-1})$; after that the stochastic kernel $\pi_n(da|h_{n-1},\xi_n,s)$ gives rise to the jumps intensity $\lambda_n(\Gamma|h_{n-1},s)$ from the current state x_{n-1} to $\Gamma\in\mathcal{B}(\mathbf{X}_\Delta)$, where

$$\lambda_n(\Gamma|h_{n-1},\xi_n,s)=\int_{\mathbf{A}}\pi_n(da|h_{n-1},\xi_n,s)q(\Gamma\setminus\{x_{n-1}\}|x_{n-1},a); \tag{3}$$

parameter $s>0$ is the time interval passed after the jump epoch t_{n-1}. After the corresponding interval θ_n, the new state $x_n\in\mathbf{X}_\Delta$ of the process $X(t)$ is realized at the jump epoch $t_n=t_{n-1}+\theta_n$. The joint distribution of (Θ_n,X_n) is given below. And so on. If $\theta_n=\infty$ then $x_n=\Delta$ and actually the process is over: the triples $(\theta=\infty,\Delta,\Delta)$ will be repeated endlessly. The same happens if $\theta_n<\infty$ and $x_n=\Delta$. Along with the intensity λ_n, we need the following integral

$$\Lambda_n(\Gamma,h_{n-1},\xi_n,t)=\int_{(0,t]}\lambda_n(\Gamma|h_{n-1},\xi_n,s)ds. \tag{4}$$

Note that, in case $q_x(a)\geq\varepsilon>0$, $\Lambda_n(\mathbf{X}_\Delta|h_{n-1},\xi_n,\infty)=\infty$ if $x_{n-1}\neq\Delta$.

Now, the distribution of $H_0=(\Xi_0,X_0)$ is given by $p_0(d\xi_0)\cdot\gamma(dx_0)$ and, for any $n\in\mathbb{N}\setminus\{0\}$, the stochastic kernel G_n on $\bar{\mathbb{R}}_+\times\boldsymbol{\Xi}_\Delta\times\mathbf{X}_\Delta$ given $\mathbf{H}_{n-1}$ is defined by formulae

$$\begin{aligned} G_n(\{\infty\}\times\{\Delta\}\times\{\Delta\}|h_{n-1}) &= \delta_{x_{n-1}}(\{\Delta\}); \\ G_n(\{\infty\}\times\Gamma_{\boldsymbol{\Xi}}\times\{\Delta\}|h_{n-1}) &= \delta_{x_{n-1}}(\mathbf{X})\int_{\Gamma_{\boldsymbol{\Xi}}}e^{-\Lambda(\mathbf{X}_\Delta,h_{n-1},\xi_n,\infty)}p_n(d\xi_n|h_{n-1}); \\ G_n(\Gamma_{\mathbb{R}}\times\Gamma_{\boldsymbol{\Xi}}\times\Gamma_{\mathbf{X}}|h_{n-1}) &= \delta_{x_{n-1}}(\mathbf{X})\int_{\Gamma_{\boldsymbol{\Xi}}}\int_{\Gamma_{\mathbb{R}}}\lambda_n(\Gamma_{\mathbf{X}}|h_{n-1},\xi_n,t) \\ &\quad \times e^{-\Lambda_n(\mathbf{X}_\Delta,h_{n-1},\xi_n,t)}dt\ p_n(d\xi_n|h_{n-1}); \\ G_n(\{\infty\}\times\boldsymbol{\Xi}_\Delta\times\mathbf{X}|h_{n-1}) &= G_n(\mathbb{R}_+\times\{\Delta\}\times\mathbf{X}_\Delta|h_{n-1})=0. \end{aligned} \tag{5}$$

Here $\Gamma_{\mathbb{R}}\in\mathcal{B}(\mathbb{R}_+)$, $\Gamma_{\boldsymbol{\Xi}}\in\mathcal{B}(\boldsymbol{\Xi})$, $\Gamma_{\mathbf{X}}\in\mathcal{B}(\mathbf{X}_\Delta)$.

It remains to apply the induction and Ionescu-Tulcea's theorem [2, Prop.7.28] or [17, p.294] to obtain the probability measure P_γ^S on $(\Omega, \mathcal{F})$ called strategic measure. A more detailed discussion and connection to the martingales, compensator etc [14] can be found in [20].

Below, when $\gamma(\cdot)$ is a Dirac measure concentrated at $x \in \mathbf{X}$, we use the 'degenerated' notation P_x^S. Expectations with respect to P_γ^S and P_x^S are denoted as E_γ^S and E_x^S, respectively. The set of all π-ξ-strategies S will be denoted as Π_S; the collections of all π- and ξ-strategies will be denoted as Π_π and Π_ξ correspondingly.

We aim to study several classes of control strategies and the associated measures. That is important for stochastic optimal control. For example, one can consider the following problem:

$$\begin{aligned} W_0(S) &= E_\gamma^S\left[\sum_{n=1}^{\infty}\int_{(T_{n-1},T_n]}\int_{\mathbf{A}}\pi_n(da|H_{n-1},\Xi_n,t-T_{n-1})c_0^+(X_{n-1},a)dt\right] \\ &\quad +E_\gamma^S\left[\sum_{n=1}^{\infty}\int_{(T_{n-1},T_n]}\int_{\mathbf{A}}\pi_n(da|H_{n-1},\Xi_n,t-T_{n-1})c_0^-(X_{n-1},a)dt\right] \\ &= E_\gamma^S\left[\int_{(0,T_\infty)}\int_{\mathbf{A}}\pi(da|t)c_0(X(t),a)\ dt\right] \to \inf_{S\in\Pi_S} \qquad (6) \\ &\quad \text{subject to} \\ W_i(S) &\le d_i, \quad i=1,2,\ldots,N, \end{aligned}$$

where all the objectives $W_i(S)$ have the form similar to $W_0(S)$ with function c_0 being replaced with other given cost rates c_i; d_i are given numbers. Here and below, $\infty-\infty \stackrel{\triangle}{=} +\infty$ and

$$\pi(da|t)=\sum_{n=1}^{\infty} I\{T_{n-1}<t\le T_n\}\pi_n(da|H_{n-1},\Xi_n,t-T_{n-1})$$

is the $\mathcal{P}(\mathbf{A})$-valued random process. The notions of optimal and ε-optimal strategies are conventional.

Remark 1 *Suppose a strategy S is such that, for some $m \ge 0$, all kernels $\{\pi_n\}_{n=1}^\infty$ for $x_{n-1} \ne \Delta$ do not depend on the ξ_m-component. Then one can omit $\xi_m \in \mathbf{\Xi}_\Delta$ and $\Xi_m \in \mathbf{\Xi}_\Delta$ from the consideration. In this case, instead of the strategic measure $P_\gamma^S(d\omega)$, we can everywhere use the marginal $\tilde{P}_\gamma^S(d\tilde{\omega}) = P_\gamma^S(d\tilde{\omega}\times\mathbf{\Xi})$. Here*

$$\tilde{\omega}=(\xi_0,x_0,\xi_1,\theta_1,\ldots,x_{m-1},\theta_m,x_m,\xi_{m+1},\theta_{m+1},\ldots)$$

and $\tilde{\omega}\times\mathbf{\Xi}=(\xi_0,x_0,\xi_1,\theta_1,\ldots,x_{m-1},\mathbf{\Xi},\theta_m,x_m,\xi_{m+1},\theta_{m+1},\ldots)$. Below, we omit the tilde and hope this will not lead to a confusion.

For example, for a purely relaxed strategy $S\in\Pi_\pi$, the strategic measure is defined on the space of sequences

$$\omega=(x_0,\theta_1,x_1,\ldots),$$

and that is standard for CTMDP [5, 6, 16, 19].

As was mentioned, the space $\mathbf{\Xi}$ can be chosen by the decision maker. Let us look at several possibilities.

Definition 3 *Suppose $\mathbf{\Xi}=\mathbf{A}$, the relaxations are absent, and the functions φ_n in (2) have the form $\varphi_n(h_{n-1},\xi_n,s)=\xi_n$, so that the argument ξ_0 never appears and thus can be omitted.*

Then such a strategy will be called a standard ξ-strategy. It will be denoted as $S = \{\mathbf{A},\ p_n,\ n = 1, 2, \ldots\}$ and below we usually write A_n (or a_n) instead of Ξ_n (or ξ_n), $n = 1, 2, \ldots$. If we consider only such strategies then we deal with the so called exponential semi-Markov decision process [5, p.498]. In case $p_n(d\xi_n|h_{n-1}) = p_n(da_n|h_{n-1}) = p_n^M(da_n|x_{n-1})$ $(n = 1, 2, \ldots)$, the standard ξ-strategy will be called Markov. The collection of all Markov standard ξ-strategies will be denoted as Π_ξ^M, they are often denoted as p^m.

According to Remark 1, slightly modified sample spaces are associated with different types of strategies which are again denoted in different ways. For the reader's convenience, we summarize the main notations in the following table.

