Exact finite approximations of average-cost countable Markov Decision Processes

Arie Leizarowitz
Faculty of Mathematics
Technion, Haifa 32000
Israel

Adam Shwartz
Faculty of Electrical Engineering
Technion, Haifa 32000
Israel

July 2007

Abstract

For a countable-state Markov decision process we introduce an embedding which produces a finite-state Markov decision process. The finite-state embedded process has the same optimal cost, and moreover, it has the same dynamics as the original process when restricting to the approximating set. The embedded process can be used as an approximation which, being finite, is more convenient for computation and implementation.

Keywords: Markov Decision Processes, countable state space, finite approximations, average cost criterion.

AMS subject classification: 60G70, 60J10, 62L20.

1 Introduction

In this paper we develop a tool that is useful in studying countable state Markov Decision Processes (MDPs). A Markov Decision Process is a controlled dynamical system with probabilistic transitions, that are influenced by the control actions (for precise definitions see §§ 1.1). We consider discrete-time MDPs with a discrete *state space* X which is either *finite* or *countably infinite*, to which we will refer in the sequel as *countable*. The cost under consideration is the long-time average cost.

Countable MDPs are obviously more difficult to study, analytically and numerically, than finite state MDPs. Several approaches were developed to deal with this issue. The first approach is to reduce the state space by clustering together "equivalent" states: see e.g. [GDG] and references therein. This approach provides a smaller state space and exact relations, but requires a very special structure of the MDP in order for the derived model to have a finite state space. Namely, equivalent states must have the same transition probability into and out of the state, under any action, and the same immediate cost. This special structure seldom exists in applications.

A second approach approximates countable MDPs by finite state MDPs using a truncation of the state space. Existing results show that as the size of the approximating MDP increases, its cost function and, under some conditions its optimal policies approach those of the original, countable MDP. See, e.g. [C, A1, A2] and references therein. This approach is applicable in greater generality, but typically provides approximations without an error estimate—thus the results are "asymptotic" in nature.

We propose a different approach, with the advantage that the optimal cost of the approximating, finite MDP agrees with that of the countable MDP. Moreover, restricted to the approximating set the optimal policies agree as well. Thus the term "exact approximations." The main idea is finite embedding. In [F] embedding techniques were used to obtain optimal policies within various classes. We apply the embedding approach for approximations with general, compact action spaces. Section 5 develops some applications, where in some cases exact closed-form expressions can be obtained.

We conclude this section with a precise statement of the problem. In §2 we introduce the main idea—the finite embedding, and prove its existence. In §3 we show that the embedding possesses the desired properties. We discuss some extensions in §4.

1.1 Problem formulation

Consider a process with state space $X \subset \{0, 1, 2, 3, ...\}$. When the system occupies state $i \in X$, then the controller can influence its behavior by choosing *an action* a from the compact *action set* A_i, $i \in X$, which is a subset of the *action space* A. Choosing an action $a \in A_i$ has a twofold effect:

(i) A *running cost* $c(i, a)$ is incurred,

(ii) The system transits from state i to j according to the *transition probability* $P(j|i, a)$.

Thus an MDP is defined in terms of a quadruplet

$$\mathcal{M} = (X, \{A_i\}, c(i, a), P(j|i, a)).$$

The state and action at time $k \geq 0$ are denoted x_k and a_k respectively, so that the system's behavior on the infinite time interval is described in terms of the stochastic process $\{(x_k, a_k)\}_{k=0}^{\infty}$.

Admissible policies. A policy π is a law which is used to choose the actions $a_k \in A_{x_k}$. It is admissible if its choice at time k depends only on the history $(x_0, a_0, ..., x_{k-1}, a_{k-1}, x_k)$ of the system up to time k. A policy can be either deterministic or randomized, so that the choice of a_k may be made according to a probability measure on A_{x_k}. A (possibly randomized) policy which depends only on x_k is called a "stationary Markov policy". Such polices generate a state process $\{x_k\}_{k=0}^{\infty}$ which is a Markov chain with stationary transition probabilities.

The cost criterion. An admissible policy π generates the stochastic processes $\{x_k\}_{k=0}^{\infty}$ and $\{a_k\}_{k=0}^{\infty}$, and the expected cost flow

$$C_N(i, \pi) = E_i^{\pi} \sum_{k=0}^{N-1} c(x_k, a_k), \; N \geq 1. \tag{1.1}$$

The expectation E_i^{π} in (1.1) is with respect to the probability measure P_i^{π} induced by π on the set of sequences $\{(x_k, a_k)\}_{k=0}^{\infty}$ with $x_0 = i$. We address the optimal control problem of minimizing the functional

$$\pi \mapsto J_i(\pi) = \liminf_{N \to \infty} \frac{1}{N} C_N(i, \pi), \; x_0 = i,$$

over all admissible policies, and call $J_i(\pi)$ *the expected long-run average cost.* The notation $g^\star(\mathcal{M})$ for the optimal cost makes explicit the model $\mathcal{M}$ under consideration. An *optimal policy* realizes the minimal long-run average cost, and an ϵ-optimal policy realizes it up to ϵ.

We need to exclude one case in which it is not possible to approximate a countable MDP by a finite one.

