J. Appl. Probab. 1–24 (2024)
*doi:*10.1017/jpr.2024.25

MAXIMIZING THE PROBABILITY OF VISITING A SET INFINITELY OFTEN FOR A MARKOV DECISION PROCESS WITH BOREL STATE AND ACTION SPACES

FRANÇOIS DUFOUR,* *Bordeaux INP*
TOMÁS PRIETO-RUMEAU,** *UNED*

Abstract

We consider a Markov control model with Borel state space, metric compact action space, and transitions assumed to have a density function with respect to some probability measure satisfying some continuity conditions. We study the optimization problem of maximizing the probability of visiting some subset of the state space infinitely often, and we show that there exists an optimal stationary Markov policy for this problem. We endow the set of stationary Markov policies and the family of strategic probability measures with adequate topologies (namely, the narrow topology for Young measures and the ws^{∞}-topology, respectively) to obtain compactness and continuity properties, which allow us to obtain our main results.

Keywords: Markov decision process; visiting a set infinitely often; non-additive optimality criterion; Young measures; ws^{∞}-topology

2020 Mathematics Subject Classification: Primary 90C40
Secondary 60J10

1. Introduction

Markov decision processes (MDPs) are a family of controlled stochastic processes which are suitable for the modeling of several sequential decision-making problems under uncertainty. They arise in many applications, such as engineering, computer science, telecommunications, finance, etc. Their study started in the late 1950s and the early 1960s with the seminal works of Bellman, Blackwell, Howard, and Veinott, to name just a few authors; see, e.g., [8, 21]. Both the theoretical foundations and applications of MDPs continue to be areas of active research. From a theoretical point of view, there are several techniques which allow one to establish the existence of optimal controls, as well as to analyze the main features of the optimal value function. Roughly speaking, there exist two families of such techniques. The first is related to the dynamic programming principle established by Bellman, and is also known as the backward induction principle for the finite-horizon case. In this framework, the optimality equation states that the value function is the fixed point of an operator (the so-called Bellman operator) which incorporates the current cost plus an expected value related to future costs.

Received 11 January 2023; accepted 9 February 2024.
* Postal address: INRIA Team Astral, 200 avenue de la Vieille Tour, 33405 Talence cedex. Email address: francois.dufour@math.u-bordeaux.fr
** Postal address: Department of Statistics, Operations Research, and Numerical Calculus, Faculty of Science, UNED, calle Juan del Rosal 10, 28040 Madrid, Spain. Email address: tprieto@ccia.uned.es

This approach and its various extensions, such as the value iteration and policy iteration algorithms, are studied in detail in the references [6, 9, 16, 18–20, 26, 32] (this is a non-exhaustive list). The second family of methods is based on an equivalent formulation of MDPs in a linear programming context. The key idea is to transform the original dynamic control problem into an infinite-dimensional static optimization problem over a space of finite measures. In this setting, the underlying variables of the linear program are given by the occupation measures of the controlled process, satisfying the so-called characteristic equation, and the value of the linear program is the integral of the one-step cost function with respect to an occupation measure. This last technique is particularly well adapted to problems with constraints; see the references [10, 18, 24–26], among others, for a detailed presentation and analysis of this approach. Although different, these two approaches share a common characteristic: namely, they are well adapted to additive-type performance criteria, but they do not allow for the study of non-additive performance criteria.

It must be emphasized that the non-additive type of criterion has undergone significant development in recent years; see for example the risk-sensitive optimality criterion [7, 12, 34] and the so-called gambling houses [33]. The distinctive feature of these non-additive criteria is that the criterion to be optimized cannot be decomposed as the sum of individual contributions (typically, the rewards earned at each single transition), nor is it linearly related to such sums (as for the classical long-run average-reward optimality criterion). Non-additive criteria introduce some nonlinearity in the objective function via, e.g., a utility function. For instance, in risk-sensitive models, the criterion to be maximized is the expected utility of the total rewards over a finite time horizon or the total discounted rewards over an infinite time horizon. Nonlinearity is introduced precisely by the utility function, which often takes the form of a power function or a logarithmic function. The usual technique for dealing with such problems is to develop an ad hoc dynamic-programming-like approach by introducing an operator which is nonlinear in the reward function. In gambling house problems, the expectation of a long-run average reward is maximized, making this again a nonlinear problem. Conditions under which this problem corresponds, in fact, to an average-reward MDP are studied. In summary, non-additive problems are usually tackled by relating them to other standard additive problems or by using operators resembling the Bellman principle. For the problem we are interested in, which is also of a non-additive nature, Schäl [31] developed a ‘vanishing discount factor’ approach based on total-expected-reward MDPs (details are given in the next paragraph).

In this paper, we study another non-additive optimality criterion which aims at maximizing the probability of visiting a certain subset of the state space infinitely often. This criterion is also asymptotic; more precisely, the performance function cannot be approximated by finite-horizon criteria, since the distribution of the state process over a finite horizon does not allow one to characterize the whole history of the process (or at least its tails) and contains no relevant information about whether a subset of the state space has been reached infinitely often. It is important to emphasize that this optimality criterion has been very little studied in the literature. To the best of our knowledge, it was first studied by Schäl in [31], who showed the existence of an optimal deterministic stationary policy in the context of an MDP with finite state space. More recently, in [15], this existence result has been generalized to models with countable state space. The approach proposed by Schäl in [31] uses properties of continuity of matrix inversion and of fundamental matrices of finite-state Markov chains and is therefore restricted to the finite-state framework. The results obtained in [15] rely heavily on the fact that the state space is countable. It should be noted that in this particular context the space of deterministic stationary policies is compact with respect to pointwise convergence.

https://doi.org/10.1017/jpr.2024.25 Published online by Cambridge University Press

The objective of the present paper is to extend these results to a model with a general Borel state space and to show the existence of an optimal randomized stationary policy. This is a continuation of the results presented in [15], in the sense that we take up the idea of introducing an auxiliary control model as a means of analyzing the original control model. This auxiliary model is an MDP with total expected reward that, roughly speaking, can be studied by finding a suitable topology on the set of strategic probability measures to ensure that this set is compact and the expected reward functional is semicontinuous.

In relation to [15], an important difficulty in our paper is to show, in the context of a general state space, the compactness of the space of strategic measures induced by randomized stationary policies; this is unlike the case of a countable state space, for which the proof was straightforward. Our approach relies on the introduction of Young measures to model randomized stationary policies and thus endow this space with a topology. In this framework, the space of stationary strategic measures can be seen as the image of the space of Young measures under a certain map. The difficulty is to show the continuity of this map. Another delicate point is to provide a sufficient condition to ensure the continuity of the criterion for the auxiliary model that is written as a function on the space of strategic measures. In [15], a condition is formulated in terms of the entrance time to the set of recurrent states. This condition is based on the fact that the countable state space of a Markov chain can be split into two classes: recurrent and transient states. In the context of a general state space, this structural property is not that simple. To overcome this difficulty, we introduce a general condition imposing a drift condition by using standard Foster–Lyapunov criteria. Under these conditions, we are able to prove that the problem of maximizing the probability of visiting a set infinitely often has an optimal policy which is Markov stationary. It is worth stressing that the result on the compactness of strategic probability measures for stationary Markov policies (Theorem 2.1) is a new result and is interesting in itself, not just in the context of the specific control model in this paper.

The problem studied in this paper (that of maximizing the probability of visiting a subset of the state space infinitely often) can be applied in population dynamics problems in which the population is controlled so as to be driven to some target set or to avoid some set. Indeed, the symmetric problem of minimizing the probability of being absorbed by some subset of the state space can be interpreted as the problem of preventing the population from being trapped in some region of the state space. Research is currently in progress on a game-theoretic approach to this problem, in which two competing species try to drive each other out of some region of the state space by maximizing or minimizing the probabilities of remaining in that region from some time onward.

Notation. We now introduce some notation used throughout the paper. We will write $\mathbb{N} = \{0, 1, 2, \ldots\}$. On a measurable space $(\mathbf{\Omega}, \mathcal{F})$ we will write $\mathcal{P}(\mathbf{\Omega})$ for the set of probability measures, assuming that there is no risk of confusion regarding the σ-algebra on $\mathbf{\Omega}$. If $(\mathbf{\Omega}', \mathcal{F}')$ is another measurable space, a stochastic kernel on $\mathbf{\Omega}'$ given $\mathbf{\Omega}$ is a mapping $Q : \mathbf{\Omega} \times \mathcal{F}' \to [0, 1]$ such that $\omega \mapsto Q(B|\omega)$ is measurable on $(\mathbf{\Omega}, \mathcal{F})$ for every $B \in \mathcal{F}'$, and $B \mapsto Q(B|\omega)$ is in $\mathcal{P}(\mathbf{\Omega}')$ for every $\omega \in \mathbf{\Omega}$. We denote by $\mathcal{B}(\mathbf{\Omega})$ the family of bounded measurable functions $f : \mathbf{\Omega} \to \mathbb{R}$ (here, $\mathbb{R}$ is endowed with its Borel σ-algebra). If Q is a stochastic kernel on $\mathbf{\Omega}'$ given $\mathbf{\Omega}$ and f is a real-valued measurable function on $\mathbf{\Omega}'$, allowed to take infinite values, we will denote by Qf the $\mathcal{F}$-measurable function defined by the following (provided that the integral is well defined):

$$Qf(\omega) = \int_{\mathbf{\Omega}'} f(z)Q(dz|\omega) \quad \text{for } \omega \in \mathbf{\Omega}.$$

https://doi.org/10.1017/jpr.2024.25 Published online by Cambridge University Press

In general, for a product of measurable spaces we will always consider the product σ-algebra. The indicator function of a set $C \in \mathcal{F}$ is written $\mathbf{I}_C$.

Throughout this paper, any metric space $\mathbf{S}$ will be endowed with its Borel σ-algebra $\mathfrak{B}(\mathbf{S})$. Also, when considering the product of a finite family of metric spaces, we will consider the product topology (which makes the product again a metric space). By $\mathcal{C}(\mathbf{S})$ we will denote the family of bounded continuous functions $f : \mathbf{S} \to \mathbb{R}$. Let $(\mathbf{\Omega}, \mathcal{F})$ be a measurable space and $\mathbf{S}$ a metric space. We say that $f : \mathbf{\Omega} \times \mathbf{S} \to \mathbb{R}$ is a *Carathéodory function* if $f(\omega, \cdot)$ is continuous on $\mathbf{S}$ for every $\omega \in \mathbf{\Omega}$, and $f(\cdot, s)$ is an $\mathcal{F}$-measurable function for every $s \in \mathbf{S}$. The family of Carathéodory functions thus defined is denoted by $Car(\mathbf{\Omega} \times \mathbf{S})$. The family of Carathéodory functions which, in addition, are bounded is denoted by $Car_b(\mathbf{\Omega} \times \mathbf{S})$. When the metric space $\mathbf{S}$ is separable, any $f \in Car(\mathbf{\Omega} \times \mathbf{S})$ is a jointly measurable function on $(\mathbf{\Omega} \times \mathbf{S}, \mathcal{F} \otimes \mathfrak{B}(\mathbf{S}))$. The Dirac probability measure $\delta_s \in \mathcal{P}(\mathbf{S})$ concentrated at some $s \in \mathbf{S}$ is given by $B \mapsto \mathbf{I}_B(s)$ for $B \in \mathfrak{B}(\mathbf{S})$.

Given $\lambda \in \mathcal{P}(\mathbf{S})$, let $L^1(\mathbf{S}, \mathfrak{B}(\mathbf{S}), \lambda)$ be the family of real-valued measurable functions on $\mathbf{S}$ which are λ-integrable (as usual, functions which are equal λ-almost everywhere are identified). With the usual definition of the 1-norm, we have that $L^1(\mathbf{S}, \mathfrak{B}(\mathbf{S}), \lambda)$ is a Banach space and convergence in norm is

$$f_n \xrightarrow{L^1} f \quad \text{if and only if} \quad \int_{\mathbf{S}} |f_n - f| \mathrm{d}\lambda \to 0.$$

The family of λ-essentially bounded measurable functions is $L^\infty(\mathbf{S}, \mathfrak{B}(\mathbf{S}), \lambda)$, and the essential supremum norm in L^∞ is denoted by $\| \cdot \|$. The weak* topology on the space $L^\infty(\mathbf{S}, \mathfrak{B}(\mathbf{S}), \lambda)$, denoted by $\sigma(L^\infty(\mathbf{S}, \mathfrak{B}(\mathbf{S}), \lambda), L^1(\mathbf{S}, \mathfrak{B}(\mathbf{S}), \lambda))$, is the coarsest topology for which the mappings

$$f \mapsto \int_{\mathbf{S}} fh \mathrm{d}\lambda \quad \text{for } h \in L^1(\mathbf{S}, \mathfrak{B}(\mathbf{S}), \lambda)$$

are continuous. We will write $f_n \xrightarrow{*} f$ when

$$\int_{\mathbf{S}} f_n h \mathrm{d}\lambda \to \int_{\mathbf{S}} fh \mathrm{d}\lambda \quad \text{for every } h \in L^1(\mathbf{S}, \mathfrak{B}(\mathbf{S}), \lambda).$$

The rest of the paper is organized as follows. In Section 2, we define the original control model $\mathcal{M}$ and state some important preliminary results, the proof of one of these results being postponed to Appendix A, at the end of the paper. A family of auxiliary MDPs $\mathcal{M}_\beta$, parametrized by $0 < \beta < 1$, is introduced in Section 3. The main results of the paper are given in Section 4.

2. Definition of the control model $\mathcal{M}$

Elements of the control model $\mathcal{M}$.