Strategy	Sample space
General (π-ξ-strategy) $S = \{\mathbf{\Xi},\ p_0,\ \langle p_n, \pi_n\rangle,\ n = 1, 2, \ldots\} \in \Pi_S$	$\Omega = \{(\xi_0, x_0, \xi_1, \theta_1, x_1, \xi_2, \theta_2, \ldots)\}$
Purely randomized (ξ-strategy) $S = \{\mathbf{\Xi},\ p_0,\ \langle p_n, \varphi_n\rangle,\ n = 1, 2, \ldots\} \in \Pi_\xi$	$\Omega = \{(\xi_0, x_0, \xi_1, \theta_1, x_1, \xi_2, \theta_2, \ldots)\}$
Purely relaxed (π-strategy) $S = \{\pi_n,\ n = 1, 2, \ldots\} \in \Pi_\pi$	$\Omega = \{(x_0, \theta_1, x_1, \theta_2, \ldots)\}$
Markov standard ξ-strategy $S = \{\mathbf{A},\ p_n^M(da_n\|x_{n-1}),\ n = 1, 2, \ldots\}$ $= p^M \in \Pi_\xi^M$	$\Omega = \{(x_0, a_1, \theta_1, x_1, a_2, \theta_2, \ldots)\}$

We introduced the new, richer set of strategies Π_S, and one of the targets is to establish the sufficiency of smaller classes Π_π and Π_ξ. More about the model in [20].

3 Occupation Measures and Sufficient Classes of Strategies

Definition 4 *Following [5, 6], for a fixed strategy $S \in \Pi_S$, we introduce the occupation measures for $n = 1, 2, \ldots$:*

$$\eta_n^S(\Gamma_{\mathbf{X}} \times \Gamma_{\mathbf{A}}) = E_\gamma^S \left[\int_{(T_{n-1}, T_n] \cap \mathbb{R}_+} I\{X_{n-1} \in \Gamma_{\mathbf{X}}\} \pi_n(\Gamma_{\mathbf{A}}|H_{n-1}, \Xi_n, t - T_{n-1}) dt \right],$$

where $\Gamma_{\mathbf{X}} \in \mathcal{B}(\mathbf{X}), \Gamma_{\mathbf{A}} \in \mathcal{B}(\mathbf{A})$.

Remark 2 *If S is a standard ξ-strategy then, for $n = 1, 2, \ldots$*

$$\eta_n^S(\Gamma_{\mathbf{X}} \times \Gamma_{\mathbf{A}}) = E_\gamma^S \left[I\{X_{n-1} \in \Gamma_{\mathbf{X}}\} I\{A_n \in \Gamma_{\mathbf{A}}\} \Theta_n \right] = E_\gamma^S [\delta_{X_{n-1}}(\Gamma_{\mathbf{X}}) \delta_{A_n}(\Gamma_{\mathbf{A}}) \Theta_n],$$

and

$$\begin{aligned} & \int_{\Gamma_{\mathbf{X}}} \int_{\Gamma_{\mathbf{A}}} q_x(a) \eta_n^S(dx, da) \\ = \ & E_\gamma^S \left[I\{X_{n-1} \in \Gamma_{\mathbf{X}}\} I\{A_n \in \Gamma_{\mathbf{A}}\} I\{q_{X_{n-1}}(A_n) > 0\} E_\gamma^S [\Theta_n | X_{n-1}, A_n] \right] \\ = \ & E_\gamma^S \left[I\{X_{n-1} \in \Gamma_{\mathbf{X}}\} I\{A_n \in \Gamma_{\mathbf{A}}\} \right] \end{aligned}$$

confirming that, e.g., if S is Markov standard then $\sum_{n=1}^{\infty} q_x(a)\eta_n^S$ coincides with the (total) occupation measure on $\mathbb{K}$ in the discrete-time MDP with the same state and action spaces $\mathbf{X}_\Delta$ and $\mathbf{A}$ and transition probability

$$Q(\Gamma_{\mathbf{X}}|x,a) = I\{q_x(a)=0\}I\{\Gamma_{\mathbf{X}} \ni \Delta\} + I\{q_x(a)>0\}\frac{q(\Gamma_{\mathbf{X}} \setminus \{x\}|x,a)}{q_x(a)},$$

under the control strategy $p_n^M(da|x_{n-1})$. We discuss the relations to the discrete-time MDP in Section 6.

For any non-negative function r, for any $S \in \Pi_S$,

$$E_\gamma^S\left[\sum_{n=1}^{\infty}\int_{(T_{n-1},T_n]\cap\mathbb{R}_+}\int_{\mathbf{A}}\pi_n(da|H_{n-1},\Xi_n,t-T_{n-1})r(X_{n-1},a)dt\right] \tag{7}$$

$$=\sum_{n=1}^{\infty}\int_{\mathbf{X}\times\mathbf{A}} r(x,a)\eta_n^S(dx,da).$$

Now, after we introduce the sets
$\mathcal{D}_S = \{\ \{\eta_n^S\}_{n=1}^{\infty},\ S \in \Pi_S\}$,
$\mathcal{D}_\pi = \{\ \{\eta_n^S\}_{n=1}^{\infty},\ S \in \Pi_\pi,\ S$ is Markov$\}$ and
$\mathcal{D}_\xi = \{\ \{\eta_n^S\}_{n=1}^{\infty},\ S \in \Pi_\xi$ with $\mathbf{\Xi} = \mathbf{A}$, ξ-strategy S is Markov standard$\}$,
the problem (6) can be reformulated as

$$\left.\begin{array}{ll} & \displaystyle\sum_{n=1}^{\infty}\int_{\mathbf{X}\times\mathbf{A}} c_0(x,a)\eta_n(dx,da) \to \inf_{\{\eta_n\}_{n=1}^{\infty}\in\mathcal{D}_S} \\ \text{subject to} & \\ & \displaystyle\sum_{n=1}^{\infty}\int_{\mathbf{X}\times\mathbf{A}} c_i(x,a)\eta_n(dx,da) \le d_i, \quad i=1,2,\dots,N. \end{array}\right\}$$

Condition 1 *(a) $q_x(a) > 0$ for all $(x,a) \in \mathbb{K}$. (b) $\exists \varepsilon > 0 :\ \forall x \in \mathbf{X}$ $\inf_{a\in\mathbf{A}(x)} q_x(a) \ge \varepsilon$.*

It is clear that the possible gap

$$\alpha(x,a) \stackrel{\triangle}{=} q_x(a) - q(\mathbf{X}\setminus\{x\}|x,a) = q(\{\Delta\}|x,a) \ge 0$$

can be understood as the discount factor depending on the current state and action. More about this in [20]. If $\alpha > 0$ is a constant then we deal with the classical discounted model [5, 6, 10, 19] satisfying the requirement 1-(b). Certainly, if $q_x(a) = 0$ for some $(x,a) \in \mathbb{K}$, and that state x cannot be reached under any control strategy S, then one can consider the state space $\mathbf{X}\setminus\{x\}$. Similarly, if $q_x(a) \equiv 0$ for all $a \in \mathbf{A}(x)$ and $\forall i = 0,1,2,\dots,N$, $\forall n = 1,2,\dots$ $c_i(x,a) \equiv 0$ for all $a \in \mathbf{A}(x)$, then one can denote that state x as Δ (meaning, the process escaped from the state space $\mathbf{X}$). The situation, when $q_x(a) = 0$ and $c_i(x,a) \ne 0$ for a reachable state x and for some i and $a \in \mathbf{A}(x)$, is more delicate.

Theorem 1 *Suppose Condition 1-(a) is satisfied. Then, for any π-ξ-strategy S, there is a Markov standard ξ-strategy S_ξ such that $\eta_n^{S_\xi} \ge \eta_n^S$ for all $n = 1,2,\dots$. Hence, Markov standard ξ-strategies are sufficient for solving optimization problem (6) with negative costs c_i.*

If Condition 1-(b) is satisfied, then $\mathcal{D}_S = \mathcal{D}_\xi$. Hence, Markov standard ξ-strategies are sufficient in the problem (6).

Theorem 2 *$\mathcal{D}_S = \mathcal{D}_\pi$. Thus, Markov π-strategies are sufficient in the problem (6).*

The proofs will appear in [20].

If $\eta_n^{S_1} = \eta_n^{S_2}$ for all $n = 1, 2, \ldots$ then, for any cost rate c_0, the expected total costs $W_0(S_1) = W_0(S_2)$ are the same. But other important objectives (e.g. the variances) may be different. Consider the following simple example: $\mathbf{X} = \{1\}$, $\mathbf{A} = \mathbf{A}(1) = \{a_1, a_2\}$, $\gamma(1) = 1$, $q_1(a_1) = \lambda_1$, $q_1(a_2) = \lambda_2$, $N = 0$. Note that $q(\mathbf{X} \setminus \{1\}|1, a) = 0$ and $q(\mathbf{X}|1, a) = -q_1(a) < 0$. After introducing the cemetery Δ with $\alpha(1, a) = q(\{\Delta\}|1, a) = q_1(a)$, we obtain the standard conservative transition rate q. In this model, we have a single sojourn time $\Theta = T$, so that the n index is omitted. Suppose the cost rate $c_0(1, a) = c^a$ is given. Let S_1 be the stationary π-strategy with the values $\pi^s(a_1|x) = \beta = 1 - \pi^s(a_2|1)$. Then $\eta^{S_1} = \eta^{S_2}$ for the Markov standard ξ-strategy S_2 defined by $p(a_1|1) = \frac{\beta\lambda_1}{\beta\lambda_1 + (1-\beta)\lambda_2} = 1 - p(a_2|1)$: see the proof of Theorem 1 in [20]. For S_1, the sample space contains only the pairs $(x_0 = 1, \theta)$, the (random) total cost is $C = [\beta c^{a_1} + (1 - \beta)c^{a_2}]\theta$,

$$E_\gamma^{S_1}[C] = \frac{\beta c^{a_1} + (1-\beta)c^{a_2}}{\beta\lambda_1 + (1-\beta)\lambda_2} \text{ and } E_\gamma^{S_1}[C^2] = \frac{2[\beta c^{a_1} + (1-\beta)c^{a_2}]^2}{[\beta\lambda_1 + (1-\beta)\lambda_2]^2}.$$

