Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

On the Asymptotic Optimality of Finite Approximations to Markov Decision Processes with Borel Spaces

Naci Saldi, Serdar Yüksel, Tamás Linder

To cite this article:
Naci Saldi, Serdar Yüksel, Tamás Linder (2017) On the Asymptotic Optimality of Finite Approximations to Markov Decision Processes with Borel Spaces. Mathematics of Operations Research

Published online in Articles in Advance 15 Mar 2017

. http://dx.doi.org/10.1287/moor.2016.0832

http://pubsonline.informs.org/journal/moor/

MATHEMATICS OF OPERATIONS RESEARCH
Articles in Advance, pp. 1–34
ISSN 0364-765X (print), ISSN 1526-5471 (online)

On the Asymptotic Optimality of Finite Approximations to Markov Decision Processes with Borel Spaces

Naci Saldi,[a] **Serdar Yüksel,**[b] **Tamás Linder**[b]

[a] Coordinated Science Laboratory, University of Illinois, Urbana, Illinois 61801, USA; [b] Department of Mathematics and Statistics, Queen's University, Kingston, Ontario, Canada, K7L 3N6

Contact: nsaldi@illinois.edu (NS); yuksel@mast.queensu.ca (SY); linder@mast.queensu.ca (TL)

Received: July 20, 2015
Revised: March 29, 2016; September 20, 2016
Accepted: October 4, 2016
Published Online in Articles in Advance: March 15, 2017

MSC2010 Subject Classification: 93E20, 90C40, 90C39
OR/MS Subject Classification: Primary: Dynamic programming/optimal control, probability; secondary: Infinite state, Markov processes

https://doi.org/10.1287/moor.2016.0832

Abstract. Calculating optimal policies is known to be computationally difficult for Markov decision processes (MDPs) with Borel state and action spaces. This paper studies finite-state approximations of discrete time Markov decision processes with Borel state and action spaces, for both discounted and average costs criteria. The stationary policies thus obtained are shown to approximate the optimal stationary policy with arbitrary precision under quite general conditions for discounted cost and more restrictive conditions for average cost. For compact-state MDPs, we obtain explicit rate of convergence bounds quantifying how the approximation improves as the size of the approximating finite state space increases. Using information theoretic arguments, the order optimality of the obtained convergence rates is established for a large class of problems. We also show that as a pre-processing step, the action space can also be finitely approximated with sufficiently large number points; thereby, well known algorithms, such as value or policy iteration, Q-learning, etc., can be used to calculate near optimal policies.

Funding: This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

Keywords: Markov decision processes • stochastic control • finite state approximation • quantization

1. Introduction

In this paper, our goal is to study the finite-state approximation problem for computing near optimal policies for discrete time Markov decision processes (MDPs) with Borel state and action spaces, under discounted and average costs criteria. Although the existence and structural properties of optimal policies have been studied extensively in the literature, computing such policies is generally a challenging problem for systems with uncountable state spaces. This situation also arises in the fully observed reduction of a partially observed Markov decision process even when the original system has finite state and action spaces (see, e.g., Yu and Bertsekas [45]).

As has been extensively studied in the literature (see, e.g., Chow and Tsitsiklis [11] and the literature review below), one way to compute approximately optimal solutions for such MDPs is to construct a reduced model with a new transition probability and a one-stage cost function by quantizing the state/action spaces, i.e., by discretizing them on a finite grid. We exhibit that under quite general continuity conditions on the one-stage cost function and the transition probability for the discounted cost and under some additional restrictions on the ergodicity properties of Markov chains induced by deterministic stationary policies for the average cost, the optimal policy for the approximating finite model applied to the original model has a cost that converges to the optimal cost as the discretization becomes finer. Moreover, under additional continuity conditions on the transition probability and the one-stage cost function we also obtain bounds for a rate of approximation in terms of the number of points used to discretize the state space, thereby providing a trade-off between the computation cost and the performance loss in the system. In particular, we study the following two problems.

(Q1) Under what conditions on the components of the MDP do the true costs corresponding to the optimal policies obtained from finite models converge to the optimal value function as the number of grid points goes to infinity? For this problem, we are only concerned with the convergence of the approximation; that is, we do not establish bounds for a rate of approximation.

(Q2) Can we obtain explicit bounds on the performance loss due to the discretization in terms of the number of grid points if we strengthen the conditions sufficient in **(Q1)**?

Combined with our recent works (Saldi et al. [33, 34]), where we investigated the asymptotic optimality of the quantization of action sets, the results in this paper lead to a constructive algorithm for obtaining approximately optimal solutions. First the action space is quantized with small error and then the state space is quantized

with small error, which results in a finite model that well approximates the original MDP. When the state space is compact, we also obtain rates of convergence for both approximations, and using information theoretic tools we establish that the obtained rates of convergence are order-optimal for a given class of MDPs. Since there exist various computational algorithms for finite-state Markov decision problems, the analysis in this paper can be considered to be *constructive*.

Various methods have been developed to compute approximate value functions and near optimal policies. A partial list of these techniques is as follows: approximate dynamic programming, approximate value or policy iteration, simulation-based techniques, neuro-dynamic programming (or reinforcement learning), state aggregation, etc. For rather complete surveys of these techniques, we refer the reader to Fox [17], White [42, 43], Langen [28], Bertsekas and Tsitsiklis [6], Ren and Krogh [32], Ortner [30], White [40, 41], Bertsekas [3], Dufour and Prieto-Rumeau [14, 15], and references therein. With the exception of Dufour and Prieto-Rumeau [15] and Ortner [30], these papers in general study either the finite horizon cost or the discounted infinite horizon cost. Also, the majority of these results are for MDPs with discrete (i.e., finite or countable) state and action spaces, or a bounded one-stage cost function (e.g., Fox [17], White [42, 43], Van Roy [37], White [40, 41], Cavazos-Cadena [9], Bertsekas and Tsitsiklis [6], Ren and Krogh [32], Ortner [30], Bertsekas [3]). Those that consider general state and action spaces (see, e.g., Dufour and Prieto-Rumeau [13, 14, 15], Bertsekas [3], Chow and Tsitsiklis [11]) assume in general Lipschitz type continuity conditions on the components of the control model to provide a rate of convergence analysis for the approximation error. Some of the results only consider approximating the value function and do not provide a procedure to compute near optimal policies (e.g., Langen [28], White [43], Dufour and Prieto-Rumeau [14]).

Our paper differs from these results in the following ways: (i) we consider a general setup, where the state and action spaces are Borel (with the action space being compact), and the one-stage cost function is possibly unbounded; (ii) since we do not aim to provide rate of convergence result in the first problem **(Q1)**, the continuity assumptions we impose on the components of the control model are weaker than the conditions imposed in prior works that considered general state and action spaces; and (iii) we also consider the challenging average cost criterion under reasonable assumptions. The price we pay for imposing weaker assumptions in **(Q1)** is that we do not obtain explicit performance bounds in terms of the number of grid points used in the approximations. However, such bounds can be obtained under further assumptions on the transition probability and the one-stage cost functions; this is considered in problem **(Q2)** for compact-state MDPs.

Our approach to solve problem **(Q1)** can be summarized as follows: (i) first, we obtain approximation results for the compact-state case, (ii) we find conditions under which a compact representation leads to near optimality for noncompact state MDPs, and (iii) we prove the convergence of the finite-state models to noncompact models. As a byproduct of this analysis, we obtain *compact-state-space approximation*s for an MDP with noncompact Borel state space. In particular, our findings directly lead to finite models if the state space is countable; similar problems in the countable context have been studied in the literature for the discounted cost; see Puterman [31, Section 6.10.2].

We note that the proposed method for solving the approximation problem for compact-state MDPs with the discounted cost is partly inspired by Van Roy [37]. Specifically, we generalize the operator proposed for an approximate value iteration algorithm in Van Roy [37] to uncountable state spaces. Next, unlike in Van Roy [37], we use this operator as a transition step between the original optimality operator and the optimality operator of the approximate model. In Ortner [30], a similar construction was given for finite-state action MDPs. Our method to obtain finite-state MDPs from the compact-state model can be regarded as a generalization of this construction. We note that a related work of Dufour and Prieto-Rumeau [15] develops a sequence of approximations using empirical distributions of an underlying probability measure with respect to which the transition probability of the MDP is absolutely continuous. By imposing Lipschitz type continuity conditions on the components of the control model, Dufour and Prieto-Rumeau [15] obtain a concentration inequality type upper bound on the accuracy of the approximation based on the Wasserstein distance of order 1 between the probability measure and its empirical estimate. These conditions are stronger than what we impose for the problem **(Q1)**. We note that Dufour and Prieto-Rumeau [15] adopts a simulation-based approximation leading to probabilistic guarantees on the approximation, whereas we adopt a quantization based approach leading to deterministic approximation guarantees. For a review of further simulation based methods, see, e.g., Chang et al. [10], Jain and Varaiya [25].

The approach developed in the paper is also useful in networked control applications where transmission of real-valued actions to an actuator is not realistic when there is an information transmission constraint between a plant, a controller, and an actuator (see, e.g., Yüksel and Başar [46]). On the other hand, the elements of a finite action set can be transmitted across a finite capacity information channel. Even though the problem of optimal

quantization for information transmission from a plant/sensor to a controller has been studied extensively (see, e.g., references in Yüksel and Başar [46]), these type of results appear to be new in the networked control literature when the problem of transmitting signals from a controller to an actuator is considered. Furthermore, tools from information theory allow for obtaining lower bounds on the approximation performance; using such an argument we show that the construction in this paper is order optimal for a large class of models.

The rest of the paper is organized as follows. In Section 2 we study the approximation problem **(Q1)** for MDPs with compact state space. In Section 3 an analogous approximation result is obtained for MDPs with noncompact state space. Discretization of the action space is considered in Section 4 for a general state space. In Section 5 we derive quantitative bounds on the approximation error in terms of the number of points used to discretize the state space for the compact-state case. In Section 6 the order optimality of the obtained bounds on the approximation errors is established. In Section 7 we present an example to numerically illustrate our results. Section 8 concludes the paper.

1.1. Notation and Conventions

For a metric space E, the Borel σ-algebra (the smallest σ-algebra that contains the open sets of E) is denoted by $\mathscr{B}(\mathsf{E})$. We let $B(\mathsf{E})$ and $C_b(\mathsf{E})$ denote the set of all bounded Borel measurable and continuous real functions on E, respectively. For any $u \in C_b(\mathsf{E})$ or $u \in B(\mathsf{E})$, let $\|u\| := \sup_{e\in\mathsf{E}} |u(e)|$, which turns $C_b(\mathsf{E})$ and $B(\mathsf{E})$ into Banach spaces. Given any Borel measurable function $w\colon \mathsf{E} \to [1,\infty)$ and any real valued Borel measurable function u on E, we define the w-norm of u as

$$\|u\|_w := \sup_{e\in\mathsf{E}} \frac{|u(e)|}{w(e)}$$

and let $B_w(\mathsf{E})$ denote the Banach space of all real valued measurable functions u on E with finite w-norm; see Hernández-Lerma and Lasserre [22]. Let $\mathscr{P}(\mathsf{E})$ denote the set of all probability measures on E. A sequence $\{\mu_n\}$ of probability measures on E is said to converge weakly (respectively, setwise; see Hernández-Lerma and Lasserre [23]) to a probability measure μ if $\int_{\mathsf{E}} g(e)\mu_n(de) \to \int_{\mathsf{E}} g(e)\mu(de)$ for all $g \in C_b(\mathsf{E})$ (respectively, for all $g \in B(\mathsf{E})$). For any $\mu, \nu \in \mathscr{P}(\mathsf{E})$, the total variation distance between μ and ν, denoted as $\|\mu-\nu\|_{\mathrm{TV}}$, is equivalently defined as

$$\|\mu-\nu\|_{\mathrm{TV}} := 2 \sup_{D\in\mathscr{B}(\mathsf{E})} |\mu(D)-\nu(D)| = \sup_{\|g\|\le 1} \left| \int_{\mathsf{E}} g(e)\mu(de) - \int_{\mathsf{E}} g(e)\nu(de) \right|.$$

Unless otherwise specified, the term "measurable" will refer to Borel measurability in the rest of the paper.

1.2. Markov Decision Processes

A discrete-time Markov decision process can be described by a five-tuple

$$(\mathsf{X}, \mathsf{A}, \{\mathsf{A}(x)\colon x \in \mathsf{X}\}, p, c),$$

where Borel spaces (i.e., Borel subsets of complete and separable metric spaces) X and A denote the *state* and *action* spaces, respectively. The collection $\{\mathsf{A}(x)\colon x \in \mathsf{X}\}$ is a family of nonempty subsets $\mathsf{A}(x)$ of A that give the admissible actions for the state $x \in \mathsf{X}$. The *stochastic kernel* $p(\cdot \mid x,a)$ denotes the *transition probability* of the next state given that previous state-action pair is (x,a); see Hernández-Lerma and Lasserre [21]. Hence, it satisfies the following: (i) $p(\cdot \mid x,a)$ is an element of $\mathscr{P}(\mathsf{X})$ for all (x,a), and (ii) $p(D \mid \cdot,\cdot)$ is a measurable function from $\mathsf{X}\times\mathsf{A}$ to $[0,1]$ for each $D \in \mathscr{B}(\mathsf{X})$. The *one-stage cost* function c is a measurable function from $\mathsf{X}\times\mathsf{A}$ to $\mathbb{R}$. In this paper, it is assumed that $\mathsf{A}(x) = \mathsf{A}$ for all $x \in \mathsf{X}$.

Define the history spaces $\mathsf{H}_0 = \mathsf{X}$ and $\mathsf{H}_t = (\mathsf{X}\times\mathsf{A})^t \times \mathsf{X}$, $t = 1,2,\ldots$ endowed with their product Borel σ-algebras generated by $\mathscr{B}(\mathsf{X})$ and $\mathscr{B}(\mathsf{A})$. A *policy* is a sequence $\pi = \{\pi_t\}$ of stochastic kernels on A given H_t. The set of all policies is denoted by Π. Let Φ denote the set of stochastic kernels φ on A given X, and let $\mathbb{F}$ denote the set of all measurable functions f from X to A. A *randomized Markov* policy is a sequence $\pi = \{\pi_t\}$ of stochastic kernels on A given X. A *deterministic Markov* policy is a sequence of stochastic kernels $\pi = \{\pi_t\}$ on A given X such that $\pi_t(\cdot \mid x) = \delta_{f_t(x)}(\cdot)$ for some $f_t \in \mathbb{F}$, where δ_z denotes the point mass at z. The set sof randomized and deterministic Markov policies are denoted by RM and M, respectively. A *randomized stationary* policy is a constant sequence $\pi = \{\pi_t\}$ of stochastic kernels on A given X such that $\pi_t(\cdot \mid x) = \varphi(\cdot \mid x)$ for all t for some $\varphi \in \Phi$. A *deterministic stationary* policy is a constant sequence of stochastic kernels $\pi = \{\pi_t\}$ on A given X such that $\pi_t(\cdot \mid x) = \delta_{f(x)}(\cdot)$ for all t for some $f \in \mathbb{F}$. The set of randomized and deterministic stationary policies are identified with the sets Φ and $\mathbb{F}$, respectively.

According to the Ionescu Tulcea theorem (see Hernández-Lerma and Lasserre [21]), an initial distribution μ on X and a policy π define a unique probability measure P_μ^π on $\mathsf{H}_\infty = (\mathsf{X}\times\mathsf{A})^\infty$. The expectation with respect

to P_μ^π is denoted by $\mathbb{E}_\mu^\pi$. If $\mu = \delta_x$, we write P_x^π and $\mathbb{E}_x^\pi$ instead of $P_{\delta_x}^\pi$ and $\mathbb{E}_{\delta_x}^\pi$. The cost functions to be minimized in this paper are the β-discounted cost and the average cost, respectively given by

$$J(\pi,x)=\mathbb{E}_x^\pi\left[\sum_{t=0}^{\infty}\beta^t c(x_t,a_t)\right],\qquad V(\pi,x)=\limsup_{T\to\infty}\frac{1}{T}\mathbb{E}_x^\pi\left[\sum_{t=0}^{T-1}c(x_t,a_t)\right].$$

With this notation, the discounted and average value functions of the control problem are defined as

$$J^*(x):=\inf_{\pi\in\Pi}J(\pi,x),\qquad V^*(x):=\inf_{\pi\in\Pi}V(\pi,x).$$

A policy π^* is said to be optimal if $J(\pi^*,x)=J^*(x)$ (or $V(\pi^*,x)=V^*(x)$ for the average cost) for all $x\in\mathsf{X}$. Under fairly mild conditions, the set $\mathbb{F}$ of deterministic stationary policies contains an optimal policy for discounted cost (see, e.g., Hernández-Lerma and Lasserre [21], Feinberg et al. [16]) and average cost optimal control problems (under somewhat stronger continuity/recurrence conditions; see, e.g., Feinberg et al. [16]).