Definition 1.1 *Let σ be a stationary Markov policy of an MDP $\mathcal{M}$, generating the state-action process $\{(x_k, a_k)\}$. We say that σ* is a drifting policy *if for every finite set $F \subset X$ and every initial condition $i \in F$,*

$$P_i^\sigma(x_k \in F) \to 0 \text{ as } k \to \infty. \tag{1.2}$$

We say that an MDP $\mathcal{M}$ is a drifting MDP *if there exists a constant $\delta > 0$ such that any stationary Markov policy of $\mathcal{M}$, say σ, that satisfies*

$$J(\sigma) < g^\star(\mathcal{M}) + \delta \tag{1.3}$$

is drifting. An MDP is called non-drifting *if it is not a drifting MDP.*

Obviously it is not possible to approximate a drifting MDP by a finite MDP while preserving both cost structure, optimality and dynamics. Moreover, if the MDP is drifting then by Definition 1.1 for a certain $\delta > 0$, for every $0 < \nu < \delta$, every stationary Markov policy with average cost smaller than $g^\star + \nu$, eventually leaves every finite set F with probability 1. But this yields the existence of a Markov policy (not necessarily stationary) with average cost $g^\star$ (that is, an optimal policy) which eventually leaves every finite set with probability 1. Assuming that this does not happen is the non drifting condition. In this paper we will therefore consider only non-drifting MDPs. Borkar's coercive condition [B] requires that the immediate cost is higher than $g^\star$ for "far states"—as in (5.1)—and thus ensures a non-drifting condition.

2 The embedding

We wish to associate with the given MDP $\mathcal{M}$ a finite state MDP $\mathcal{M}_0$,

$$\mathcal{M}_0 = (X_0, A_0, Q_0(j|i, a), c_0(i, a))$$

with cost flow $C_N^0(i, \pi)$ defined as in (1.1), in such a manner that the two MDPs will share common optimality properties and, in some sense, will share transition probabilities. To this end we introduce the notion of embedding, whose exact definition is presented below. In §3 we prove that this definition implies the desired exact approximation property.

We will next define the embedding notion, which will be followed by a construction of an embedding $\mathcal{M}_0$. Denote by $\{(x_k, a_k)\}_{k=0}^{\infty}$ the generic state-action process of $\mathcal{M}$ and by $\{(\xi_k, \alpha_k)\}_{k=0}^{\infty}$ the generic state-action process of $\mathcal{M}_0$. Given a finite subset $Z \subset X$ define

$$\eta = \inf\{k \geq 0 : x_k \notin Z\} \qquad \tau = \inf\{k > \eta : x_k \in Z\} \ . \tag{2.1}$$

Thus if the initial state is in Z then η is the first exit time and τ is the first return time, while if the initial state is not in Z then $\eta = 0$ and again τ is the first return time. For a stopping time $\nu > 0$ define

$$C_\nu(i, \pi) = E_i^\pi \sum_{k=0}^{\nu-1} c(x_k, a_k)$$

and note that if $\nu = N$, a deterministic integer, this definition agrees with (1.1). (If, however, ν is a random variable then $C_\nu(i, \pi) \neq C_N(i, \pi) \mid_{N=\nu}$, the latter being a random variable.) Given a finite subset $Z_0 \subset X_0$, define η_0, τ_0 and C_ν^0 analogously.

Definition 2.1 *We say that $\mathcal{M}_0$ is embedded in $\mathcal{M}$ if there exist subsets $Z_0 \subset X_0$ and $Z \subset X$, and a one-to-one mapping $e : Z_0 \mapsto Z$ from Z_0 onto Z such that, for any stationary Markov policy σ of $\mathcal{M}$ under which Z contains at least one recurrent state, the following holds.*

There exists a stationary Markov policy σ_0 of $\mathcal{M}_0$, such that if the processes start with initial states $x_0 \in Z$ and $\xi_0 = e^{-1}(x_0) \in Z_0$ respectively, then x_τ and ξ_{τ_0} are identically distributed (under the probability measures $P_{x_0}^{\sigma}$ and $P_{\xi_0}^{\sigma_0}$ respectively) and, if the term on the right is finite,

$$C_\tau(x_0, \sigma) = C_{\tau_0}^0(\xi_0, \sigma_0).$$

Embedding means that if we restrict attention only to the states Z in X and to the corresponding states Z_0 in X_0, then the performance of any stationary Markov policy σ on $\mathcal{M}$ can be imitated by the performance of some stationary Markov policy σ_0 on $\mathcal{M}_0$. This imitation can be achieved when considering finite subsets F of the state space X on which σ generates a nontrivial dynamics, namely the states in F are not all transient under σ.

Remark 2.2 *The idea of embedding is a natural generalization of Kac's Theorem and the "chain on Z" idea [MT, Thm. 10.2.3, §10.4.2 and 10.6]. Kac's Theorem gives the value of the stationary probability of a Markov chain in terms of the "cycle times." Moreover, the stationary distribution of a chain can be obtained from that of the chain restricted to a subset Z, again in terms of excursion times outside Z. In our case we need to account also for the time and costs accrued during the excursion outside of Z.*

We now establish that for certain finite sets $Z \subset X$ there exists a finite state $\mathcal{M}_0 = (X_0, A_0, P_0, c_0)$ and an embedding $e(\cdot)$ of $\mathcal{M}_0$ in $\mathcal{M}$. This embedding will be useful and significant for stationary Markov policies σ which induce a nontrivial dynamics on Z. The embedding is such that if $e(\cdot)$ is defined on Z_0, then $Z = e(Z_0)$ and

$$\#(X_0) = 2\#(Z_0).$$

In §3 we will employ the embedding result to establish existence and characterize optimal policies for certain countable state MDPs. As we shall see there, we need the embedding to be such that, in addition to the cost flow, the expected return times agree, that is, $E\tau = E\tau_0$.