For S_2, the sample space contains the triplets $(x_0 = 1, \xi = a, \theta)$, the total (random) cost is $C = I\{a = a_1\}c^{a_1}\theta + I\{a = a_2\}c^{a_2}\theta$,

$$E_\gamma^{S_2}[C] = \frac{\beta\lambda_1}{\beta\lambda_1 + (1-\beta)\lambda_2} \cdot \frac{c^{a_1}}{\lambda_1} + \frac{(1-\beta)\lambda_2}{\beta\lambda_1 + (1-\beta)\lambda_2} \cdot \frac{c^{a_2}}{\lambda_2} = E_\gamma^{S_1}[C],$$

but

$$E_\gamma^{S_2}[C^2] = \frac{\beta\lambda_1}{\beta\lambda_1 + (1-\beta)\lambda_2} \cdot \frac{2(c^{a_1})^2}{\lambda_1^2} + \frac{(1-\beta)\lambda_2}{\beta\lambda_1 + (1-\beta)\lambda_2} \cdot \frac{2(c^{a_2})^2}{\lambda_2^2}.$$

The difference

$$E_\gamma^{S_2}[C^2] - E_\gamma^{S_2}[C^2] = \frac{2}{\beta\lambda_1 + (1-\beta)\lambda_2} \left\{ \frac{\beta(c^{a_1})^2}{\lambda_1} + \frac{(1-\beta)(c^{a_2})^2}{\lambda_2} \right.$$

$$\left. - \frac{[\beta c^{a_1} + (1-\beta)c^{a_2}]^2}{\beta\lambda_1 + (1-\beta)\lambda_2} \right\} = \frac{2\beta(1-\beta)[\lambda_2 c^{a_1} - \lambda_1 c^{a_2}]^2}{\lambda_1\lambda_2[\beta\lambda_1 + (1-\beta)\lambda_2]^2}$$

is non-negative, so that the variance of the total cost is bigger for the ξ-strategy S_2.

4 Sufficiency of ξ-strategies, General Case

Example presented in Section 5 shows that, if Condition 1 is not satisfied, then it can happen that, for a π-strategy S, there is no equivalent Markov standard ξ-strategy having the same occupation measures. Below, we describe a more general class of ξ-strategies which turns to be sufficient in the general case.

Definition 5 *A Poisson-related ξ-strategy*

$$S = \{\mathbf{\Xi}, \varepsilon, \tilde{p}_{n,k}(da|x_{n-1}),\ n = 1, 2, \ldots,\ k = 1, 2, \ldots\}$$

is defined by a constant $\varepsilon > 0$ and a sequence of stochastic kernels $\tilde{p}_{n,k}(da|x)$ from $\mathbf{X}_\Delta$ to $\mathbf{A}$ with $\tilde{p}_{n,k}(\mathbf{A}(x)|x) = 1$. Here $\mathbf{\Xi} = (\mathbf{A} \times \mathbb{R})^\infty = \{(\alpha_1, \tau_1, \alpha_2, \tau_2, \ldots)\}$, and for $n = 1, 2, \ldots$ the distribution p_n of $\Xi_n = (A_1^n, T_1^n, A_2^n, \ldots)$ given $\mathbf{H}_{n-1}$ is defined as follows:

- *for all* $k \geq 1$, $p_n(A_k^n \in \Gamma_{\mathbf{A}} | h_{n-1}) = \tilde{p}_{n,k}(\Gamma_{\mathbf{A}} | x_{n-1})$;
- *for all* $k \geq 1$, $p_n(T_k^n \leq t | h_{n-1}) = 1 - e^{-\varepsilon t}$; *random variables* T_k^n *are mutually independent and also independent of* $\mathcal{F}_{T_{n-1}} = \mathcal{B}(\mathbf{H}_{n-1})$;
- *finally,*

$$\varphi_n(\xi_0, x_0, \xi_1, \theta_1, \ldots, x_{n-1}, \xi_n, s) = \sum_{k=1}^{\infty} I\{\tau_1^n + \ldots + \tau_{k-1}^n < s \leq \tau_1^n + \ldots + \tau_k^n\} \alpha_k^n,$$

and the mapping φ_n *in fact depends only on* ξ_n.

The Ξ_0 *component plays no role and is omitted.*

Such a strategy means that, after any jump of the controlled process $X(t)$, we simulate a Poisson process and apply different randomized controls during the different sojourn times of that Poisson process.

Theorem 3 *For any control strategy* S, *there is a Poisson-related* ξ*-strategy* S^P *such that* $\{\eta_n^S\}_{n=1}^{\infty} = \{\eta_n^{S^P}\}_{n=1}^{\infty}$. *The value of* $\varepsilon > 0$ *can be chosen arbitrarily.*

The proof can be found in [20]. The explicit form of the S^P strategy is given by the following expressions. Suppose $S = \{\boldsymbol{\Xi}, p_0, \langle p_n, \pi_n \rangle, n = 1, 2, \ldots\} \in \Pi_S$ is a given control strategy and fix an arbitrary $\varepsilon > 0$. The (standard) space $\widetilde{\boldsymbol{\Xi}} = (\mathbf{A} \times \mathbb{R})^{\infty}$ which appears in the definition of a Poisson-related strategy, is equipped with tilde. It has no concern to the calculations. For a fixed $n \geq 1$, we introduce random functions $Q_k(w)$ depending on $\omega \in \Omega$:

$$Q_k(w) \stackrel{\triangle}{=} \frac{\varepsilon(\varepsilon w)^{k-1}}{(k-1)!} e^{-\varepsilon w - \Lambda_n(\mathbf{X}_{\Delta}, H_{n-1}, \Xi_n, w)}, \quad k = 1, 2, \ldots, \ w \in \mathbb{R}_+^0$$

and (random) function $f_w(t)$:

$$f_w(t) \stackrel{\triangle}{=} [\lambda_n(\mathbf{X}_{\Delta} | H_{n-1}, \Xi_n, w + t) + \varepsilon] e^{-\Lambda_n(\mathbf{X}_{\Delta}, H_{n-1}, \Xi_n, w+t) + \Lambda_n(\mathbf{X}_{\Delta}, H_{n-1}, \Xi_n, w) - \varepsilon t},$$

$w, t \in \mathbb{R}_+^0$.

Now, the Poisson-related ξ-strategy S^P of our interest is defined by

$$\begin{aligned}
\tilde{p}_{n,1}(\Gamma_{\mathbf{A}} | x_{n-1}) &\stackrel{\triangle}{=} E_{\gamma}^S \left[\int_{(0,\infty)} f_0(t) \int_{(0,t]} \int_{\Gamma_{\mathbf{A}}} \pi_n(da | H_{n-1}, \Xi_n, u) \right. \\
&\quad \left. \times [q_{X_{n-1}}(a) + \varepsilon] du \ dt | X_{n-1} = x_{n-1} \right]; \\
\tilde{p}_{n,k}(\Gamma_{\mathbf{A}} | x_{n-1}) &\stackrel{\triangle}{=} \frac{1}{E_{\gamma}^S \left[\int\limits_{(0,\infty)} Q_{k-1}(w) dw | X_{n-1} = x_{n-1} \right]} \\
&\quad \times E_{\gamma}^S \left[\int\limits_{(0,\infty)} Q_{k-1}(w) \int\limits_{(0,\infty)} f_w(t) \int\limits_{(0,t]} \int\limits_{\Gamma_{\mathbf{A}}} \pi_n(da | H_{n-1}, \Xi_n, w + u) \right. \\
&\quad \left. \times [q_{X_{n-1}}(a) + \varepsilon] du \ dt \ dw | X_{n-1} = x_{n-1} \right],
\end{aligned}$$

for $k \geq 2$.

By the way, the normalizing denominator $E_\gamma^S\left[\int\limits_{(0,\infty)} Q_{k-1}(w)dw|X_{n-1}=x_{n-1}\right]$ equals the P_γ^S-probability and also the $P_\gamma^{S^P}$-probability that Θ_n is bigger than the $Erlang(\varepsilon, k-1)$ RV, i.e. that the action A_k is actually applied when using the S^P strategy.

5 Example

This example illustrates that Markov standard strategies (as well as stationary standard ξ-strategies and stationary π-strategies) are not sufficient in optimization problems.

Consider the following CTMDP, very similar to the one described in [9, Ex.3.1]. $\mathbf{X}=\{1\}$, $\mathbf{A}=\mathbf{A}(1)=(0,1]$, $\gamma(\{1\})=1$, $q_1(a)=a$, $c_0(x,a)=a$, $N=0$. Note that $q(\mathbf{X}\setminus\{1\}|1,a)=0$ and $q(\mathbf{X}|1,a)=-q_1(a)=-a<0$. After introducing the cemetery Δ with $\alpha(1,a)=q(\{\Delta\}|1,a)=q_1(a)$, we obtain the standard conservative transition rate q. In this model, we have a single sojourn time $\Theta=T$, so that the n index is omitted.