Remark 1.1. We note that the pathwise infinite sum $\sum_{t=0}^{\infty}\beta^t c(x_t,a_t)$ may not be well defined in the definition of J if c is only assumed to be measurable. However, further assumptions that will be imposed in later sections ensure that J is a well-defined function.

1.3. Auxiliary Results

To avoid measurability problems associated with the operators that will be defined for the approximation problem in the discounted cost case, it is necessary to enlarge the set of functions on which these operators can act. To this end, in this section we review the notion of analytic sets and lower semi-analytic functions, and state the main results that will be used in the sequel to tackle these measurability problems. For a detailed treatment of analytic sets and lower semi-analytic functions, we refer the reader to Shreve and Bertsekas [36], Blackwell et al. [7], Kuratowski [27, Chapter 39], and Bertsekas and Shreve [5, Chapter 7].

Let $\mathbb{N}^\infty$ be the set of sequences of natural numbers endowed with the product topology. With this topology, $\mathbb{N}^\infty$ is a complete and separable metric space. A subset A of a Borel space E is said to be *analytic* if it is a continuous image of $\mathbb{N}^\infty$. Note that Borel sets are always analytic.

A function $g\colon \mathsf{E}\to\mathbb{R}$ is said to be *universally measurable* if for any $\mu\in\mathcal{P}(\mathsf{E})$, there is a Borel measurable function $g_\mu\colon \mathsf{E}\to\mathbb{R}$ such that $g=g_\mu$ μ almost everywhere. It is said to be *lower semi-analytic* if the set $\{e\colon g(e)<c\}$ is analytic for any $c\in\mathbb{R}$. Any Borel measurable function is lower semi-analytic and any lower semi-analytic function is universally measurable. The latter property implies that the integral of any lower semi-analytic function with respect to any probability measure is well defined. We let $B^l(\mathsf{E})$ and $B_w^l(\mathsf{E})$ denote the set of all bounded lower semi-analytic functions and lower semi-analytic functions with finite w-norm, respectively. Since any pointwise limit of a sequence of lower semi-analytic functions is lower semi-analytic (see Kuratowski [27, Theorem 1, p. 512]), $(B^l(\mathsf{E}),\|\cdot\|)$ and $(B_w^l(\mathsf{E}),\|\cdot\|_w)$ are Banach spaces.

We now state the results that will be used in the sequel.

Proposition 1.1 (Bertsekas and Shreve [5, Proposition 7.47, p. 179]). *Suppose E_1 and E_2 are Borel spaces. Let $g\colon \mathsf{E}_1\times\mathsf{E}_2\to\mathbb{R}$ be lower semi-analytic. Then, $g^*(e_1):=\inf_{e_2\in\mathsf{E}_2}g(e_1,e_2)$ is also lower semi-analytic.*

Proposition 1.2 (Bertsekas and Shreve [5, Proposition 7.48, p. 180]). *Suppose E_1 and E_2 as in Proposition 1.1. Let $g\colon \mathsf{E}_1\times\mathsf{E}_2\to\mathbb{R}$ be lower semi-analytic and $q(de_2\mid e_1)$ be a stochastic kernel on E_2 given E_1; then the function*

$$h(e_1):=\int_{\mathsf{E}_2}g(e_2)\,q(de_2\mid e_1).$$

is lower semi-analytic.

2. Finite-State Approximations of MDPs with Compact State Space

In this section we consider **(Q1)** for the MDPs with compact state space. To distinguish compact-state MDPs from noncompact ones, the state space of the compact-state MDPs will be denoted by Z instead of X. We impose the assumptions below on the components of the Markov decision process; additional new assumptions will be made for the average cost problem in Section 2.2.

Assumption 2.1. (a) *The one-stage cost function c is in $C_b(\mathsf{Z}\times\mathsf{A})$.*

(b) *The stochastic kernel $p(\cdot\mid z,a)$ is weakly continuous in (z,a); i.e., for all z and a, $p(\cdot\mid z_k,a_k)\to p(\cdot\mid z,a)$ weakly when $(z_k,a_k)\to(z,a)$.*

(c) *Z and A are compact.*

Before proceeding with the main results, we first describe the procedure used to obtain finite-state models. Let d_{Z} denote the metric on Z. Since the state space Z is assumed to be compact and thus totally bounded, one can find a sequence $(\{z_{n,i}\}_{i=1}^{k_n})_{n\geq 1}$ of finite grids in Z such that for all n,

$$\min_{i\in\{1,\ldots,k_n\}} d_{\mathsf{Z}}(z,z_{n,i}) < 1/n, \quad \text{for all } z\in\mathsf{Z}.$$

The finite grid $\{z_{n,i}\}_{i=1}^{k_n}$ is called an $(1/n)$-net in Z. Let $\mathsf{Z}_n := \{z_{n,1},\ldots,z_{n,k_n}\}$ and define function Q_n mapping Z to Z_n by

$$Q_n(z) := \arg\min_{z_{n,i}\in\mathsf{Z}_n} d_{\mathsf{Z}}(z,z_{n,i}),$$

where ties are broken so that Q_n is measurable. In the literature, Q_n is often called a nearest neighborhood quantizer with respect to distortion measure d_{Z}; see Gray and Neuhoff [19]. For each n, Q_n induces a partition $\{\mathcal{S}_{n,i}\}_{i=1}^{k_n}$ of the state space Z given by

$$\mathcal{S}_{n,i} = \{z\in\mathsf{Z} : Q_n(z) = z_{n,i}\},$$

with diameter $\operatorname{diam}(\mathcal{S}_{n,i}) := \sup_{z,\,y\in\mathcal{S}_{n,i}} d_{\mathsf{Z}}(z,y) < 2/n$. Let $\{\nu_n\}$ be a sequence of probability measures on Z satisfying

$$\nu_n(\mathcal{S}_{n,i}) > 0, \quad \text{for all } i,n. \tag{1}$$

We let $\nu_{n,i}$ be the restriction of ν_n to $\mathcal{S}_{n,i}$ defined by

$$\nu_{n,i}(\cdot) := \frac{\nu_n(\cdot)}{\nu_n(\mathcal{S}_{n,i})}.$$

The measures $\nu_{n,i}$ will be used to define a sequence of finite-state MDPs, denoted as MDP_n $(n\geq 1)$, to approximate the original model. To this end, for each n define the one-stage cost function $c_n\colon \mathsf{Z}_n\times\mathsf{A}\to\mathbb{R}$ and the transition probability p_n on Z_n given $\mathsf{Z}_n\times\mathsf{A}$ by

$$c_n(z_{n,i},a) := \int_{\mathcal{S}_{n,i}} c(z,a)\nu_{n,i}(dz), \qquad p_n(\cdot\mid z_{n,i},a) := \int_{\mathcal{S}_{n,i}} Q_n\times p(\cdot\mid z,a)\nu_{n,i}(dz),$$

where $Q_n\times p(\cdot\mid z,a)\in\mathcal{P}(\mathsf{Z}_n)$ is the pushforward of the measure $p(\cdot\mid z,a)$ with respect to Q_n; that is,

$$Q_n\times p(z_{n,j}\mid z,a) = p(\mathcal{S}_{n,j}\mid z,a)$$

for all $z_{n,j}\in\mathsf{Z}_n$. For each n, we define MDP_n as a Markov decision process with the following components: Z_n is the state space, A is the action space, p_n is the transition probability, and c_n is the one-stage cost function. History spaces, policies, and cost functions are defined in a similar way as in the original model.

2.1. Discounted Cost

Here we consider **(Q1)** for the discounted cost criterion with a discount factor $\beta\in(0,1)$. Throughout this section, it is assumed that Assumption 2.1 holds.

Define the operator T on $B(\mathsf{Z})$ by

$$Tu(z) := \min_{a\in\mathsf{A}}\left[c(z,a) + \beta\int_{\mathsf{Z}} u(y)p(dy\mid z,a)\right]. \tag{2}$$

In the literature T is called the *Bellman optimality operator*. It can be proved that under Assumption 2.1(a) and 2.1(b), T is a contraction operator with modulus β mapping $C_b(\mathsf{Z})$ into itself (see Hernández-Lerma [20, Theorem 2.8, p. 23]); that is, $Tu\in C_b(\mathsf{Z})$ for all $u\in C_b(\mathsf{Z})$ and

$$\|Tu - Tv\| \leq \beta\|u-v\|, \quad \text{for all } u,v\in C_b(\mathsf{Z}).$$

The following theorem is a widely known result in the theory of Markov decision processes (see again Hernández-Lerma [20, Theorem 2.8, p. 23]) that also holds without a compactness assumption on the state space.

Theorem 2.1. *The value function J^* is the unique fixed point in $C_b(\mathsf{Z})$ of the contraction operator T; i.e.,*

$$J^* = TJ^*.$$

Furthermore, a deterministic stationary policy f^* is optimal if and only if it satisfies the optimality equation; i.e.,

$$J^*(z) = c(z, f^*(z)) + \beta \int_{\mathsf{Z}} J^*(y)\, p(dy \mid z, f^*(z)). \tag{3}$$

Finally, there exists a deterministic stationary policy f^* that is optimal, so it satisfies (3).

Define, for all $n \geq 1$, the operator T_n, which is the Bellman optimality operator for MDP$_n$, by

$$T_n u(z_{n,i}) := \min_{a \in \mathsf{A}} \left[c_n(z_{n,i}, a) + \beta \sum_{j=1}^{k_n} u(z_{n,j}) p_n(z_{n,j} \mid z_{n,i}, a) \right];$$

equivalently,

$$T_n u(z_{n,i}) = \min_{a \in \mathsf{A}} \int_{\mathcal{S}_{n,i}} \left[c(z,a) + \beta \int_{\mathsf{Z}} \hat{u}(y) p(dy \mid z, a) \right] \nu_{n,i}(dz),$$

where $u\colon \mathsf{Z}_n \to \mathbb{R}$ and $\hat{u}$ is the piecewise constant extension of u to Z given by $\hat{u}(z) = u \circ Q_n(z)$. For each n, under Assumption 2.1, Hernández-Lerma [20, Theorem 2.8, p. 23] implies the following: (i) T_n is a contraction operator with modulus β mapping $B(\mathsf{Z}_n)$ $(= C_b(\mathsf{Z}_n))$ into itself, (ii) the fixed point of T_n is the value function J_n^* of MDP$_n$, and (iii) there exists an optimal stationary policy f_n^* for MDP$_n$, which therefore satisfies the optimality equation. Hence, we have

$$J_n^* = T_n J_n^* = T_n J_n(f_n^*, \cdot) = J_n(f_n^*, \cdot),$$

where J_n denotes the discounted cost for MDP$_n$. Let us extend the optimal policy f_n^* for MDP$_n$ to X by letting $\hat{f}_n(z) = f_n^* \circ Q_n(z) \in \mathbb{F}$.

The following theorem is the main result of this section. It states that the cost function of the policy $\hat{f}_n$ converges to the value function J^* as $n \to \infty$.

Theorem 2.2. *The discounted cost of the policy $\hat{f}_n$, obtained by extending the optimal policy f_n^* of MDP$_n$ to Z, converges to the optimal value function J^* of the original MDP*

$$\lim_{n \to \infty} \| J(\hat{f}_n, \cdot) - J^* \| = 0.$$

Hence, to find a near optimal policy for the original MDP, it is sufficient to compute the optimal policy of MDP$_n$ for sufficiently large n and then extend this policy to the original state space.

To prove Theorem 2.2 we need a series of technical results. We first define an operator $\hat{T}_n$ on $B^l(\mathsf{Z})$ by extending T_n to $B^l(\mathsf{Z})$:

$$\hat{T}_n u(z) := \inf_{a \in \mathsf{A}} \int_{\mathcal{S}_{n,i_n(z)}} \left[c(x,a) + \beta \int_{\mathsf{Z}} u(y) p(dy \mid x, a) \right] \nu_{n,i_n(z)}(dx), \tag{4}$$

where $i_n\colon \mathsf{Z} \to \{1, \ldots, k_n\}$ maps z to the index of the partition $\{\mathcal{S}_{n,i}\}$ it belongs to. To see that this operator is well defined, let the stochastic kernel $r_n(dx \mid z)$ on Z given Z be defined as

$$r_n(dx \mid z) := \sum_{i=1}^{k_n} \nu_{n,i}(dx) 1_{\mathcal{S}_{n,i}}(z),$$

where 1_B denotes the indicator function of the set B. Next, we can write the right-hand side of (4) as

$$\inf_{a \in \mathsf{A}} \int_{\mathsf{Z}} \left[c(x,a) + \beta \int_{\mathsf{Z}} u(y) p(dy \mid x, a) \right] r_n(dx \mid z).$$

Therefore, by Propositions 1.1 and 1.2, we can conclude that $\hat{T}_n$ maps $B^l(\mathsf{Z})$ into $B^l(\mathsf{Z})$. Furthermore, it is a contraction operator with modulus β that can be shown using Hernández-Lerma [20, Proposition A.2, p. 122]. Hence, it has a unique fixed point $\hat{J}_n^*$ that belongs to $B(\mathsf{Z})$, and this fixed point must be constant over the sets $\mathcal{S}_{n,i}$ because of the averaging operation on each $\mathcal{S}_{n,i}$. Furthermore, since $\hat{T}_n(u \circ Q_n) = (T_n u) \circ Q_n$ for all $u \in B(\mathsf{Z}_n)$, we have

$$\hat{T}_n(J_n^* \circ Q_n) = (T_n J_n^*) \circ Q_n = J_n^* \circ Q_n.$$

Hence, the fixed point of $\hat{T}_n$ is the piecewise constant extension of the fixed point of T_n; i.e.,

$$\hat{J}_n^* = J_n^* \circ Q_n.$$

Remark 2.1. In the rest of this paper, when we take the integral of any function with respect to $\nu_{n,i_n(z)}$, it is tacitly assumed that the integral is taken over all set $\mathcal{S}_{n,i_n(z)}$. Hence, we can drop $\mathcal{S}_{n,i_n(z)}$ in the integral for the ease of notation.

We now define another operator F_n on $B^l(\mathsf{Z})$ by simply interchanging the order of the infimum and the integral in (4); i.e.,

$$\begin{aligned} F_n u(z) &:= \int \inf_{a\in\mathsf{A}}\left[c(x,a)+\beta\int_{\mathsf{Z}} u(y)\,p(dy\mid x,a)\right]\nu_{n,i_n(z)}(dx)\\ &= \Gamma_n T u(z), \end{aligned}$$

where

$$\Gamma_n u(z) := \int u(x)\nu_{n,i_n(z)}(dx).$$

We note that F_n is the extension (to infinite state spaces) of the operator defined in Van Roy [37, p. 236] for the proposed approximate value iteration algorithm. However, unlike in Van Roy [37], F_n will serve here as an intermediate point between T and $\hat{T}_n$ (or T_n) to solve **(Q1)** for the discounted cost. To this end, we first note that F_n is a contraction operator on $B^l(\mathsf{Z})$ with modulus β. Indeed, it is clear that F_n maps $B^l(\mathsf{Z})$ into itself by Propositions 1.1 and 1.2. Furthermore, for any $u, v \in B^l(\mathsf{Z})$, we clearly have $\|\Gamma_n u - \Gamma_n v\| \le \|u - v\|$. Hence, since T is a contraction operator on $B^l(\mathsf{Z})$ with modulus β, F_n is also a contraction operator on $B^l(\mathsf{Z})$ with modulus β.

Remark 2.2. Since we only assume that the stochastic kernel p is weakly continuous, it is not true that $\hat{T}_n$ and F_n map $B(\mathsf{Z})$ into itself (see Hernández-Lerma and Lasserre [21, Proposition D.5, p. 182]). This is the point where we need to enlarge the set of functions on which these operators act.

The following theorem states that the fixed point, say u_n^*, of F_n converges to the fixed point J^* (i.e., the value function) of T as n goes to infinity. Note that although T is originally defined on $C_b(\mathsf{Z})$, it can be proved that T, when acting on $B^l(\mathsf{Z})$, maps $B^l(\mathsf{Z})$ into itself.

Theorem 2.3. *If u_n^* is the unique fixed point of F_n, then* $\lim_{n\to\infty}\|u_n^* - J^*\| = 0$.

The proof of Theorem 2.3 requires two lemmas.

Lemma 2.1. *For any $u \in B^l(\mathsf{Z})$, we have*

$$\|u - \Gamma_n u\| \le 2 \inf_{r\in\mathsf{Z}^{k_n}} \|u - \Phi_r\|,$$

where $\Phi_r(z) = \Sigma_{i=1}^{k_n} r_i 1_{S_{n,i}}(z)$, $r = (r_1, \ldots, r_{k_n})$.