The MDP under consideration consists of the following elements:

- The state space $\mathbf{X}$ and the action space $\mathbf{A}$ are both Borel spaces endowed with their respective Borel σ-algebras $\mathfrak{B}(\mathbf{X})$ and $\mathfrak{B}(\mathbf{A})$.
- For each $x \in \mathbf{X}$, the nonempty measurable set $\mathbf{A}(x) \subseteq \mathbf{A}$ stands for the set of available actions in state x. Let $\mathbf{K} = \{(x, a) \in \mathbf{X} \times \mathbf{A} : a \in \mathbf{A}(x)\}$ be the family of feasible state–action pairs. Assumption (A.1) below will ensure that $\mathbf{K} \in \mathfrak{B}(\mathbf{X}) \otimes \mathfrak{B}(\mathbf{A})$.

https://doi.org/10.1017/jpr.2024.25 Published online by Cambridge University Press

- The initial distribution of the system is the probability measure $\nu \in \mathcal{P}(\mathbf{X})$, and the transitions of the system are given by a stochastic kernel Q on $\mathbf{X}$ given $\mathbf{K}$.
- Finally, regarding the performance criterion to be optimized (which we will define later), a nontrivial measurable subset of the state space $G \subseteq \mathbf{X}$ is given.

Our main conditions on the control model $\mathcal{M}$ are stated next. In Assumption (A.2) we introduce a *reference probability measure* $\lambda \in \mathcal{P}(\mathbf{X})$.

Assumption A.

(A.1) The action set $\mathbf{A}$ is compact, and the correspondence from $\mathbf{X}$ to subsets of $\mathbf{A}$ defined by $x \mapsto \mathbf{A}(x)$ is weakly measurable with nonempty compact values (see Definition 18.1 in [1]).
(A.2) There exist a reference probability measure $\lambda \in \mathcal{P}(\mathbf{X})$ and a measurable function $q : \mathbf{K} \times \mathbf{X} \to \mathbb{R}^+$ such that for every $D \in \mathfrak{B}(\mathbf{X})$ and every $(x, a) \in \mathbf{K}$,

$$Q(D|x, a) = \int_D q(x, a, y)\lambda(dy). \tag{2.1}$$

Also, for every $x \in \mathbf{X}$ we have the following continuity condition:

$$\lim_{k\to\infty} \int_{\mathbf{X}} |q(x, b_k, y) - q(x, b, y)|\lambda(dy) = 0 \quad \text{whenever } b_k \to b \text{ in } \mathbf{A}(x).$$

(A.3) The initial distribution ν is absolutely continuous with respect to λ in Assumption (A.2).

Remark 2.1.

(a) Under Assumption (A.1), and as a consequence of Theorem 18.6 in [1], we have that $\mathbf{K}$ is indeed a measurable subset of $\mathbf{X} \times \mathbf{A}$.

(b) The continuity condition in Assumption (A.2) can equivalently be written as

$$q(x, b_k, \cdot) \xrightarrow{L^1} q(x, b, \cdot) \quad \text{for any } x \in \mathbf{X} \text{ when } b_k \to b \text{ in } \mathbf{A}(x).$$

(c) The condition in Assumption (A.3) is not restrictive at all. Indeed, if it were not true that $\nu \ll \lambda$, then we would consider the reference probability measure $\eta = \frac{1}{2}(\nu + \lambda)$ in Assumption (A.2), so that (2.1) would become

$$Q(D|x, a) = \int_D q(x, a, y)\frac{d\lambda}{d\eta}(y)\, \eta(dy)$$

and the continuity condition would be satisfied as well.

(d) The fact that $\mathbf{X}$ is a Borel space will be used to ensure that the ws^∞-topology on the set of strategic measures (details will be given below) coincides with the weak topology—see [3, 23]—and that the set of strategic probability measures is compact with this topology. Otherwise, no topological properties of $\mathbf{X}$ will be used.

Control policies. The space of admissible histories up to time n for the controlled process is denoted by $\mathbf{H}_n$ for $n \in \mathbb{N}$. It is defined recursively by $\mathbf{H}_0 = \mathbf{X}$ and $\mathbf{H}_n = \mathbf{K} \times \mathbf{H}_{n-1}$ for any

https://doi.org/10.1017/jpr.2024.25 Published online by Cambridge University Press

$n \geq 1$, and it is endowed with the corresponding product σ-algebra. A control policy π is a sequence $\{\pi_n\}_{n\geq 0}$ of stochastic kernels on $\mathbf{A}$ given $\mathbf{H}_n$, denoted by $\pi_n(da|h_n)$, such that for each $n \geq 0$ and $h_n = (x_0, a_0, \ldots, x_n) \in \mathbf{H}_n$ we have $\pi_n(\mathbf{A}(x_n)|h_n) = 1$. The set of all policies is denoted by $\mathbf{\Pi}$.

Dynamics of the control model. The canonical sample space of all possible state and actions is $\mathbf{\Omega} = (\mathbf{X} \times \mathbf{A})^\infty$, and it is endowed with the product σ-algebra $\mathcal{F}$. Thus, a generic element $\omega \in \mathbf{\Omega}$ consists of a sequence of the form $(x_0, a_0, x_1, a_1, \ldots)$, where $x_n \in \mathbf{X}$ and $a_n \in \mathbf{A}$ for any $n \in \mathbb{N}$. For $n \in \mathbb{N}$, the projection functions X_n and A_n from $\mathbf{\Omega}$ to $\mathbf{X}$ and $\mathbf{A}$ respectively are defined by $X_n(\omega) = x_n$ and $A_n(\omega) = a_n$ for $\omega = (x_0, a_0, \ldots, x_n, a_n, \ldots)$, and they are called the state and control variables at time n. The history process up to time n is denoted by $H_n = (X_0, A_0, \ldots, X_n)$.

For any policy $\pi \in \mathbf{\Pi}$, there exists a unique probability measure $\mathbb{P}_\nu^\pi$ on $(\mathbf{\Omega}, \mathcal{F})$ supported on $\mathbf{K}^\infty$ (that is, $\mathbb{P}_\nu^\pi(\mathbf{K}^\infty) = 1$) which satisfies the following conditions:

(i) We have $\mathbb{P}_\nu^\pi\{X_0 \in D\} = \nu(D)$ for any $D \in \mathfrak{B}(\mathbf{X})$.

(ii) For any $n \geq 0$ and $C \in \mathfrak{B}(\mathbf{A})$ we have $\mathbb{P}_\nu^\pi\{A_n \in C|H_n\} = \pi_n(C|H_n)$.

(iii) For any $n \geq 0$ and $D \in \mathfrak{B}(\mathbf{X})$ we have $\mathbb{P}_\nu^\pi\{X_{n+1} \in D|H_n, A_n\} = Q(D|X_n, A_n)$.

The probability measure $\mathbb{P}_\nu^\pi$ is usually referred to as a *strategic probability measure*. The corresponding expectation operator is $\mathbb{E}_\nu^\pi$. The set of all strategic probability measures $\{\mathbb{P}_\nu^\pi\}_{\pi\in\mathbf{\Pi}}$ is denoted by $\mathcal{S}_\nu \subseteq \mathcal{P}(\mathbf{\Omega})$.

We will also denote by $\mathbb{P}_{\nu,n}^\pi$ the distribution of the random variable H_n; more formally, for any $n \geq 0$ we define $\mathbb{P}_{\nu,n}^\pi$ as the pushforward probability measure on $(\mathbf{X} \times \mathbf{A})^n \times \mathbf{X}$ given by $\mathbb{P}_\nu^\pi \circ H_n^{-1}$. Consequently, we have $\mathbb{P}_{\nu,n}^\pi(\mathbf{H}_n) = 1$; note also that $\mathbb{P}_{\nu,0}^\pi = \nu$.

Optimality criterion. The performance criterion that we want to maximize is the probability of visiting the set G infinitely often. Let us denote by $N_G : \mathbf{\Omega} \to \mathbb{N} \cup \{\infty\}$ the integer-valued process that counts the total number of visits of the state process to the set G; that is,

$$N_G(\omega) = \sum_{n=0}^{\infty} \mathbf{I}_G(X_n(\omega)).$$

For notational convenience, if $I \subseteq \mathbb{N}$ is a set of decision epochs, we let $N_G(I) = \sum_{n\in I} \mathbf{I}_G(X_n)$, which stands for the number of visits to G at times in I.

For the initial distribution ν, the value function of the control model $\mathcal{M}$ is given by

$$V^*(\nu) = \sup_{\pi\in\mathbf{\Pi}} \mathbb{P}_\nu^\pi\{N_G = \infty\}. \tag{2.2}$$

A policy attaining the above supremum is said to be optimal. In the sequel, the model $\mathcal{M}$ introduced in this section will be referred to as the primary or original control problem.

Stationary Markov policies. Let $\mathbf{\Pi}_s$ be the set of stochastic kernels π on $\mathbf{A}$ given $\mathbf{X}$ such that $\pi(\mathbf{A}(x)|x) = 1$ for each $x \in \mathbf{X}$. We have that $\mathbf{\Pi}_s$ is nonempty as a consequence of our hypotheses. Indeed, from the Kuratowski–Ryll-Nardzewski selection theorem [1, Theorem 18.13] we derive the existence of a measurable selector for the correspondence $x \mapsto \mathbf{A}(x)$; that is, there exists a measurable $f : \mathbf{X} \to \mathbf{A}$ such that $f(x) \in \mathbf{A}(x)$ for every $x \in \mathbf{X}$. Then we have that $\pi(da|x) = \delta_{f(x)}(da)$ indeed belongs to $\mathbf{\Pi}_s$.

https://doi.org/10.1017/jpr.2024.25 Published online by Cambridge University Press

We say that $\{\pi_n\}_{n\in\mathbb{N}} \in \mathbf{\Pi}$ is a stationary Markov policy if there is some $\pi \in \mathbf{\Pi}_s$ such that

$$\pi_n(\,\cdot\,|x_0, a_0, b_0, \ldots, x_n) = \pi(\,\cdot\,|x_n) \quad \text{for all } n \geq 0 \text{ and } h_n = (x_0, a_0, \ldots, x_n) \in \mathbf{H}_n.$$

Without risk of confusion, we will identify the set of all stationary Markov policies with $\mathbf{\Pi}_s$ itself. The set of all strategic probability measures for the stationary Markov policies $\{\mathbb{P}_\nu^\pi\}_{\pi\in\mathbf{\Pi}_s}$ is denoted by $\boldsymbol{\mathcal{S}}_\nu^s$.

Lemma 2.1. *If $\pi, \pi' \in \mathbf{\Pi}_s$ coincide λ-almost everywhere (that is, $\pi(da|x) = \pi'(da|x)$ for all x in a set of λ-probability one), then $\mathbb{P}_\nu^\pi = \mathbb{P}_\nu^{\pi'}$.*

Proof. Suppose that $Y \in \mathfrak{B}(\mathbf{X})$ with $\lambda(Y) = 1$ is such that $\pi(da|x) = \pi'(da|x)$ for every $x \in Y$. We will prove by induction on $n \geq 0$ that $(X_0, A_0, \ldots, X_n, A_n)$ has the same distribution under $\mathbb{P}_\nu^\pi$ and under $\mathbb{P}_\nu^{\pi'}$. Letting $n = 0$, and given $D \in \mathfrak{B}(\mathbf{X})$ and $C \in \mathfrak{B}(\mathbf{A})$, we have

$$\begin{aligned}\mathbb{P}_\nu^\pi\{X_0 \in D, A_0 \in C\} &= \int_D \pi(C|x)\frac{d\nu}{d\lambda}(x)\,\lambda(dx) \;=\; \int_{D\cap Y} \pi(C|x)\frac{d\nu}{d\lambda}(x)\,\lambda(dx)\\ &= \int_D \pi'(C|x)\frac{d\nu}{d\lambda}(x)\,\lambda(dx) \;=\; \mathbb{P}_\nu^{\pi'}\{X_0 \in D, A_0 \in C\}.\end{aligned}$$

Assuming the result is true for $n \geq 0$, and again letting $D \in \mathfrak{B}(\mathbf{X})$ and $C \in \mathfrak{B}(\mathbf{A})$, we have

$$\mathbb{P}_\nu^\pi\{X_{n+1} \in D, A_{n+1} \in C \mid X_n = x, A_n = a\} = \int_D \pi(C|y)q(x, a, y)\lambda(dy),$$

which, by the same argument, is equal to $\mathbb{P}_\nu^{\pi'}\{X_{n+1} \in D, A_{n+1} \in C \mid X_n = x, A_n = a\}$. The conditional distributions and the distribution of (X_n, A_n) being the same under π and π', we conclude the result. □

Remark 2.2. It turns out that two stationary Markov policies under the conditions of Lemma 2.1, i.e., which coincide λ-almost surely, are indistinguishable, since they yield the same strategic probability measure; in addition, they have the same performance (recall (2.2)), since the criterion to be maximized depends on the strategic probability measures.

Young measures. Let $\boldsymbol{\mathcal{R}} \supseteq \mathbf{\Pi}_s$ be the set of stochastic kernels γ on $\mathbf{A}$ given $\mathbf{X}$ such that $\gamma(\mathbf{A}(x)|x) = 1$ for λ-almost every $x \in \mathbf{X}$. On $\boldsymbol{\mathcal{R}}$ we consider the equivalence relation $\approx$ defined as $\gamma \approx \gamma'$ when $\gamma(\,\cdot\,|x) = \gamma'(\,\cdot\,|x)$ for λ-almost all $x \in \mathbf{X}$. The corresponding quotient space $\boldsymbol{\mathcal{R}} / \approx$, denoted by $\boldsymbol{\mathcal{Y}}$, is the so-called set of *Young measures*. As a consequence of Lemma 2.1 and Remark 2.2, there is no loss of generality in identifying the set of stationary Markov policies $\mathbf{\Pi}_s$ with $\boldsymbol{\mathcal{Y}}$, since all the elements in the same $\approx$-equivalence class have the same strategic probability measure. Thus we will indiscriminately use the symbols $\mathbf{\Pi}_s$ and $\boldsymbol{\mathcal{Y}}$ to refer to both the stationary Markov policies and the set of Young measures. Hence, we will write

$$\boldsymbol{\mathcal{S}}_\nu^s = \{\mathbb{P}_\nu^\pi\}_{\pi\in\boldsymbol{\mathcal{Y}}} \subseteq \boldsymbol{\mathcal{S}}_\nu.$$

The set of Young measures $\boldsymbol{\mathcal{Y}}$ will be equipped with the narrow (stable) topology. This is the coarsest topology which makes the following mappings continuous:

$$\pi \mapsto \int_{\mathbf{X}}\int_{\mathbf{A}} f(x, a)\pi(da|x)\lambda(dx),$$

https://doi.org/10.1017/jpr.2024.25 Published online by Cambridge University Press

for any $f \in Car(\mathbf{X} \times \mathbf{A})$ such that for some Φ in $L^1(\mathbf{X}, \mathfrak{B}(\mathbf{X}), \lambda)$ we have $|f(x, a)| \leq \Phi(x)$ for every $(x, a) \in \mathbf{X} \times \mathbf{A}$; see [2, Theorem 2.2]. By [4, Lemma 1], the set $\boldsymbol{\mathcal{Y}}$ endowed with the narrow topology becomes a compact metric space. We state this result in the next lemma.