It is obvious that, for any Markov standard ξ-strategy p^M (which is also stationary),

$$\eta^{p^M}(\{1\}\times\Gamma_{\mathbf{A}})=E_\gamma^{p^M}\left[\int_{(0,T]\cap\mathbb{R}_+} I\{A(t)\in\Gamma_{\mathbf{A}}\}dt\right]=\int_{\Gamma_{\mathbf{A}}} p^M(da|1)\cdot\frac{1}{a}$$

and

$$W_0(p^M)=E_\gamma^{p^M}\left[\int_{(0,T]\cap\mathbb{R}_+} A(t)dt\right]=\int_{\mathbf{A}} a\ \eta^{p^M}(\{1\}\times da)=\int_{\mathbf{A}} a\frac{1}{a}p^M(da|1)=1.$$

For an arbitrary stationary π-strategy S_π, we similarly obtain

$$\eta^{S_\pi}(\{1\}\times\Gamma_{\mathbf{A}})=\pi(\Gamma_{\mathbf{A}})\Big/\int_{\mathbf{A}} a\ \pi(da)$$

and

$$W_0(S_\pi)=\int_{\mathbf{A}} a\ \eta^{S_\pi}(\{1\}\times da)=1.$$

On the other hand, under an arbitrarily fixed $\kappa>0$, for the purely deterministic strategy $\varphi(1,s)=e^{-\kappa s}$, the (first) sojourn time $\Theta=T$ has the cumulative distribution function (CDF) $1-e^{\frac{-1+e^{-\kappa\theta}}{\kappa}}$, so that $P_\gamma^\varphi(\Theta=\infty)=e^{-\frac{1}{\kappa}}$. Under an arbitrarily fixed $U\in(0,1]$ we have

$$\eta^\varphi(\{1\}\times(U,1])=\int_U^1 \frac{e^{\frac{-1+a}{\kappa}}}{\kappa a}da. \tag{8}$$

The detailed calculation is given in [20]. The measure $\eta^\varphi(\{1\}\times da)$ is absolutely continuous w.r.t. the Lebesgue measure, the density being $\frac{e^{\frac{-1+a}{\kappa}}}{\kappa a}$ and

$$W_0(\varphi)=\int_{\mathbf{A}} a\ \eta^\varphi(\{1\}\times da)=1-e^{-\frac{1}{\kappa}}. \tag{9}$$

It is clear that $\inf_{S\in\Pi_S} W_0(S)=0$: see (9) with $\kappa\to\infty$, but the optimal strategy does not exist because $\Theta>0$ and $c_0(x,a)>0$. Note also that, if we extend the action space to $[0,1]$ and keep q_1 and c_0 continuous, i.e., $q_1(0)=c_0(0)=0$, then stationary deterministic strategy $\varphi^*(x)=0$ is optimal with $W_0(\varphi^*)=0$.

According to Theorem 1, there is a Markov standard ξ-strategy S_ξ such that $\eta^{S_\xi} \geq \eta^\varphi$. It is given by the following formula:

$$P^M((U,1]|1) = \frac{E^\varphi_\gamma \left[\int_{(0,\Theta]} I\{e^{-\kappa t} \in (U,1]\} e^{-\kappa t} dt\right]}{E^\varphi_\gamma \left[\int_{(0,\Theta]} e^{-\kappa t} dt\right]}.$$

After the change of variables $y = e^{-\kappa t}$, the numerator becomes

$$E^\varphi_\gamma \left[\int_{[e^{-\kappa\Theta},1)} I\{y \in (U,1]\} \frac{dy}{\kappa}\right] = 1 - e^{\frac{U-1}{\kappa}}$$

and

$$P^M((U,1]|1) = \frac{1 - e^{\frac{U-1}{\kappa}}}{1 - e^{-\frac{1}{\kappa}}} = \int_U^1 \frac{\frac{1}{\kappa} e^{\frac{a-1}{\kappa}}}{1 - e^{-\frac{1}{\kappa}}} da.$$

Now, since for any $a \in \mathbf{A}$ the expectation of Θ is $\frac{1}{a}$,

$$\eta^{S_\xi}(\{1\} \times \Gamma_{\mathbf{A}}) = \int_{\Gamma_{\mathbf{A}}} \frac{\frac{1}{\kappa a} e^{\frac{a-1}{\kappa}}}{1 - e^{-\frac{1}{\kappa}}} da \geq \int_{\Gamma_{\mathbf{A}}} e^{\frac{a-1}{\kappa}} \frac{1}{\kappa a} da = \eta^\varphi(\{1\} \times \Gamma_{\mathbf{A}}).$$

Let us construct the Poisson-related ξ-strategy S^P such that $\eta^{S^P} = \eta^\varphi$, using the expressions given at the end of Section 4.

As usual, we omit index $n = 1$. Now $\Lambda(\mathbf{X}_\Delta, h_0, \xi_0, t) = \int_0^t e^{-\kappa s} ds = \frac{1 - e^{-\kappa t}}{\kappa}$ and, using the formulae for $Q_k(w)$ and $f_w(t)$, we obtain

$$\begin{aligned}
\tilde{p}_1((U,1]|1) &= \int_0^\infty \left[\int_0^t I\{e^{-\kappa s} \in (U,1]\}[e^{-\kappa s} + \varepsilon] ds\right] (e^{-\kappa t} + \varepsilon) e^{\frac{-1+e^{-\kappa t}}{\kappa} - \varepsilon t} dt \\
&= \int_0^{-\frac{\ln U}{\kappa}} \left[\frac{1}{\kappa}(1 - e^{-\kappa t}) + \varepsilon t\right] (e^{-\kappa t} + \varepsilon) e^{\frac{-1+e^{-\kappa t}}{\kappa} - \varepsilon t} dt \\
&\quad + \frac{1}{\kappa} \int_{-\frac{\ln U}{\kappa}}^\infty [1 - U - \varepsilon \ln U](e^{-\kappa t} + \varepsilon) e^{\frac{-1+e^{-\kappa t}}{\kappa} - \varepsilon t} dt,
\end{aligned}$$

and the density of the $\tilde{p}_1$ distribution is given by

$$-\frac{d\tilde{p}_1((u,1]|1)}{du} = \frac{1}{\kappa}\left(1 + \frac{\varepsilon}{u}\right) u^{\frac{\varepsilon}{\kappa}} e^{\frac{-1+u}{\kappa}}.$$

The starting point for the description of the desired Poisson-related strategy S^P is as follows.

- On the interval $(0, T_1]$ one should choose the action A_1 using the CDF $a^{\frac{\varepsilon}{\kappa}} e^{\frac{-1+a}{\kappa}}$, $a \in (0,1] = \mathbf{A}$.

- The expected cost on the interval $(0, T_1 \wedge \Theta]$ equals

$$\frac{1}{\kappa} \int_0^1 \left(1 + \frac{\varepsilon}{a}\right) a^{\frac{\varepsilon}{\kappa}} e^{\frac{a-1}{\kappa}} \cdot \frac{a}{a + \varepsilon} da = \frac{1}{\kappa} \int_0^1 a^{\frac{\varepsilon}{\kappa}} e^{\frac{a-1}{\kappa}} da.$$

For $k \geq 2$, we have

$$\tilde{p}_k((U,1]|1) = \frac{\int_0^{\frac{-\ln U}{\kappa}} \frac{\varepsilon(\varepsilon w)^{k-2}}{(k-2)!} e^{-\varepsilon w} e^{\frac{-1+e^{-\kappa w}}{\kappa}} [1 - e^{\frac{U-e^{-\kappa w}}{\kappa} + \varepsilon w + \frac{\varepsilon}{\kappa}\ln U}] dw}{\int_0^{\infty} \frac{\varepsilon(\varepsilon w)^{k-2}}{(k-2)!} e^{-\varepsilon w} e^{\frac{-1+e^{-\kappa w}}{\kappa}} dw},$$

and the desired Poisson-related strategy S^P is as follows.