Proof. Fix any $r \in \mathsf{Z}^{k_n}$. Next, using the identity $\Gamma_n\Phi_r = \Phi_r$, we obtain

$$\|u - \Gamma_n u\| \le \|u - \Phi_r\| + \|\Phi_r - \Gamma_n u\| = \|u - \Phi_r\| + \|\Gamma_n\Phi_r - \Gamma_n u\| \le \|u - \Phi_r\| + \|\Phi_r - u\|.$$

Since r is arbitrary, this completes the proof. □

Notice that because of the operator Γ_n, the fixed point u_n^* of F_n must be constant over the sets $\mathcal{S}_{n,i}$. We use this property to prove the next lemma.

Lemma 2.2. *We have*

$$\|u_n^* - J^*\| \le \frac{2}{1-\beta} \inf_{r\in\mathsf{Z}^{k_n}} \|J^* - \Phi_r\|.$$

Proof. Note that $\Gamma_n u_n^* = u_n^*$ since u_n^* is constant over the sets $\mathcal{S}_{n,i}$. Thus, we have

$$\begin{aligned} \|u_n^* - J^*\| &\le \|u_n^* - \Gamma_n J^*\| + \|\Gamma_n J^* - J^*\|\\ &= \|F_n u_n^* - \Gamma_n T J^*\| + \|\Gamma_n J^* - J^*\|\\ &= \|\Gamma_n T u_n^* - \Gamma_n T J^*\| + \|\Gamma_n J^* - J^*\| \quad \text{(by the definition of } F_n\text{)}\\ &\le \|T u_n^* - T J^*\| + \|\Gamma_n J^* - J^*\| \quad \text{(since } \|\Gamma_n u - \Gamma_n v\| \le \|u - v\|\text{)}\\ &\le \beta\|u_n^* - J^*\| + \|\Gamma_n J^* - J^*\|. \end{aligned}$$

Hence, we obtain $\|u_n^* - J^*\| \le (1/(1-\beta))\|\Gamma_n J^* - J^*\|$. The result now follows from Lemma 2.1. □

Proof of Theorem 2.3. Recall that since Z is compact, the function J^* is uniformly continuous and $\operatorname{diam}(\mathcal{S}_{n,i}) < 2/n$ for all $i = 1,\dots,k_n$. Hence, $\lim_{n\to\infty} \inf_{r\in\mathsf{Z}^{k_n}} \|J^* - \Phi_r\| = 0$, which completes the proof in view of Lemma 2.2. □

The next step is to show that the fixed point $\hat{J}_n^*$ of $\hat{T}_n$ converges to the fixed point J^* of T. To this end, we first prove the following result.

Lemma 2.3. *For any $u \in C_b(\mathsf{Z})$, $\|\hat{T}_n u - F_n u\| \to 0$ as $n \to \infty$.*

Proof. Note that since $\int_{\mathsf{Z}} u(x)p(dx \mid y,a)$ is continuous as a function of (y,a) by Assumption 2.1(b), it is sufficient to prove that for any $l \in C_b(\mathsf{Z}\times\mathsf{A})$

$$\left\|\min_a \int l(y,a)\nu_{n,i_n(z)}(dy) - \int \min_a l(y,a)\nu_{n,i_n(z)}(dy)\right\| := \sup_{z\in\mathsf{Z}}\left|\min_a \int l(y,a)\nu_{n,i_n(z)}(dy) - \int \min_a l(y,a)\nu_{n,i_n(z)}(dy)\right| \to 0$$

as $n \to \infty$. Fix any $\varepsilon > 0$. Define $\{z_i\}_{i=1}^{\infty} := \bigcup_n \mathsf{Z}_n$ and let $\{a_i\}_{i=1}^{\infty}$ be a sequence in A such that $\min_{a\in\mathsf{A}} l(z_i,a) = l(z_i,a_i)$; such a_i exists for each z_i because $l(z_i,\cdot)$ is continuous and A is compact. Define $g(y) := \min_{a\in\mathsf{A}} l(y,a)$, which can be proved to be continuous and therefore uniformly continuous since Z is compact. Thus by the uniform continuity of l, there exists $\delta > 0$ such that $d_{\mathsf{Z}\times\mathsf{A}}((y,a),(y',a')) < \delta$ implies $|g(y) - g(y')| < \varepsilon/2$ and $|l(y,a) - l(y',a')| < \varepsilon/2$. Choose n_0 such that $2/n_0 < \delta$. Then for all $n \ge n_0$, $\max_{i\in\{1,\dots,k_n\}} \operatorname{diam}(\mathcal{S}_{n,i}) < 2/n < \delta$. Hence, for all $y \in \mathcal{S}_{n,i}$ we have $|l(y,a_i) - \min_{a\in\mathsf{A}} l(y,a)| \le |l(y,a_i) - l(z_i,a_i)| + |\min_{a\in\mathsf{A}} l(z_i,a) - \min_{a\in\mathsf{A}} l(y,a)| = |l(y,a_i) - l(z_i,a_i)| + |g(z_i) - g(y)| < \varepsilon$. This implies

$$\begin{aligned}\left\|\min_a \int l(y,a)\nu_{n,i_n(z)}(dy) - \int \min_a l(y,a)\nu_{n,i_n(z)}(dy)\right\| &\le \left\|\int l(y,a_i)\nu_{n,i_n(z)}(dy) - \int \min_a l(y,a)\nu_{n,i_n(z)}(dy)\right\| \\ &\le \sup_{z\in\mathsf{Z}} \int \sup_{y\in\mathcal{S}_{n,i_n(z)}} \left|l(y,a_i) - \min_a l(y,a)\right| \nu_{n,i_n(z)}(dy) < \varepsilon.\end{aligned}$$

This completes the proof. □

Theorem 2.4. *The fixed point $\hat{J}_n^*$ of $\hat{T}_n$ converges to the fixed point J^* of T.*

Proof. We have

$$\begin{aligned}\|\hat{J}_n^* - J^*\| &\le \|\hat{T}_n\hat{J}_n^* - \hat{T}_n J^*\| + \|\hat{T}_n J^* - F_n J^*\| + \|F_n J^* - F_n u_n^*\| + \|F_n u_n^* - J^*\| \\ &\le \beta\|\hat{J}_n^* - J^*\| + \|\hat{T}_n J^* - F_n J^*\| + \beta\|J^* - u_n^*\| + \|u_n^* - J^*\|.\end{aligned}$$

Hence

$$\|\hat{J}_n^* - J^*\| \le \frac{\|\hat{T}_n J^* - F_n J^*\| + (1+\beta)\|J^* - u_n^*\|}{1-\beta}.$$

The theorem now follows from Theorem 2.3 and Lemma 2.3. □

Recall the optimal stationary policy f_n^* for MDP_n and its extension $\hat{f}_n(z) = f_n^* \circ Q_n(z)$ to Z. Since $\hat{J}_n^* = J_n^* \circ Q_n$, it is straightforward to prove that $\hat{f}_n$ is the optimal selector of $\hat{T}_n\hat{J}_n^*$; that is,

$$\hat{T}_n\hat{J}_n^* = \hat{J}_n^* = \hat{T}_{\hat{f}_n}\hat{J}_n^*,$$

where $\hat{T}_{\hat{f}_n}$ is defined as

$$\hat{T}_{\hat{f}_n} u(z) := \int \left[c(x,\hat{f}_n(x)) + \beta\int_{\mathsf{Z}} u(y)p(dy \mid x,\hat{f}_n(x))\right]\nu_{n,i_n(z)}(dx).$$

Define analogously

$$T_{\hat{f}_n} u(z) := c(z,\hat{f}_n(z)) + \beta\int_{\mathsf{Z}} u(y)p(dy \mid z,\hat{f}_n(z)).$$

It can be proved that both $\hat{T}_{\hat{f}_n}$ and $T_{\hat{f}_n}$ are contraction operators on $B^l(\mathsf{Z})$ with modulus β, and it is known that the fixed point of $T_{\hat{f}_n}$ is the true cost function of the stationary policy $\hat{f}_n$ (i.e., $J(\hat{f}_n,z)$).

Lemma 2.4. $\|\hat{T}_{\hat{f}_n} u - T_{\hat{f}_n} u\| \to 0$ *as* $n \to \infty$, *for any* $u \in C_b(\mathsf{Z})$.

Proof. The statement follows from the uniform continuity of the function $c(z,a)+\beta \int_{\mathsf{Z}} u(y)\, p(dy \mid z,a)$ and that $\hat{f}_n$ is constant over the sets $\mathcal{S}_{n,i}$. □

Now, we prove the main result of this section.

Proof of Theorem 2.2. We have

$$\begin{aligned}\|J(\hat{f}_n,\cdot) - J^*\| &\le \|T_{\hat{f}_n} J(\hat{f}_n,\cdot) - T_{\hat{f}_n} J^*\| + \|T_{\hat{f}_n} J^* - \hat{T}_{\hat{f}_n} J^*\| + \|\hat{T}_{\hat{f}_n} J^* - \hat{T}_{\hat{f}_n} \hat{J}_n^*\| + \|\hat{J}_n^* - J^*\| \\ &\le \beta \|J(\hat{f}_n,\cdot) - J^*\| + \|T_{\hat{f}_n} J^* - \hat{T}_{\hat{f}_n} J^*\| + \beta \|J^* - \hat{J}_n^*\| + \|\hat{J}_n^* - J^*\|.\end{aligned}$$

Hence, we obtain

$$\|J(\hat{f}_n,\cdot) - J^*\| \le \frac{\|T_{\hat{f}_n} J^* - \hat{T}_{\hat{f}_n} J^*\| + (1+\beta)\|\hat{J}_n^* - J^*\|}{1-\beta}.$$

The result follows from Lemma 2.4 and Theorem 2.4. □

2.2. Average Cost

In this section we impose some new conditions on the components of the original MDP in addition to Assumption 2.1 to solve **(Q1)** for the average cost. A version of the first two conditions was imposed in Vega-Amaya [38] and Jaśkiewicz and Nowak [26] to show the existence of the solution to the Average Cost Optimality Equation (ACOE) and the optimal stationary policy.

Assumption 2.2. *Suppose Assumption 2.1 holds with item* (b) *replaced by condition* (f) *below. In addition, there exist a nontrivial finite measure* ζ *on* Z, *a nonnegative measurable function* θ *on* $\mathsf{Z}\times\mathsf{A}$, *and a constant* $\lambda \in (0,1)$ *such that for all* $(z,a) \in \mathsf{Z}\times\mathsf{A}$,

(d) $p(B \mid z,a) \ge \zeta(B)\theta(z,a)$ *for all* $B \in \mathcal{B}(\mathsf{Z})$,
(e) $(1-\lambda)/(\zeta(\mathsf{Z})) \le \theta(z,a)$, *and*
(f) *the stochastic kernel* $p(\cdot \mid z,a)$ *is continuous in* (z,a) *with respect to the total variation distance.*

Throughout this section, it is assumed that Assumption 2.2 holds. Observe that any deterministic stationary policy f defines a stochastic kernel $p(\cdot \mid z, f(z))$ on Z given Z which is the transition probability of the Markov chain $\{z_t\}_{t=1}^{\infty}$ (state process) induced by f. For any $t \ge 1$, let us write $p^t(\cdot \mid z, f(z))$ to denote the t-step transition probability of this Markov chain given the initial point z; that is, $p^t(\cdot \mid z, f(z))$ is recursively defined as

$$p^{t+1}(\cdot \mid z, f(z)) = \int_{\mathsf{Z}} p(\cdot \mid x, f(x)) p^t(dx \mid z, f(z)).$$

To study average cost optimal control problems, it is in general assumed that there exists an invariant distribution under any stationary control policy, so that the average cost of any stationary policy can be written as an integral of the one-stage cost function with respect to this invariant distribution. With this representation, one can then deduce the optimality of stationary policies using the linear programming or the convex analytic methods (see Hernández-Lerma and Lasserre [21], Borkar [8]). However, to solve the approximation problem for the average cost, we need, in addition to the existence of an invariant distribution, the convergence of t-step transition probabilities to the invariant distribution, at some rate, for both the original and the reduced problems. Therefore, it is crucial to impose proper conditions on the original model so that, on the one hand, they guarantee the convergence of t-step transition probabilities to the invariant distribution for all stationary policies for the original system and, on the other hand, one is able to show that similar conditions are satisfied by the reduced problems. Conditions (d) and (e) in Assumption 2.2 are examples of such conditions that were also used in the literature extensively. Indeed, if we define the weight function $w \equiv 1$, then condition (e) corresponds to the so-called "drift inequality": for all $(z,a) \in \mathsf{Z}\times\mathsf{A}$,

$$\int_{\mathsf{Z}} w(y)\, p(dy \mid z,a) \le \lambda w(z) + \zeta(w)\theta(z,a),$$

and condition (d) corresponds to the so-called "minorization" condition, both of which were used in literature for studying geometric ergodicity of Markov chains (see Hernández-Lerma and Lasserre [22], Meyn and Tweedie [29], and references therein).

The following theorem is a consequence of Vega-Amaya [38, Theorem 3.3], Gordienko and Hernandez-Lerma [18, Lemma 3.4], and Jaśkiewicz and Nowak [26, Theorem 3] and also holds with Assumption 2.2(f) replaced by Assumption 2.1(b).

Theorem 2.5. *For any $f \in \mathbb{F}$, the stochastic kernel $p(\cdot \mid z, f(z))$ is positive Harris recurrent with unique invariant probability measure μ_f. Therefore, we have*

$$V(f,z) = \int_{\mathsf{Z}} c(z, f(z))\mu_f(dz) =: \rho_f.$$

The Markov chain $\{z_t\}_{t=1}^{\infty}$ induced by f is geometrically ergodic; that is, there exist positive real numbers R and $\kappa < 1$ such that for every $z \in \mathsf{Z}$

$$\sup_{f\in\mathbb{F}} \|p^t(\cdot \mid z, f(z)) - \mu_f\|_{\mathrm{TV}} \le R\kappa^t,$$

where R and κ continuously depend on $\zeta(\mathsf{Z})$ and λ. Finally, there exist $f^* \in \mathbb{F}$ and $h^* \in B(\mathsf{Z})$ such that the triplet (h^*, f^*, ρ_{f^*}) satisfies the average cost optimality equality (ACOE); i.e.,

$$\rho_{f^*} + h^*(z) = \min_{a\in\mathsf{A}}\left[c(z,a) + \int_{\mathsf{Z}} h^*(y)p(dy \mid z,a)\right] = c(z, f^*(z)) + \int_{\mathsf{Z}} h^*(y)p(dy \mid z, f^*(z)),$$

and therefore,

$$\inf_{\pi\in\Pi} V(\pi, z) =: V^*(z) = \rho_{f^*}.$$

For each n, define the one-stage cost function $b_n\colon \mathsf{Z}\times\mathsf{A} \to [0,\infty)$ and the stochastic kernel q_n on Z given $\mathsf{Z}\times\mathsf{A}$ as

$$b_n(z,a) := \int c(x,a)\nu_{n,i_n(z)}(dx), \qquad q_n(\cdot \mid z,a) := \int p(\cdot \mid x,a)\nu_{n,i_n(z)}(dx).$$

Observe that c_n (i.e., the one-stage cost function of MDP_n) is the restriction of b_n to Z_n, and p_n (i.e., the stochastic kernel of MDP_n) is the pushforward of the measure q_n with respect to Q_n; that is, $c_n(z_{n,i}, a) = b_n(z_{n,i}, a)$ for all $i = 1, \ldots, k_n$ and $p_n(\cdot \mid z_{n,i}, a) = Q_n \times q_n(\cdot \mid z_{n,i}, a)$.

For each n, let $\widehat{\mathrm{MDP}}_n$ be defined as a Markov decision process with the following components: Z is the state space, A is the action space, q_n is the transition probability, and c is the one-stage cost function. Similarly, let $\widetilde{\mathrm{MDP}}_n$ be defined as a Markov decision process with the following components: Z is the state space, A is the action space, q_n is the transition probability, and b_n is the one-stage cost function. History spaces, policies, and cost functions are defined in a similar way as before. The models $\widehat{\mathrm{MDP}}_n$ and $\widetilde{\mathrm{MDP}}_n$ are used as transitions between the original MDP and MDP_n in a similar way as the operators F_n and $\hat{T}_n$ were used as transitions between T and T_n for the discounted cost. We note that a similar technique was used in the proof of Ortner [30, Theorem 2], which studied the approximation problem for finite state-action MDPs. In Ortner [30] the one-stage cost function is first perturbed and then the transition probability is perturbed. We first perturb the transition probability and then the cost function. However, our proof method is otherwise quite different from that of Ortner [30, Theorem 2] since Ortner [30] assumes finite state and action spaces.

We note that a careful analysis of $\widetilde{\mathrm{MDP}}_n$ reveals that its Bellman optimality operator is essentially the operator $\hat{T}_n$. Hence, the value function of $\widetilde{\mathrm{MDP}}_n$ is the piecewise constant extension of the value function of MDP_n for the discounted cost. A similar conclusion will be made for the average cost in Lemma 2.5.