2.1 The existence of embedding

Theorem 2.3 *Let $\mathcal{M}$ be a countable state MDP and let $Z \subset X$ be a finite set. Then there exists a finite state MDP $\mathcal{M}_0$ with state space X_0 and an embedding $e : Z_0 \mapsto Z$ of $\mathcal{M}_0$ in $\mathcal{M}$ such that*

$$\#(X_0) = 2\#(Z_0).$$

The result has non-trivial content provided that $Z \subset X$ contains a recurrent state of some stationary Markov policy of $\mathcal{M}$.

Proof: We will define an MDP $\mathcal{M}_0$, and will then establish that it has the properties asserted in the theorem. Denote

$$Z = \{z_1, ..., z_n\},$$

let Z_0 be a finite set

$$Z_0 = \{s_1, ..., s_n\} \tag{2.2}$$

and let $e : Z_0 \mapsto Z$ be defined by

$$e(s_i) = z_i,\ \ i = 1, 2, ..., n. \tag{2.3}$$

With each state s_i in Z_0 we associate a state ω_i, and we then define

$$X_0 = \{s_1, ..., s_n\} \cup \{\omega_1, ..., \omega_n\}. \tag{2.4}$$

We have to specify $\mathcal{M}_0$ as a quadruplet

$$(X_0, (A_0)_s, c_0(s, a), P_0(s'|s, a))$$

and define explicitly A_0, c_0 and P_0. We define $A_0(s_i)$ $(= (A_0)_{s_i})$ by

$$A_0(s_i) = A(e(s_i)) \text{ for every } 1 \leq i \leq n. \tag{2.5}$$

The definition of the action sets $A_0(\omega_i)$ will be given below.

We next define the transition probabilities $P_0(s|s_i, a)$, $s \in X_0$, $a \in A_0(s_i)$. First, for $s_i \in Z_0$

$$P_0(s|s_i, a) = \begin{cases} P(e(s_j)|e(s_i), a) & \text{if } s = s_j \in Z_0 \\ 1 - \sum_{j=1}^{n} P_0(s_j|s_i, a) & \text{if } s = \omega_i \\ 0 & \text{if } s = w_j,\ j \neq i. \end{cases} \tag{2.6}$$

The corresponding cost is defined by

$$c_0(s_i, a) = c(e(s_i), a) \tag{2.7}$$

for $a \in A_0(s_i) = A(e(s_i))$.

We now fix $1 \leq i \leq n$ and define the action sets and the transition probabilities for the state $\omega_i \in X_0 \setminus Z_0$. Let σ be a stationary Markov policy of $\mathcal{M}$ such that Z contains at least one recurrent state. Recall the definitions (2.1) and let

$$q_j(\sigma) = P^{\sigma}\{x_\tau = z_j \mid x_{\eta-1} = z_i\}. \tag{2.8}$$

This is the probability that the process $\{x_k\}$ will first enter Z through state z_j, conditioned on having left Z from z_i while employing the action $a \in A(z_i)$ specified by σ. *The action for state* ω_i induced by σ is the collection of $n+1$ nonnegative numbers

$$\alpha(\sigma) = (\lambda q_1(\sigma), \ldots, \lambda q_n(\sigma), c(\sigma)) \tag{2.9}$$

where the constant λ satisfies $0 < \lambda \leq 1$. These two parameters λ and $c(\sigma)$ which define the action $\alpha(\sigma)$ will be specified below. The quantities $q_1(\sigma), ..., q_n(\sigma)$ are the probabilities associated in (2.8) with the fixed state z_i and the stationary Markov policy σ. We define the action set of ω_i to be

$$A_0(\omega_i) = \bigcup_{\sigma} \alpha(\sigma), \tag{2.10}$$

where the union is over all the stationary Markov policies under which Z contains a recurrent state. Of course, two different policies σ_1 and σ_2 may give rise to the same action, that is $\alpha(\sigma_1) = \alpha(\sigma_2)$.

We now define the transition probabilities from ω_i. If $\alpha \in A_0(\omega)$, then

$$\alpha = (\alpha_1, \ldots, \alpha_n, \alpha_{n+1}) = (\lambda q_1, \ldots, \lambda q_n, c) \tag{2.11}$$

for some constant $0 < \lambda \leq 1$, and it follows that $\sum_{j=1}^{n} \alpha_j = \lambda$. We then define $P_0(s|\omega_i, \alpha)$ by

$$P_0(s|\omega_i, \alpha) = \begin{cases} \alpha_j & \text{if } s = s_j \in Z_0 \\ 1 - \lambda & \text{if } s = \omega_i \\ 0 & \text{if } s \notin Z_0 \cup \{\omega_i\}. \end{cases} \tag{2.12}$$

Thus for every choice of $0 < \lambda \leq 1$, the conditional probability to enter Z_0 through s_j, given that the process did enter Z_0, is q_j, independent of λ. However, the value of λ determines the expected time that would elapse until entrance, and we choose λ in such manner that this expected time turns out to be equal to the corresponding time for $\mathcal{M}$. Namely, if σ_0 is the policy that uses $\alpha(\sigma)$ in state ω_i, then λ is chosen such that

$$1 + E^{\sigma_0}_{\omega_i} \tau_0 = E^{\sigma}_{z_i}[\tau \mid \eta = 1] \ . \tag{2.13}$$