Lemma 2.2. *Under Assumption A, the sets $\boldsymbol{\mathcal{Y}}$ of Young measures and $\boldsymbol{\Pi}_s$ of stationary Markov policies are compact metric spaces when endowed with the narrow topology.*

The ws^∞-topology. Finally, we introduce the so-called ws^∞-topology on $\boldsymbol{\mathcal{S}}_\nu$. Let

$$\boldsymbol{\mathcal{U}}_0 = \boldsymbol{\mathcal{B}}(\mathbf{X}) \quad \text{and} \quad \boldsymbol{\mathcal{U}}_n = Car_b(\mathbf{X}^{n+1} \times \mathbf{A}^n) \quad \text{for } n \geq 1,$$

where the arguments of $g \in \boldsymbol{\mathcal{U}}_n$ will be sorted as $(x_0, a_0, \ldots, x_{n-1}, a_{n-1}, x_n)$. The ws^∞-topology on the set $\boldsymbol{\mathcal{S}}_\nu$—see [28]—is the smallest topology that makes the mappings

$$\mathbb{P} \mapsto \int_{\boldsymbol{\Omega}} g(H_n) \mathrm{d}\mathbb{P}$$

continuous for any $n \geq 0$ and $g \in \boldsymbol{\mathcal{U}}_n$. Therefore, a sequence $\{\mathbb{P}_k\}_{k \in \mathbb{N}} \subseteq \boldsymbol{\mathcal{S}}_\nu$ converges to $\mathbb{P} \in \boldsymbol{\mathcal{S}}_\nu$ for the ws^∞-topology, denoted by $\mathbb{P}_k \Rightarrow \mathbb{P}$, if and only if

$$\lim_{k \to \infty} \int_{\boldsymbol{\Omega}} g(H_n) \mathrm{d}\mathbb{P}_k = \int_{\boldsymbol{\Omega}} g(H_n) \mathrm{d}\mathbb{P} \tag{2.3}$$

for any $n \in \mathbb{N}$ and $g \in \boldsymbol{\mathcal{U}}_n$.

The proof of Lemmas 2.3 and Theorem 2.1 below are fairly involved; they are presented in Appendix A. Note that Lemma 2.3 simply states that the mapping that associates to a Young measure its corresponding strategic probability measure is continuous.

Lemma 2.3. *Suppose that Assumption A holds. If $\{\pi_k\}_{k \in \mathbb{N}}$ and π in $\boldsymbol{\mathcal{Y}}$ are such that $\pi_k \to \pi$, then $\mathbb{P}_\nu^{\pi_k} \Rightarrow \mathbb{P}_\nu^{\pi}$.*

Theorem 2.1. *Under Assumption A, the sets $\boldsymbol{\mathcal{S}}_\nu$ and $\boldsymbol{\mathcal{S}}_\nu^s$ are compact metric spaces when endowed with the ws^∞-topology.*

It is worth mentioning that compactness of $\boldsymbol{\mathcal{S}}_\nu$ is a known result (see [3] or [28]), whereas the new result that we establish here is compactness of $\boldsymbol{\mathcal{S}}_\nu^s$.

3. The control model $\mathcal{M}_\beta$

By its definition, the performance criterion introduced above is of asymptotic type, and it cannot be written in an additive form. Therefore, it appears difficult to study this type of control problem directly by using the traditional techniques, such as the dynamic programming or linear programming approach. To address this difficulty, we introduced in [15], in the particular context of a countable state space, an auxiliary model $\mathcal{M}_\beta$ as a tool for analyzing the original control problem.

Elements of the control model $\mathcal{M}_\beta$. Given $0 < \beta < 1$, the auxiliary control model $\mathcal{M}_\beta$, which is based on the primary control model $\mathcal{M}$, is a 'total-expected-reward' MDP defined as follows:

- The state and action spaces are augmented with isolated points,

$$\hat{\mathbf{X}} = \mathbf{X} \cup \{\Delta\} \quad \text{and} \quad \hat{\mathbf{A}} = \mathbf{A} \cup \{a_\Delta\},$$

so that $\hat{\mathbf{X}}$ and $\hat{\mathbf{A}}$ are again Borel spaces.

https://doi.org/10.1017/jpr.2024.25 Published online by Cambridge University Press

- The set of available actions when the state variable is $x \in \hat{\mathbf{X}}$ is the set $\hat{\mathbf{A}}(x) \subseteq \hat{\mathbf{A}}$, where

$$\hat{\mathbf{A}}(x) = \mathbf{A}(x) \quad \text{for } x \in \mathbf{X} \quad \text{and} \quad \hat{\mathbf{A}}(\Delta) = \{a_\Delta\}.$$

  Let $\hat{\mathbf{K}} \subseteq \hat{\mathbf{X}} \times \hat{\mathbf{A}}$ be the set of feasible state–action pairs; hence $\hat{\mathbf{K}} = \mathbf{K} \cup \{(\Delta, a_\Delta)\}$.

- The dynamics of this model is governed by the stochastic kernel Q_β defined by

$$\begin{aligned} Q_\beta(\Gamma|x, a) = \mathbf{I}_G(x)\big[\beta Q(\Gamma \cap \mathbf{X}|x, a) + (1-\beta)\delta_\Delta(\Gamma)\big] \\ + \mathbf{I}_{\mathbf{X}-G}(x) Q(\Gamma \cap \mathbf{X}|x, a) + \mathbf{I}_{\{\Delta\}}(x)\delta_\Delta(\Gamma) \end{aligned}$$

  for any $\Gamma \in \mathfrak{B}(\hat{\mathbf{X}})$ and $(x, a) \in \hat{\mathbf{K}}$.
  The initial probability distribution is the same as for $\mathcal{M}$, that is, $\nu \in \mathcal{P}(\mathbf{X})$. With a slight abuse of notation, we will also denote by ν the measure $\nu(\,\cdot \cap \mathbf{X})$ on $\hat{\mathbf{X}}$.

- The reward function is $\mathbf{I}_G(x)$ for $(x, a) \in \hat{\mathbf{K}}$, where $G \subseteq \mathbf{X}$ is the measurable set given in the definition of $\mathcal{M}$.

Hence, the model $\mathcal{M}_\beta$ is indeed a standard total-expected-reward MDP parametrized by $0 < \beta < 1$. The transitions generated by Q_β can be interpreted very simply as follows. If the state process is in $\mathbf{X} - G$, then the transition is made according to Q, the stochastic kernel of the original model. If the state process is in G, then either the transition is made according to the original model $\mathcal{M}$ (with probability β), or the state process moves to Δ (with probability $1 - \beta$). Finally, it is easy to see that Δ is an absorbing state under a_Δ, the unique action available at Δ. However, it should be observed that $\mathcal{M}_\beta$ is not necessarily an absorbing MDP in the terminology of [17, Section 2], because absorption in Δ may not occur in a finite expected time.

The definition of $\mathcal{M}_\beta$ is somewhat inspired by the well-known equivalent formulation of a discounted MDP, with discount factor β, as a total-reward MDP in which the transition probabilities are multiplied by the discount factor β, while $1 - \beta$ is the probability of killing the process at each transition; see, e.g., [17, p. 132]. In our case, however, the β factor is incorporated only when the process is in G.

Control policies. The construction and definition of the control policies for $\mathcal{M}_\beta$ is very similar to what was presented for the original model in Section 2. Therefore, we will just present the key elements and will skip some details. Letting $\hat{\mathbf{H}}_0 = \hat{\mathbf{X}}$, for $n \geq 1$ we have

$$\hat{\mathbf{H}}_n = \mathbf{H}_n \cup \bigcup_{j=0}^{n} \big(\mathbf{K}^j \times \{(\Delta, a_\Delta)\}^{n-j} \times \{\Delta\}\big),$$

which includes the histories of the original control model, plus those paths reaching Δ at time $0 \leq j \leq n$ and remaining in Δ thereafter. The set of all admissible histories is given by

$$\mathbf{K}^\infty \cup \bigcup_{n=0}^{\infty} \big(\mathbf{K}^n \times \{(\Delta, a_\Delta)\}^\infty\big). \tag{3.1}$$

The family of control policies for $\mathcal{M}_\beta$ is denoted by $\hat{\mathbf{\Pi}}$, and the family of stationary Markov policies is $\hat{\mathbf{\Pi}}_s$. These are defined similarly to those of $\mathcal{M}$, where we replace $\mathbf{X}$, $\mathbf{A}$, and $\mathbf{H}_n$ with $\hat{\mathbf{X}}$, $\hat{\mathbf{A}}$, and $\hat{\mathbf{H}}_n$, respectively.

https://doi.org/10.1017/jpr.2024.25 Published online by Cambridge University Press

Dynamics of the control model. Let $\hat{\Omega} = (\hat{\mathbf{X}} \times \hat{\mathbf{A}})^{\infty}$ be the state–action canonical sample space endowed with its product σ-algebra $\hat{\mathcal{F}}$. The state, action, and history process projections are denoted by $\hat{X}_n$, $\hat{A}_n$, and $\hat{H}_n$, and they are defined as in Section 2. The counting processes associated to the visits of the state process to the set G are $\hat{N}_G = \sum_{n=0}^{\infty} \mathbf{I}_G(\hat{X}_n)$ and $\hat{N}_G(I) = \sum_{n \in I} \mathbf{I}_G(\hat{X}_k)$ for $I \subseteq \mathbb{N}$.

Recall that we were considering a reference probability measure $\lambda \in \mathcal{P}(\mathbf{X})$ for the control model $\mathcal{M}$. For the auxiliary control model, we will use the reference probability measure $\hat{\lambda} \in \mathcal{P}(\hat{\mathbf{X}})$ defined by

$$\hat{\lambda}(\Gamma) = \frac{1}{2}\big(\lambda(\Gamma \cap \mathbf{X}) + \delta_{\Delta}(\Gamma)\big) \quad \text{for } \Gamma \in \mathfrak{B}(\hat{\mathbf{X}}).$$

Given $\hat{\pi} \in \hat{\mathbf{\Pi}}$ and the initial distribution ν, there exists a unique probability measure $\mathbb{P}_{\nu}^{\beta,\hat{\pi}}$ on $(\hat{\Omega}, \hat{\mathcal{F}})$ supported on the set (3.1) of histories that models the dynamics of $\mathcal{M}_{\beta}$. The corresponding expectation operator is denoted by $\mathbb{E}_{\nu}^{\beta,\hat{\pi}}$.

Optimality criterion. For the model $\mathcal{M}_{\beta}$, the objective is the maximization of the total-expected-reward criterion given by the expected number of visits to the set G: for any $\hat{\pi} \in \hat{\mathbf{\Pi}}$,

$$\mathbb{E}_{\nu}^{\beta,\hat{\pi}}\Big[\sum_{n=0}^{\infty} \mathbf{I}_G(\hat{X}_n)\Big] = \mathbb{E}_{\nu}^{\beta,\hat{\pi}}[\hat{N}_G],$$

and we let

$$V_{\beta}^{*}(\nu) = \sup_{\hat{\pi} \in \hat{\mathbf{\Pi}}} \mathbb{E}_{\nu}^{\beta,\hat{\pi}}[\hat{N}_G]$$

be the value function. A policy $\pi^* \in \hat{\mathbf{\Pi}}$ satisfying $\mathbb{E}_{\nu}^{\beta,\pi^*}[\hat{N}_G] = V_{\beta}^{*}(\nu)$ is said to be optimal.

Remark 3.1. Given $\pi = \{\pi_n\}_{n \in \mathbb{N}}$ in $\mathbf{\Pi}$, we can define $\pi^{\Delta} = \{\pi_n^{\Delta}\}_{n \in \mathbb{N}}$ in $\hat{\mathbf{\Pi}}$ by setting $\pi_n^{\Delta}(\,\cdot\,|h_n) = \pi_n(\,\cdot\,|h_n)$ whenever $h_n \in \mathbf{H}_n$ and $\pi_n^{\Delta}(\,\cdot\,|h_n) = \delta_{a_{\Delta}}(\,\cdot\,)$ for $h_n = (x_0, a_0, \ldots, \Delta) \in \hat{\mathbf{H}}_n - \mathbf{H}_n$. Conversely, given a control policy $\hat{\pi} = \{\hat{\pi}_n\}_{n \in \mathbb{N}}$ in $\hat{\mathbf{\Pi}}$, we can restrict it to the sample paths in $\mathbf{H}_n$ to define a control policy $\hat{\pi}^{|\mathbf{X}} = \{\hat{\pi}_n^{|\mathbf{X}}\}_{n \in \mathbb{N}}$ in $\mathbf{\Pi}$ given by $\hat{\pi}^{|\mathbf{X}}(\,\cdot\,|h_n) = \hat{\pi}_n(\,\cdot\,|h_n)$ when $h_n \in \mathbf{H}_n$.