- On the interval $\left(\sum_{i=1}^{k-1} T_i, \sum_{i=1}^{k} T_i\right]$ one should choose the action A_k using probability density

$$\begin{aligned} -\frac{d\tilde{p}_k((a,1]|1)}{da} &= \frac{\varepsilon \int_0^{-\frac{\ln a}{\kappa}} \frac{(\varepsilon w)^{k-2}}{(k-2)!} e^{-\frac{1}{\kappa}} \left(\frac{1}{\kappa} + \frac{\varepsilon}{\kappa a}\right) e^{\frac{a}{\kappa} + \frac{\varepsilon}{\kappa}\ln a} dw}{\int_0^{\infty} \frac{\varepsilon(\varepsilon w)^{k-2}}{(k-2)!} e^{-\varepsilon w} e^{\frac{-1+e^{-\kappa w}}{\kappa}} dw} \\ &= \frac{\frac{a+\varepsilon}{a\kappa(k-1)!} \left(\frac{-\varepsilon \ln a}{\kappa}\right)^{k-1} e^{\frac{a}{\kappa} + \frac{\varepsilon \ln a}{\kappa} - \frac{1}{\kappa}}}{\int_0^{\infty} \frac{\varepsilon(\varepsilon w)^{k-2}}{(k-2)!} e^{-\varepsilon w} e^{\frac{-1+e^{-\kappa w}}{\kappa}} dw}. \end{aligned}$$

- If the action A_k is actually applied then the duration $\tilde{T}_k$ is the smallest RV between the sojourn time (in state 1 under the action A_k) and the independent $\exp(\varepsilon)$ random variable T_k. If A_k is not applied, we put $\tilde{T}_k = 0$. The expected length of that interval $\tilde{T}_k$ (if positive, that is, with probability $\int_0^{\infty} \frac{\varepsilon(\varepsilon w)^{k-2}}{(k-2)!} e^{-\varepsilon w} e^{\frac{-1+e^{-\kappa w}}{\kappa}} dw$: see the proof of Th.5 in [20]) equals $\frac{1}{A_k+\varepsilon}$ and

$$E_\gamma^{S^P}\left[\int_{(0,\tilde{T}_k]} I\{A_k \in (U,1]\} dt\right] = \int_U^1 \frac{1}{\kappa a} e^{\frac{-1+\varepsilon\ln a + a}{\kappa}} \frac{\left(\frac{-\varepsilon\ln a}{\kappa}\right)^{k-1}}{(k-1)!} da.$$

- The expected cost on that interval equals $\int_0^1 \frac{1}{\kappa} e^{\frac{-1+\varepsilon\ln a + a}{\kappa}} \frac{\left(\frac{-\varepsilon\ln a}{\kappa}\right)^{k-1}}{(k-1)!} da.$

One can easily compute

$$\begin{aligned} \eta^{S^P}(\{1\}\times(U,1]) &= \sum_{k=1}^{\infty} E_\gamma^{S^P}\left[\int_{(0,\tilde{T}_k]} I\{A_k \in (U,1]\} dt\right] \\ &= \int_U^1 \frac{1}{\kappa} a^{\frac{\varepsilon}{\kappa}-1} e^{\frac{a-1}{\kappa}} da + \int_U^1 \frac{1}{\kappa} a^{\frac{\varepsilon}{\kappa}-1} e^{\frac{a-1}{\kappa}} \left(a^{-\frac{\varepsilon}{\kappa}} - 1\right) da \\ &= \int_U^1 \frac{e^{\frac{a-1}{\kappa}}}{\kappa a} da = \eta^{\varphi}(\{1\}\times(U,1]) : \end{aligned}$$

see (8).

Similarly,

$$\begin{aligned} W_0(S^P) &= \frac{1}{\kappa}\int_0^1 a^{\frac{\varepsilon}{\kappa}} e^{\frac{a-1}{\kappa}} da + \int_0^1 \frac{1}{\kappa} a^{\frac{\varepsilon}{\kappa}} e^{\frac{a-1}{\kappa}} \left(a^{-\frac{\varepsilon}{\kappa}} - 1\right) da \\ &= \int_0^1 \frac{1}{\kappa} e^{\frac{a-1}{\kappa}} da = 1 - e^{-\frac{1}{\kappa}} = W_0(\varphi) : \end{aligned}$$

see (9).

6 Continuous and Discrete-Time MDP

6.1 Non-Zero Jumps Intensity

Suppose Condition 1-(b) is satisfied (or Condition 1-(a) if the cost rates c_i are negative). Then, according to Theorem 1, Markov standard ξ-strategies are sufficient in problem (6). Formula (6) takes the form

$$\begin{aligned} W_0(S) &= \sum_{n=1}^{\infty} E_\gamma^S \left[I\{X_{n-1} \neq \Delta\} E_\gamma^S \left[c_0(X_{n-1}, A_n)\Theta_n | \mathcal{F}_{T_{n-1}} \right] \right] \\ &= \sum_{n=1}^{\infty} E_\gamma^S \left[I\{X_{n-1} \neq \Delta\} \int_{\mathbf{A}} \frac{c_0(X_{n-1}, a)}{q_{X_{n-1}}(a)} p_n^M(da|X_{n-1}) \right] \to \inf_{S \in \Pi_\xi^M}. \end{aligned} \tag{10}$$

It remains to notice that $\forall \Gamma_{\mathbf{X}} \in \mathcal{B}(\mathbf{X}_\Delta)$

$$P_\gamma^S(X_n \in \Gamma_{\mathbf{X}} | \mathcal{F}_{T_{n-1}}) = \int_{\mathbf{A}} \frac{q(\Gamma_{\mathbf{X}} \setminus \{X_{n-1}\}|X_{n-1}, a)}{q_{X_{n-1}}(a)} p_n^M(da|X_{n-1}) \tag{11}$$

to deduce that actually we deal with a discrete-time MDP in the class of randomized Markov control strategies p^M. Indeed, at any one time moment n, having the current state $x_{n-1} \in \mathbf{X}$ and choosing action $a \in \mathbf{A}(x_{n-1})$, we face the one-step cost $c_0(x_{n-1}, a)/q_{x_{n-1}}(a)$, and the process moves to a state $x_n \in \Gamma_{\mathbf{X}} \in \mathcal{B}(\mathbf{X}_\Delta)$ with probability $q(\Gamma_{\mathbf{X}} \setminus \{x_{n-1}\}|x_{n-1}, a)/q_{x_{n-1}}(a)$. State Δ is absorbing with zero one-step cost. It is known that randomized Markov control strategies are sufficient for solving discrete-time problems with the total expected cost [17, Lemma 2].

The optimality equation looks as follows

$$\inf_{a \in \mathbf{A}(x)} \left\{ c_0(x, a)/q_x(a) + \int_{\mathbf{X} \setminus \{x\}} v(y) q(dy|x, a)/q_x(a) - v(x) \right\} = 0, \quad x \in \mathbf{X}_\Delta, \tag{12}$$

and all the theory of discrete-time MDP is applicable.

Remark 3 *If $c_0(x, a) \geq 0$ and the model is semi-continuous (see [2, Def.8.7], or [3, Ass.2.1], or [17, Con.5]) then the Bellman function $v(x) = \inf_{p^M \in \Pi_\xi^M} W_0(p^M)$, where x is the initial state (i.e. $\gamma(dy) = \delta_x(dy)$), is the minimal non-negative solution to (12). Moreover, there exists a stationary deterministic uniformly (or persistently) optimal strategy φ^*, that is, a strategy satisfying $v(x) = W_0(\varphi^*)$ for all initial states $x \in \mathbf{X}$. The (measurable) mapping $\varphi^* : \mathbf{X}_\Delta \to \mathbf{A}$ provides the infimum in (12) [2, Prop.9.12, Cor.9.17.2]. One can find more about the total-cost MDP in [1, 2, 7, 13] and other monographs and articles.*

Note that MDP with total (undiscounted) expected cost is a challenging area, full of unexpected: strategy and value iterations may be unsuccessful, a conserving strategy (providing the infimum in (12)) may be not optimal, and so on: see the corresponding counter-examples in [18, Ch.2]. At the same time, particular cases, like transient and discounted models are well studied [1, 2, 13, 17]. For instance, in the standard discounted case, if $\alpha(x, a) = \alpha > 0$, the model is semi-continuous, and the cost rate c_0 is bounded, then equation (12) has a single bounded solution on $\mathbf{X}$ with $v(\Delta) = 0$, and the stationary conserving strategy exists and is optimal. By the way, here equation (12) takes the form

$$\inf_{a \in \mathbf{A}(x)} \left\{ c_0(x, a) + \int_{\mathbf{X} \setminus \{x\}} v(y) q(dy|x, a) - q(\mathbf{X} \setminus \{x\}|x, a) v(x) - \alpha v(x) \right\} = 0, \quad x \in \mathbf{X},$$

coincident with the Bellman equation investigated in many works on CTMDP [10, 19, 21]. Note that the discount factor was state -dependent in [23]. One can investigate also the case when the cost rate c_0 is not necessarily bounded, working in the spaces with 'weighted' norms. Similar approach was demonstrated in [10, 19, 21] for CTMDP and in [1, 13], [22, §6.10] for discrete-time MDP.

In the cited works on CTMDP, many efforts were made to ensure that the controlled process is non-explosive, that is, $P_\gamma^S(T_\infty = \infty) = 1$ for all strategies S. We underline here that explosions are not excluded in the current article: we simply consider the $X(t)$ process up to the moment T_∞ which may be finite.

Let us apply the recent results on constrained discrete-time MDP with total expected cost [4] to the problem (6).

Condition 2 *(a) There exists a dominating probability measure m on $\mathbf{X}$: $\forall (x,a) \in \mathbb{K}$ $q(\cdot|x,a) \ll m$. (Here the measure $q(\cdot|x,a)$ is considered to be defined on $\mathbf{X} \setminus \{x\}$ for $(x,a) \in \mathbb{K}$.)*

(b) $\mathbf{A}$ is compact, $\forall x \in \mathbf{X}$ $\mathbf{A}(x) = \mathbf{A}$; for any $\Gamma_{\mathbf{X}} \in \mathcal{B}(\mathbf{X}_\Delta)$ and $x \in \mathbf{X}$ function $q(\Gamma_{\mathbf{X}} \setminus \{x\}|x,\cdot)/q_x(\cdot)$ is continuous on $\mathbf{A}$; functions $c_i(x,\cdot)$, $i = 0,1,2,\dots,N$ are continuous on $\mathbf{A}$ for any $x \in \mathbf{X}$.