Remark 2.4 *If z_i is not accessible from the state in Z which is recurrent under σ then we do not need to specify actions for ω_i, while if it is accessible then necessarily it is recurrent, so that the expectation on the right-hand side of (2.13) is finite. An appropriate value of λ can clearly be chosen since for $\lambda = 1$ the left-hand side equals 1, while as $\lambda \to 0$ this expression diverges.*

We note that if it is possible to leave Z from z_i in one step then

$$E^{\sigma}_{z_i}[\tau \mid \eta = 1] = 1 + \left[\sum_{x_j \notin Z} P(x_j|z_i, \sigma(z_i)) \right]^{-1} \sum_{x_j \notin Z} P(x_j|z_i, \sigma(z_i)) E^{\sigma}_{x_j} \tau.$$

If, however, this is not possible, then we can define the action in $A_0(\omega_i)$ that corresponds to σ in an arbitrary manner. Finally note that the normalizing constant (in square bracket) above is just $P_0(\omega_i|s_i, \alpha(\sigma))$.

We next consider the cost associated with the action α, namely $c_0(\omega_i, \alpha)$, which is chosen to be such that

$$C_\tau(z_i, \sigma)) = C^0_{\tau_0}(e^{-1}(z_i), \sigma_0)) \tag{2.14}$$

holds for all $z_i \in Z$. We distinguish between two cases: If under σ the process $\{x_k\}$ starting at z_i does not leave Z, then equality holds in (2.14) by definition. If the process does leave Z, than

$$C_\tau(z_i, \sigma) = E^\sigma_{z_i} \sum_{k=0}^{\tau-1} c(x_k, a_k)$$

can be expressed in the form

$$c(z_i, \sigma(z_i)) + \sum_{x_j \notin Z} P(x_j|z_i, \sigma(z_i)) C_\tau(x_j, \sigma) + \sum_{z_j \in Z} P(z_j|z_i, \sigma(z_i)) C_\tau(z_j, \sigma).$$

We recall that $c_0(\omega_i, \alpha)$ is actually the $n+1$ component of $\alpha(\sigma)$ in (2.9), denoted $c(\sigma)$. It follows that setting

$$c_0(\omega_i, \alpha) = [P(\omega_i|e^{-1}(z_i), \sigma_0)]^{-1} \sum_{x_j \notin Z} P(x_j|z_i, \sigma(z_i)) C_\tau(x_j, \sigma) [E^{\sigma_0}_{\omega_i} \tau_0]^{-1} \tag{2.15}$$

ensures the desired equality (2.14). Thus to each σ there corresponds an action $\alpha = \alpha(\sigma)$, and a cost

$$c_0(\omega_i, \alpha) = \alpha_{n+1} = c(\sigma)$$

which in view of the explicit expression (2.15), indeed depends on the state and action alone.

The definition of $\mathcal{M}_0$ is thus complete, and it follows from the definition that $\mathcal{M}_0$ indeed has the properties asserted in the theorem. □

Remark 2.5 *Note that the calculation of C_τ is very similar to (and is as complicated as) that of the relative value function in the optimality equation. More precisely, since we are dealing with a single policy, this is related to the solution of the Poisson equation: see e.g. [MS, Thm. 9.5].*

3 Existence of optimal policies

In order to use the embedding result it is needed that some optimal policy will have a recurrent state within some finite set. The following is a simple condition under which there exists a finite subset Z as required in the theorem presented in the previous section. Suppose that we have an estimate

$$g^\star(\mathcal{M}) < \gamma, \tag{3.1}$$

for some γ, and moreover, for some ordering of the states $\{x_j\}$ the following holds:

$$\liminf_{j\to\infty} \{\min\{c(x_j, a) : a \in A(x_j)\}\} > \gamma. \tag{3.2}$$

It clearly follows from (3.1) and (3.2) that if $J(\sigma) < \gamma$, then some finite set Z contains a recurrent state of σ. Such a set is, e.g.,

$$Z = \{x : \min\{c(x_j, a) : a \in A(x_j)\} < \gamma\}. \tag{3.3}$$

An estimate as in (3.1) does not require a computation of the optimal policy, but can be obtained by restricting to a special type of policies.

Fix some state z and denote $s = e^{-1}(z)$. Define $\nu = \inf\{k > 0 : x_k = z\}$, and let ν_0 be defined analogously. A sum such as the cycle cost $C_\nu(z, \sigma)$ below is well-defined if it is finite when c is replaced by its absolute value $|c|$.

Theorem 3.1 *Fix a stationary Markov policy σ such that $z \in Z$ is recurrent under σ, and let $\mathcal{M}_0$ be an embedding as above. Let σ_0 be the stationary Markov policy of $\mathcal{M}_0$ associated with σ. If*

$$C_\nu(z, \sigma) = E_z^\sigma \sum_{k=0}^{\nu-1} c(x_k, a_k)$$

is well defined then

$$\lim_{N\to\infty} \frac{1}{N} C_N(z, \sigma) = \lim_{N\to\infty} \frac{1}{N} C_N^0(e^{-1}(z), \sigma_0).$$

Remark 3.2 *Conditions under which the average cost does not depend on the initial state are standard, and therefore we shall not elaborate on this point.*

The proof applies without change when the condition that $C_\nu(z, \sigma)$ is well defined is replaced by the condition that the immediate costs $c(x, a)$ are all nonnegative.

Proof: We denote $s = e^{-1}(z)$, and by construction s is recurrent under σ_0. We note that s_j is accessible from s under σ_0 if and only if $z_j = e(s_j)$ is accessible from z under σ. Therefore we may assume that all states in Z (resp. Z_0) communicate, and ignore transient states. It is also convenient to ignore (or remove from Z) states from which z can be reached only by leaving Z.