First, notice that if we define

$$\theta_n(z,a) := \int \theta(y,a)\nu_{n,i_n(z)}(dy),$$
$$\zeta_n := Q_n \times \zeta \quad \text{(i.e., pushforward of } \zeta \text{ with respect to } Q_n),$$

then it is straightforward to prove that for all n, both $\widehat{\mathrm{MDP}}_n$ and $\widetilde{\mathrm{MDP}}_n$ satisfy Assumption 2.2(d), (e) when θ is replaced by θ_n, and Assumption 2.2(d), (e) is true for MDP_n when θ and ζ are replaced by the restriction of θ_n to Z_n and ζ_n, respectively.

Hence, Theorem 2.5 holds (with the same R and κ) for $\widehat{\mathrm{MDP}}_n$, $\widetilde{\mathrm{MDP}}_n$, and MDP_n for all n. Therefore, we denote by $\hat{f}_n^*$, $\tilde{f}_n^*$ and f_n^* the optimal stationary policies of $\widehat{\mathrm{MDP}}_n$, $\widetilde{\mathrm{MDP}}_n$, and MDP_n with the corresponding average costs $\hat{\rho}^n_{\hat{f}_n^*}$, $\tilde{\rho}^n_{\tilde{f}_n^*}$, and $\rho^n_{f_n^*}$, respectively.

Furthermore, we also write $\hat{\rho}^n_f$, $\tilde{\rho}^n_f$, and ρ^n_f to denote the average cost of any stationary policy f for $\widehat{\mathrm{MDP}}_n$, $\widetilde{\mathrm{MDP}}_n$, and MDP_n, respectively. The corresponding invariant probability measures are also denoted in a similar manner, with μ replacing ρ.

The following lemma essentially says that MDP_n and $\widetilde{\mathrm{MDP}}_n$ are not very different.

Lemma 2.5. *The stationary policy given by the piecewise constant extension of the optimal policy f_n^* of MDP_n to Z (i.e., $f_n^* \circ Q_n$) is optimal for $\widetilde{\mathrm{MDP}}_n$ with the same cost function $\rho^n_{f_n^*}$. Hence, $\tilde{f}_n^* = f_n^* \circ Q_n$ and $\tilde{\rho}^n_{\tilde{f}_n^*} = \rho^n_{f_n^*}$.*

Proof. Note that by Theorem 2.5 there exists $h_n^* \in B(\mathsf{Z}_n)$ such that the triplet $(h_n^*, f_n^*, \rho^n_{f_n^*})$ satisfies the ACOE for MDP_n. But it is straightforward to show that the triplet $(h_n^* \circ Q_n, f_n^* \circ Q_n, \rho^n_{f_n^*})$ satisfies the ACOE for $\widetilde{\mathrm{MDP}}_n$. By Gordienko and Hernandez-Lerma [18, Lemma 5.2], this implies that $f_n^* \circ Q_n$ is an optimal stationary policy for $\widetilde{\mathrm{MDP}}_n$ with cost function $\rho^n_{f_n^*}$. Hence $\tilde{f}_n^* = f_n^* \circ Q_n$ and $\tilde{\rho}^n_{\tilde{f}_n^*} = \rho^n_{f_n^*}$. □

The following theorem is the main result of this section. It states that if one applies the piecewise constant extension of the optimal stationary policy of MDP_n to the original MDP, the resulting cost function will converge to the value function of the original MDP.

Theorem 2.6. *The average cost of the optimal policy $\tilde{f}_n^*$ for $\widetilde{\mathrm{MDP}}_n$, obtained by extending the optimal policy f_n^* of MDP_n to Z, converges to the optimal value function $J^* = \rho_{f^*}$ of the original MDP; i.e.,*

$$\lim_{n\to\infty} |\rho_{\tilde{f}_n^*} - \rho_{f^*}| = 0.$$

Hence, to find a near optimal policy for the original MDP, it is sufficient to compute the optimal policy of MDP_n for sufficiently large n, and then extend this policy to the original state space.

To show the statement of Theorem 2.6 we will prove a series of auxiliary results.

Lemma 2.6. *For all $t \geq 1$ we have*

$$\lim_{n\to\infty} \sup_{(y,f)\in\mathsf{Z}\times\mathbb{F}} \|p^t(\cdot \mid y, f(y)) - q_n^t(\cdot \mid y, f(y))\|_{\mathrm{TV}} = 0.$$

Proof. We will prove the lemma by induction. Note that if one views the stochastic kernel $p(\cdot \mid z, a)$ as a mapping from $\mathsf{Z}\times\mathsf{A}$ to $\mathscr{P}(\mathsf{Z})$, then Assumption 2.2(f) implies that this mapping is continuous, and therefore uniformly continuous, when $\mathscr{P}(\mathsf{Z})$ is equipped with the metric induced by the total variation distance.

For $t = 1$ the claim holds by the following argument:

$$\begin{aligned}
\sup_{(y,f)\in\mathsf{Z}\times\mathbb{F}} \|p(\cdot \mid y, f(y)) - q_n(\cdot \mid y, f(y))\|_{\mathrm{TV}} &:= 2 \sup_{(y,f)\in\mathsf{Z}\times\mathbb{F}} \sup_{D\in\mathscr{B}(\mathsf{Z})} |p(D \mid y, f(y)) - q_n(D \mid y, f(y))| \\
&\leq 2 \sup_{(y,f)\in\mathsf{Z}\times\mathbb{F}} \sup_{D\in\mathscr{B}(\mathsf{Z})} \int |p(D \mid y, f(y)) - p(D \mid z, f(y))| \nu_{n,i_n(y)}(dz) \\
&\leq \sup_{(y,f)\in\mathsf{Z}\times\mathbb{F}} \int \|p(\cdot \mid y, f(y)) - p(\cdot \mid z, f(y))\|_{\mathrm{TV}} \nu_{n,i_n(y)}(dz) \\
&\leq \sup_{y\in\mathsf{Z}} \sup_{(z,a)\in\mathscr{S}_{n,i_n(y)}\times\mathsf{A}} \|p(\cdot \mid y, a) - p(\cdot \mid z, a)\|_{\mathrm{TV}}.
\end{aligned}$$

As the mapping $p(\cdot \mid z, a)$: $\mathsf{Z}\times\mathsf{A} \to \mathscr{P}(\mathsf{Z})$ is uniformly continuous with respect to the total variation distance and $\max_{n,i} \mathrm{diam}(\mathscr{S}_{n,i}) \to 0$ as $n \to \infty$, the result follows. Assume the claim is true for $t \geq 1$; then we have

$$\begin{aligned}
&\sup_{(y,f)\in\mathsf{Z}\times\mathbb{F}} \|p^{t+1}(\cdot \mid y, f(y)) - q_n^{t+1}(\cdot \mid y, f(y))\|_{\mathrm{TV}} \\
&\quad := \sup_{(y,f)\in\mathsf{Z}\times\mathbb{F}} \sup_{\|g\|\leq 1} \left| \int_{\mathsf{Z}} g(x) p^{t+1}(dx \mid y, f(y)) - \int_{\mathsf{Z}} g(x) q_n^{t+1}(dx \mid y, f(y)) \right| \\
&\quad \leq \sup_{(y,f)\in\mathsf{Z}\times\mathbb{F}} \left(\sup_{\|g\|\leq 1} \left| \int_{\mathsf{Z}} \int_{\mathsf{Z}} g(x) p(dx \mid z, f(z)) p^t(dz \mid y, f(y)) - \int_{\mathsf{Z}} \int_{\mathsf{Z}} g(x) p(dx \mid z, f(z)) q_n^t(dz \mid y, f(y)) \right| \right. \\
&\qquad \left. + \sup_{\|g\|\leq 1} \left| \int_{\mathsf{Z}} \int_{\mathsf{Z}} g(x) p(dx \mid z, f(z)) q_n^t(dz \mid y, f(y)) - \int_{\mathsf{Z}} \int_{\mathsf{Z}} g(x) q_n(dx \mid z, f(z)) q_n^t(dz \mid y, f(y)) \right| \right) \\
&\quad \leq \sup_{(y,f)\in\mathsf{Z}\times\mathbb{F}} \|p^t(\cdot \mid y, f(y)) - q_n^t(\cdot \mid y, f(y))\|_{\mathrm{TV}} + \sup_{(z,f)\in\mathsf{Z}\times\mathbb{F}} \|p(\cdot \mid z, f(z)) - q_n(\cdot \mid z, f(z))\|_{\mathrm{TV}} \qquad (5)
\end{aligned}$$

where the last inequality follows from the following property of the total variation distance: for any $h \in \mathscr{B}(\mathsf{Z})$ and $\mu, \nu \in \mathscr{P}(\mathsf{Z})$ we have $|\int_{\mathsf{Z}} h(z)\mu(dz) - \int_{\mathsf{Z}} h(z)\nu(dz)| \leq \|h\| \|\mu - \nu\|_{\mathrm{TV}}$. By the first step of the proof and the induction hypothesis, the last term converges to zero as $n \to \infty$. This completes the proof. □

Remark 2.3. This is the point where we need the continuity of the transition probability p with respect to the total variation distance. If we assume that the stochastic kernel p is only weakly or setwise continuous, then it does not seem possible to prove a result similar to Lemma 2.6 for the weak and the setwise topologies.

Using Lemma 2.6 we prove the following result.

Lemma 2.7. *We have* $\sup_{f\in\mathbb{F}}|\hat{\rho}^n_f-\rho_f|\to 0$ *as* $n\to\infty$, *where* $\hat{\rho}^n_f$ *is the cost function of the policy* f *for* $\widehat{\mathrm{MDP}}_n$ *and* ρ_f *is the cost function of the policy* f *for the original MDP.*

Proof. For any $t\geq 1$ and $y\in\mathsf{Z}$ we have

$$\begin{aligned}\sup_{f\in\mathbb{F}}|\hat{\rho}^n_f-\rho_f| &= \sup_{f\in\mathbb{F}}\left|\int_{\mathsf{Z}}c(z,f(z))\hat{\mu}^n_f(dz)-\int_{\mathsf{Z}}c(z,f(z))\mu_f(dz)\right|\\ &\leq \sup_{f\in\mathbb{F}}\left|\int_{\mathsf{Z}}c(z,f(z))\hat{\mu}^n_f(dz)-\int_{\mathsf{Z}}c(z,f(z))\,q^t_n(dz\mid y,f(y))\right|\\ &\quad+\sup_{f\in\mathbb{F}}\left|\int_{\mathsf{Z}}c(z,f(z))\,q^t_n(dz\mid y,f(y))-\int_{\mathsf{Z}}c(z,f(z))\,p^t(dz\mid y,f(y))\right|\\ &\quad+\sup_{f\in\mathbb{F}}\left|\int_{\mathsf{Z}}c(z,f(z))\,p^t(dz\mid y,f(y))-\int_{\mathsf{Z}}c(z,f(z))\mu_f(dz)\right|\\ &\leq 2R\kappa^t\|c\|+\|c\|\sup_{(y,f)\in\mathsf{Z}\times\mathbb{F}}\|q^t_n(\cdot\mid y,f(y))-p^t(\cdot\mid y,f(y))\|_{\mathrm{TV}}\quad\text{(by Theorem 2.5(ii)),}\end{aligned}$$

where R and κ are the constants in Theorem 2.5. Thus, the result follows from Lemma 2.6. □

The following theorem states that the value function of $\widehat{\mathrm{MDP}}_n$ converges to the value function of the original MDP.

Lemma 2.8. *We have* $|\hat{\rho}^n_{\hat{f}^*_n}-\rho_{f^*}|\to 0$ *as* $n\to\infty$.

Proof. Notice that

$$|\hat{\rho}^n_{\hat{f}^*_n}-\rho_{f^*}|=\max(\hat{\rho}^n_{\hat{f}^*_n}-\rho_{f^*},\rho_{f^*}-\hat{\rho}^n_{\hat{f}^*_n})\leq\max(\hat{\rho}^n_{f^*}-\rho_{f^*},\rho_{\hat{f}^*_n}-\hat{\rho}^n_{\hat{f}^*_n})\leq\sup_f|\hat{\rho}^n_f-\rho_f|.$$

Thus, the result follows from Lemma 2.7. □

Lemma 2.9. *We have* $\sup_{f\in\mathbb{F}}|\tilde{\rho}^n_f-\hat{\rho}^n_f|\to 0$ *as* $n\to\infty$.

Proof. It is straightforward to show that $b_n\to c$ uniformly. Since the probabilistic structure of $\widetilde{\mathrm{MDP}}_n$ and $\widehat{\mathrm{MDP}}_n$ are the same (i.e., $\hat{\mu}^n_f=\tilde{\mu}^n_f$ for all f), we have

$$\begin{aligned}\sup_{f\in\mathbb{F}}|\tilde{\rho}^n_f-\hat{\rho}^n_f| &= \sup_{f\in\mathbb{F}}\left|\int_{\mathsf{Z}}b_n(z,f(z))\hat{\mu}^n_f(dz)-\int_{\mathsf{Z}}c(z,f(z))\hat{\mu}^n_f(dz)\right|\leq\sup_{f\in\mathbb{F}}\int_{\mathsf{Z}}|b_n(z,f(z))-c(z,f(z))|\hat{\mu}^n_f(dz)\\ &\leq\|b_n-c\|.\end{aligned}$$

This completes the proof. □

The next lemma states that the difference between the value functions of $\widetilde{\mathrm{MDP}}_n$ and $\widehat{\mathrm{MDP}}_n$ converges to zero.

Lemma 2.10. *We have* $|\tilde{\rho}^n_{\tilde{f}^*_n}-\hat{\rho}^n_{\hat{f}^*_n}|\to 0$ *as* $n\to\infty$.

Proof. See the proof of Lemma 2.8. □

The following result states that if we apply the optimal policy of $\widetilde{\mathrm{MDP}}_n$ to $\widehat{\mathrm{MDP}}_n$, then the resulting cost converges to the value function of $\widehat{\mathrm{MDP}}_n$.

Lemma 2.11. *We have* $|\hat{\rho}^n_{\tilde{f}^*_n}-\hat{\rho}^n_{\hat{f}^*_n}|\to 0$ *as* $n\to\infty$.

Proof. Since $|\hat{\rho}^n_{\tilde{f}^*_n}-\hat{\rho}^n_{\hat{f}^*_n}|\leq|\hat{\rho}^n_{\tilde{f}^*_n}-\tilde{\rho}^n_{\tilde{f}^*_n}|+|\tilde{\rho}^n_{\tilde{f}^*_n}-\hat{\rho}^n_{\hat{f}^*_n}|$, the result follows from Lemmas 2.9 and 2.10. □

Now we are ready to prove the main result of this section.

Proof of Theorem 2.6. We have $|\rho_{\tilde{f}^*_n}-\rho_{f^*}|\leq|\rho_{\tilde{f}^*_n}-\hat{\rho}^n_{\tilde{f}^*_n}|+|\hat{\rho}^n_{\tilde{f}^*_n}-\hat{\rho}^n_{\hat{f}^*_n}|+|\hat{\rho}^n_{\hat{f}^*_n}-\rho^n_{f^*}|$. The result now follows from Lemmas 2.7, 2.11, and 2.8. □

3. Finite State Approximations of MDPs with Noncompact State Space

In this section we consider **(Q1)** for noncompact state MDPs with unbounded one-stage cost. To solve **(Q1)**, we use the following strategy: (i) first, we define a sequence of compact-state MDPs to approximate the original MDP; (ii) we use Theorems 2.2 and 2.6 to approximate the compact-state MDPs by finite-state models; and (iii) we prove the convergence of the finite-state models to the original model. In fact, steps (ii) and (iii) will be accomplished simultaneously.

We impose the assumptions below on the components of the Markov decision process; additional assumptions will be imposed for the average cost problem. With the exception of the local compactness of the state space, these are the usual assumptions used in the literature for studying Markov decision processes with unbounded cost.

Assumption 3.1. (a) *The one-stage cost function c is continuous.*

(b) *The stochastic kernel $p(\cdot \mid x,a)$ is weakly continuous in (x,a).*

(c) X *is locally compact and* A *is compact.*

(d) *There exist nonnegative real numbers M and $\alpha \in [1, 1/\beta)$ and a continuous weight function $w\colon \mathsf{X} \to [1,\infty)$ such that for each $x \in \mathsf{X}$, we have*

$$\sup_{a\in\mathsf{A}} |c(x,a)| \le Mw(x), \tag{6}$$

$$\sup_{a\in\mathsf{A}} \int_{\mathsf{X}} w(y)\, p(dy \mid x,a) \le \alpha w(x), \tag{7}$$

and $\int_{\mathsf{X}} w(y)p(dy \mid x,a)$ is continuous in (x,a).

Since X is locally compact separable metric space, there exists a nested sequence of compact sets $\{K_n\}$ such that $K_n \subset \operatorname{int} K_{n+1}$ and $\mathsf{X} = \bigcup_{n=1}^{\infty} K_n$ Aliprantis and Border [1, Lemma 2.76, p. 58].