The linear program $\mathbb{LP}$ associated with the constrained problem (6) looks as follows:

$$W_0 = \int_{\mathbf{X}\times\mathbf{A}} \frac{c_0(x,a)}{q_x(a)} \eta(dx,da) \to \inf \tag{13}$$

subject to $\eta \in \mathbb{L}_C$, where $\mathbb{L}_C$ is the space of (possibly infinite-valued) feasible measures, that is, satisfying equation

$$\eta(\Gamma_{\mathbf{X}} \times \mathbf{A}) = \gamma(\Gamma_{\mathbf{X}}) + \int_{\mathbf{X}\times\mathbf{A}} \frac{q(\Gamma_{\mathbf{X}} \setminus \{x\}|x,a)}{q_x(a)} \eta(dx,da), \tag{14}$$

and such that, for any $i = 0,1,2,\dots,N$, the integral

$$W_i = \int_{\mathbf{X}\times\mathbf{A}} \frac{c_i(x,a)}{q_x(a)} \eta(dx,da) \tag{15}$$

is well defined and satisfies the constraints

$$W_i \le d_i, \quad i = 1,2,\dots,N. \tag{16}$$

Note that, for any Markov standard ξ-strategy p^M, the total sum of (slightly modified) occupation measures $\sum_{n=1}^{\infty} q_x(a)\eta_n^{p^M}(dx,da)$ satisfies equation (14).

We also need auxiliary linear programs $\mathbb{LP}_i$, $i = 0,1,2,\dots,N$

$$W_i \to \inf$$

subject to $\eta \in \mathbb{L}_i$, where $\mathbb{L}_i$ is the space of measures satisfying (14) and such that the integral (15) is well defined.

Proposition 1 *Let Condition 2 be satisfied. Suppose, for any $i = 0,1,2,\dots,N$, the minimal value of the linear program $\mathbb{LP}_i$ is finite and let η^* be the optimal solution of the constrained linear program $\mathbb{LP}$ (13),(14),(15),(16). Then η^* gives rise to the so called 'induced' stochastic kernel $p^*(da|x)$ which defines the stationary standard ξ-strategy solving problem (6).*

The proof follows from [4, Th.5.2]. Generally speaking, after the measure η^* is obtained, the state space $\mathbf{X}$ is split into two disjoint parts $\mathbf{X} = V \cup V^c$. The subset V^c is the largest (in some sense) such that the measure $\eta^*(dx, da)$ is σ-finite on it and hence can be disintegrated: $\eta^*(dx, da) = p^*(da|x)\eta^*(dx \times \mathbf{A})$. On the set V, $p^*(da|x) = \delta_{f(x)}(da)$, where $f(x)$ is a specially constructed function: see Lemma 5.1, Prop.5.1 and Def.5.1 in [4]. Easier constructions can be found in [4, §4], where the set $\mathbf{A}$ is finite.

Note that in [4] the number of constraints N was not necessarily finite.

If all cost rates $c_i \geq 0$ are non-negative, then a stronger version of Proposition 1 is valid (see [3]): if the model is semi-continuous and the objective (13) is finite for some feasible measure η, then there is a stationary standard ξ-strategy solving problem (6).

6.2 General Case

Now we investigate the general case when Condition 1 is not necessarily fulfilled. For an arbitrarily fixed $\varepsilon > 0$, consider the discrete-time MDP $\mathcal{M}$ with the same state and action spaces $\mathbf{X}_\Delta$ and $\mathbf{A}$ and the same set $\mathbb{K}$ of admissible state-action pairs. Transition probability on $\mathbf{X}_\Delta$ is defined by

$$Q(\Gamma_{\mathbf{X}}|y, b) = \frac{q(\Gamma_{\mathbf{X}} \setminus \{y\}|y, b) + \varepsilon I\{\Gamma_{\mathbf{X}} \ni y\}}{q_y(b) + \varepsilon};$$

the initial distribution is γ. Here and below, it is convenient to denote the states and actions in the $\mathcal{M}$ model as y and b. The notions of a control strategy p and the corresponding strategic measure ${}^{\mathcal{M}}P^p_\gamma$ in $\mathcal{M}$ are conventional [12, 17]. The (total) occupation measure ${}^{\mathcal{M}}\eta^p$ on $\mathbb{K}$ is defined by the standard formula (see [13, §9.4]):

$${}^{\mathcal{M}}\eta^p(\Gamma_{\mathbf{X}} \times \Gamma_{\mathbf{A}}) = \sum_{m=1}^{\infty} {}^{\mathcal{M}}\eta^p_m(\Gamma_{\mathbf{X}} \times \Gamma_{\mathbf{A}}) = \sum_{m=1}^{\infty} {}^{\mathcal{M}}E^p_\gamma[I\{Y_{m-1} \in \Gamma_{\mathbf{X}},\ B_m \in \Gamma_{\mathbf{A}}\}]. \tag{17}$$

Let $\mathcal{D}$ be the full collection of such occupation measures under different Markov strategies p. Other strategies do not extend the set $\mathcal{D}$ [17, Lemma 2].

Lemma 1 *$\mathcal{D}$ coincides with the space of all (total) occupation measures $\sum_{n=1}^{\infty}(q_x(a) + \varepsilon)\eta^S_n$ under different strategies $S \in \Pi_S$ in the original continuous-time model. (See Section 3.)*

Now it is clear that solving the original constrained problem (6) is equivalent to solving the corresponding discrete-time MDP with one-step costs $c_i(y, b)/[q_y(b) + \varepsilon]$. By the way, in the unconstrained case, the optimality equation takes the form

$$\inf_{b \in \mathbf{A}(y)} \left\{ c_0(y, b)/[q_x(a) + \varepsilon] + \int_{\mathbf{X}\setminus\{y\}} v(z)q(dz|y, b)/[q_y(b) + \varepsilon] \right.$$
$$\left. + \varepsilon v(y)/[q_y(b) + \varepsilon] - v(y) \right\} = 0, \qquad y \in \mathbf{X}_\Delta$$

yielding

$$\inf_{b \in \mathbf{A}(y)} \left\{ c_0(y, b) + \int_{\mathbf{X}\setminus\{y\}} v(z)q(dz|y, b) - q_y(b)v(y) \right\} = 0 \tag{18}$$

for such $y \in \mathbf{X}_\Delta$ that $v(y) \in \mathbb{R}$. The last equation is well known for the problems with total (undiscounted) cost [11].

All the assertions in Remark 3 hold true. Note also that, for any stationary deterministic strategy φ^s ${}^{\mathcal{M}}\eta^{\varphi^s} = \sum_{n=1}^{\infty}(q_x(a)+\varepsilon)\eta_n^{\varphi^s}$: see the proof of Lemma 1.

Under Condition 2 one can investigate the linear program similar to (13), (14),(15),(16) and apply Proposition 1. If, for instance, the problem is unconstrained ($N = 0$) and the value of $\mathbb{LP}$

$$\int_{\mathbf{X}\times\mathbf{A}} \frac{c_0(x,a)}{q_x(a)+\varepsilon}\eta(dx,da) \to \inf_{\eta} \tag{19}$$

subject to

$$\hat{\eta}(\Gamma_{\mathbf{X}}) = \eta(\Gamma_{\mathbf{X}}\times\mathbf{A}) = \gamma(\Gamma_{\mathbf{X}}) + \int_{\mathbf{X}\times\mathbf{A}}\left[\frac{q(\Gamma_{\mathbf{X}}\setminus\{x\}|x,a)+\varepsilon I\{x\in\Gamma_{\mathbf{X}}\}}{q_x(a)+\varepsilon}\right]\eta(dx,da)$$

is finite, then, in case the optimal marginal $\hat{\eta}^*(\cdot)$ is σ-finite, one can disintegrate the optimal measure $\eta^*(dx,da) = \hat{\eta}^*(dx)p^*(da|x)$ and obtain an optimal stationary Poisson-related ξ-strategy with (n,k)-independent stochastic kernels $\tilde{p}_{n,k}(da|x) = p^*(da|x)$. Here we assumed that the $\mathbb{LP}$ (19) has an optimal solution η^*.

In the example presented in Section 5 all the conditions 2 were satisfied except for the compactness of the action space $\mathbf{A}$. Remember, the set of Markov standard ξ-strategies was not sufficient there in the problem $W_0(S) \to \inf_{S\in\Pi_S}$. In the corresponding discrete-time MDP, there is no optimal stationary strategy. The $\mathbb{LP}$ looks as follows:

$$\int_{(0,1]}\frac{a}{a+\varepsilon}\eta(\{1\}\times da) \to \inf_{\eta} \tag{20}$$

subject to

$$\hat{\eta}(\{1\}) = \eta(\{1\}\times(0,1]) = 1 + \int_{(0,1]}\frac{\varepsilon}{a+\varepsilon}\eta(\{1\}\times da). \tag{21}$$

If $\hat{\eta}(\{1\}) < \infty$ (note that $\hat{\eta}^p(\{1\}) < \infty$ for any stationary strategy p), one can write η in the form $\eta(\{1\}\times da) = \hat{\eta}(\{1\})\times p(da)$ and, for any probability measure p, we have

$$\hat{\eta}(\{1\}) = \left[1 - \int_{(0,1]}\frac{\varepsilon}{a+\varepsilon}p(da)\right]^{-1} = \left[\int_{(0,1]}\frac{a}{a+\varepsilon}p(da)\right]^{-1}$$

$$\Longrightarrow \text{Total cost equals } \int_{(0,1]}\frac{a}{a+\varepsilon}p(da)\left[1-\int_{(0,1]}\frac{\varepsilon}{a+\varepsilon}p(da)\right]^{-1} = 1$$

as expected. Proposition 1 does not help because the solution to $\mathbb{LP}$ (20) does not exist. The infimum equals zero because, e.g. for the measure $\eta(\{1\}\times da) = (a+\varepsilon)\frac{e^{\frac{a-1}{\kappa}}}{\kappa a}da$ with $\kappa > 0$ we have that $\hat{\eta}((\{1\}) = \infty$ and $\int_{(0,1]}\frac{a}{a+\varepsilon}\eta(\{1\}\times da) = 1 - e^{-\frac{1}{\kappa}}$. On the other hand, to obtain zero in (20) we must have $\eta(\{1\}\times da) = 0$ which violates the requirement (21).