Hence, there is a bijection between $\hat{\mathbf{\Pi}}$ and $\mathbf{\Pi}$. Note that this establishes a bijection between $\mathbf{\Pi}_s$ and $\hat{\mathbf{\Pi}}_s$ as well.

For a complete overview of techniques for solving total-expected-reward MDPs, we refer to [19, Chapter 9]. In that reference, the value iteration algorithm, the policy iteration algorithm, and the linear programming approach are discussed. Other related results on value iteration are described in Theorems 3.5 and 3.7 in [15], and the linear programming (or convex analytic) approach is also studied in [13, 14].

The relation between $\mathcal{M}$ and $\mathcal{M}_{\beta}$. Next we state a result that allows us to establish a correspondence between the performance functions of the two models $\mathcal{M}$ and $\mathcal{M}_{\beta}$. Roughly, this will imply that for any pair of control policies on correspondence (recall Remark 3.1), the performance criterion of the model $\mathcal{M}_{\beta}$ multiplied by $1 - \beta$ will converge, as the parameter β tends towards 1, to the performance criterion of the original model $\mathcal{M}$. This very important fact

https://doi.org/10.1017/jpr.2024.25 Published online by Cambridge University Press

establishes the link from $\mathcal{M}$ to a criterion of an additive nature, for which it will be possible to use standard methods of analysis. In our paper, we use the so-called direct approach, along the lines developed in [3, 5, 23, 28, 29]; roughly speaking, this consists in finding a suitable topology on the set of strategic probability measures to ensure that it is compact and that the reward functional is semicontinuous.

The next result is proved in [15] in the context of a countable state space. It can easily be generalized to the framework of our paper, namely, a general state space of Borel type. In order to avoid unnecessary repetition we will omit the proof here; we refer the reader to Section 3 of [15]. In the sequel we make the convention that $\beta^\infty = 0$ for $\beta \in (0, 1)$.

Proposition 3.1. *For every* $0 < \beta < 1$, $\pi \in \mathbf{\Pi}$, *and* $n \in \mathbb{N}$ *we have*

$$\frac{1}{1-\beta} \cdot \mathbb{E}_\nu^\pi[1-\beta^{N_G([0,n])}] = \mathbb{E}_\nu^{\beta,\pi^\Delta}[\hat{N}_G([0,n])] \quad \textit{and} \quad \frac{1}{1-\beta} \cdot \mathbb{E}_\nu^\pi[1-\beta^{N_G}] = \mathbb{E}_\nu^{\beta,\pi^\Delta}[\hat{N}_G].$$

Remark 3.2.

(a) Since $\hat{\mathbf{\Pi}} = \{\pi^\Delta : \pi \in \mathbf{\Pi}\}$, for every $0 < \beta < 1$, $\hat{\pi} \in \hat{\mathbf{\Pi}}$, and $n \in \mathbb{N}$ we have

$$\mathbb{E}_\nu^{\beta,\hat{\pi}}[\hat{N}_G([0,n])] = \frac{1}{1-\beta} \cdot \mathbb{E}_\nu^{\hat{\pi}|\mathbf{X}}[1-\beta^{N_G([0,n])}]$$

and

$$\mathbb{E}_\nu^{\beta,\hat{\pi}}[\hat{N}_G] = \frac{1}{1-\beta} \cdot \mathbb{E}_\nu^{\hat{\pi}|\mathbf{X}}[1-\beta^{N_G}].$$

(b) Since $1-\beta^{N_G}$ decreases to $\mathbf{I}_{\{N_G=\infty\}}$ as β increases to 1, by dominated convergence we have for any $\pi \in \mathbf{\Pi}$

$$\lim_{\beta\uparrow 1}(1-\beta)\mathbb{E}_\nu^{\beta,\pi^\Delta}[\hat{N}_G] = \lim_{\beta\uparrow 1}\mathbb{E}_\nu^\pi[1-\beta^{N_G}] = \mathbb{P}_\nu^\pi\{N_G=\infty\}.$$

So the performance criterion of $\mathcal{M}_\beta$ multiplied by $1-\beta$ converges as $\beta \uparrow 1$ to that of $\mathcal{M}$ for each single control policy.

Assumptions on $\mathcal{M}_\beta$. We define the time of entrance of the state process $\{\hat{X}_n\}_{n\in\mathbb{N}}$ into a subset $\mathcal{C} \in \mathfrak{B}(\hat{\mathbf{X}})$ of the state space as the random variable $\tau_\mathcal{C}$ defined on $\hat{\mathbf{\Omega}}$ by

$$\tau_\mathcal{C}(\omega) = \min\{n \in \mathbb{N} : \hat{X}_n(\omega) \in \mathcal{C}\},$$

with the usual convention that the minimum over the empty set is ∞.

For a stationary Markov policy $\pi \in \mathbf{\Pi}_s$, we define the stochastic kernel Q_π on $\mathbf{X}$ given $\mathbf{X}$ as

$$Q_\pi(\Gamma|x) = \int_\mathbf{X} Q(\Gamma|x,a)\pi(da|x) \quad \text{for } x \in \mathbf{X} \text{ and } \Gamma \in \mathfrak{B}(\mathbf{X}).$$

For $n \geq 1$, we denote by Q_π^n the nth composition of Q_π with itself, and we make the convention that $Q_\pi^0(\,\cdot\,|x) = \delta_x(\,\cdot\,)$. Similarly, for a stationary Markov policy $\hat{\pi} \in \hat{\mathbf{\Pi}}_s$ of the auxiliary control model, we define the stochastic kernel $Q_{\beta,\hat{\pi}}$ on $\hat{\mathbf{X}}$ given $\hat{\mathbf{X}}$ and $Q_{\beta,\hat{\pi}}^n$ for $n \geq 0$.

https://doi.org/10.1017/jpr.2024.25 Published online by Cambridge University Press

Assumption B. For each $0 < \beta < 1$ and $\hat{\pi} \in \hat{\mathbf{\Pi}}_s$, there exists a measurable set $\mathcal{C}^{\beta,\hat{\pi}} \subseteq \hat{\mathbf{X}}$ satisfying $\hat{\lambda}(G \cap \mathcal{C}^{\beta,\hat{\pi}}) = 0$ and $Q_{\beta,\hat{\pi}}(\mathcal{C}^{\beta,\hat{\pi}}|x) = 1$ for $\hat{\lambda}$-almost all $x \in \mathcal{C}^{\beta,\hat{\pi}}$. In addition, there exists a finite constant K_β such that for every $\hat{\pi} \in \hat{\mathbf{\Pi}}_s$,

$$\mathbb{E}_x^{\beta,\hat{\pi}}[\tau_{\mathcal{C}^{\beta,\hat{\pi}}}] \leq K_\beta \quad \text{for } \hat{\lambda}\text{-almost all } x \in \hat{\mathbf{X}} - \mathcal{C}^{\beta,\hat{\pi}}. \tag{3.2}$$

Observe that, under Assumption B, the control model $\mathcal{M}_\beta$ is not necessarily an absorbing MDP in the terminology of [17, Section 2], since the set $\mathcal{C}^{\beta,\hat{\pi}}$ depends on the control policy. The stability Assumption B can be informally and very loosely explained as follows. Given any fixed $0 < \beta < 1$, for each stationary Markov policy $\hat{\pi} \in \hat{\mathbf{\Pi}}_s$ there exists a set $\mathcal{C}^{\beta,\hat{\pi}}$ which is closed (not in the topological sense, but in the terminology of Markov chains) and disjoint from G, and which is reached in a bounded expected time when starting from outside it (formally, all these statements hold almost surely with respect to the reference measure $\hat{\lambda}$).

We now state some consequences of Assumption B.

Lemma 3.1. *Let Assumption B be satisfied. Then for any $x \in \hat{\mathbf{X}}$, $\hat{\pi} \in \hat{\mathbf{\Pi}}_s$, and $n \geq 1$ we have $Q^n_{\beta,\hat{\pi}}(\,\cdot\,|x) \ll \hat{\lambda}$. In particular, the distribution of $\hat{X}_n$ under $\mathbb{P}_\nu^{\beta,\hat{\pi}}$ is absolutely continuous with respect to $\hat{\lambda}$ for any $n \in \mathbb{N}$.*

Proof. Indeed, if $\Gamma \in \mathfrak{B}(\hat{\mathbf{X}})$ is such that $\hat{\lambda}(\Gamma) = 0$, then necessarily $\Gamma \subseteq \mathbf{X}$ with $\lambda(\Gamma) = 0$. By definition of the stochastic kernel Q_β, and recalling (2.1), it is then clear that $Q_{\beta,\hat{\pi}}(\Gamma|x) = 0$ for every $x \in \hat{\mathbf{X}}$. Now, supposing again that $\Gamma \in \mathfrak{B}(\hat{\mathbf{X}})$ is such that $\hat{\lambda}(\Gamma) = 0$, and letting $n \geq 1$ and $x \in \mathbf{X}$, we have

$$Q^{n+1}_{\beta,\hat{\pi}}(\Gamma|x) = \int_{\hat{\mathbf{X}}} Q_{\beta,\hat{\pi}}(\Gamma|y) Q^n_{\beta,\hat{\pi}}(dy|x),$$

which indeed equals 0 because $Q_{\beta,\hat{\pi}}(\Gamma|y) = 0$.

The second statement is a straightforward consequence of the above results except for the case $n = 0$. In that case, however, the distribution of $\hat{X}_0$ is $\nu \ll \lambda$. □

The fact that $\mathcal{C}^{\beta,\hat{\pi}}$ is closed $\hat{\lambda}$-almost surely under $\hat{\pi}$ is extended to further transitions.

Lemma 3.2. *Suppose that Assumption B holds. Given arbitrary $\hat{\pi} \in \hat{\mathbf{\Pi}}_s$ and $n \geq 1$, we have that $Q^n_{\beta,\hat{\pi}}(\mathcal{C}^{\beta,\hat{\pi}}|x) = 1$ for $\hat{\lambda}$-almost all $x \in \mathcal{C}^{\beta,\hat{\pi}}$.*

Proof. Let us prove the result by induction. By Assumption B, it holds for $n = 1$. Now assume that the claim holds for some $n \geq 1$. Let $C_n \subseteq \mathcal{C}^{\beta,\hat{\pi}}$, with $\hat{\lambda}(\mathcal{C}^{\beta,\hat{\pi}} - C_n) = 0$, be the set of $x \in \hat{\mathbf{X}}$ for which $Q_{\beta,\hat{\pi}}(\mathcal{C}^{\beta,\hat{\pi}}|x) = 1$ and $Q^n_{\beta,\hat{\pi}}(\mathcal{C}^{\beta,\hat{\pi}}|x) = 1$. If $x \in C_n$ we have

$$Q^{n+1}_{\beta,\hat{\pi}}(\mathcal{C}^{\beta,\hat{\pi}}|x) \geq \int_{C_n} Q^n_{\beta,\hat{\pi}}(\mathcal{C}^{\beta,\hat{\pi}}|y) Q_{\beta,\hat{\pi}}(dy|x) = Q_{\beta,\hat{\pi}}(C_n|x).$$

Since $Q_{\beta,\hat{\pi}}(\,\cdot\,|x) \ll \hat{\lambda}$ (recall Lemma 3.1), it follows that

$$1 = Q_{\beta,\hat{\pi}}(C_n|x) + Q_{\beta,\hat{\pi}}(\mathcal{C}^{\beta,\hat{\pi}} - C_n|x) = Q_{\beta,\hat{\pi}}(C_n|x),$$

giving the result. □

We conclude this section by proposing a sufficient condition for Assumption B. It imposes a drift towards a closed set with $\hat{\lambda}$-null intersection with G by using some standard Foster–Lyapunov criteria. Such conditions are well known in the field of Markov chains. They provide

https://doi.org/10.1017/jpr.2024.25 Published online by Cambridge University Press

easily verifiable conditions for testing stability (existence of invariant measure, recurrence, ergodicity, etc.).