Lemma 3.1. *For any compact subset K of X and for any $\varepsilon > 0$, there exists a compact subset K_ε of X such that*

$$\sup_{(x,a)\in K\times\mathsf{A}} \int_{K_\varepsilon^c} w(y)\, p(dy \mid x,a) < \varepsilon,$$

where D^c denotes the complement of the set D.

Proof. We prove the lemma by contradiction. Assume the claim is wrong. Since every compact subset K of X is a subset of K_n for some n, the negation of the above lemma is equivalent to the following statement: there exists a compact set $K \subset \mathsf{X}$ and $\varepsilon > 0$ such that for all $n \ge 1$ we have

$$\sup_{(x,a)\in K\times\mathsf{A}} \int_{K_n^c} w(y)\, p(dy \mid x,a) \ge \varepsilon.$$

Note that w is integrable with respect to the probability measures in the set $\{p(\cdot \mid x,a)\colon (x,a) \in K \times \mathsf{A}\}$ since

$$\sup_{(x,a)\in K\times\mathsf{A}} \int_{\mathsf{X}} w(y)\, p(dy \mid x,a) \le \alpha \sup_{x\in K} w(x) < \infty.$$

For each n, we prove that $\int_{(\operatorname{int} K_n)^c} w(y)p(dy \mid x,a)$ is an upper semicontinuous function on $K \times \mathsf{A}$. Recall that $\int_{\mathsf{X}} w(y)p(dy \mid x,a)$ is a continuous function of (x,a). Let $(x_k,a_k) \to (x,a)$ in $K \times \mathsf{A}$; then $p(\cdot \mid x_k,a_k) \to p(\cdot \mid x,a)$ weakly and $\int_{\mathsf{X}} w(y)p(dy \mid x_k,a_k) \to \int_{\mathsf{X}} w(y)p(dy \mid x,a)$ by our assumption. If we take $f_k = g_k = f = g = w$ in Serfozo [35, Theorem 3.3], this result implies that $\nu_k(\cdot) \to \nu(\cdot)$ weakly, where

$$\nu_k(D) = \int_D w(y)p(dy \mid x_k,a_k) \qquad \nu(D) = \int_D w(y)p(dy \mid x,a),$$

for all $D \in \mathcal{B}(\mathsf{X})$. Thus, by Bartoszynski [2, Theorem A] we have

$$\int_{(\operatorname{int} K_n)^c} w(y)\, p(dy \mid x,a) := \nu((\operatorname{int} K_n)^c) \ge \limsup_{k\to\infty} \nu_k((\operatorname{int} K_n)^c) := \limsup_{k\to\infty} \int_{(\operatorname{int} K_n)^c} w(y)\, p(dy \mid x_k,a_k).$$

Hence, $\int_{(\text{int}\, K_n)^c} w(y)\, p(dy \mid x,a)$ is upper semicontinuous. Since $K \times \mathsf{A}$ is compact, there exists $(x_n, a_n) \in K \times \mathsf{A}$ such that

$$\sup_{(x,a)\in K\times \mathsf{A}} \int_{(\text{int}\, K_n)^c} w(y)\, p(dy \mid x,a) = \int_{(\text{int}\, K_n)^c} w(y)\, p(dy \mid x_n, a_n).$$

The sequence $\{(x_n, a_n)\}$ (being a sequence in a compact set $K \times \mathsf{A}$) has an converging subsequence $\{(x_{n_k}, a_{n_k})\}$ with the limit $(x,a) \in K \times \mathsf{A}$. Thus, for all $m \geq 2$, we have

$$\begin{aligned}\int_{K_{m-1}^c} w(y)\, p(dy \mid x,a) &\geq \int_{(\text{int}\, K_m)^c} w(y)\, p(dy \mid x,a)\\ &\geq \limsup_{k\to\infty} \int_{(\text{int}\, K_m)^c} w(y)\, p(dy \mid x_{n_k}, a_{n_k})\\ &\geq \limsup_{k\to\infty} \int_{(\text{int}\, K_{n_k})^c} w(y)\, p(dy \mid x_{n_k}, a_{n_k}) \geq \varepsilon,\end{aligned}$$

where the third inequality follows from the fact that $(\text{int}\, K_m)^c \supset (\text{int}\, K_{n_k})^c$ for k sufficiently large. But this is a contradiction because w is $p(\cdot \mid x,a)$ integrable. □

Let $\{\nu_n\}$ be a sequence of probability measures such that for each $n \geq 1$, $\nu_n \in \mathcal{P}(K_n^c)$ and

$$\gamma_n := \int_{K_n^c} w(x)\nu_n(dx) < \infty, \tag{8}$$

$$\gamma = \sup_n \tau_n := \sup_n \max\left\{0, \sup_{(x,a)\in \mathsf{X}\times\mathsf{A}} \int_{K_n^c} (\gamma_n - w(y))p(dy \mid x,a)\right\} < \infty. \tag{9}$$

For example, such probability measures can be constructed by choosing $x_n \in K_n^c$ such that $w(x_n) < \inf_{x\in K_n^c} w(x) + 1/n$ and letting $\nu_n(\cdot) = \delta_{x_n}(\cdot)$.

Similar to the finite-state MDP construction in Section 2, we define a sequence of compact-state MDPs, denoted as c-MDP$_n$, to approximate the original model. To this end, for each n let $\mathsf{X}_n = K_n \cup \{\Delta_n\}$, where $\Delta_n \in K_n^c$ is a so-called pseudostate. We define the transition probability p_n on X_n given $\mathsf{X}_n \times \mathsf{A}$ and the one-stage cost function $c_n \colon \mathsf{X}_n \times \mathsf{A} \to [0,\infty)$ by

$$p_n(\cdot \mid x,a) = \begin{cases} p(\cdot \cap K_n \mid x,a) + p(K_n^c \mid x,a)\delta_{\Delta_n}, & \text{if } x \in K_n,\\ \displaystyle\int_{K_n^c} (p(\cdot \cap K_n \mid z,a) + p(K_n^c \mid z,a)\delta_{\Delta_n})\nu_n(dz), & \text{if } x = \Delta_n,\end{cases}$$

$$c_n(x,a) = \begin{cases} c(x,a), & \text{if } x \in K_n,\\ \displaystyle\int_{K_n^c} c(z,a)\nu_n(dz), & \text{if } x = \Delta_n.\end{cases}$$

With these definitions, c-MDP$_n$ is defined as a Markov decision process with the components $(\mathsf{X}_n, \mathsf{A}, p_n, c_n)$. History spaces, policies, and cost functions are defined in a similar way as in the original model. Let Π_n, Φ_n, and $\mathbb{F}_n$ denote the set of all policies, randomized stationary policies, and deterministic stationary policies of c-MDP$_n$, respectively. For each policy $\pi \in \Pi_n$ and initial distribution $\mu \in \mathcal{P}(\mathsf{X}_n)$, we denote the cost functions for c-MDP$_n$ by $J_n(\pi,\mu)$ and $V_n(\pi,\mu)$.

To obtain the main result of this section, we introduce, for each n, another MDP, denoted by $\overline{\text{MDP}}_n$, with the components $(\mathsf{X}, \mathsf{A}, q_n, b_n)$ where

$$q_n(\cdot \mid x,a) = \begin{cases} p(\cdot \mid x,a), & \text{if } x \in K_n,\\ \displaystyle\int_{K_n^c} p(\cdot \mid z,a)\nu_n(dz), & \text{if } x \in K_n^c,\end{cases} \qquad b_n(x,a) = \begin{cases} c(x,a), & \text{if } x \in K_n,\\ \displaystyle\int_{K_n^c} c(z,a)\nu_n(dz), & \text{if } x \in K_n^c.\end{cases}$$

For each policy $\pi \in \Pi$ and initial distribution $\mu \in \mathcal{P}(\mathsf{X})$, we denote the cost functions for $\overline{\text{MDP}}_n$ by $\bar{J}_n(\pi,\mu)$ and $\bar{V}_n(\pi,\mu)$.

3.1. Discounted Cost

In this section we consider **(Q1)** for the discounted cost criterion with a discount factor $\beta \in (0,1)$. Throughout this section, it is assumed that Assumption 3.1 holds. The following result states that c-MDP$_n$ and $\overline{\text{MDP}}_n$ are equivalent for the discounted cost.

Lemma 3.2. *We have*

$$\bar{J}_n^*(x) = \begin{cases} J_n^*(x), & \text{if } x \in K_n, \\ J_n^*(\Delta_n), & \text{if } x \in K_n^c, \end{cases} \tag{10}$$

where $\bar{J}_n^*$ is the discounted value function of $\overline{\text{MDP}}_n$ and J_n^* is the discounted value function of c-MDP$_n$, provided that there exist optimal deterministic stationary policies for $\overline{\text{MDP}}_n$ and c-MDP$_n$. Furthermore, if, for any deterministic stationary policy $f \in \mathbb{F}_n$, we define $\bar{f}(x) = f(x)$ on K_n and $\bar{f}(x) = f(\Delta_n)$ on K_n^c, then

$$\bar{J}_n(\bar{f}, x) = \begin{cases} J_n(f, x), & \text{if } x \in K_n, \\ J_n(f, \Delta_n), & \text{if } x \in K_n^c. \end{cases} \tag{11}$$

In particular, if the deterministic stationary policy $f_n^* \in \mathbb{F}_n$ is optimal for c-MDP$_n$, then its extension $\bar{f}_n^*$ to X is also optimal for $\overline{\text{MDP}}_n$.

Proof. The proof of (11) is a consequence of the following facts: $b_n(x,a) = b_n(y,a)$ and $q_n(\cdot \mid x,a) = q_n(\cdot \mid y,a)$ for all $x, y \in K_n^c$ and $a \in \mathsf{A}$. In other words, K_n^c in $\overline{\text{MDP}}_n$ behaves like the pseudostate Δ_n in c-MDP$_n$ when $\bar{f}$ is applied to $\overline{\text{MDP}}_n$.

Let $\bar{\mathbb{F}}_n$ denote the set of all deterministic stationary policies in $\mathbb{F}$ that are obtained by extending policies in $\mathbb{F}_n$ to X. If we can prove that $\min_{f \in \mathbb{F}} \bar{J}_n(f,x) = \min_{f \in \bar{\mathbb{F}}_n} \bar{J}_n(f,x)$ for all $x \in \mathsf{X}$, then (10) follows from (11). Let $f \in \mathbb{F} \backslash \bar{\mathbb{F}}_n$. We have two cases: (i) $\bar{J}_n(f,z) = \bar{J}_n(f,y)$ for all $z, y \in K_n^c$ or (ii) there exists $z, y \in K_n^c$ such that $\bar{J}_n(f,z) < \bar{J}_n(f,y)$.

For the case (i), if we define the deterministic Markov policy π^0 as $\pi^0 = \{f_0, f, f, \ldots\}$, where $f_0(x) = f(z)$ on K_n^c for some fixed $z \in K_n^c$ and $f_0(x) = f(x)$ on K_n, then using the expression

$$\bar{J}_n(\pi^0, x) = b_n(x, f_0(x)) + \beta \int_{\mathsf{X}} \bar{J}_n(f, x')\, q_n(dx' \mid x, f_0(x)), \tag{12}$$

it is straightforward to show that $\bar{J}_n(\pi^0,x) = \bar{J}_n(f,x)$ on K_n and $\bar{J}_n(\pi^0,x) = \bar{J}_n(f,z)$ on K_n^c. Therefore, $\bar{J}_n(\pi^0,x) = \bar{J}_n(f,x)$ for all $x \in \mathsf{X}$ since $\bar{J}_n(f,x) = \bar{J}_n(f,z)$ for all $x \in K_n^c$. For all $t \geq 1$ define the deterministic Markov policy π^t as $\pi^t = \{f_0, \pi^{t-1}\}$. Analogously, one can prove that $\bar{J}_n(\pi^t,x) = \bar{J}_n(\pi^{t+1},x)$ for all $x \in \mathsf{X}$. Since $\bar{J}_n(\pi^t,x) \to \bar{J}_n(f_0,x)$ as $t \to \infty$, we have $\bar{J}_n(f_0,x) = \bar{J}_n(f,x)$ for all $x \in \mathsf{X}$, where $f_0 \in \bar{\mathbb{F}}_n$.

For the second case, if we again consider the deterministic Markov policy $\pi^0 = \{f_0, f, f, \ldots\}$, then by (12) we have $\bar{J}_n(\pi^0, y) = \bar{J}_n(f,z) < \bar{J}_n(f,y)$. Since $\min_{f \in \mathbb{F}} \bar{J}_n(f,y) \leq \bar{J}_n(\pi^0, y)$, this completes the proof. □

For each n, let us define w_n by letting $w_n(x) = w(x)$ on K_n and $w_n(x) = \int_{K_n^c} w(z) \nu_n(dz) =: \gamma_n$ on K_n^c. Hence, $w_n \in B(\mathsf{X})$ by (8).

Lemma 3.3. *For all n and $x \in \mathsf{X}$, the components of $\overline{\text{MDP}}_n$ satisfy the following*:

$$\sup_{a \in \mathsf{A}} |b_n(x,a)| \leq M w_n(x), \tag{13}$$

$$\sup_{a \in \mathsf{A}} \int_{\mathsf{X}} w_n(y)\, q_n(dy \mid x,a) \leq \alpha w_n(x) + \gamma, \tag{14}$$

where γ is the constant in (9).

Proof. It is straightforward to prove (13) by using the definitions of b_n and w_n and Equation (6). To prove (14), we have to consider two cases: $x \in K_n$ and $x \in K_n^c$. For the first case, $q_n(\cdot \mid x,a) = p(\cdot \mid x,a)$, and therefore, we have

$$\begin{aligned} \sup_{a \in \mathsf{A}} \int_{\mathsf{X}} w_n(y)\, p(dy \mid x,a) &= \sup_{a \in \mathsf{A}} \left\{ \int_{\mathsf{X}} w(y)\, p(dy \mid x,a) + \int_{K_n^c} (\gamma_n - w(y))\, p(dy \mid x,a) \right\} \\ &\leq \sup_{a \in \mathsf{A}} \int_{\mathsf{X}} w(y)\, p(dy \mid x,a) + \gamma \quad \text{(by (9))} \\ &\leq \alpha w(x) + \gamma = \alpha w_n(x) + \gamma \quad \text{(as } w_n = w \text{ on } K_n\text{)}. \end{aligned}$$

For $x \in K_n^c$, we have

$$\begin{aligned}\sup_{a\in\mathsf{A}} \int_{\mathsf{X}} w_n(y)\, q_n(dy \mid x,a) &= \sup_{a\in\mathsf{A}} \int_{K_n^c} \left(\int_{\mathsf{X}} w_n(y)\, p(dy \mid z,a) \right) \nu_n(dz) \\ &\le \int_{K_n^c} \left(\sup_{a\in\mathsf{A}} \int_{\mathsf{X}} w_n(y)\, p(dy \mid z,a) \right) \nu_n(dz) \\ &\le \int_{K_n^c} (\alpha w(z) + \gamma)\, \nu_n(dz) \qquad (15) \\ &= \alpha w_n(x) + \gamma,\end{aligned}$$

where (15) can be proved following the same arguments as for the case $x \in K_n$. This completes the proof. □

Note that if we define $c_{n,0}(x) = 1 + \sup_{a\in\mathsf{A}} |b_n(x,a)|$ and $c_{n,t}(x) = \sup_{a\in\mathsf{A}} \int_{\mathsf{X}} c_{n,t-1}(y)\, q_n(dy \mid x,a)$, by (13) and (14) and an induction argument, we obtain (see Hernández-Lerma and Lasserre [22, p. 46])

$$c_{n,t}(x) \le L w_n(x) \alpha^t + L\gamma \sum_{j=0}^{t-1} \alpha^j \quad \text{for all } x \in \mathsf{X}, \qquad (16)$$

where $L = 1 + M$. Let $\beta_0 > \beta$ be such that $\alpha\beta_0 < 1$ and let $C_n \colon \mathsf{X} \to [1,\infty)$ be defined by

$$C_n(x) = \sum_{t=0}^{\infty} \beta_0^t c_{n,t}(x).$$

For all $x \in \mathsf{X}$, by (16), we then have

$$C_n(x) := \sum_{t=0}^{\infty} \beta_0^t c_{n,t}(x) \le \frac{L}{1-\beta_0\alpha} w_n(x) + \frac{L\beta_0}{(1-\beta_0)(1-\beta_0\alpha)}\gamma =: L_1 w_n(x) + L_2. \qquad (17)$$

Hence $C_n \in B(\mathsf{X})$ as $w_n \in B(\mathsf{X})$. Moreover, for all $(x,a) \in \mathsf{X} \times \mathsf{A}$, C_n satisfies (see Hernández-Lerma and Lasserre [22, p. 45])

$$\int_{\mathsf{X}} C_n(y)\, q_n(dy \mid x,a) = \sum_{t=0}^{\infty} \beta_0^t \int_{\mathsf{X}} c_{n,t}(y)\, q_n(dy \mid x,a) \le \sum_{t=0}^{\infty} \beta_0^t c_{n,t+1}(x) \le \frac{1}{\beta_0} \sum_{t=0}^{\infty} \beta_0^t c_{n,t}(x) = \alpha_0 C_n(x),$$

where $\alpha_0 := 1/\beta_0$ and $\alpha_0\beta < 1$ since $\beta_0 > \beta$. Therefore, for all $x \in \mathsf{X}$, components of $\overline{\mathrm{MDP}}_n$ satisfy

$$\sup_{a\in\mathsf{A}} |b_n(x,a)| \le C_n(x) \qquad (18)$$

$$\sup_{a\in\mathsf{A}} \int_{\mathsf{X}} C_n(y)\, q_n(dy \mid x,a) \le \alpha_0 C_n(x). \qquad (19)$$

Since $C_n \in B(\mathsf{X})$, the Bellman optimality operator $\bar{T}_n$ of $\overline{\mathrm{MDP}}_n$ maps $B^l(\mathsf{X})$ into $B^l(\mathsf{X})$ and is given by

$$\bar{T}_n u(x) = \inf_{a\in\mathsf{A}} \left[b_n(x,a) + \beta \int_{\mathsf{X}} u(y)\, q_n(dy \mid x,a) \right] = \begin{cases} \inf_{a\in\mathsf{A}} \left[c(x,a) + \beta \int_{\mathsf{X}} u(y)\, p(dy \mid x,a) \right], & \text{if } x \in K_n, \\ \inf_{a\in\mathsf{A}} \int_{K_n^c} \left[c(z,a) + \beta \int_{\mathsf{X}} u(y)\, p(dy \mid z,a) \right] \nu_n(dz), & \text{if } x \in K_n^c. \end{cases}$$

Then successive approximations to the discounted value function of $\overline{\mathrm{MDP}}_n$ are given by $v_n^0 = 0$ and $v_n^{t+1} = \bar{T}_n v_n^t$ ($t \ge 1$). Since $\alpha_0\beta < 1$, it can be proved as in Hernández-Lerma and Lasserre [22, Theorem 8.3.6, p. 47] and Hernández-Lerma and Lasserre [22, (8.3.34), p. 52] that

$$|v_n^t(x)|, |\bar{J}_n^*(x)| \le \frac{C_n(x)}{1-\sigma_0}, \quad \text{for all } x, \qquad (20)$$

$$\|v_n^t - \bar{J}_n^*\|_{C_n} \le \frac{\sigma_0^t}{1-\sigma_0^t}, \qquad (21)$$

where $\sigma_0 = \beta\alpha_0 < 1$.