Let the random times ν and ν_0 be as in the sentence that precedes Theorem 3.1. Under the recurrence assumption the limit of the average cost flow exists. It follows from Theorem 17.2.1 of [MT] that under σ the process pair $\{x_k, a_k\}$ possesses an invariant probability measure which we denote by π. Moreover, almost surely under P_z^σ we have

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} c(x_k, a_k) = \frac{E_z^\sigma \sum_{k=0}^{\nu-1} c(x_k, a_k)}{E_z^\sigma \nu} = E_\pi c(x, a),$$

and analogously for $\mathcal{M}_0$:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} c_0(\xi_k, \alpha_k) = \frac{E_s^{\sigma_0} \sum_{k=0}^{\nu_0-1} c_0(\xi_k, \alpha_k)}{E_s^{\sigma_0} \nu_0} = E_{\pi_0} c_0(\xi, \alpha).$$

In view of our assumptions concerning recurrence and existence of cycle costs, this implies that

$$\lim_{N\to\infty} \frac{1}{N} C_N(z, \sigma) = \frac{E_z^\sigma \sum_{k=0}^{\nu-1} c(x_k, a_k)}{E_z^\sigma \nu}$$

and similarly for C^0. It therefore suffices to establish that

$$\frac{E_z^\sigma \sum_{k=0}^{\nu-1} c(x_k, a_k)}{E_z^\sigma \nu} = \frac{E_s^{\sigma_0} \sum_{k=0}^{\nu_0-1} c_0(\xi_k, \alpha_k)}{E_s^{\sigma_0} \nu_0}. \tag{3.4}$$

We note that the numerators in (3.4) are the cycle costs corresponding to z and s, assumed to be well defined. We first deal with the numerator in the left-hand side of (3.4). Let I_A denote the indicator of the set A, that is

$$I_A(x) = \begin{cases} 1 & \text{if} \quad x \in A \\ 0 & \text{if} \quad x \notin A. \end{cases}$$

Recalling the definitions of η and τ we have

$$E_z^\sigma \sum_{k=0}^{\nu-1} c(x_k, a_k)$$

$$= E_z^\sigma \sum_{k=0}^{\min(\nu,\eta)-1} c(x_k, a_k) + E_z^\sigma I_{\{\eta<\nu\}} \left(\sum_{k=\eta}^{\tau-1} c(x_k, a_k) + \sum_{k=\tau}^{\nu-1} c(x_k, a_k) \right) .$$

By the construction of the embedding,

$$E_z^\sigma \sum_{k=0}^{\min(\nu,\eta)-1} c(x_k, a_k) = E_s^{\sigma_0} \sum_{k=0}^{\min(\nu_0,\eta_0)-1} c(\xi_k, \alpha_k)$$

since, while x_k is in Z, both transition probabilities and immediate costs agree. Also

$$E_z^\sigma I_{\{\eta<\nu\}} \sum_{k=\eta}^{\tau-1} c(x_k, a_k) = E_s^{\sigma_0} I_{\{\eta_0<\nu_0\}} \sum_{k=\eta_0}^{\tau_0-1} c(\xi_k, \alpha_k)$$

by the definition of the costs $c_0(\omega_i, \alpha)$. Finally, using the Markov property,

$$E_z^\sigma I_{\{\eta<\nu\}} \sum_{k=\tau}^{\nu-1} c(x_k, a_k) = \sum_{z_j \in Z} P_z^\sigma(\eta < \nu,\ x_\tau = z_j) C_\nu(z_j, \sigma) \ . \qquad (3.5)$$

Now write

$$P_z^\sigma(\eta < \nu,\ x_\tau = z_j) = \sum_{z_i \in Z} P_z^\sigma(\eta < \nu,\ x_\tau = z_j \mid x_{\eta-1} = z_i) P_z^\sigma(x_{\eta-1} = z_i) \ .$$

Recalling that ν is the return time to state z, we express the first probability on the right-hand side as

$$P_z^\sigma(\eta < \nu,\ x_\tau = z_j \mid x_{\eta-1} = z_i) = P_z^\sigma(x_t \neq z,\ 1 \leq t < \eta,\ x_\tau = z_j \mid x_{\eta-1} = z_i).$$

We now observe that the right hand side describes the conditional probability of two events: one before the conditioning, one after. Since this is a Markov process we have conditional independence and so

$$P_z^\sigma(\eta < \nu,\ x_\tau = z_j \mid x_{\eta-1} = z_i) = P_z^\sigma(\eta < \nu \mid x_{\eta-1} = z_i) \cdot P_z^\sigma(x_\tau = z_j \mid x_{\eta-1} = z_i).$$

It follows from (3.5) and the above computation that

$$E_z^\sigma I_{\{\eta<\nu\}} \sum_{k=\tau}^{\nu-1} c(x_k, a_k)$$

$$= \sum_{z_j, z_i \in Z} P_z^\sigma(\eta < \nu \mid x_{\eta-1} = z_i) \cdot P_z^\sigma(x_\tau = z_j \mid x_{\eta-1} = z_i) \cdot P_z^\sigma(x_{\eta-1} = z_i) \cdot C_\nu(z_j, \sigma). \tag{3.6}$$