Proposition 3.2. *Suppose that for each* $0 < \beta < 1$ *and* $\hat{\pi} \in \hat{\mathbf{\Pi}}_s$ *there exist a constant* $\gamma^{\beta,\hat{\pi}} \geq 0$ *and the following:*

- *a measurable function* $W^{\beta,\hat{\pi}} : \hat{\mathbf{X}} \to [0, +\infty]$ *satisfying*

$$Q_{\beta,\hat{\pi}} W^{\beta,\hat{\pi}}(x) \leq W^{\beta,\hat{\pi}}(x) + \gamma^{\beta,\hat{\pi}} \quad \text{for } \hat{\lambda} - \text{almost all } x \in \hat{\mathbf{X}} \tag{3.3}$$

  and such that the set $\mathcal{C}^{\beta,\hat{\pi}}$ *defined as* $\mathcal{C}^{\beta,\hat{\pi}} = \{x \in \hat{\mathbf{X}} : W^{\beta,\hat{\pi}}(x) < +\infty\}$ *satisfies* $\hat{\lambda}(\mathcal{C}^{\beta,\hat{\pi}}) > 0$ *and* $\hat{\lambda}(G \cap \mathcal{C}^{\beta,\hat{\pi}}) = 0$*;*

- *a measurable function* $V^{\beta,\hat{\pi}} : \hat{\mathbf{X}} \to [0, +\infty]$ *and a constant* $K_\beta < +\infty$ *satisfying*

$$V^{\beta,\hat{\pi}}(x) \leq K_\beta \quad \text{for } \hat{\lambda} - \text{almost all } x \in \hat{\mathbf{X}} - \mathcal{C}^{\beta,\hat{\pi}} \tag{3.4}$$

  and

$$Q_{\beta,\hat{\pi}} V^{\beta,\hat{\pi}}(x) \leq V^{\beta,\hat{\pi}}(x) - 1 + \gamma^{\beta,\hat{\pi}} \mathbf{I}_{\mathcal{C}^{\beta,\hat{\pi}}}(x) \quad \text{for } \hat{\lambda} - \text{almost all } x \in \hat{\mathbf{X}}. \tag{3.5}$$

Under these conditions, Assumption B *is satisfied.*

Proof. Let $0 < \beta < 1$ and $\hat{\pi} \in \hat{\mathbf{\Pi}}_s$. The inequality (3.3) implies that

$$\int_{\hat{\mathbf{X}} - \mathcal{C}^{\beta,\hat{\pi}}} W^{\beta,\hat{\pi}}(y) Q_{\beta,\hat{\pi}}(dy|x) \leq W^{\beta,\hat{\pi}}(x) + \gamma^{\beta,\hat{\pi}} \quad \text{for } \hat{\lambda}\text{-almost all } x \in \hat{\mathbf{X}}.$$

But since $W^{\beta,\hat{\pi}}(y)$ is infinite when $y \notin \mathcal{C}^{\beta,\hat{\pi}}$ and $W^{\beta,\hat{\pi}}(x)$ is finite on $\mathcal{C}^{\beta,\hat{\pi}}$, this implies that $Q_{\beta,\hat{\pi}}(\hat{\mathbf{X}} - \mathcal{C}^{\beta,\hat{\pi}}|x) = 0$ and so $Q_{\beta,\hat{\pi}}(\mathcal{C}^{\beta,\hat{\pi}}|x) = 1$ for $\hat{\lambda}$-almost all $x \in \mathcal{C}^{\beta,\hat{\pi}}$. This establishes the first part of Assumption B. Now, following the proof of Lemma 11.3.6 in [22] and combining (2.1) and (3.5), we get $\mathbb{E}_x^{\beta,\hat{\pi}}[\tau_{\mathcal{C}^{\beta,\hat{\pi}}}] \leq V^{\beta,\hat{\pi}}(x)$ for $\hat{\lambda}$-almost all $x \in \hat{\mathbf{X}} - \mathcal{C}^{\beta,\hat{\pi}}$. The stated result follows from (3.4). □

We note that the sufficient conditions given in Proposition 3.2 above are, in general, more easily verifiable than Assumption B. Indeed, Proposition 3.2 involves just the *one-step transitions* of the system (see (3.3) and (3.5)). On the other hand, checking Assumption B directly requires one to compute the expected hitting time of the set $\mathcal{C}^{\beta,\hat{\pi}}$, which in principle requires knowledge of the distribution of the Markov chain over the *whole time horizon*, and then check that these expected hitting times are bounded uniformly in the initial state of the system.

It is worth mentioning that the condition (3.3) yields, loosely, the partition between recurrent and transient states of the Markov chain, while (3.5) is closely related to the computation of the expected absorption time by the set of recurrent states of the Markov chain.

Next, in Example 3.1, we illustrate the verification of the conditions in Proposition 3.2; see also Example 4.1 below.

Example 3.1. (Taken from [15, Example 4.6].) Consider the state space $\mathbf{X} = \{0, 1, 2\}$ with no actions at states 0 and 2, and $\mathbf{A}(1) = \{a_1, a_2\}$. The transition probabilities from state 0 are $q(0, 0) = q(0, 2) = 1/2$, and from state 2 they are $q(2, 2) = 1$ (we do not make any action explicit in this notation, since there are no actions at 0 and 2). The transitions from state 1 are $q(1, a_1, 0) = 2/3$ and $q(1, a_1, 1) = 1/3$ under a_1, and $q(1, a_2, 0) = q(1, a_2, 1) =$

https://doi.org/10.1017/jpr.2024.25 Published online by Cambridge University Press

$q(1, a_2, 2) = 1/3$ under a_2. We let $G = \{0\}$. We can identify $\hat{\mathbf{\Pi}}_s$ with the interval $[0, 1]$, where $a \in [0, 1]$ is the probability of selecting a_1 and $1 - a$ is the probability of choosing action a_2. Let 1 be the initial state of the system.

In order to verify Assumption B, we will use the sufficient conditions in Proposition 3.2. After adding the cemetery state Δ, the transition matrix for the stationary policy $a \in [0, 1]$ in the model $\mathcal{M}_\beta$ is

$$\begin{pmatrix} \beta/2 & 0 & \beta/2 & 1-\beta \\ (1+a)/3 & 1/3 & (1-a)/3 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

where the states are ordered 0, 1, 2, Δ.

Letting $W^{\beta,a} = (+\infty, +\infty, 1, 1)$ and $\gamma^{\beta,a} = 1$, we observe that (3.3) in Proposition 3.2 holds with, therefore, $\mathcal{C}^{\beta,a} = \{2, \Delta\}$ (roughly, W equals 1 on the recurrent states of the Markov chain, and it is infinite on the transient states). Regarding the inequalities (3.5), we obtain their minimal nonnegative solution, which is

$$V^{\beta,a}(0) = \frac{2}{2-\beta}, \qquad V^{\beta,a}(1) = \frac{(1+a)}{2-\beta} + \frac{3}{2}, \qquad V^{\beta,a}(2) = V^{\beta,a}(\Delta) = 0.$$

This minimal nonnegative solution is precisely the expected time the Markov chain spends in the transient states.

Solving the total-expected-reward MDP model $\mathcal{M}_\beta$ yields the optimal policy $\hat{\pi}^*_\beta = 1$ with $V^*_\beta(1) = \frac{2}{2-\beta} + \frac{3}{2}$ for every $0 < \beta < 1$.

4. Main results

In this section we prove our results on the solution of the control model $\mathcal{M}$.

Proposition 4.1. *Under Assumptions A and B, for each* $0 < \beta < 1$ *the mapping*

$$P \mapsto \int_{\mathbf{\Omega}} (1 - \beta^{N_G}) \mathrm{d}P$$

is lower semicontinuous on $\boldsymbol{\mathcal{S}}_\nu$ *and continuous on* $\boldsymbol{\mathcal{S}}^s_\nu$.

Proof. First we prove the following preliminary fact:

$$\lim_{n\to\infty} \sup_{\hat{\pi}\in\hat{\mathbf{\Pi}}_s} \mathbb{E}^{\beta,\hat{\pi}}_\nu[\hat{N}_G([n, \infty))] = 0 \tag{4.1}$$

for any $0 < \beta < 1$. Observe that for arbitrary $\hat{\pi} \in \hat{\mathbf{\Pi}}_s$ and $m \geq n \geq 0$,

$$\begin{aligned} \mathbb{P}^{\beta,\hat{\pi}}_\nu\{\tau_{\mathcal{C}^{\beta,\hat{\pi}}} = n, \hat{X}_m \in G\} &\leq \mathbb{P}^{\beta,\hat{\pi}}_\nu\{\hat{X}_n \in \mathcal{C}^{\beta,\hat{\pi}}, \hat{X}_m \in G\} \\ &= \mathbb{E}^{\beta,\hat{\pi}}_\nu[Q^{m-n}_{\beta,\hat{\pi}}(G|\hat{X}_n)\mathbf{I}_{\mathcal{C}^{\beta,\hat{\pi}}}(\hat{X}_n)]. \end{aligned}$$

Recalling that $\hat{\lambda}(G \cap \mathcal{C}^{\beta,\hat{\pi}}) = 0$, Lemmas 3.1 and 3.2 imply that the above expectation vanishes, and so $\mathbb{P}^{\beta,\hat{\pi}}_\nu\{\tau_{\mathcal{C}^{\beta,\hat{\pi}}} = n, \hat{X}_m \in G\} = 0$. Therefore, we have

$$\hat{N}_G([n, \infty)) = \mathbf{I}_{\{\tau_{\mathcal{C}^{\beta,\hat{\pi}}} > n\}} \sum_{m=n}^{\tau_{\mathcal{C}^{\beta,\hat{\pi}}}-1} \mathbf{I}_G(\hat{X}_m) \leq (\tau_{\mathcal{C}^{\beta,\hat{\pi}}} - n)^+ \quad \mathbb{P}^{\beta,\hat{\pi}}_\nu\text{-almost surely,}$$

https://doi.org/10.1017/jpr.2024.25 Published online by Cambridge University Press

and so

$$\begin{aligned}\mathbb{E}_\nu^{\beta,\hat\pi}[\hat N_G([n,\infty))] &\le \sum_{m>n}^{\infty}(m-n)\mathbb{P}_\nu^{\beta,\hat\pi}\{\tau_{\mathcal{C}^{\beta,\hat\pi}}=m\}\\ &= \sum_{m>n}^{\infty}(m-n)\mathbb{E}_\nu^{\beta,\hat\pi}\big[\mathbf{I}_{\hat{\mathbf{X}}-\mathcal{C}^{\beta,\hat\pi}}(\hat X_n)\,\mathbb{P}_{\hat X_n}^{\beta,\hat\pi}\{\tau_{\mathcal{C}^{\beta,\hat\pi}}=m-n\}\big]\end{aligned}$$

by the Markov property. Then it follows that

$$\begin{aligned}\mathbb{E}_\nu^{\beta,\hat\pi}[\hat N_G([n,\infty))] &\le \mathbb{E}_\nu^{\beta,\hat\pi}\big[\mathbf{I}_{\hat{\mathbf{X}}-\mathcal{C}^{\beta,\hat\pi}}(\hat X_n)\cdot\sum_{m>n}^{\infty}(m-n)\mathbb{P}_{\hat X_n}^{\beta,\hat\pi}\{\tau_{\mathcal{C}^{\beta,\hat\pi}}=m-n\}\big]\\ &= \mathbb{E}_\nu^{\beta,\hat\pi}\big[\mathbf{I}_{\hat{\mathbf{X}}-\mathcal{C}^{\beta,\hat\pi}}(\hat X_n)\cdot\mathbb{E}_{\hat X_n}^{\beta,\hat\pi}[\tau_{\mathcal{C}^{\beta,\hat\pi}}]\big]\\ &\le \mathbb{E}_\nu^{\beta,\hat\pi}\big[\mathbf{I}_{\hat{\mathbf{X}}-\mathcal{C}^{\beta,\hat\pi}}(\hat X_n)\cdot K_\beta\big],\end{aligned}$$

where the last inequality is obtained because $\mathbb{E}_x^{\beta,\hat\pi}[\tau_{\mathcal{C}^{\beta,\hat\pi}}]\le K_\beta$ for $\hat\lambda$-almost every $x\in\hat{\mathbf{X}}-\mathcal{C}^{\beta,\hat\pi}$—recall (3.2)—and because the distribution of $\hat X_n$ is absolutely continuous with respect to $\hat\lambda$—see Lemma 3.1. From the Markov inequality we derive that

$$\mathbb{E}_\nu^{\beta,\hat\pi}[\hat N_G([n,\infty))]\le K_\beta\cdot\mathbb{P}_\nu^{\beta,\hat\pi}\{\tau_{\mathcal{C}^{\beta,\hat\pi}}>n\}\le K_\beta^2/n;$$

note that this bound does not depend on $\hat\pi\in\hat{\mathbf{\Pi}}_s$. This shows (4.1).

We now proceed with the proof. For each fixed $n\ge 0$, the mapping on $(\mathbf{X}\times\mathbf{A})^{n-1}\times\mathbf{X}$ given by

$$(x_0,a_0,\dots,x_n)\mapsto\prod_{0\le j\le n}\beta^{\mathbf{I}_G(x_j)}=\beta^{N_G([0,n])}$$

is in $\boldsymbol{\mathcal{U}}_n$. Therefore, by the definition of the ws^∞-topology, the function

$$P\mapsto\int_{\mathbf{\Omega}}(1-\beta^{N_G([0,n])})\mathrm{d}P \tag{4.2}$$

is continuous on $\boldsymbol{\mathcal{S}}_\nu$, and so the function $P\mapsto\int_{\mathbf{\Omega}}(1-\beta^{N_G})\mathrm{d}P$ is lower semicontinuous on $\boldsymbol{\mathcal{S}}_\nu$, since it is the increasing limit as $n\to\infty$ of continuous functions.

Now we prove upper semicontinuity on $\boldsymbol{\mathcal{S}}_\nu^s$. Fix arbitrary $0\le n\le m$ and $P\in\boldsymbol{\mathcal{S}}_\nu^s$ such that $P=\mathbb{P}_\nu^\pi$ for some $\pi\in\mathbf{\Pi}_s$. Using Proposition 3.1,

$$\begin{aligned}\int_{\mathbf{\Omega}}(1-\beta^{N_G([0,m])})\mathrm{d}\mathbb{P}_\nu^\pi &= (1-\beta)\mathbb{E}_\nu^{\beta,\pi^\Delta}[\hat N_G([0,m])]\\ &\le (1-\beta)\mathbb{E}_\nu^{\beta,\pi^\Delta}[\hat N_G([0,n])]\\ &\quad+(1-\beta)\sup_{\hat\pi\in\hat{\mathbf{\Pi}}_s}\mathbb{E}_\nu^{\beta,\hat\pi}[\hat N_G([n+1,\infty))]\\ &= \int_{\mathbf{\Omega}}(1-\beta^{N_G([0,n])})\mathrm{d}\mathbb{P}_\nu^\pi+\epsilon_n,\end{aligned}$$

where we have defined

$$\epsilon_n=(1-\beta)\sup_{\hat\pi\in\hat{\mathbf{\Pi}}_s}\mathbb{E}_\nu^{\beta,\hat\pi}[\hat N_G([n+1,\infty))],$$

https://doi.org/10.1017/jpr.2024.25 Published online by Cambridge University Press

with $\epsilon_n \to 0$ as a consequence of (4.1). By the result in [27, Proposition 10.1], it follows that the limit as $n \to \infty$ of the function (4.2) is an upper semicontinous function on $\boldsymbol{\mathcal{S}}_\nu^s$. This completes the proof of the continuity statement on $\boldsymbol{\mathcal{S}}_\nu^s$. □

Proposition 4.2. *Suppose that Assumptions A and B are satisfied and fix $0 < \beta < 1$. There exists a stationary policy $\hat{\pi}^*$ in $\hat{\mathbf{\Pi}}_s$ that is optimal for the control model $\mathcal{M}_\beta$, i.e.,*

$$V_\beta^*(\nu) = \sup_{\hat{\pi} \in \hat{\mathbf{\Pi}}} \mathbb{E}_\nu^{\beta,\hat{\pi}}[\hat{N}_G] = \mathbb{E}_\nu^{\beta,\hat{\pi}^*}[\hat{N}_G].$$

Proof. Using the result in [30, p. 368], we have

$$\sup_{\hat{\pi} \in \hat{\mathbf{\Pi}}} \mathbb{E}_\nu^{\beta,\hat{\pi}}[\hat{N}_G] = \sup_{\hat{\pi} \in \hat{\mathbf{\Pi}}_s} \mathbb{E}_\nu^{\beta,\hat{\pi}}[\hat{N}_G].$$

Consequently, from Proposition 3.1 we obtain

$$\sup_{\pi \in \mathbf{\Pi}} \mathbb{E}_\nu^{\pi}[1-\beta^{N_G}] = \sup_{\pi \in \mathbf{\Pi}_s} \mathbb{E}_\nu^{\pi}[1-\beta^{N_G}] = \sup_{P \in \boldsymbol{\mathcal{S}}_\nu^s} \int (1-\beta^{N_G}) \mathrm{d}P.$$

But the latter supremum is attained at some $P^* \in \boldsymbol{\mathcal{S}}_\nu^s$, because it is the maximum of a continuous function (Proposition 4.1) on a compact set (Theorem 2.1). If $P^* = \mathbb{P}_\nu^{\pi_*}$ for some $\pi_* \in \mathbf{\Pi}_s$, then we have that $\pi_*^\Delta \in \hat{\mathbf{\Pi}}_s$ is an optimal stationary Markov policy for the control problem $\mathcal{M}_\beta$. □

We are now in position to prove the main result of this paper, on maximizing the probability of visiting the set G infinitely often.