Let us show that, for any $\delta > 0$, there is a non-stationary Markov strategy in the assiciated discrete-time MDP $\mathcal{M}$ with the total expected cost smaller than δ. Take a_1 such that $\frac{a_1}{a_1+\varepsilon} < \frac{\delta}{2}$, take a_2 such that $\frac{a_2}{a_2+\varepsilon} < \frac{\delta}{4}$ and so on: take a_m such that $\frac{a_m}{a_m+\varepsilon} < \frac{\delta}{2^m}$. Then, for this Markov deterministic non-stationary strategy $p_m^*(da|x) = \delta_{a_m}(da)$, we have

$$\begin{aligned}{}^{\mathcal{M}}E_\gamma^{p^*}\left[\sum_{m=1}^{\infty}\frac{c_0(X_{m-1},A_m)}{q_{X_{m-1}}(A_m)+\varepsilon}\right] &= \frac{a_1}{a_1+\varepsilon} + \frac{\varepsilon}{a_1+\varepsilon}\left[\frac{a_2}{a_2+\varepsilon}+\frac{\varepsilon}{a_2+\varepsilon}\right.\\ &\left.\times\left[\frac{a_3}{a_3+\varepsilon}+\ldots\right]\right] < \frac{\delta}{2}+\frac{\delta}{4}+\frac{\delta}{8}+\ldots = \delta.\end{aligned}$$

This strategy p^* gives rise to the corresponding Poisson-related δ-optimal strategy in the original CTMDP, with degenerate probabilities $\tilde{p}_k$: $\tilde{p}_k(da|1) = \delta_{a_k}(da)$. More examples of discrete-time MDP, where only non-stationary strategies can be δ-optimal, in [18, §2.2.11].

Remark 4 *In the case of unconstrained problem with $N = 0$ and $c_0 \geq 0$, in the associated discrete-time MDP $\mathcal{M}$, there is a δ-optimal non-randomized Markov strategy for any $\delta > 0$ [2, Prop.9.19]. That strategy gives rise to the δ-optimal Poisson-related strategy with degenerate probabilities $\tilde{p}_{n,k}$. Many other known statements from the discrete-time theory can be directly applied to the Poisson-related strategies in the framework of CTMDP. See for example the transient and absorbing MDPs in [1].*

7 Acknowledgement

The author is thankful to Dr. Y.Zhang for fruitful discussions and careful reading of the draft of this article.

8 Appendix

Proof of Lemma 1 (sketch). Assume $\varepsilon > 0$ is fixed.

1. For the proof of inclusion $\{\sum_{n=1}^{\infty}(q_x(a)+\varepsilon)\eta_n^S,\ S \in \Pi_S\} \subset \mathcal{D}$, according to Theorem 3, it is sufficient to consider only Poisson-related strategies S^P. Suppose such a strategy $S^P = \{\mathbf{\Xi}, \varepsilon, \tilde{p}_{n,k}(da|x_{n-1}),\ n,k = 1,2,\ldots\}$ is given. The elements of $\xi_n \in \mathbf{\Xi}$ are denoted as $\xi_n = (\alpha_1^n, \tau_1^n, \alpha_2^n, \tau_2^n, \ldots)$. We intend to build a control strategy $p = \{p_{m+1}(da|\ {}^{\mathcal{M}}h_m)\}_{m=0}^{\infty}$ in the $\mathcal{M}$ model such that

$$ {}^{\mathcal{M}}\eta^p(dx,da) = \sum_{n=1}^{\infty}(q_x(a)+\varepsilon)\eta_n^{S^P}(dx,da). \tag{22}$$

The elements relevant to the $\mathcal{M}$ model are equipped with the left upper index $\mathcal{M}$. It will be convenient to denote trajectories in $\mathcal{M}$ as ${}^{\mathcal{M}}\omega = (y_0, b_1, y_1, \ldots)$.

For a given history ${}^{\mathcal{M}}h_m = (y_0, b_1, y_1, \ldots, b_m, y_m)$ with $y_m \neq \Delta$ we define $l_1({}^{\mathcal{M}}h_m) = \min\{l \geq 1:\ l \leq m;\ y_l \neq y_{l-1}\} \wedge (m+1)$.

For $k \geq 1$, if $l_k({}^{\mathcal{M}}h_m) = m + 1 - \sum_{i=1}^{k-1} l_i({}^{\mathcal{M}}h_m)$ then $n({}^{\mathcal{M}}h_m) = k$; otherwise

$$\begin{aligned} l_{k+1}({}^{\mathcal{M}}h_m) \ &= \ \min\Bigg\{ l \geq 1:\ l \leq m - \sum_{i=1}^{k} l_i({}^{\mathcal{M}}h_m); \\ &\qquad y_{\sum_{i=1}^{k} l_i({}^{\mathcal{M}}h_m)+l} \neq y_{\sum_{i=1}^{k} l_i({}^{\mathcal{M}}h_m)+l-1} \Bigg\} \wedge (m+1-\sum_{i=1}^{k} l_i({}^{\mathcal{M}}h_m)). \end{aligned}$$

After that, $\sum_{i=1}^{n({}^{\mathcal{M}}h_m)} l_i({}^{\mathcal{M}}h_m) = m+1$, and we put $k({}^{\mathcal{M}}h_m) \stackrel{\triangle}{=} l_{n({}^{\mathcal{M}}h_m)}({}^{\mathcal{M}}h_m)$ and apply the randomized action according to the distribution

$$p_{m+1}(da|{}^{\mathcal{M}}h_m) = \tilde{p}_{n({}^{\mathcal{M}}h_m),k({}^{\mathcal{M}}h_m)}(da|y_m).$$

This past-dependent randomized strategy p in $\mathcal{M}$ is the desired one. Figure 1 illustrates this construction and the connection between (random) histories ${}^{\mathcal{M}}H_m$ and trajectories of the original control process $X(t)$.

For an (infinite) trajectory ${}^{\mathcal{M}}\omega$, the values $l_i({}^{\mathcal{M}}\omega) \in \mathbb{N} \cup \{\infty\}$, $i = 1, 2, \ldots$ are defined in the similar way: $l_1({}^{\mathcal{M}}\omega) = \min\{l \geq 1 : \ y_l \neq y_{l-1}\}$ and for $k \geq 1$ such that $l_k({}^{\mathcal{M}}\omega) < \infty$ and $y_{\sum_{i=1}^{k} l_i({}^{\mathcal{M}}\omega)+l-1} \neq \Delta$, we put

$$l_{k+1}({}^{\mathcal{M}}\omega) = \min\left\{l \geq 1 : \ y_{\sum_{i=1}^{k} l_i({}^{\mathcal{M}}\omega)+l} \neq y_{\sum_{i=1}^{k} l_i({}^{\mathcal{M}}\omega)+l-1}\right\}.$$

Below, these functions on the sample space of the $\mathcal{M}$ model are, as usual, denoted by capital letters L_k (random variables).

For any $n = 1, 2, \ldots$ for arbitrary $\Gamma_{\mathbf{X}} \in \mathcal{B}(\mathbf{X})$, $\Gamma_{\mathbf{A}} \in \mathcal{B}(\mathbf{A})$

$$\int_{\Gamma_{\mathbf{X}}} \int_{\Gamma_{\mathbf{A}}} (q_x(a) + \varepsilon)\eta_n^{S^P}(dx, da) = {}^{\mathcal{M}}E_\gamma^p \left[\sum_{m=\sum_{i=1}^{n-1} L_i+1}^{\sum_{i=1}^{n} L_i} I\{Y_{m-1} \in \Gamma_{\mathbf{X}}\} I\{B_m \in \Gamma_{\mathbf{A}}\} \right].$$