Similarly to (3.5) we have the following expression for σ_0:

$$E_s^\sigma I_{\{\eta_0<\nu_0\}} \sum_{k=\tau_0}^{\nu_0-1} c(\xi_k, \alpha_k) = \sum_{s_j \in Z_0} P_s^{\sigma_0}(\xi_{\tau_0} = s_j) C_{\nu_0}^0(s_j, \sigma_0). \tag{3.7}$$

We now repeat the discussion that appears in the text between equations (3.5) and (3.6) for the embedded process. Since all the probabilities and conditional probabilities in (3.6) agree with the corresponding quantities of the embedded process, in view of (3.5) and (3.7) it remains to establish that

$$C_\nu(z_j, \sigma) = C_{\nu_0}^0(s_j, \sigma_0). \tag{3.8}$$

We proceed as before to consider two cases. We compute $C_\nu(z_j, \sigma)$ as follows:

$$C_\nu(z_j, \sigma) = E_{z_j}^\sigma \sum_{k=0}^{\nu-1} c(x_k, a_k)$$

$$= E_{z_j}^\sigma \sum_{k=0}^{\min(\nu,\eta)-1} c(x_k, a_k) + E_{z_j}^\sigma \left\{ I_{\{\eta<\nu\}} \sum_{k=\eta}^{\nu-1} c(x_k, a_k) \right\}. \tag{3.9}$$

The first term once again agrees with the embedded chain. Conditioning on the exit point we can write the second term as

$$E_{z_j}^\sigma \left\{ I_{\{\eta<\nu\}} \sum_{k=\eta}^{\nu-1} c(x_k, a_k) \right\} = \sum_{z_i \in Z} P_{z_j}^\sigma(\eta < \nu, x_{\eta-1} = z_i) C_\nu(z_i, \sigma), \tag{3.10}$$

and similarly

$$E_{s_j}^{\sigma_0} \left\{ I_{\{\eta_0<\nu_0\}} \sum_{k=\eta_0}^{\nu_0-1} c_0(\xi_k, \alpha_k) \right\} = \sum_{s_i \in Z_0} P_{s_j}^{\sigma_0}(\eta_0 < \nu_0, \xi_{\eta_0-1} = s_i) C_{\nu_0}^0(s_i, \sigma_0). \tag{3.11}$$

By construction

$$P^{\sigma}_{z_j}(\eta < \nu, x_{\eta-1} = z_i) = P^{\sigma_0}_{s_j}(\eta_0 < \nu_0, \xi_{\eta_0-1} = s_i),$$

and since z is accessible from z_j, $P^{\sigma}_{z_j}(\eta < \nu) < 1$. Iterating equations (3.9), (3.10) and (3.11) we see that the costs for both models agree, up to a last term that goes to zero geometrically fast with the number of iterations. It follows that the numerators of both sides in (3.4) agree, and the proof for the denominators is similar. Thus (3.4) is established, and the proof of the theorem is complete. □

Theorem 3.3 *Let $\mathcal{M}$ be a Markov Decision Process, and suppose that the state z is recurrent under a stationary Markov optimal policy σ, and that the cycle cost is finite. Then for any embedding such that $z \in Z$, the optimal cost of $\mathcal{M}_0$ agrees with that of $\mathcal{M}$. Moreover, $\mathcal{M}_0$ has an optimal policy σ_0 that agrees with σ on corresponding states of Z and Z_0.*

The theorem assumes explicitly that there exists a stationary Markov optimal policy. This holds for most applications: for conditions see for example [P] and references therein.

4 Extensions

First, note that the requirement that the cycle cost is finite holds whenever the immediate costs are bounded, since we assume recurrence. Moreover, as noted in Remark 3.2, this requirement is not needed when the immediate costs are all of the same sign (positive or negative).

The following result is immediate, but nonetheless useful.

Theorem 4.1 *Fix some i. Suppose the stationary policies σ and σ' are such that the associated actions $\alpha(\sigma)$ and $\alpha(\sigma')$ have costs starting at ω_i that satisfy*

$$\alpha_{n+1}(\sigma) = c(\sigma) > c(\sigma') = \alpha_{n+1}(\sigma'),$$

while

$$\alpha_i(\sigma) = \alpha_i(\sigma') \text{ for every } i = 1, \ldots, n.$$

Then the action $\alpha(\sigma)$ may be eliminated from $A_0(\omega_i)$.

This is quite clear from the definition, and in fact this follows from standard results of action elimination in MDPs [L].

Next note that, even if the excursion costs C_τ are difficult to calculate, any approximation of C_τ and of the mean excursion times leads to a non-exact, approximate embedding, in the sense that optimal costs are not equal anymore. However, it is easy to see that the approximation is continuous in the sense that as the approximations of C_τ and $E\tau$ improve, the costs (including optimal costs) of the embedded model approach those of the original MDP.

We now outline the extension to constrained MDPs, where a detailed description of the model may be found in [A2]. In addition to the usual four components of an MDP we define a collection of immediate cost functions $\{d_k(x,a),\ k=1,\ldots,K\}$. Define $J_i^k(\pi)$ in the same way that $J_i(\pi)$ is defined, but with d_k replacing the immediate cost c. The constrained optimization problem is to minimize the functional $J_i(\pi)$, subject to the constraints

$$J_i^k(\pi) \leq V_k$$

for some prescribed constants V_k, $1 \leq k \leq K$.

Standard approximations of constrained problems are more difficult to handle and establish than unconstrained approximations of optimization problem. The reason for this is that when we require the approximate model to satisfy the hard constraints

$$J_i^{0,k}(\pi) \leq V_k,\ k = 1, 2, ..., K,$$

then clearly we may lose continuity, in the sense that even if the original problem is feasible (that is, there exist policies satisfying the constraints), an approximation of the required type may not be feasible [A2]. However, using our exact approximation, this difficulty does not arise.