Theorem 4.1. *Suppose that Assumptions A and B hold.*

(i) There is some $\pi \in \mathbf{\Pi}_s$ that is optimal for the control model $\mathcal{M}$, that is,

$$\mathbb{P}_\nu^{\pi}\{N_G = \infty\} = \sup_{\gamma \in \mathbf{\Pi}} \mathbb{P}_\nu^{\gamma}\{N_G = \infty\} = V^*(\nu).$$

(ii) Consider a sequence $\{\beta_k\}_{k\in\mathbb{N}}$ increasing to 1 and let $\hat{\pi}_k \in \hat{\mathbf{\Pi}}_s$ be an optimal policy for the model $\mathcal{M}_{\beta_k}$. If $\hat{\pi}_k^{|\mathbf{X}} \to \pi \in \mathbf{\Pi}_s$ in the narrow topology, then π is optimal for $\mathcal{M}$.

(iii) We have $\lim_{\beta\uparrow 1}(1-\beta)V_\beta^*(\nu) = V^*(\nu)$.

Proof. (i) Let $\{\beta_k\}_{k\in\mathbb{N}}$ be a sequence in (0, 1) with $\beta_k \uparrow 1$, and let the policy $\hat{\pi}_k \in \hat{\mathbf{\Pi}}_s$ be optimal for $\mathcal{M}_{\beta_k}$. Such a policy exists by Proposition 4.2. Consider now the sequence $\{\pi_k\}_{k\in\mathbb{N}}$ in $\mathbf{\Pi}_s$ where $\pi_k = \hat{\pi}_k^{|\mathbf{X}}$ for $k \in \mathbb{N}$. The set $\mathbf{\Pi}_s$ being compact (Lemma 2.2), we have that $\{\pi_k\}_{k\in\mathbb{N}}$ has a convergent subsequence, and we can assume without loss of generality that the whole sequence $\{\pi_k\}_{k\in\mathbb{N}}$ converges to some $\pi \in \mathbf{\Pi}_s$. In particular, by Lemma 2.3, we also have $\mathbb{P}_\nu^{\pi_k} \Rightarrow \mathbb{P}_\nu^{\pi}$.

By Proposition 3.1 it follows that for arbitrary $\gamma \in \mathbf{\Pi}$,

$$\begin{aligned}\mathbb{E}_\nu^{\pi_k}[1-\beta_k^{N_G}] &= (1-\beta_k)\mathbb{E}_\nu^{\beta_k,\hat{\pi}_k}[\hat{N}_G] \;=\; (1-\beta_k)V_{\beta_k}^*(\nu)\\ &\geq (1-\beta_k)\mathbb{E}_\nu^{\beta_k,\gamma^\Delta}[\hat{N}_G] \;=\; \mathbb{E}_\nu^{\gamma}[1-\beta_k^{N_G}].\end{aligned}$$

Now, by the first statement in Remark 3.2(b), observe that

$$\mathbb{E}_\nu^{\gamma}[1-\beta_k^{N_G}] \geq \mathbb{P}_\nu^{\gamma}\{N_G = \infty\},$$

https://doi.org/10.1017/jpr.2024.25 Published online by Cambridge University Press

and so for any $j \le k$ in $\mathbb{N}$ we obtain the inequalities

$$\mathbb{E}_\nu^{\pi_k}[1-\beta_j^{N_G}] \ge \mathbb{E}_\nu^{\pi_k}[1-\beta_k^{N_G}] = (1-\beta_k)V_{\beta_k}^*(\nu) \ge \mathbb{P}_\nu^\gamma\{N_G=\infty\}.$$

Taking the limit as $k\to\infty$ with j fixed in the previous inequalities, we get from Proposition 4.1 that

$$\mathbb{E}_\nu^{\pi}[1-\beta_j^{N_G}] \ge \limsup_{k\to\infty}(1-\beta_k)V_{\beta_k}^*(\nu) \ge \liminf_{k\to\infty}(1-\beta_k)V_{\beta_k}^*(\nu) \ge \mathbb{P}_\nu^\gamma\{N_G=\infty\}.$$

In these inequalities, we take the limit as $j\to\infty$ and the supremum in $\gamma\in\mathbf{\Pi}$ to obtain

$$\mathbb{P}_\nu^{\pi}\{N_G=\infty\} \ge \limsup_{k\to\infty}(1-\beta_k)V_{\beta_k}^*(\nu) \ge \liminf_{k\to\infty}(1-\beta_k)V_{\beta_k}^*(\nu) \ge \sup_{\gamma\in\mathbf{\Pi}}\mathbb{P}_\nu^\gamma\{N_G=\infty\}.$$

Consequently, the above inequalities are all, in fact, equalities:

$$\mathbb{P}_\nu^{\pi}\{N_G=\infty\} = V^*(\nu) = \lim_{k\to\infty}(1-\beta_k)V_{\beta_k}^*(\nu). \tag{4.3}$$

This establishes that the stationary Markov policy $\pi\in\mathbf{\Pi}_s$ is optimal for the control model $\mathcal{M}$ and that $(1-\beta_k)V_{\beta_k}^*(\nu)\to V^*(\nu)$.

(ii) We can obtain the proof of this statement from the proof of (i), just by assuming that the whole sequence $\{\pi_k\}$ converges (and not invoking a convergent subsequence).

(iii) Using Proposition 3.1 we obtain that

$$(1-\beta)V_\beta^*(\nu) = \sup_{\gamma\in\mathbf{\Pi}}\mathbb{E}_\nu^\gamma[1-\beta^{N_G}].$$

This implies that $(1-\beta)V_\beta^*(\nu)$ decreases as $\beta\uparrow 1$, and so $\lim_{\beta\uparrow 1}(1-\beta)V_\beta^*(\nu)$ exists. Since we have convergence through some sequence of $\beta_k\uparrow 1$ (recall (4.3)), the stated result follows. □

Summarizing, there indeed exists a stationary Markov policy in $\mathbf{\Pi}_s$ that maximizes the probability of visiting the set G infinitely often. This policy is obtained as an accumulation/limit point as $\beta\uparrow 1$ in the sense of convergence of Young measures of optimal stationary Markov policies for the total-expected-reward problem $\sup_{\hat\pi\in\hat{\mathbf{\Pi}}}\mathbb{E}_\nu^{\beta,\hat\pi}[\hat N_G]$ of the control model $\mathcal{M}_\beta$.

Example 4.1. Consider a control model with state space $\mathbf{X}=\{0, c, g, g_0, g_1, g_2, \ldots\}$. There are two actions available at state 0, i.e. $\mathbf{A}(0)=\{a_1, a_2\}$, and there are no actions at the other states. The transition probabilities are $q(0, a_1, g_0)=1$, $q(0, a_2, c)=q(0, a_2, g)=1/2$. From state g_i, the transition probabilities are

$$q(g_i, g_{i+1}) = \frac{1}{i+1} \quad\text{and}\quad q(g_i, c) = \frac{i}{i+1} \quad \text{for } i\ge 0,$$

while c and g are absorbing. Let $G=\{g, g_0, g_1, \ldots\}$ and let 0 be the initial state of the system. It should be clear that Assumption A holds.

Consider now the control model $\mathcal{M}_\beta$. We identify $\hat{\mathbf{\Pi}}_s$ with the interval $[0, 1]$, where $0\le a\le 1$ stands for the probability of choosing action a_1 at state 0. Let us verify the conditions in Proposition 3.2. Concerning (3.3), we put $W^{\beta,a}(c)=W^{\beta,a}(\Delta)=1$ and we let $W^{\beta,a}$ be $+\infty$ in the remaining states, thus obtaining $\mathcal{C}^{\beta,a}=\{c, \Delta\}$, with $\gamma^{\beta,a}=1$. This identifies the transient

https://doi.org/10.1017/jpr.2024.25 Published online by Cambridge University Press

and recurrent states of the Markov chain. For (3.5), we let $V^{\beta,a}(c) = V^{\beta,a}(\Delta) = 0$, and we try to find the minimal solution of (3.5). These inequalities become

$$\frac{1}{2}(1-a)V^{\beta,a}(g) + aV^{\beta,a}(g_0) \leq V^{\beta,a}(0) - 1, \quad \beta V^{\beta,a}(g) \leq V^{\beta,a}(g) - 1,$$

and

$$\frac{\beta}{i+1} V^{\beta,a}(g_{i+1}) \leq V^{\beta,a}(g_i) - 1 \quad \text{for } i \geq 0.$$

By iteration of the latter inequalities we obtain

$$\frac{\beta^{i+1}}{(i+1)!} V^{\beta,a}(g_{i+1}) \leq V^{\beta,a}(g_0) - \sum_{j=0}^{i} \frac{\beta^j}{j!}.$$

The minimal solution is attained when $V^{\beta,a}(0) = e^{\beta}$, and thus for any $i \geq 0$

$$V^{\beta,a}(g_i) = 1 + \frac{\beta}{k+1} + \frac{\beta^2}{(k+2)(k+1)} + \frac{\beta^3}{(k+3)(k+2)(k+1)} + \dots,$$

$V^{\beta,a}(g) = \frac{1}{1-\beta}$, and

$$V^{\beta,a}(0) = 1 + \frac{1-a}{2(1-\beta)} + ae^{\beta}.$$

We note that this minimal solution $V^{\beta,a}(x)$ is precisely the expected time the process spends in the transient states of the chain when starting from $x \in \mathbf{X}$. We also note that $V^{\beta,a}(x)$ is bounded in x (recall (3.4)): to see this, observe that $V^{\beta,a}(g_i) \leq e^{\beta}$ for every $i \geq 0$. It should be clear that the optimal value function of the control model $\mathcal{M}_\beta$ for the initial state 0 is

$$V^*_\beta(0) = \max\left\{\frac{1}{2(1-\beta)}, e^{\beta}\right\}$$

and that the optimal policy is $\hat{\pi}^*_\beta = 1$ for $0 < \beta \leq \hat{\beta}$ and $\hat{\pi}^*_\beta = 0$ for $\hat{\beta} \leq \beta < 1$, where $\hat{\beta} \simeq 0.76804$ is the unique solution in (0, 1) of $\frac{1}{2(1-z)} = e^z$. The interesting feature of this example is that the optimal policy π^*_β of $\mathcal{M}_\beta$ varies with β (cf. Example 3.1).

From Theorem 4.1(ii)–(iii) we obtain that

$$0 = \pi^* = \lim_{\beta \uparrow 1} \hat{\pi}^{*|\mathbf{X}}_\beta$$

is an optimal policy for $\mathcal{M}$ and that $V^*(1) = \lim_{\beta\uparrow 1}(1-\beta)V^*_\beta(1) = 1/2$.

Appendix A. Proof of Theorem 2.1

This section is mainly devoted to proving the compactness of $\boldsymbol{\mathcal{S}}^s_v$, the compactness of $\boldsymbol{\mathcal{S}}_v$ being a known result (see the proof of Theorem 2.1 below). In what follows, we suppose that Assumption A holds.

https://doi.org/10.1017/jpr.2024.25 Published online by Cambridge University Press

Lemma A.1. *Suppose that* $\mathbf{Z}$ *is a Borel space, and let* $f : \mathbf{Z} \times \mathbf{X} \times \mathbf{A} \times \mathbf{X} \to \mathbb{R}$ *be bounded, measurable, and continuous on* $\mathbf{A}$ *when the remaining variables* $z \in \mathbf{Z}$ *and* $x, y \in \mathbf{X}$ *are fixed. Define the real-valued functions* g *and* h *on* $\mathbf{X} \times \mathbf{K}$ *and* $\mathbf{Z} \times \mathbf{X}$ *respectively by*

$$g(z, x, a) = \int_{\mathbf{X}} f(z, x, a, y) q(x, a, y) \lambda(dy) \quad \text{and} \quad h(z, x) = \max_{a \in \mathbf{A}(x)} g(z, x, a).$$

Under these conditions, the following hold:

(i) The function g *is bounded and measurable on* $\mathbf{Z} \times \mathbf{K}$*, and* $g(z, x, \cdot)$ *is continuous on* $\mathbf{A}(x)$ *for fixed* $(z, x) \in \mathbf{Z} \times \mathbf{X}$*.*

(ii) The function h *is bounded and measurable on* $\mathbf{Z} \times \mathbf{X}$*.*

Proof. (i) It is clear that g is bounded and measurable. To establish the continuity property, we fix $(z, x) \in \mathbf{Z} \times \mathbf{X}$ and consider a converging sequence $a_n \to a$ in $\mathbf{A}$. By the dominated convergence theorem, we have $f(z, x, a_n, \cdot) \xrightarrow{*} f(z, x, a, \cdot)$ in the weak* topology $\sigma(L^\infty(\mathbf{X}, \mathfrak{B}(\mathbf{X}), \lambda), L^1(\mathbf{X}, \mathfrak{B}(\mathbf{X}), \lambda))$. We also have (recall Remark 2.1(b)) the convergence $q(x, a_n, \cdot) \xrightarrow{L^1} q(x, a, \cdot)$ in norm in $L^1(\mathbf{X}, \mathfrak{B}(\mathbf{X}), \lambda)$. Using the result in, e.g., [11, Proposition 3.13.iv], we conclude that

$$g(z, x, a_n) = \int_{\mathbf{X}} f(z, x, a_n, y) q(x, a_n, y) \lambda(dy) \to g(z, x, a).$$

This completes the proof of the statement (i).