Similar to v_n^t, let us define $v^0 = 0$ and $v^{t+1} = Tv^t$, where T: $B_w(\mathsf{X}) \to B_w(\mathsf{X})$, the Bellman optimality operator for the original MDP, is given by

$$Tu(x) = \inf_{a\in\mathsf{A}}\left[c(x,a) + \beta\int_{\mathsf{X}} u(y)\,p(dy \mid x,a)\right].$$

Again by Hernández-Lerma and Lasserre [22, Theorem 8.3.6, p. 47] and Hernández-Lerma and Lasserre [22, (8.3.34), p. 52], we have

$$|v^t(x)|, |J^*(x)| \le M\frac{w(x)}{1-\sigma}, \quad \text{for all } x, \tag{22}$$

$$\|v^t - J^*\|_w \le M\frac{\sigma^t}{1-\sigma}, \tag{23}$$

where $\sigma = \beta\alpha < 1$.

Lemma 3.4. *For any compact set $K \subset \mathsf{X}$, we have*

$$\lim_{n\to\infty}\sup_{x\in K}|v_n^t(x) - v^t(x)| = 0 \tag{24}$$

for all $t \ge 1$.

Proof. We prove (24) by induction on t. For $t = 1$, the claim trivially holds since any compact set $K \subset \mathsf{X}$ is inside K_n for sufficiently large n, and therefore, $b_n = c$ on K for sufficiently large n (recall $v_n^0 = v^0 = 0$). Assume the claim is true for $t \ge 1$. Fix any compact set K. Recall the definition of compact subsets K_ε of X in Lemma 3.1. By definition of q_n, b_n, and w_n, there exists $n_0 \ge 1$ such that for all $n \ge n_0$, $q_n = p$, $b_n = c$, and $w_n = w$ on K. With these observations, for each $n \ge n_0$ we have

$$\begin{aligned}
\sup_{x\in K}|v_n^{t+1}(x) - v^{t+1}(x)| &= \sup_{x\in K}\left|\inf_{\mathsf{A}}\left[c(x,a) + \beta\int_{\mathsf{X}} v_n^t(y)p(dy \mid x,a)\right] - \min_{\mathsf{A}}\left[c(x,a) + \beta\int_{\mathsf{X}} v^t(y)p(dy \mid x,a)\right]\right| \\
&\le \beta\sup_{(x,a)\in K\times\mathsf{A}}\left|\int_{\mathsf{X}} v_n^t(y)p(dy \mid x,a) - \int_{\mathsf{X}} v^t(y)p(dy \mid x,a)\right| \\
&= \beta\sup_{(x,a)\in K\times\mathsf{A}}\left|\int_{K_\varepsilon}(v_n^t(y) - v^t(y))p(dy \mid x,a) + \int_{K_\varepsilon^c}(v_n^t(y) - v^t(y))p(dy \mid x,a)\right| \\
&\le \beta\left\{\sup_{x\in K_\varepsilon}|v_n^t(x) - v^t(x)| + \sup_{(x,a)\in K\times\mathsf{A}}\left|\int_{K_\varepsilon^c}(v_n^t(y) - v^t(y))p(dy \mid x,a)\right|\right\}
\end{aligned}$$

Note that we have $|v^t| \le M(w/(1-\sigma))$ by (22). Since $w_n \le \gamma_{\max} w$, where $\gamma_{\max} := \max\{1, \gamma\}$, we also have $|v_n^t| \le (L_1\gamma_{\max} w + L_2)/(1-\sigma_0) \le (L_1\gamma_{\max} + L_2)w/(1-\sigma_0)$ by (17) and (20) (as $w \ge 1$). Let us define

$$R := \frac{L_1\gamma_{\max} + L_2}{1-\sigma_0} + \frac{M}{1-\sigma}.$$

Thus by Lemma 3.1 we have

$$\sup_{x\in K}|v_n^{t+1}(x) - v^{t+1}(x)| \le \beta\sup_{x\in K_\varepsilon}|v_n^t(x) - v^t(x)| + \beta R\varepsilon.$$

Since the first term converges to zero as $n \to \infty$ by the induction hypothesis, and ε is arbitrary, the claim is true for $t+1$. This completes the proof. □

The following theorem states that the discounted value function of $\overline{\text{MDP}}_n$ converges to the discounted value function of the original MDP uniformly on each compact set $K \subset \mathsf{X}$.

Theorem 3.1. *For any compact set $K \subset \mathsf{X}$ we have*

$$\lim_{n\to\infty}\sup_{x\in K}|\bar{J}_n^*(x) - J^*(x)| = 0. \tag{25}$$

Proof. Fix any compact set $K \subset \mathsf{X}$. Since w is continuous and therefore bounded on K, it is sufficient to prove $\lim_{n\to\infty} \sup_{x\in K}(|\bar{J}_n^*(x) - J^*(x)|/(w(x)))$. Let n be chosen such that $K \subset K_n$, and so $w_n = w$ on K. We then have

$$\begin{aligned}
\sup_{x\in K} \frac{|\bar{J}_n^*(x) - J^*(x)|}{w(x)} &\le \sup_{x\in K} \frac{|\bar{J}_n^*(x) - v_n^t(x)|}{w(x)} + \sup_{x\in K} \frac{|v_n^t(x) - v^t(x)|}{w(x)} + \sup_{x\in K} \frac{|v^t(x) - J^*(x)|}{w(x)} \\
&\le \sup_{x\in K} \frac{|\bar{J}_n^*(x) - v_n^t(x)|}{C_n(x)} \frac{C_n(x)}{w(x)} + \sup_{x\in K} \frac{|v_n^t(x) - v^t(x)|}{w(x)} + M\frac{\sigma^t}{1-\sigma^t} \quad \text{(by (23))} \\
&\le \sup_{x\in K} \frac{|\bar{J}_n^*(x) - v_n^t(x)|}{C_n(x)} \frac{(L_1 w_n(x) + L_2)}{w(x)} + \sup_{x\in K} \frac{|v_n^t(x) - v^t(x)|}{w(x)} + \frac{M\sigma^t}{1-\sigma^t} \quad \text{(by (17))} \\
&\le (L_1 + L_2) \sup_{x\in K} \frac{|\bar{J}_n^*(x) - v_n^t(x)|}{C_n(x)} + \sup_{x\in K} \frac{|v_n^t(x) - v^t(x)|}{w(x)} + \frac{M\sigma^t}{1-\sigma^t} \quad (w_n = w \text{ on } K) \\
&\le (L_1 + L_2)\frac{\sigma_0^t}{1-\sigma_0} + \sup_{x\in K} \frac{|v_n^t(x) - v^t(x)|}{w(x)} + \frac{M\sigma^t}{1-\sigma^t} \quad \text{(by (21))}.
\end{aligned}$$

Since $w \ge 1$ on X, $\sup_{x\in K}(|v_n^t(x) - v^t(x)|/(w(x))) \to 0$ as $n \to \infty$ for all t by Lemma 3.4. Hence, the last expression can be made arbitrarily small. This completes the proof. □

In the remainder of this section, we use the above results and Theorem 2.2 to compute a near optimal policy for the original MDP. It is straightforward to check that for each n, c-MDP$_n$ satisfies the assumptions in Theorem 2.2. Let $\{\varepsilon_n\}$ be a sequence of positive real numbers such that $\lim_{n\to\infty} \varepsilon_n = 0$.

By Theorem 2.2, for each $n \ge 1$, there exists a deterministic stationary policy $f_n \in \mathbb{F}_n$, obtained from the finite state approximations of c-MDP$_n$, such that

$$\sup_{x\in \mathsf{X}_n} |J_n(f_n, x) - J_n^*(x)| \le \varepsilon_n,$$

where for each n, finite-state models are constructed replacing $(\mathsf{Z}, \mathsf{A}, p, c)$ with the components $(\mathsf{X}_n, \mathsf{A}, p_n, c_n)$ of c-MDP$_n$ in Section 2. By Lemma 3.2, for each $n \ge 1$ we also have

$$\sup_{x\in \mathsf{X}} |\bar{J}_n(f_n, x) - \bar{J}_n^*(x)| \le \varepsilon_n, \tag{26}$$

where, with an abuse of notation, we also denote the extended (to X) policy by f_n. Let us define operators $\bar{R}_n\colon B_{C_n}(\mathsf{X}) \to B_{C_n}(\mathsf{X})$ and $R_n\colon B_w(\mathsf{X}) \to B_w(\mathsf{X})$ by

$$\bar{R}_n u(x) = \begin{cases} c(x, f_n(x)) + \beta \int_{\mathsf{X}} u(y) p(dy \mid x, f_n(x)), & \text{if } x \in K_n, \\ \int_{K_n^c} [c(z, f_n(z)) + \beta \int_{\mathsf{X}} u(y) p(dy \mid z, f_n(z))] \nu_n(dz), & \text{if } x \in K_n^c, \end{cases}$$

$$R_n u(x) = c(x, f_n(x)) + \beta \int_{\mathsf{X}} u(y) p(dy \mid x, f_n(x)).$$

By Hernández-Lerma and Lasserre [22, Remark 8.3.10, p. 54], $\bar{R}_n$ is a contraction operator with modulus σ_0 and R_n is a contraction operator with modulus σ. Furthermore, the fixed point of $\bar{R}_n$ is $\bar{J}_n(f_n, x)$ and the fixed point of R_n is $J(f_n, x)$. For each $n \ge 1$, let us define $\bar{u}_n^0 = u_n^0 = 0$ and $\bar{u}_n^{t+1} = \bar{R}_n \bar{u}_n^t$, $u_n^{t+1} = R_n u_n^t$ $(t \ge 1)$. One can prove the following (see the proof of Hernández-Lerma and Lasserre [22, Theorem 8.3.6, p. 51]):

$$|\bar{u}_n^t(x)|, |\bar{J}_n(f_n, x)| \le \frac{C_n(x)}{1-\sigma_0}, \quad \|\bar{u}_n^t - \bar{J}_n(f_n, \cdot)\|_{C_n} \le \frac{\sigma_0^t}{1-\sigma_0}, \quad |u_n^t(x)|, |J(f_n, x)| \le M\frac{w(x)}{1-\sigma}, \quad \|u_n^t - J(f_n, \cdot)\|_w \le M\frac{\sigma^t}{1-\sigma}.$$

Lemma 3.5. *For any compact set $K \subset \mathsf{X}$, we have*

$$\lim_{n\to\infty} \sup_{x\in K} |\bar{u}_n^t(x) - u_n^t(x)| = 0.$$

Proof. The lemma can be proved using the same arguments as in the proof of Lemma 3.4 and so we omit the details. □

Lemma 3.6. *For any compact set $K \subset \mathsf{X}$, we have*

$$\lim_{n\to\infty} \sup_{x\in K} |\bar{J}_n(f_n, x) - J(f_n, x)| = 0. \tag{27}$$

Indeed, this is true for all sequences of policies in $\mathbb{F}$.

Proof. The lemma can be proved using the same arguments as in the proof of Theorem 3.1. □

The following theorem is the main result of this section that states that the true cost functions of the policies obtained from finite state models converge to the value function of the original MDP. Hence, to obtain a near optimal policy for the original MDP, it is sufficient to compute the optimal policy for the finite state model that has sufficiently large number of grid points.

Theorem 3.2. *For any compact set $K \subset \mathsf{X}$, we have*

$$\lim_{n\to\infty} \sup_{x\in K} |J(f_n, x) - J^*(x)| = 0.$$

Therefore,

$$\lim_{n\to\infty} |J(f_n, x) - J^*(x)| = 0 \quad \text{for all } x \in \mathsf{X}.$$

Proof. The result follows from (25)–(27). □

3.2. Average Cost

In this section we obtain approximation results, analogous to Theorems 3.1 and 3.2, for the average cost criterion. To do this, we impose some new assumptions on the components of the original MDP in addition to Assumption 3.1. These assumptions are the unbounded counterpart of Assumption 2.2. With the exception of Assumption 3.2(j), versions of these assumptions were imposed in Vega-Amaya [38], Gordienko and Hernandez-Lerma [18], and Jaśkiewicz and Nowak [26] to study the existence of the solution to the Average Cost Optimality Equality (ACOE) and Inequality (ACOI). In what follows, for any finite signed measure ϑ and measurable function h on X, we let $\vartheta(h) := \int_{\mathsf{X}} h(x)\vartheta(dx)$ and

$$\|\vartheta\|_w := \sup_{\|g\|_w \le 1} \left| \int_{\mathsf{X}} g(x)\vartheta(dx) \right|.$$

Here $\|\vartheta\|_w$ is called the w-norm of ϑ.

Assumption 3.2. *Suppose Assumption 3.1 holds with item* (b) *and* (7) *replaced by conditions* (j) *and* (e) *below, respectively. In addition, there exist a probability measure η on X and a positive measurable function ϕ: $\mathsf{X}\times\mathsf{A} \to (0,\infty)$ such that for all $(x,a) \in \mathsf{X}\times\mathsf{A}$*

(e) $\int_{\mathsf{X}} w(y)p(dy \mid x,a) \le \alpha w(x) + \eta(w)\phi(x,a)$, *where* $\alpha \in (0,1)$.
(f) $p(D \mid x,a) \ge \eta(D)\phi(x,a)$ *for all* $D \in \mathcal{B}(\mathsf{X})$.
(g) *The weight function w is η-integrable; i.e.,* $\eta(w) < \infty$.
(h) *For each* $n \ge 1$, $\inf_{(x,a)\in K_n\times\mathsf{A}} \phi(x,a) > 0$.
(j) *The stochastic kernel $p(\cdot \mid x,a)$ is continuous in (x,a) with respect to the w-norm.*

Throughout this section, it is assumed that Assumption 3.2 holds. Conditions (e), (f), and (g) of Assumption 3.2 are unbounded counterparts of conditions (d) and (e) in Assumption 2.2. Recall that condition (e) corresponds to the so-called "drift inequality" and condition (f) corresponds to the so-called "minorization" condition that guarantees the geometric ergodicity of Markov chains induced by stationary policies (see Hernández-Lerma and Lasserre [22], Meyn and Tweedie [29], and references therein). These assumptions are quite general for studying average cost problems with unbounded one-stage costs. In addition, they are proper for the approximation problem in the sense that it can be shown that if the original problem satisfies these, then the reduced models constructed in the sequel satisfy similar conditions. There is only one minor difference between Assumption 3.2(f) and the standard minorization condition: in the literature ϕ is in general required to be nonnegative instead of positive.