The embedding results hold for this model, with the following minor modification. Recall the definitions (2.11) and (2.15) of the action in $\mathcal{M}_0$. Define $d^k(\sigma)$ as in (2.15) and define $\alpha(\sigma)$ by

$$\alpha = (\alpha_1, \ldots, \alpha_n, \alpha_{n+1}, \alpha_{n+2}, \ldots, \alpha_{n+K+1}) = (\lambda q_1, \ldots, \lambda q_n, c, d^1, \ldots, d^K). \tag{4.1}$$

Then the same arguments show that for the embedded chain, all costs agree with those of the original model, so that we may approximate the countable chain by a finite chain.

Note, however, that this model is much less robust. Whereas small errors in the calculation of the cycle cost for the optimization problem may, in the worst case, lead to sub-optimality, in the constrained case such errors may lead to infeasibility, that is, violation of the constraints. Thus, if cycle costs can not be computed exactly, this approximation shares the infeasibility problem with other, more traditional approximation methods.

5 Examples

Example 5.1 Markov Decision Processes serve as a common model in the control of dams and reservoirs [LB]. It is standard to discretize both space and time, in order to arrive at a manageable model. So, let us model the inflow into the water reservoir by a Markov chain D_t and let L_t denote the water level at the reservoir. At each epoch (usually month), the decisions are to use Z_t of the water for electricity generation, and evacuate Y_t through spillways. The water level is then given by

$$L_{t+1} = L_t + D_t - (Y_t + Z_t).$$

Denote $x_t = (L_t, D_t)$ and $a_t = (Y_t, Z_t)$. The revenue (negative cost) in the model arises from the sale of electrical power. The amount of power produced is a linear function of the amount of water Z_t used for generation, but also depends non-linearly on the water level L_t (since higher water level entails higher energy per unit of water). Thus the cost $c = c(l, z)$ depends on the state and action taken. Since D_t is a Markov chain, we obtain an MDP with discrete state process x_t and discrete action process a_t (assuming that the discretization of D_t is compatible with that of Y_t and Z_t). Both control variables are positive and bounded due to practical considerations (limited capacity of the generators, and of the spill mechanisms), so that the number of control actions is finite. Denote the maximal allowed value of Y_t (resp. Z_t) by $\bar{y}$ (resp. $\bar{z}$).

Since water levels which are too high may pose danger, if water level and inflow rates are too high, the maximal value of the control variables must be used, that is $Y_t = \bar{y}$ and $Z_t = \bar{z}$. This may be formulated by defining a finite set $F \subset X$ so that if the state (l, d) is outside this set, the allowed action is only $(\bar{y}, \bar{z})$. It is therefor natural to use our embedding results so that only states in F need be considered. We note that since the state space is two dimensional, the number of states that we ignore (outside F) may be significant.

The case where D_t is i.i.d. is particularly simple, since in this case the state process is one-dimensional, and outside F the process behaves exactly like a random walk. Therefor, results on random walks may be used to calculate the entrance distribution and mean time. In fact, in the simple case that the cost satisfies $c(l, \bar{z}) = \bar{c}$, a constant, for water level outside of F, the cycle cost is just a constant multiple of the average return time. For an explicit calculation in a simple case, see the next example.

In water reservoir applications, the object for optimization is often a group of reservoirs: there could be dozens of reservoirs, all connected. In this case the "curse of dimensionality" makes it impossible to solve such models, and additional approximations are required. For a step in this direction see the multi-dimensional queueing problem below.

Example 5.2 In the simplest case where D_t are Bernoulli, and where the release $Y_t + Z_t$ can only take the values 0 or 1 (with some probability which we can choose), we arrive at a generic model of operations research. Consider the problem of controlling a single queue. New jobs arrive according to an i.i.d. sequence of Bernoulli random variables with mean λ, and join an infinite queue. The job at the head of the queue is served, and (independently) the probability of completion of service (representing the speed of service) is the control variable a. Assume that there exists some $I > 0$ and $\mu > \lambda > \delta$ so that

$$A(x) = [\delta, \mu] \text{ for } x < I \quad \text{and } A(x) = \{\mu\} \text{ for } x \geq I.$$

That is, service rate is controlled for small queue size, but maximal rate must be used if the queue is large. If we assume further that, for $x \geq I$ the immediate cost is sub-linear, that is, it satisfies $c(x, a) = c(x, \mu) \leq c \cdot x$ for some positive c, then all our assumptions hold: state 0 is recurrent under any policy and cycle costs are finite. The embedded model $\mathcal{M}_0$ can be computed as follows. Set $X_0 = \{0, 1, \ldots, I - 1\}$. Since the process moves at most one step per unit time, the only way to exit this set it through state I, so we can set $\omega = I$. To calculate c_0, we need

$$C_\tau(I, \sigma) = E_I \sum_{k=0}^{\tau-1} c(x_k, \mu)$$

which in fact does not depend on σ, since the only available action is μ. This expression is identical to a standard queue, without control, with arrival rates