(iii) By Assumption (A.1) and the maximum measurable theorem [1, Theorem 18.19], we conclude that h is measurable. □

Given $k \geq 1$, we say a function $f : (\mathbf{X} \times \mathbf{A})^k \times \mathbf{X} \to \mathbb{R}$ is in $\mathcal{R}_k$ if it is of the form

$$f(x_0, a_0, \ldots, x_{k-1}, a_{k-1}, x_k) = \bar{f}(x_0, \ldots, x_k) \bar{f}^{(0)}(a_0) \cdots \bar{f}^{(k-1)}(a_{k-1}),$$

where $\bar{f} \in \mathcal{B}(\mathbf{X}^{k+1})$ and $\bar{f}^{(i)} \in \mathcal{C}(\mathbf{A})$ for $0 \leq i < k$. For $k = 0$ we simply let $\mathcal{R}_0 = \mathcal{B}(\mathbf{X})$. Clearly, $\mathcal{R}_k \subseteq \mathcal{U}_k$. For any integer $k \geq 1$, define the operator $\mathfrak{A}_k : \mathcal{R}_k \to \mathcal{R}_{k-1}$ as follows. If $f \in \mathcal{R}_k$ then $\mathfrak{A}_k f$ is given by

$$\begin{aligned}(\mathfrak{A}_k f)(x_0, a_0, \ldots, x_{k-1}) &= \max_{a \in \mathbf{A}(x_{k-1})} \int_{\mathbf{X}} f(x_0, a_0, \ldots, x_{k-1}, a, y) q(x_{k-1}, a, y) \lambda(dy) \\ &= \prod_{j=0}^{k-2} \bar{f}^{(j)}(a_j) \max_{a \in \mathbf{A}(x_{k-1})} \int_{\mathbf{X}} \bar{f}(x_0, \ldots, x_{k-1}, y) \bar{f}^{(k-1)}(a) q(x_{k-1}, a, y) \lambda(dy) \qquad \text{(A.1)}\end{aligned}$$

for $(x_0, a_0, \ldots, x_{k-1}) \in (\mathbf{X} \times \mathbf{A})^{k-1} \times \mathbf{X}$, where by convention $\prod_{j=0}^{-1} \bar{f}^{(j)}(a_j) = 1$. Note that Lemma A.1(ii) ensures that $\mathfrak{A}_k$ indeed maps $\mathcal{R}_k$ into $\mathcal{R}_{k-1}$ because the max in the right-hand-side of (A.1) is a measurable function of the variables $(x_0, \ldots, x_{k-1}) \in \mathbf{X}^k$. In addition, we have that $\|\mathfrak{A}_k f\| \leq \|f\|$ in their respective norms in L^∞. The successive composition of these operators is defined as

$$\mathfrak{T}_k = \mathfrak{A}_1 \circ \ldots \circ \mathfrak{A}_{k-1} \circ \mathfrak{A}_k \quad \text{for } k \geq 1,$$

so that $\mathfrak{T}_k : \mathcal{R}_k \to \mathcal{R}_0$. By convention, $\mathfrak{T}_0$ will be the identity operator on $\mathcal{R}_0$.

https://doi.org/10.1017/jpr.2024.25 Published online by Cambridge University Press

Lemma A.2. *Given any $k \geq 0$, and arbitrary $f \in \mathcal{R}_k$ and $\pi \in \mathbf{\Pi}$, we have*

$$\int_{(\mathbf{X}\times\mathbf{A})^k\times\mathbf{X}} f(h_k)\mathbb{P}^{\pi}_{\nu,k}(dh_k) \leq \int_{\mathbf{X}} (\mathfrak{T}_k f)(x_0)\nu(dx_0).$$

Proof. The stated result is trivial for $k = 0$. If $k \geq 1$, observe that we can write

$$\begin{aligned}
&\int_{(\mathbf{X}\times\mathbf{A})^k\times\mathbf{X}} f(h_k)\mathbb{P}^{\pi}_{\nu,k}(h_k) = \int_{\mathbf{H}_k} f(h_k)\mathbb{P}^{\pi}_{\nu,k}(h_k)\\
&= \int_{\mathbf{H}_{k-1}}\int_{\mathbf{A}(x_{k-1})}\int_{\mathbf{X}} f(h_{k-1}, a, y)q(x_{k-1}, a, y)\lambda(dy)\pi_{k-1}(da|h_{k-1})\mathbb{P}^{\pi}_{\nu,k-1}(dh_{k-1})\\
&\leq \int_{\mathbf{H}_{k-1}} \mathfrak{A}_k f(h_{k-1})\mathbb{P}^{\pi}_{\nu,k-1}(dh_{k-1}).
\end{aligned}$$

By iterating this inequality and noting that $\mathbb{P}^{\pi}_{\nu,0} = \nu$, we obtain the desired result. □

Lemma A.3. *For $k \geq 1$, suppose that we are given a sequence $\{v_n\}_{n\in\mathbb{N}}$ in $\mathcal{R}_k$ which is of the form*

$$v_n(x_0, a_0, \ldots, x_k) = \bar{v}_n(x_0, \ldots, x_k)\cdot\bar{v}^{(0)}(a_0)\cdots\bar{v}^{(k-1)}(a_{k-1}),$$

where $\bar{v}_n \in \mathcal{B}(\mathbf{X}^{k+1})$ satisfies $\sup_n \|\bar{v}_n\| < \infty$ in L^∞ and

$$\bar{v}_n(x_0, \ldots, x_{k-1}, \cdot) \overset{*}{\rightharpoonup} 0 \quad \textit{for each } (x_0, \ldots, x_{k-1}) \in \mathbf{X}^k,$$

and the functions $\bar{v}^{(i)} \in \mathcal{C}(\mathbf{A})$ for $0 \leq i < k$ do not depend on n. Under these conditions, $g_n = \mathfrak{A}_k v_n \in \mathcal{R}_{k-1}$ can be written as

$$g_n(x_0, a_0, \ldots, x_{k-1}) = \bar{g}_n(x_0, \ldots, x_{k-1})\cdot\bar{v}^{(0)}(a_0)\cdots\bar{v}^{(k-2)}(a_{k-2}),$$

where $\{\bar{g}_n\}_{n\in\mathbb{N}}$ is a sequence of functions in $\mathcal{B}(\mathbf{X}^k)$ with $\sup_n \|\bar{g}_n\| < \infty$ which satisfy that $\bar{g}_n(x_0, \ldots, x_{k-1}) \to 0$ for any $(x_0, \ldots, x_{k-1}) \in \mathbf{X}^k$ as $n \to \infty$ and, in particular,

$$\bar{g}_n(x_0, \ldots, x_{k-2}, \cdot) \overset{*}{\rightharpoonup} 0 \quad \textit{for any } (x_0, \ldots, x_{k-2}) \in \mathbf{X}^{k-1}.$$

Proof. The expression given for g_n is easily deduced from (A.1). Moreover, the fact that $\{\bar{g}_n\}$ is bounded in the L^∞ norm is also straightforward. To prove the limit property, we proceed by contradiction: we suppose that for some $(x_0, \ldots, x_{k-1})$ and some subsequence $\{n'\}$ there exists $\epsilon > 0$ with

$$\left|\max_{a\in\mathbf{A}(x_{k-1})}\int_{\mathbf{X}} \bar{v}_{n'}(x_0, \ldots, x_{k-1}, y)\bar{v}^{(k-1)}(a)q(x_{k-1}, a, y)\lambda(dy)\right| \geq \epsilon.$$

Assuming that the above maximum is attained (recall Lemma A.1(ii)) at some $a^*_{n'} \in \mathbf{A}(x_{k-1})$, we may also suppose that $a^*_{n'} \to a^*$ for some $a^* \in \mathbf{A}(x_{k-1})$. Then we have the following convergences:

$$\bar{v}_{n'}(x_0, \ldots, x_{k-1}, \cdot) \overset{*}{\rightharpoonup} 0 \quad \text{and} \quad \bar{v}^{(k-1)}(a^*_{n'})q(x_{k-1}, a^*_{n'}, \cdot) \xrightarrow{L^1} \bar{v}^{(k-1)}(a^*)q(x_{k-1}, a^*, \cdot).$$

https://doi.org/10.1017/jpr.2024.25 Published online by Cambridge University Press

To check the latter convergence, just note that

$$\begin{aligned}
&\int_{\mathbf{X}} \left|\bar{v}^{(k-1)}(a^*_{n'})q(x_{k-1}, a^*_{n'}, y) - \bar{v}^{(k-1)}(a^*)q(x_{k-1}, a^*, y)\right|\lambda(dy)\\
&\leq \left|\bar{v}^{(k-1)}(a^*_{n'}) - \bar{v}^{(k-1)}(a^*)\right| \int_{\mathbf{X}} q(x_{k-1}, a^*_{n'}, y)\lambda(dy)\\
&+ \left|\bar{v}^{(k-1)}(a^*)\right| \int_{\mathbf{X}} |q(x_{k-1}, a^*_{n'}, y) - q(x_{k-1}, a^*, y)|\lambda(dy),
\end{aligned}$$

which indeed converges to 0 as $n' \to \infty$. Applying [11, Proposition 3.13.iv], we obtain

$$\int_{\mathbf{X}} \bar{v}_{n'}(x_0, \ldots, x_{k-1}, y)\bar{v}^{(k-1)}(a^*_{n'})q(x_{k-1}, a^*_{n'}, y)\lambda(dy) \to 0,$$

which is a contradiction. The last statement concerning the weak* convergence follows easily from dominated convergence, since $\{\bar{v}_n\}$ is uniformly bounded in the L^∞ norm. This completes the proof. □

To summarize this lemma informally, note that if we have a sequence of the form

$$v_n = \bar{v}_n \cdot \bar{v}^{(0)} \cdots \bar{v}^{(k-1)} \in \boldsymbol{\mathcal{R}}_k$$

with $\{\bar{v}_n\}$ bounded and $v_n(x_0, \ldots, x_{k-1}, \cdot) \overset{*}{\rightharpoonup} 0$, then $g_n = \mathfrak{A}_k v_n$ again satisfies the same properties, since

$$g_n = \bar{g}_n \cdot \bar{v}^{(0)} \cdots \bar{v}^{(k-2)} \in \boldsymbol{\mathcal{R}}_{k-1}$$

with $\{\bar{g}_n\}$ bounded and $g_n(x_0, \ldots, x_{k-2}, \cdot) \overset{*}{\rightharpoonup} 0$.

Proof of Lemma 2.3. Suppose that we have a sequence $\{\gamma_n\}_{n\in\mathbb{N}}$ in $\boldsymbol{\mathcal{Y}}$ converging to some $\gamma \in \boldsymbol{\mathcal{Y}}$. We will show that $\mathbb{P}^{\gamma_n}_{\nu} \Rightarrow \mathbb{P}^{\gamma}_{\nu}$. To prove this result, it suffices to show that for any $k \geq 0$ and $f \in \boldsymbol{\mathcal{U}}_k$, we have (recall (2.3))

$$\lim_{n\to\infty} \int_{\mathbf{H}_k} f(h_k)\mathbb{P}^{\gamma_n}_{\nu,k}(dh_k) = \int_{\mathbf{H}_k} f(h_k)\mathbb{P}^{\gamma}_{\nu,k}(dh_k). \tag{A.2}$$

Also, by [28, Theorem 3.7(ii)], it suffices to show (A.2) for functions f which are of the form

$$f(x_0, a_0, \ldots, x_{k-1}, a_{k-1}, x_k) = \bar{f}(x_0, \ldots, x_{k-1}, x_k)\hat{f}(a_0, \ldots, a_{k-1}) \tag{A.3}$$

for $\bar{f} \in \boldsymbol{\mathcal{B}}(\mathbf{X}^{k+1})$ and $\hat{f} \in \boldsymbol{\mathcal{C}}(\mathbf{A}^k)$. As a preliminary step, we will establish the limit (A.2) for functions in $\boldsymbol{\mathcal{R}}_k$, that is, those for which, in addition, the function $\hat{f}$ in (A.3) is the product of k functions in $\boldsymbol{\mathcal{C}}(\mathbf{A})$, denoted by $\bar{f}^{(i)}$ for $0 \leq i \leq k-1$.