Note that although Assumption 3.2(j) seems to be restrictive, it is weaker than the assumptions imposed in the literature for studying approximation of average cost problems with unbounded cost (see Dufour and Prieto-Rumeau [15]). Indeed, it is assumed in Dufour and Prieto-Rumeau [15] that the transition probability p is Lipschitz continuous in (x,a) with respect to w-norm. The reason for imposing such a strong condition on

the transition probability is to obtain convergence rate for the approximation problem. Since we do not aim to provide rate of convergence result in this section, it is natural to impose continuity instead of Lipschitz continuity of the transition probability. However, it does not seem possible to replace continuity with respect to the w-norm by a weaker convergence notion. One reason is that with a weaker continuity notion it is not possible to prove that the transition probability of c-MDP$_n$ is continuous with respect to the total variation distance, which is needed if one wants to use Theorem 2.6 and cannot be relaxed as explained in Remark 2.3.

Analogous with Theorem 2.5, the following theorem is a consequence of Vega-Amaya [38, Theorems 3.3]; Gordienko and Hernandez-Lerma [18, Lemma 3.4] (see also Hernández-Lerma and Lasserre [22, Proposition 10.2.5, p. 126]); and Jaśkiewicz and Nowak [26, Theorem 3], which also holds with Assumption 3.2(j) replaced by Assumption 3.1(b).

Theorem 3.3. *For each $f \in \mathbb{F}$, the stochastic kernel $p(\cdot \mid x, f(x))$ is positive Harris recurrent with unique invariant probability measure μ_f. Furthermore, w is μ_f-integrable, and therefore, $\rho_f := \int_{\mathsf{X}} c(x,f)\mu_f(dx) < \infty$. There exist positive real numbers R and $\kappa < 1$ such that*

$$\sup_{f\in\mathbb{F}} \|p^t(\cdot \mid x, f(x)) - \mu_f\|_w \le Rw(x)\kappa^t \tag{28}$$

for all $x \in \mathsf{X}$, where R and κ continuously depend on α, $\eta(w)$, and $\inf_{f\in\mathbb{F}} \eta(\phi(y, f(y)))$. Finally, there exist $f^* \in \mathbb{F}$ and $h^* \in B_w(\mathsf{X})$ such that the triplet (h^*, f^*, ρ_{f^*}) satisfies the average cost optimality equality (ACOE), and therefore,

$$\inf_{\pi\in\Pi} V(\pi, x) := V^*(x) = \rho_{f^*},$$

for all $x \in \mathsf{X}$.

Note that (28) implies that for each $f \in \mathbb{F}$, the average cost is given by $V(f,x) = \int_{\mathsf{X}} c(y, f(y))\mu_f(dy)$ for all $x \in \mathsf{X}$ (instead of μ_f-a.e.); that is, the average cost is independent of the initial point.

Recall that V_n and $\bar{V}_n$ denote the average costs of c-MDP$_n$ and $\overline{\text{MDP}}_n$, respectively. The value functions for average cost are denoted analogously to the discounted cost case. Similar to Lemma 3.2, the following result states that MDP$_n$ and $\overline{\text{MDP}}_n$ are not too different for the average cost.

Lemma 3.7. *Suppose Theorem 3.3 holds for $\overline{\text{MDP}}_n$ and Theorem 2.5 holds for MDP$_n$; then we have*

$$\bar{V}_n^*(x) = \begin{cases} V_n^*(x), & \text{if } x \in K_n, \\ V_n^*(\Delta_n), & \text{if } x \in K_n^c. \end{cases} \tag{29}$$

Furthermore, if for any deterministic stationary policy $f \in \mathbb{F}_n$, we define $\bar{f}(x) = f(x)$ on K_n and $\bar{f}(x) = f(\Delta_n)$ on K_n^c, then

$$\bar{V}_n(\bar{f}, x) = \begin{cases} V_n(f, x), & \text{if } x \in K_n, \\ V_n(f, \Delta_n), & \text{if } x \in K_n^c. \end{cases} \tag{30}$$

In particular, if the deterministic stationary policy $f_n^* \in \mathbb{F}_n$ is optimal for MDP$_n$, then its extension $\bar{f}_n^*$ to X is also optimal for $\overline{\text{MDP}}_n$.

Proof. Let the triplet $(h_n^*, f_n^*, \rho_{f_n^*}^n)$ satisfy the ACOE for c-MDP$_n$ so that f_n^* is an optimal policy and $\rho_{f_n^*}^n$ is the average value function for c-MDP$_n$. It is straightforward to show that the triplet $(\bar{h}_n^*, \bar{f}_n^*, \rho_{f_n^*}^n)$ satisfies the ACOE for $\widetilde{\text{MDP}}_n$, where

$$\bar{h}_n^*(x) = \begin{cases} h_n^*(x), & \text{if } x \in K_n, \\ h_n^*(\Delta_n), & \text{if } x \in K_n^c, \end{cases}$$

and

$$\bar{f}_n^*(x) = \begin{cases} f_n^*(x), & \text{if } x \in K_n, \\ f_n^*(\Delta_n), & \text{if } x \in K_n^c. \end{cases}$$

By Gordienko and Hernandez-Lerma [18, Lemma 5.2] (see also Hernández-Lerma and Lasserre [21, Section 5.2]), this implies that $\bar{f}_n^*$ is an optimal stationary policy for $\overline{\text{MDP}}_n$ with cost function $\rho_{f_n^*}^n$. This completes the proof of the first part.

For the second part, let $f \in \mathbb{F}_n$ with an unique invariant probability measure $\mu_f \in \mathcal{P}(\mathsf{X}_n)$ and let $\bar{f} \in \mathbb{F}$ denote its extension to X with an unique invariant probability measure $\mu_{\bar{f}}$. It can be proved that

$$\mu_f(\cdot) = \mu_{\bar{f}}(\cdot \cap K_n) + \mu_{\bar{f}}(K_n^c)\delta_{\Delta_n}(\cdot).$$

We then have

$$\bar{V}_n(f,x) = \int_{\mathsf{X}} b_n(x,\bar{f}(x))\mu_{\bar{f}}(dx) = \int_{K_n} c_n(x,\bar{f}(x))\mu_{\bar{f}}(dx) + \mu_{\bar{f}}(K_n^c)c_n(\Delta_n,\bar{f}(\Delta_n)) = \int_{\mathsf{X}_n} c_n(x,f(x))\mu_f(dx) = V_n(f,x).$$

This completes the proof. □

By Lemma 3.7, in the remainder of this section we need only consider $\overline{\mathrm{MDP}}_n$ in place of MDP_n. Later we will show that Theorem 3.3 holds for $\overline{\mathrm{MDP}}_n$ for n sufficiently large and that Theorem 2.5 holds for c-MDP_n for all n.

Recall the definition of constants γ_n and τ_n from (8) and (9). For each $n \geq 1$, we define $\phi_n\colon \mathsf{X}\times\mathsf{A} \to (0,\infty)$ and $\varsigma_n \in \mathbb{R}$ as

$$\phi_n(x,a) := \begin{cases} \phi(x,a), & \text{if } x \in K_n, \\ \displaystyle\int_{K_n^c} \phi(y,a)\nu_n(dy), & \text{if } x \in K_n^c, \end{cases} \qquad \varsigma_n := \int_{K_n^c} w(y)\eta(dy).$$

Since $\eta(w) < \infty$ and τ_n can be made arbitrarily small by properly choosing ν_n, we assume, without loss of generality, the following.

Assumption 3.3. *The sequence of probability measures $\{\nu_n\}$ is chosen such that the following holds*

$$\lim_{n\to\infty}(\tau_n + \varsigma_n) = 0. \tag{31}$$

Let $\alpha_n := \alpha + \varsigma_n + \tau_n$.

Lemma 3.8. *For all n and $(x,a) \in \mathsf{X}\times\mathsf{A}$, the components of $\overline{\mathrm{MDP}}_n$ satisfy the following*:

$$\begin{aligned} &\sup_{a\in\mathsf{A}} |b_n(x,a)| \leq M w_n(x), \\ &\int_{\mathsf{X}} w_n(y)\, q_n(dy \mid x,a) \leq \alpha_n w_n(x) + \eta(w_n)\phi_n(x,a), \\ &q_n(D \mid x,a) \geq \eta(D)\phi_n(x,a), \quad \text{for all } D \in \mathcal{B}(\mathsf{X}). \end{aligned} \tag{32}$$

Proof. The proof of the first inequality follows from Assumption 3.2 and definitions of b_n and w_n. To prove the remaining two inequalities, we have to consider the cases $x \in K_n$ and $x \in K_n^c$ separately.

Let $x \in K_n$, and therefore $q_n(\cdot \mid x,a) = p(\cdot \mid x,a)$. The second inequality holds since

$$\begin{aligned} \int_{\mathsf{X}} w_n(y)p(dy \mid x,a) &= \int_{\mathsf{X}} w(y)\, p(dy \mid x,a) + \int_{K_n^c} (\gamma_n - w(y))\, p(dy \mid x,a) \\ &\leq \int_{\mathsf{X}} w(y)\, p(dy \mid x,a) + \tau_n \\ &\leq \alpha w(x) + \eta(w)\phi(x,a) + \tau_n \\ &\leq \alpha w_n(x) + \eta(w_n)\phi_n(x,a) + \varsigma_n\phi_n(x,a) + \tau_n \quad (\text{as } w_n = w \text{ and } \phi_n = \phi \text{ on } K_n) \\ &\leq \alpha_n w_n(x) + \eta(w_n)\phi_n(x,a), \quad (\text{as } \phi_n \leq 1 \text{ and } w_n \geq 1). \end{aligned}$$

For the last inequality, for all $D \in \mathcal{B}(\mathsf{X})$, we have

$$q_n(D \mid x,a) = p(D \mid x,a) \geq \eta(D)\phi(x,a) = \eta(D)\phi_n(x,a) \quad (\text{as } \phi_n = \phi \text{ on } K_n).$$

Hence, inequalities hold for $x \in K_n$.

For $x \in K_n^c$, we have

$$\begin{aligned}\int_{\mathsf{X}} w_n(y) q_n(dy \mid x,a) &= \int_{K_n^c} \left(\int_{\mathsf{X}} w_n(y) p(dy \mid z,a) \right) \nu_n(dz) \\ &\le \int_{K_n^c} (\alpha w(z) + \eta(w_n)\phi(x,a) + \varsigma_n \phi(x,a) + \tau_n)\nu_n(dz) \\ &= \alpha w_n(x) + \eta(w_n)\phi_n(x,a) + \varsigma_n \phi_n(x,a) + \tau_n \\ &\le \alpha_n w_n(x) + \eta(w_n)\phi_n(x,a), \quad \text{(since } \phi_n \le 1 \text{ and } w_n \ge 1\text{)} \end{aligned} \tag{33}$$

where (33) can be obtained following the same arguments as for the case $x \in K_n$. The last inequality holds for $x \in K_n^c$ since

$$q_n(D \mid x,a) = \int_{K_n^c} p(D \mid z,a)\nu_n(dz) \ge \int_{K_n^c} \eta(D)\phi(z,a)\nu_n(dz) = \eta(D)\phi_n(x,a).$$

This completes the proof. □

We note that by (31), there exists $n_0 \ge 1$ such that $\alpha_n < 1$ for $n \ge n_0$. Hence, for each $n \ge n_0$, Theorem 3.3 holds for $\overline{\mathrm{MDP}}_n$ with w replaced by w_n for some $R_n > 0$ and $\kappa_n \in (0,1)$, and we have $R_{\max} := \sup_{n \ge n_0} R_n < \infty$ and $\kappa_{\max} := \sup_{n \ge n_0} \kappa_n < 1$.

In the remainder of this section, it is assumed that $n \ge n_0$.

Lemma 3.9. *Let $g\colon \mathsf{X} \times \mathsf{A} \to \mathbb{R}$ be any measurable function such that $\sup_{a \in \mathsf{A}} |g(x,a)| \le M_g w(x)$ for some $M_g \in \mathbb{R}$. For all $t \ge 1$ and any compact set $K \subset \mathsf{X}$, we then have*

$$\sup_{(y,f) \in K \times \mathbb{F}} \left| \int_{\mathsf{X}} g_n(x,f(x))\, q_n^t(dx \mid y, f(y)) - \int_{\mathsf{X}} g(x,f(x))\, p^t(dx \mid y, f(y)) \right| \to 0$$

as $n \to \infty$, where $g_n(x,a) = g(x,a)$ on $K_n \times \mathsf{A}$ and $g_n(x,a) = \int_{K_n^c} g(z,a)\nu_n(dz)$ on $K_n^c \times \mathsf{A}$.

Proof. We will prove the lemma by induction. Fix any compact set $K \subset \mathsf{X}$. We note that in the inequalities below, we repeatedly use that $\phi, \phi_n \le 1$ without explicitly referring to this fact. Recall the definition of the compact subsets K_ε of X in Lemma 3.1 and the constant $\gamma_{\max} = \max\{1, \gamma\}$. Note that $\sup_{a \in \mathsf{A}} |g_n(x,a)| \le M_g w_n(x) \le M_g \gamma_{\max} w(x)$ for all $x \in \mathsf{X}$.

The claim holds for $t = 1$ by the following argument:

$$\begin{aligned} &\sup_{(y,f) \in K \times \mathbb{F}} \left| \int_{\mathsf{X}} g_n(x,f(x))\, q_n(dx \mid y, f(y)) - \int_{\mathsf{X}} g(x,f(x))\, p(dx \mid y, f(y)) \right| \\ &\quad = \sup_{(y,f) \in K \times \mathbb{F}} \left| \int_{\mathsf{X}} g_n(x,f(x))\, p(dx \mid y, f(y)) - \int_{\mathsf{X}} g(x,f(x))\, p(dx \mid y, f(y)) \right| \quad \text{(for } n \text{ sufficiently large)} \\ &\quad = \sup_{(y,f) \in K \times \mathbb{F}} \left| \int_{K_\varepsilon^c} g_n(x,f(x))\, p(dx \mid y, f(y)) - \int_{K_\varepsilon^c} g(x,f(x))\, p(dx \mid y, f(y)) \right| \quad \text{(for } n \text{ sufficiently large)} \\ &\quad \le M_g(1 + \gamma_{\max})\varepsilon, \end{aligned}$$

where the last inequality follows from Lemma 3.1. Since ε is arbitrary, the result follows.

Assume the claim is true for $t \ge 1$. Let us define $l_f(z) := \int_{\mathsf{X}} g(x,f(x))\, p^t(dx \mid z, f(z))$ and $l_f^n(z) := \int_{\mathsf{X}} g_n(x,f(x))\, q_n^t(dx \mid z, f(z))$. By recursively applying the inequalities in Assumption 3.2(e) and in (32) we obtain

$$\sup_{f \in \mathbb{F}} |l_f(z)| \le M_g \alpha^t w(z) + M_g \eta(w) \sum_{j=0}^{t-1} \alpha^j$$

and

$$\sup_{f \in \mathbb{F}} |l_f^n(z)| \le M_g \alpha_n^t w_n(z) + M_g \eta(w_n) \sum_{j=0}^{t-1} \alpha_n^j \le M_g \alpha_{\max}^t \gamma_{\max} w(z) + M_g \eta(w) \gamma_{\max} \sum_{j=0}^{t-1} \alpha_{\max}^j,$$

where $\alpha_{\max} := \sup_{n\geq n_0} \alpha_n < 1$. We then have

$$\begin{aligned}
&\sup_{(y,f)\in \mathsf{K}\times\mathbb{F}} \left| \int_{\mathsf{X}} g_n(x,f(x)) q_n^{t+1}(dx \mid y, f(y)) - \int_{\mathsf{X}} g(x,f(x)) p^{t+1}(dx \mid y, f(y)) \right| \\
&\quad = \sup_{(y,f)\in K\times\mathbb{F}} \left| \int_{\mathsf{X}} l_f^n(z) q_n(dz \mid y, f(y)) - \int_{\mathsf{X}} l_f(z) p(dz \mid y, f(y)) \right| \\
&\quad = \sup_{(y,f)\in K\times\mathbb{F}} \left| \int_{\mathsf{X}} l_f^n(z) p(dz \mid y, f(y)) - \int_{\mathsf{X}} l_f(z) p(dz \mid y, f(y)) \right| \quad \text{(for } n \text{ sufficiently large)} \\
&\quad \leq \sup_{(y,f)\in K\times\mathbb{F}} \left| \int_{K_\varepsilon^c} l_f^n(z) p(dz \mid y, f(y)) - \int_{K_\varepsilon^c} l_f(z) p(dz \mid y, f(y)) \right| + \sup_{(z,f)\in K_\varepsilon\times\mathbb{F}} |l_f^n(z) - l_f(z)| \\
&\quad \leq R\varepsilon + \sup_{(z,f)\in K_\varepsilon\times\mathbb{F}} |l_f^n(z) - l_f(z)|, \qquad (34)
\end{aligned}$$

where R is given by

$$R := M_g \left(\alpha^t + \alpha_{\max}^t \gamma_{\max} + \eta(w) \sum_{j=0}^{t-1} \alpha^j + \eta(w)\gamma_{\max} \sum_{j=0}^{t-1} \alpha_{\max}^j \right)$$

and the last inequality follows from Lemma 3.1. Since the claim holds for t and K_ε, the second term in (34) goes to zero as $n \to \infty$. Since ε is arbitrary, the result follows. □

In the remainder of this section, the above results are used to compute a near optimal policy for the original MDP. Let $\{\varepsilon_n\}$ be a sequence of positive real numbers converging to zero.