λ and service rate μ. However, for the standard queue transition probabilities for $x \neq 0$ do not depend on the state x. So, denote by E^0 expectation for the standard queue, and we can calculate

$$C_\tau(I, \sigma) = E_0^0 \sum_{k=0}^{\tau-1} c(x_k - I, \mu) \ .$$

But this equals $E_0^0 \tau \cdot J$ where J is the average cost for a standard queue. This can be calculated in terms of the stationary distribution π of the standard queue, that is

$$C_\tau(I, \sigma) = [E_0^0 \tau]^{-1} \sum_{x=0}^{\infty} \pi_x c(x, \mu) \ .$$

Define $\rho = \lambda(1-\mu)/\mu(1-\lambda)$. Then it is easy to check that for the standard queue, $\pi_x = (1-\rho)\rho^x$. On the other hand, by Kac's Theorem [MT, Thm. 10.2.3], $E_0^0 \tau = [\pi_0]^{-1} = 1/(1-\rho)$ so that we have an explicit expression for $C_\tau(I, \sigma)$. Clearly, this can be calculated analytically for various costs functions c, and in particular for linear costs.

This example can be extended as follows. Suppose we do not assume that the action space is restricted for $x \geq I$. Instead assume that

$$c(x, a) < \min\{c(y, a) : a \in A_y, \ y \geq I\} \tag{5.1}$$

for all $x < I$ and all a. It follows that μ is the optimal action at $x \geq I$, and we can apply Theorem 4.1 so that the previous conclusions apply.

In general, since this is a skip-free, one-dimensional problem, our results allows an easy decoupling of the behavior for $x < I$ from that for $x \geq I$. The situation is more complicated if the skip-free assumption is violated, namely either batch arrivals or batch service or both are allowed. However, as is clear from the proof of Theorem 3.3, we can write an implicit expression for the cost using the cost flows until the first hitting time of $\{0, 1, \ldots, I-1\}$, to obtain an explicit expression for the embedded model.

Example 5.3 Consider now a multi-dimensional queueing problem. Jobs of type $1, \ldots, K$ arrive according to a K-dimensional process $B(t)$ of i.i.d. vectors. The kth coordinate represents arrivals of customers of type k. Customers join infinite queues, one queue for each type. A single server chooses at each point in time which queue to serve, and serves the job at the head of the line. If job of type k is served, then the service will succeed with probability μ_k, and then the job will leave the queue.

If we impose the condition that some queue must be served as long as not all queues are empty, then the empty state will be recurrent under mild conditions. For example, it is sufficient to assume that the total number of arrivals at any unit time interval is bounded, and that the condition

$$\sum_{k=1}^{K} \frac{EB_k}{\mu_k} < 1 \ .$$

Let $Q(t)$ be the vector of queue-sizes at time t. This is the state of our MDP, and the control is the choice of queue to serve. This is easily seen to be an MDP, once immediate costs $c(q,a)$ are chosen. It is then natural to choose

$$Z = \left\{ Q : \sum_{k=1}^{K} Q_k(t) \geq Q_0 \right\}$$

and approximate the infinite model with a finite one.

Suppose for some Q_0 we have that if

$$\sum_{k=1}^{K} Q_k(t) \geq Q_0$$

then

$$c(x,a) = \sum_{k=1}^{K} c_k Q_k$$

for some positive coefficients $\{c_k\}$. Then the system simplifies considerably, and the computation of the hitting distributions and costs, required for our approximation, become feasible [BMM, W].

Acknowledgements:
We would like to thank the reviewers for their insightful comments.
This research was supported in part by the fund for promotion of research at the Technion and the fund for promotion of sponsored research at the Technion.
Adam Shwartz holds the Julius M. and Bernice Naiman Chair in Engineering at the Technion.

References

[A1] E. Altman, Denumerable constrained Markov decision processes and finite approximations *MOR* **19** 161–191, 1994.

[A2] E. Altman, *Constrained Markov decision processes*, Chapman Hall/CRC, 1999.

[BMM] J.S. Baras, D.-J. Ma and A.M. Makowski, K competing queues with geometric service requirements and linear costs: the mu-c rule is always optimal, *Systems and Control Letters* **6** 173–180, 1985.

[B] V. Borkar, Topics in controlled Markov chains, *Pitman Lecture Notes in Math.* **240** Longman Scientific, 1991.

[C] R. Cavazos-Cadena, Finite state approximations for denumerable state discounted Markov decision processes, *J. Applied Math. and Optimization* **14** 27–47, 1986.

[F] E.A. Feinberg, Sufficient classes of strategies in discrete dynamic programming I: decomposition of random strategies and embedded models. *Theory Prob. Appl..* **31** 658–668, 1986.

[FS] E.A. Feinberg and A. Shwartz, Editors. *Handbook of Markov Decision Processes*, Kluwer, 2002.

[GDG] Givan, R., T. Dean and M. Greig, Equivalence notions and model minimization in Markov decision processes, *Artificial Intelligence* **147** 163–223, 2003.

[LB] B.F. Lamond, A. Boukhtouta, Water reservoir applications of Markov Decision Processes, pp. 537–558 in [FS], 2002.

[L] J.B. Lasserre, Detecting optimal and non-optimal actions in average cost Markov decision processes, *J. Appl. Prob.* 1995.

[MS] A.M. Makowski, A. Shwartz, The Poisson equation for countable Markov chains: probabilistic methods and interpretations, pp. 269–304, In [FS], 2002.

[MT] S.P. Meyn, R.L. Tweedie, *Markov chains and stochastic stability*, Springer Verlag, New York, 1993.

[P] M. Puterman *Markov Decision processes*, Elsevier, 1990.

[W] J. Walrand, *An introduction to queueing networks.* Prentice Hall, NJ 1988.