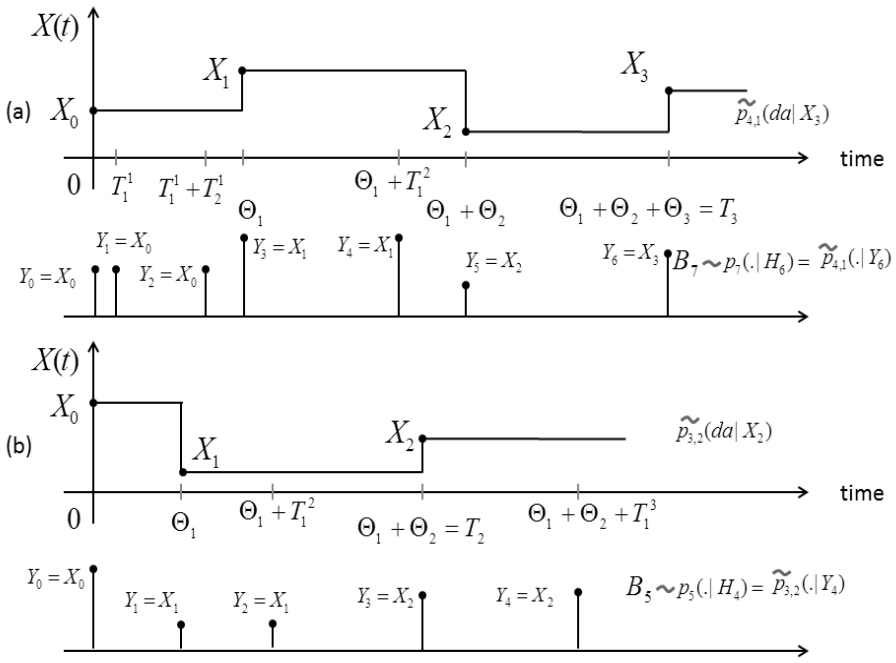

Figure 1: Two scenarios illustrating the construction of the $\mathcal{M}$ model:
(a) ${}^{\mathcal{M}}H_6 = (Y_0, B_1, Y_1, \ldots, B_6, Y_6)$; $l_1({}^{\mathcal{M}}H_6) = 3$, $l_2({}^{\mathcal{M}}H_6) = 2$, $l_3({}^{\mathcal{M}}H_6) = 1$, $l_4({}^{\mathcal{M}}H_6) = 1$, $n({}^{\mathcal{M}}H_6) = 4$, $k({}^{\mathcal{M}}H_6) = 1$;
(b) ${}^{\mathcal{M}}H_4 = (Y_0, B_1, Y_1, \ldots, B_4, Y_4)$; $l_1({}^{\mathcal{M}}H_4) = 1$, $l_2({}^{\mathcal{M}}H_4) = 2$, $l_3({}^{\mathcal{M}}H_4) = 2$, $n({}^{\mathcal{M}}H_4) = 3$, $k({}^{\mathcal{M}}H_4) = 2$.

This equality is based on the formulae

$$E_\gamma^{S^P}\left[I\{X_{n-1} \neq \Delta\} \int_{(\sum_{i=1}^{k} T_i^n, \sum_{i=1}^{k+1} T_i^n \wedge T_n]} (q_{X_{n-1}}(A_{k+1}^n) + \varepsilon)dt \middle| \sum_{i=1}^{k} T_i^n \right.$$
$$\left. < T_n, X_{n-1}, A_{k+1}^n \right] = E_\gamma^{S^P}\left[I\{X_{n-1} \neq \Delta\} \right];$$

$$P_\gamma^{S^P}(T_n<\infty,\ X_n\in\Gamma_{\mathbf{X}})=\ {}^{\mathcal{M}}P_\gamma^p\left(\sum_{i=1}^n L_i<\infty,\ Y_{\sum_{i=1}^n L_i}\in\Gamma_{\mathbf{X}}\right)$$

valid for all $n=1,2,\ldots$; $k=0,1,2,\ldots$. Therefore, (22) follows.

2. For the inverse inclusion $\{\sum_{n=1}^\infty (q_x(a)+\varepsilon)\eta_n^S,\ S\in\Pi_S\}\supset\mathcal{D}$, suppose a Markov strategy $p_m(da|x)$ in $\mathcal{M}$ is fixed and construct a past-dependent version S of a Poisson-related strategy such that

$$\sum_{n=1}^\infty (q_x(a)+\varepsilon)\eta_n^S(dx,da)=\ {}^{\mathcal{M}}\eta^p(dx,da). \tag{23}$$

Past-dependent means that the stochastic kernels $\tilde{p}_{n,k}$ will depend on the histories h_{n-1} rather than on the current states x_{n-1}.

Let $\tilde{p}_{1,k}(da|x_0)=p_k(da|x_0)$. For any history h_n with $x_n\neq\Delta$ we compute

$$k_n(h_n)\stackrel{\triangle}{=}\min\{k\geq 1:\ \sum_{i=1}^k \tau_i^n\geq\theta_n\}$$

and, in case $k_n(h_n)<\infty$, we put

$$\tilde{p}_{n+1,k}(da|h_n)=p_{\sum_{i=1}^n k_i(h_n)+k}(da|x_n).$$

If $k_n(h_n)=\infty$, the stochastic kernels $\tilde{p}_{n+1,k}$ can be defined arbitrarily.

For this strategy S, similarly to the ideas described above, one can prove equality

$$\int_{\Gamma_{\mathbf{X}}}\int_{\Gamma_{\mathbf{A}}}(q_x(a)+\varepsilon)\eta_n^S(dx,da)=\ {}^{\mathcal{M}}E_\gamma^p\left[\sum_{m=\sum_{i=1}^{n-1}L_i+1}^{\sum_{i=1}^n L_i} I\{Y_{m-1}\in\Gamma_{\mathbf{X}}\}I\{B_m\in\Gamma_{\mathbf{A}}\}\right]$$

for all $n=1,2,\ldots$, $\Gamma_{\mathbf{X}}\in\mathcal{B}(\mathbf{X})$, $\Gamma_{\mathbf{A}}\in\mathcal{B}(\mathbf{A})$.

After that, equality (23) is obvious. ■

References

[1] Altman, E. *Constrained Markov Decision Processes.* Chapman and Hall/CRC, Boca Raton, 1999.

[2] Bertsekas, D. and Shreve, S. *Stochastic Optimal Control.* Academic Press, NY, 1978.

[3] Dufour, F., Horiguchi, M. and Piunovskiy, A.: The expected total cost criterion for Markov decision processes under constraints: a convex analytic approach. *Adv. Appl. Prob.* **44** (2012) 774-793.

[4] Dufour, F. and Piunovskiy, A.: The expected total cost criterion for Markov decision processes under constraints. *Adv. Appl. Prob.* **45** (2013) 837-859.

[5] Feinberg, E.: Continuous time discounted jump Markov decision processes: a discrete-event approach. *Math. Oper. Res.* **29** (2004) 492-524.

[6] Feinberg, E.: Reduction of discounted continuous-time MDPs with unbounded jump and reward rates to discrete-time total-reward MDPs. In *Optimization, Control, and Applications of Stochastic Systems* (D.Hernandez-Hernandez and J.A.Minjares-Sosa ed.), Birkhauser, 2012, 77-97.

[7] Feinberg, E.: Total reward criteria. In *Handbook of Markov Decision Processes.* (E.Feinberg and A.Shwartz ed.), Kluwer, Boston/Dordrecht/London, 2002, 173-207.

[8] Ghosh, M. and Saha, S.: Non-stationary semi-Markov decision processes on a finite horizon. *Stoch. Anal. Appl.* **31** (2013) 183-190.

[9] Guo, X. and Zhang, Y.: Constrained total undiscounted continuous-time Markov decision processes. *Bernoulli*, accepted; http://arxiv.org/pdf/1304.3314v5.pdf

[10] Guo, X. and Piunovskiy, A.: Discounted continuous-time Markov decision processes with constraints: unbounded transition and loss rates. *Math. Oper. Res.* **36** (2011) 105-132.

[11] Guo, X., Vykertas, M., Zhang, Y.: Absorbing continuous-time Markov decision processes with total cost criteria. *Adv. Appl. Prob.* **45** (2013) 490-519.

[12] Hernández-Lerma, O. and Lasserre, J.B. *Discrete-Time Markov Control Processes.* Springer-Verlag, NY, 1996.

[13] Hernández-Lerma, O. and Lasserre, J.B. *Further Topics on Discrete-Time Markov Control Processes.* Springer-Verlag, NY, 1999.

[14] Jacod, J.: Multivariate point processes: predictable projection, Radon-Nykodym derivatives, representation of martingales. *Z. Wahrscheinlichkeitstheorie verw. Gebite.* **31** (1975) 235-253.

[15] Kitayev, M.: Semi-Markov and jump Markov controlled models: average cost criterion. *Theory. Probab. Appl.* **30** (1986) 272-288.

[16] Kitaev, M and Rykov, V. *Controlled Queueing Systems.* CRC Press, Boca Raton, 1995.

[17] Piunovskiy, A. *Optimal Control of Random Sequences in Problems with Constraints.* Kluwer, Dordrecht, 1997. 51-71.

[18] Piunovskiy, A. *Examples in Markov Decision Processes.* Imperial College Press, London, 2013.

[19] Piunovskiy, A. and Zhang, Y.: Discounted continuous-time Markov decision processes with unbounded rates: the convex analytic approach. *SIAM J. Control Optim.*, **49** (2011) 2032-2061.

[20] Piunovskiy, A.: Randomized and relaxed strategies in continuous-time Markov decision processes. *SIAM J. Control Optim.*, in press.

[21] Prieto-Rumeau, T. and Hernandez-Lerma, O. *Selected Topics on Continuous-Time Controlled Markov Chains and Markov Games.* Imperial College Press, London, 2012.

[22] Puterman, M. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* Wiley, NY, 1994. 215-235.

[23] Zhang, Y.: Convex analytic approach to constrained discounted Markov decision processes with non-constant discount factors. *TOP*, **21** (2013) 378-408.