We will prove the statement by induction. The limit in (A.2) trivially holds for $k = 0$ because $\mathbb{P}^{\pi}_{\nu,0} = \nu$ for any $\pi \in \boldsymbol{\Pi}$. Suppose that (A.2) is satisfied for some $k \geq 0$ and every function in $\boldsymbol{\mathcal{R}}_k$, and let us prove it for $k+1$ and any $f \in \boldsymbol{\mathcal{R}}_{k+1}$. We have

$$\begin{aligned}
&\int_{\mathbf{H}_{k+1}} f(h_{k+1})\mathbb{P}^{\gamma_n}_{\nu,k+1}(dh_{k+1})\\
&= \int_{\mathbf{H}_k} \Big[\int_{\mathbf{A}}\int_{\mathbf{X}} f(h_k, a, y)q(x_k, a, y)\lambda(dy)\gamma(da|x_k)\Big]\mathbb{P}^{\gamma_n}_{\nu,k}(dh_k)\\
&\quad + \int_{\mathbf{H}_k}\int_{\mathbf{A}}\int_{\mathbf{X}} f(h_k, a, y)q(x_k, a, y)\lambda(dy)\big[\gamma_n(da|x_k) - \gamma(da|x_k)\big]\mathbb{P}^{\gamma_n}_{\nu,k}(dh_k).
\end{aligned} \tag{A.4}$$

https://doi.org/10.1017/jpr.2024.25 Published online by Cambridge University Press

We study the limit of the first term in the right-hand-side of (A.4) as $n \to \infty$. Since $f \in \mathcal{R}_{k+1}$, we deduce from Lemma A.1(i) that the mapping defined on $(\mathbf{X} \times \mathbf{A})^k \times \mathbf{X}$ by

$$(x_0, a_0, \ldots, x_k) \mapsto \int_{\mathbf{A}} \int_{\mathbf{X}} f(h_k, a, y) q(x_k, a, y) \lambda(dy) \gamma(da|x_k)$$

is in $\mathcal{R}_k$. By the induction hypothesis, we can take the limit as $n \to \infty$ and so obtain

$$\begin{aligned}
&\lim_{n\to\infty} \int_{\mathbf{H}_k} \Big[\int_{\mathbf{A}} \int_{\mathbf{X}} f(h_k, a, y) q(x_k, a, y) \lambda(dy) \gamma(da|x_k) \Big] \mathbb{P}^{\gamma_n}_{\nu,k}(dh_k) \\
&\quad = \int_{\mathbf{H}_k} \Big[\int_{\mathbf{A}} \int_{\mathbf{X}} f(h_k, a, y) q(x_k, a, y) \lambda(dy) \gamma(da|x_k) \Big] \mathbb{P}^{\gamma}_{\nu,k}(dh_k) \\
&\quad = \int_{\mathbf{H}_{k+1}} f(h_{k+1}) \mathbb{P}^{\gamma}_{\nu,k+1}(dh_{k+1}).
\end{aligned}$$

Our goal now is to show that the second term in the right-hand-side of (A.4) converges to zero as $n \to \infty$. Observe that the function v_n on $(\mathbf{X} \times \mathbf{A})^k \times \mathbf{X}$ defined as

$$v_n(h_k) = \int_{\mathbf{A}} \int_{\mathbf{X}} f(h_k, a, y) q(x_k, a, y) \lambda(dy) \big[\gamma_n(da|x_k) - \gamma(da|x_k) \big]$$

takes the particular form

$$\bar{f}^{(0)}(a_0) \cdots \bar{f}^{k-1}(a_{k-1}) \cdot \bar{v}_n(x_0, \ldots, x_k)$$

(here, recall that $f = \bar{f} \cdot \bar{f}^{(0)} \cdots \bar{f}^{(k)} \in \mathcal{R}_{k+1}$), with

$$\bar{v}_n(x_0, \ldots, x_k) = \int_{\mathbf{A}} \int_{\mathbf{X}} \bar{f}(x_0, \ldots, x_k, y) \bar{f}^{(k)}(a) q(x_k, a, y) \lambda(dy) \big[\gamma_n(da|x_k) - \gamma(da|x_k) \big].$$

Then $\{\bar{v}_n\}_{n\in\mathbb{N}}$ is a bounded sequence in $\mathcal{B}(\mathbf{X}^{k+1})$, and also $\bar{v}_n(x_0, \ldots, x_{k-1}, \cdot) \overset{*}{\rightharpoonup} 0$ for any $(x_0, \ldots, x_{k-1}) \in \mathbf{X}^k$. Indeed, given arbitrary $\Phi \in L^1(\mathbf{X}, \mathfrak{B}(\mathbf{X}), \lambda)$ we have

$$\begin{aligned}
&\int_{\mathbf{X}} \bar{v}_n(x_0, \ldots, x_{k-1}, z) \Phi(z) \lambda(dz) \\
&= \int_{\mathbf{X}} \int_{\mathbf{A}} \Phi(z) \Big[\int_{\mathbf{X}} \bar{f}(x_0, \ldots, z, y) \bar{f}^{(k)}(a) q(z, a, y) \lambda(dy) \Big] \big[\gamma_n(da|z) - \gamma(da|z) \big] \lambda(dz).
\end{aligned}$$

Noting that the integral within brackets is a bounded Carathéodory function in $z \in \mathbf{X}$ and $a \in \mathbf{A}$ (recall Lemma A.1(i)), convergence to zero of the above integral follows from the definition of $\gamma_n \to \gamma$ in $\mathcal{Y}$.

We can therefore apply Lemma A.3 repeatedly to obtain that $\{\mathfrak{T}_k v_n\}_{n\in\mathbb{N}}$ is a bounded sequence in $\mathcal{B}(\mathbf{X})$ such that $\mathfrak{T}_k v_n(x_0) \to 0$ for every $x_0 \in \mathbf{X}$. By Lemma A.2 we obtain that the expression (A.4) satisfies the inequality

$$\begin{aligned}
&\int_{\mathbf{H}_k} \int_{\mathbf{A}} \int_{\mathbf{X}} f(h_k, a, y) q(x_k, a, y) \lambda(dy) \big[\gamma_n(da|x_k) - \gamma(da|x_k) \big] \mathbb{P}^{\gamma_n}_{\nu,k}(dh_k) \\
&= \int_{\mathbf{H}_k} v_n(h_k) \mathbb{P}^{\gamma_n}_{\nu,k}(dh_k) \leq \int_{\mathbf{X}} \mathfrak{T}_k v_n d\nu,
\end{aligned}$$

https://doi.org/10.1017/jpr.2024.25 Published online by Cambridge University Press

where the right-hand term converges to 0 as $n \to \infty$. Using a symmetric argument for the function $-f$, we obtain that the above left-hand term converges to zero, and so (A.4) indeed tends to zero.

So far, we have established the limit (A.2) for functions in $\mathcal{R}_k$. Note, however, that we can apply the Stone–Weierstrass theorem to the vector space spanned by the functions of the form $\hat{f}(a_0, \ldots, a_{k-1}) = \bar{f}^{(0)}(a_0) \cdots \bar{f}^{(k-1)}(a_{k-1})$, which indeed separates points in $\mathbf{A}^k$, to obtain that the functions in $\mathcal{C}(\mathbf{A}^k)$ can be uniformly approximated by functions in the above-mentioned vector space. Hence it is easy to establish that the limit in (A.2) holds for any function of the form (A.3). This completes the proof of Lemma 2.3.

Proof of Theorem 2.1. By virtue of Assumption A, the control model $\mathcal{M}$ satisfies the conditions (S1) and (S2) in [3]. Therefore, combining Theorem 2.2, Lemma 2.4, and the proof of Proposition 3.2 in [3], it can be shown that the w-topology and the ws^∞-topology are identical on $\boldsymbol{\mathcal{S}}_\nu$. Consequently, the set $\boldsymbol{\mathcal{S}}_\nu$ is compact for the ws$^\infty$-topology by Theorem 2.1 in [3], and it is metrizable in this topology.

To prove compactness of $\boldsymbol{\mathcal{S}}_\nu^s \subseteq \boldsymbol{\mathcal{S}}_\nu$, we show that it is a closed subset of $\boldsymbol{\mathcal{S}}_\nu$. Suppose that we have a sequence $\{\mathbb{P}_\nu^{\gamma_n}\}_{n\in\mathbb{N}} \subseteq \boldsymbol{\mathcal{S}}_\nu^s$, where $\gamma_n \in \boldsymbol{\mathcal{Y}}$ for each $n \in \mathbb{N}$, converging to some probability measure $\mathbb{P}_\nu^\pi \in \boldsymbol{\mathcal{S}}_\nu$ for some $\pi \in \boldsymbol{\Pi}$. But $\boldsymbol{\mathcal{Y}}$ being a compact metric space, there exists a subsequence $\{n'\}$ of $\{\gamma_n\}$ converging to some $\gamma \in \boldsymbol{\mathcal{Y}}$. Since $\mathbb{P}_\nu^{\gamma_{n'}} \Rightarrow \mathbb{P}_\nu^\gamma$ (Lemma 2.3), it follows that the limiting strategic probability measure $\mathbb{P}_\nu^\gamma = \mathbb{P}_\nu^\pi$ is in $\boldsymbol{\mathcal{S}}_\nu^s$. This completes the proof of Theorem 2.1.

Funding information

This work was supported by grant PID2021-122442NB-I00 from the Spanish Ministerio de Ciencia e Innovación.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article

References

[1] Aliprantis, C. D. and Border, K. C. (2006). *Infinite Dimensional Analysis*, 3rd edn. Springer, Berlin.

[2] Balder, E. J. (1988). Generalized equilibrium results for games with incomplete information. *Math. Operat. Res.* **13**, 265–276.

[3] Balder, E. J. (1989). On compactness of the space of policies in stochastic dynamic programming. *Stoch. Process. Appl.* **32**, 141–150.

[4] Balder, E. J. (1991). On Cournot–Nash equilibrium distributions for games with differential information and discontinuous payoffs. *Econom. Theory* **1**, 339–354.

[5] Balder, E. J. (1992). Existence without explicit compactness in stochastic dynamic programming. *Math. Operat. Res.* **17**, 572–580.

[6] Bäuerle, N. and Rieder, U. (2011). *Markov Decision Processes with Applications to Finance*. Springer, Heidelberg.

[7] Bäuerle, N. and Rieder, U. (2014). More risk-sensitive Markov decision processes. *Math. Operat. Res.* **39**, 105–120.

[8] Bellman, R. (1957). *Dynamic Programming*. Princeton University Press.

[9] Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-dynamic Programming*. Athena Scientific, Belmont, MA.

[10] Borkar, V. S. (2002). Convex analytic methods in Markov decision processes. In *Handbook of Markov Decision Processes*, Kluwer Academic Publishers, Boston, pp. 347–375.

https://doi.org/10.1017/jpr.2024.25 Published online by Cambridge University Press

[11] BREZIS, H. (2011). *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, New York.
[12] CAVAZOS-CADENA, R. (2018). Characterization of the optimal risk-sensitive average cost in denumerable Markov decision chains. *Math. Operat. Res.* **43**, 1025–1050.
[13] DUFOUR, F., HORIGUCHI, M. AND PIUNOVSKIY, A. B. (2012). The expected total cost criterion for Markov decision processes under constraints: a convex analytic approach. *Adv. Appl. Prob.* **44**, 774–793.
[14] DUFOUR, F. AND PIUNOVSKIY, A. B. (2013). The expected total cost criterion for Markov decision processes under constraints. *Adv. Appl. Prob.* **45**, 837–859.
[15] DUFOUR, F. AND PRIETO-RUMEAU, T. (2022). Maximizing the probability of visiting a set infinitely often for a countable state space Markov decision process. *J. Math. Anal. Appl.* 505, paper no. 125639.
[16] DYNKIN, E. B. AND YUSHKEVICH, A. A. (1979). *Controlled Markov Processes*. Springer, Berlin.
[17] FEINBERG, E. A. AND ROTHBLUM, U. G. (2012). Splitting randomized stationary policies in total-reward Markov decision processes. *Math. Operat. Res.* **37**, 129–153.
[18] HERNÁNDEZ-LERMA, O. AND LASSERRE, J. B. (1996). *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer, New York.
[19] HERNÁNDEZ-LERMA, O. AND LASSERRE, J. B. (1999). *Further Topics on Discrete-Time Markov Control Processes*. Springer, New York.
[20] HINDERER, K. (1970). *Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter*. Springer, Berlin.
[21] HOWARD, R. A. (1960). *Dynamic Programming and Markov Processes*. Technology Press of MIT.
[22] MEYN, S. P. AND TWEEDIE, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer, London.
[23] NOWAK, A. S. (1988). On the weak topology on a space of probability measures induced by policies. *Bull. Polish Acad. Sci. Math.* **36**, 181–186.
[24] PIUNOVSKIY, A. B. (1998). Controlled random sequences: methods of convex analysis and problems with functional constraints. *Uspekhi Mat. Nauk* **53**, 129–192.
[25] PIUNOVSKIY, A. B. (2004). Multicriteria impulsive control of jump Markov processes. *Math. Meth. Operat. Res.* **60**, 125–144.
[26] PUTERMAN, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley, New York.
[27] SCHÄL, M. (1975). Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal. *Z. Wahrscheinlichkeitsth.* **32**, 179–196.
[28] SCHÄL, M. (1975). On dynamic programming: compactness of the space of policies. *Stoch. Process. Appl.* **3**, 345–364.
[29] SCHÄL, M. (1979). On dynamic programming and statistical decision theory. *Ann. Statist.* **7**, 432–445.
[30] SCHÄL, M. (1983). Stationary policies in dynamic programming models under compactness assumptions. *Math. Operat. Res.* **8**, 366–372.
[31] SCHÄL, M. (1990). On the chance to visit a goal set infinitely often. *Optimization* **21**, 585–592.
[32] SENNOTT, L. I. (1999). *Stochastic Dynamic Programming and the Control of Queueing Systems*. John Wiley, New York.
[33] VENEL, X. AND ZILIOTTO, B. (2016). Strong uniform value in gambling houses and partially observable Markov decision processes. *SIAM J. Control Optimization* **54**, 1983–2008.
[34] ZHANG, Y. (2017). Continuous-time Markov decision processes with exponential utility. *SIAM J. Control Optimization* **55**, 2636–2660.

https://doi.org/10.1017/jpr.2024.25 Published online by Cambridge University Press