For each $f \in \mathbb{F}$, let μ_f^n denote the unique invariant probability measure of the transition kernel $q_n(\cdot \mid x, f(x))$ and let ρ_f^n denote the associated average cost; that is, $\rho_f^n := \bar{V}_n(f,x) = \int_{\mathsf{X}} b_n(y,f(y)) \mu_f^n(dy)$ for all initial points $x \in \mathsf{X}$. Therefore, the value function of $\overline{\text{MDP}}_n$, denoted by $\bar{V}_n^*$, is given by $V_n^*(x) = \inf_{f\in\mathbb{F}} \rho_f^n$; i.e., it is constant on X.

Before making the connection with Theorem 2.6, we prove the following result.

Lemma 3.10. *The transition probability p_n of c-MDP$_n$ is continuous in (x,a) with respect to the total variation distance.*

Proof. To ease the notation, we define $M(\mathsf{X}_n)$, $M(\mathsf{X})$, and $M_w(\mathsf{X})$ as the subsets of $B(\mathsf{X}_n)$, $B(\mathsf{X})$, and $B_w(\mathsf{X})$, respectively, whose elements have (corresponding) norm less than one. Let $(x_k, a_k) \to (x,a)$ in $\mathsf{X}_n \times \mathsf{A}$. Since the pseudostate Δ_n is isolated and K_n is compact, we have two cases: (i) $x_k = x = \Delta_n$ for all k large enough or (ii) $x_k \to x$ in K_n.

For the first case we have

$$\begin{aligned}
\|p_n(\cdot \mid \Delta_n, a_k) - p_n(\cdot \mid \Delta_n, a)\|_{\text{TV}} &= \sup_{g\in M(\mathsf{X}_n)} \left| \int_{\mathsf{X}_n} g(y)\, p_n(dy \mid \Delta_n, a_k) - \int_{\mathsf{X}_n} g(y)\, p_n(dy \mid \Delta_n, a) \right| \\
&\leq \sup_{g\in M(\mathsf{X})} \left| \int_{\mathsf{X}} g(y)\, q_n(dy \mid \Delta_n, a_k) - \int_{\mathsf{X}} g(y)\, q_n(dy \mid \Delta_n, a) \right| \qquad (35) \\
&= \sup_{g\in M(\mathsf{X})} \left| \int_{K_n^c} \left(\int_{\mathsf{X}} g(y)\, p(dy \mid z, a_k) - \int_{\mathsf{X}} g(y)\, p(dy \mid z, a) \right) \nu_n(dz) \right| \\
&\leq \int_{K_n^c} \sup_{g\in M(\mathsf{X})} \left| \int_{\mathsf{X}} g(y)\, p(dy \mid z, a_k) - \int_{\mathsf{X}} g(y)\, p(dy \mid z, a) \right| \nu_n(dz) \\
&\leq \int_{K_n^c} \sup_{g\in M_w(\mathsf{X})} \left| \int_{\mathsf{X}} g(y)\, p(dy \mid z, a_k) - \int_{\mathsf{X}} g(y)\, p(dy \mid z, a) \right| \nu_n(dz) \\
&= \int_{K_n^c} \|p(\cdot \mid z, a_k) - p(\cdot \mid z, a)\|_w \nu_n(dz), \qquad (36)
\end{aligned}$$

where (35) follows since if for any $g \in M(\mathsf{X}_n)$ we define $\bar{g} = g$ on K_n and $\bar{g} = g(\Delta_n)$ on K_n^c, then we have $\bar{g} \in M(\mathsf{X})$ and $\int_{\mathsf{X}_n} g(y) p_n(dy \mid x, a) = \int_{\mathsf{X}} \bar{g}(y) q_n(dy \mid x, a)$ for all $(x,a) \in \mathsf{X}_n \times \mathsf{A}$. Note that we have

$$\begin{aligned}
\sup_{g\in M_w(\mathsf{X})} \left| \int_{\mathsf{X}} g(y) p(dy \mid z, a_k) - \int_{\mathsf{X}} g(y) p(dy \mid z, a) \right| &\leq \int_{\mathsf{X}} w(y) p(dy \mid z, a_k) + \int_{\mathsf{X}} w(y) p(dy \mid z, a) \\
&\leq 2(\alpha + \eta(w)) w(z)
\end{aligned}$$

by Assumption 3.2(e), $\phi \le 1$, and $w \ge 1$. Since w (restricted to K_n^c) is ν_n-integrable, by the dominated convergence theorem (36) goes to zero as $k \to \infty$.

For the second case we have

$$\begin{aligned}
\|p_n(\cdot \mid x_k, a_k) - p_n(\cdot \mid x, a)\|_{\mathrm{TV}} &= \sup_{g \in M(\mathsf{X}_n)} \left| \int_{\mathsf{X}_n} g(y) p_n(dy \mid x_k, a_k) - \int_{\mathsf{X}_n} g(y) p_n(dy \mid x, a) \right| \\
&\le \sup_{g \in M(\mathsf{X})} \left| \int_{\mathsf{X}} g(y) q_n(dy \mid x_k, a_k) - \int_{\mathsf{X}} g(y) q_n(dy \mid x, a) \right| \\
&= \sup_{g \in M(\mathsf{X})} \left| \int_{\mathsf{X}} g(y) p(dy \mid x_k, a_k) - \int_{\mathsf{X}} g(y) p(dy \mid x, a) \right| \quad (\text{since } x_k, x \in K_n) \\
&\le \sup_{g \in M_w(\mathsf{X})} \left| \int_{\mathsf{X}} g(y) p(dy \mid x_k, a_k) - \int_{\mathsf{X}} g(y) p(dy \mid x, a) \right| \\
&= \|p(\cdot \mid x_k, a_k) - p(\cdot \mid x, a)\|_w.
\end{aligned}$$

By Assumption 3.2(j) the last term goes to zero as $k \to \infty$. □

Thus we obtain that for each $n \ge 1$, c-MDP$_n$ satisfies the assumption in Theorem 2.6 for

$$\zeta(\cdot) = \eta(\cdot \cap K_n) + \eta(K_n^c)\delta_{\Delta_n}(\cdot), \qquad \theta(x,a) = \begin{cases} \phi(x,a), & \text{if } x \in K_n, \\ \displaystyle\int_{K_n^c} \phi(y,a)\nu_n(dy), & \text{if } x = \Delta_n, \end{cases}$$

and some $\lambda \in (0,1)$, where the existence of λ follows from Assumption 3.2(h) and the fact that $\phi > 0$.

Consequently, there exists a deterministic stationary policy $f_n \in \mathbb{F}_n$, obtained from the finite state approximations of c-MDP$_n$, such that

$$\sup_{x \in \mathsf{X}_n} |V_n(f_n, x) - V_n^*(x)| \le \varepsilon_n, \tag{37}$$

where finite-state models are constructed replacing $(\mathsf{Z}, \mathsf{A}, p, c)$ with the components $(\mathsf{X}_n, \mathsf{A}, p_n, c_n)$ of c-MDP$_n$ in Section 2. By Lemma 3.7, we also have

$$|\rho_{f_n}^n - \bar{V}_n^*| \le \varepsilon_n, \tag{38}$$

where, by an abuse of notation, we also denote the policy extended to X by f_n.

Lemma 3.11. *We have*

$$\sup_{f \in \mathbb{F}} |\rho_f^n - \rho_f| \to 0 \tag{39}$$

as $n \to \infty$.

Proof. Fix any compact set $K \subset \mathsf{X}$. For any $t \ge 1$ and $y \in K$, we have

$$\begin{aligned}
\sup_{f \in \mathbb{F}} |\rho_f^n - \rho_f| &= \sup_{f \in \mathbb{F}} \left| \int_{\mathsf{X}} b_n(x, f(x)) \mu_f^n(dx) - \int_{\mathsf{X}} c(x, f(x)) \mu_f(dx) \right| \\
&\le \sup_{f \in \mathbb{F}} \left| \int_{\mathsf{X}} b_n(x, f(x)) \mu_f^n(dx) - \int_{\mathsf{X}} b_n(x, f(x)) q_n^t(dx \mid y, f(y)) \right| \\
&\quad + \sup_{f \in \mathbb{F}} \left| \int_{\mathsf{X}} b_n(x, f(x)) q_n^t(dx \mid y, f(y)) - \int_{\mathsf{X}} c(x, f(x)) p^t(dx \mid y, f(y)) \right| \\
&\quad + \sup_{f \in \mathbb{F}} \left| \int_{\mathsf{X}} c(x, f(x)) p^t(dx \mid y, f(y)) - \int_{\mathsf{X}} c(x, f(x)) \mu_f(dx) \right| \\
&\le M R_{\max} w(y) \kappa_{\max}^t + M R w(y) \kappa^t \\
&\quad + \sup_{(y,f) \in K \times \mathbb{F}} \left| \int_{\mathsf{X}} b_n(x, f(x)) q_n^t(dx \mid y, f(y)) - \int_{\mathsf{X}} c(x, f(x)) p^t(dx \mid y, f(y)) \right|,
\end{aligned}$$

where the last inequality follows from Theorem 3.3(ii) and (6) in Assumption 3.1. The result follows from Lemma 3.9. □

Theorem 3.4. *The value function of* $\overline{\text{MDP}}_n$ *converges to the value function of the original MDP; i.e.,* $|\bar{V}_n^* - V^*| \to 0$, *as* $n \to \infty$.

Proof. Since

$$|\bar{V}_n^* - V^*| = \left| \inf_{f\in\mathbb{F}} \rho_f^n - \inf_{f\in\mathbb{F}} \rho_f \right| \le \sup_{f\in\mathbb{F}} |\rho_f^n - \rho_f|,$$

the result follows from Lemma 3.11. □

The following is this section'se main result, which states that the true average cost of the policies f_n obtained from finite state approximations of c-MDP$_n$ converges to the average value function V^* of the original MDP.

Theorem 3.5. *We have* $|\rho_{f_n} - V^*| \to 0$, *as* $n \to \infty$.

Proof. We have

$$|\rho_{f_n} - V^*| \le |\rho_{f_n} - \rho_{f_n}^n| + |\rho_{f_n}^n - \bar{V}_n^*| + |\bar{V}_n^* - V^*| \le \sup_{f\in\mathbb{F}} |\rho_f - \rho_f^n| + \varepsilon_n + |\bar{V}_n^* - V^*| \quad \text{(by (38))}$$

The result follows from Lemma 3.11 and Theorem 3.4. □

4. Discretization of the Action Space

For computing near optimal policies using well-known algorithms, such as value iteration, policy iteration, and Q-learning, the action space must be finite. In this section, we show that, as a preprocessing step, the action space can taken to be finite if it has a sufficiently large number of points for accurate approximation. Throughout this section, it is assumed that Assumption 3.1 holds for the discounted cost and Assumption 3.2 holds for the average cost.

It was shown in Saldi et al. [33, 34] that any MDP with (infinite) compact action space can be well approximated by an MDP with finite action space under assumptions that are satisfied by c-MDP$_n$ for both the discounted cost and the average cost cases. Specifically, let d_{A} denote the metric on A. Since A is compact, one can find a sequence of finite subsets $\{\Lambda_k\}$ of A such that for all k

$$\min_{\hat{a}\in\Lambda_k} d_{\mathsf{A}}(a, \hat{a}) < 1/k, \quad \text{for all } a \in \mathsf{A}.$$

We define c-MDP$_{n,k}$ as the Markov decision process having the components $\{\mathsf{X}_n, \Lambda_k, p_n, c_n\}$ and we let $\mathbb{F}_n(\Lambda_k)$ denote the set of all deterministic stationary policies for c-MDP$_{n,k}$. Note that $\mathbb{F}_n(\Lambda_k)$ is the set of policies in $\mathbb{F}_n$ taking values only in Λ_k. Therefore, in a sense, c-MDP$_{n,k}$ and c-MDP$_n$ can be viewed as the same MDP, where the former has constraints on the set of policies. For each n and k, by an abuse of notation, let f_n^* and $f_{n,k}^*$ denote the optimal stationary policies of c-MDP$_n$ and c-MDP$_{n,k}$, respectively, for both the discounted and average costs. Saldi et al. [33, 34, Theorem 3.2] show that for all n, we have

$$\lim_{k\to\infty} J_n(f_{n,k}^*, x) = J_n(f_n^*, x) := J_n^*(x), \qquad \lim_{k\to\infty} V_n(f_{n,k}^*, x) = V_n(f_n^*, x), := V_n^*(x)$$

for all $x \in \mathsf{X}_n$. In other words, the discounted and average value functions of c-MDP$_{n,k}$ converge to the discounted and average value functions of c-MDP$_n$ as $k \to \infty$. We note that although Saldi et al. [34, Theorem 3.2]; [33, Theorem 3.2] are proved for nonnegative one-stage cost function, it is straightforward to check that these theorems are also valid for any real valued one-stage cost function.

Theorem 4.1. *For any* $x \in \mathsf{X}$, *there exists a subsequence* $\{k_n\}$ *such that*

$$\lim_{n\to\infty} J(f_{n,k_n}^*, x) = J^*(x), \qquad \lim_{n\to\infty} V(f_{n,k_n}^*, x) = V^*(x),$$

where $f_{n,k_n}^* \in \mathbb{F}(\Lambda_{k_n})$ *is the optimal stationary policy of c-MDP*$_{n,k_n}$.

Proof. Let us fix $x \in \mathsf{X}$. For n sufficiently large (so $x \in K_n$), we choose k_n such that $|J_n(f_{n,k_n}^*, x) - J_n(f_n^*, x)| < 1/n$ (or $|V_n(f_{n,k_n}^*, x) - V_n(f_n^*, x)| < 1/n$ for the average cost). We note that if A is a compact subset of a finite dimensional Euclidean space, then by using Saldi et al. [33, Theorems 4.1 and 4.2] one can obtain an explicit expression for k_n in terms of n under further continuity conditions on c and p. By Lemmas 3.6 and 3.11, we have $|\bar{J}_n(f_{n,k_n}^*, x) - J(f_{n,k_n}^*, x)| \to 0$ and $|\bar{V}_n(f_{n,k_n}^*, x) - V(f_{n,k_n}^*, x)| \to 0$ as $n \to \infty$, where again by an abuse of notation, the policies extended to X are also denoted by f_{n,k_n}^*. Since $\bar{J}_n(f_{n,k_n}^*, x) = J_n(f_{n,k_n}^*, x)$ and $\bar{V}_n(f_{n,k_n}^*, x) = V_n(f_{n,k_n}^*, x)$, using Theorems 3.1 and 3.4 one can immediately obtain

$$\lim_{n\to\infty} J(f_{n,k_n}^*, x) = J^*(x), \qquad \lim_{n\to\infty} V(f_{n,k_n}^*, x) = V^*(x). \qquad \square$$

Theorem 4.1 implies that before discretizing the state space to compute the near optimal policies, one can discretize, without loss of generality, the action space A in advance on a finite grid using a sufficiently large number of grid points.

5. Rate of Convergence Analysis for Compact-State MDPs

In this section we consider **(Q2)** for MDPs with compact state space; that is, we derive an upper bound on the performance loss due to discretization in terms of the cardinality of the set Z_n (i.e., number of grid points) . To do this, we will impose some new assumptions on the components of the MDP in addition to Assumptions 2.1 and 2.2. First, we present some definitions that are needed in the development.

For each $g \in C_b(\mathsf{Z})$, let

$$\|g\|_{\mathrm{Lip}} := \sup_{(z,y)\in \mathsf{Z}\times\mathsf{Z}} \frac{|g(z)-g(y)|}{d_{\mathsf{Z}}(z,y)}.$$

If $\|g\|_{\mathrm{Lip}}$ is finite, then g is called Lipschitz continuous with Lipschitz constant $\|g\|_{\mathrm{Lip}}$. $\mathrm{Lip}(\mathsf{Z})$ denotes the set of all Lipschitz continuous functions on Z; i.e.,

$$\mathrm{Lip}(\mathsf{Z}) := \{g \in C_b(\mathsf{Z}) \colon \|g\|_{\mathrm{Lip}} < \infty\}$$

and $\mathrm{Lip}(\mathsf{Z}, K)$ denotes the set of all $g \in \mathrm{Lip}(\mathsf{Z})$ with $\|g\|_{\mathrm{Lip}} \le K$. The *Wasserstein distance of order* 1 Villani [39, p. 95] between two probability measures ζ and ξ over Z is defined as

$$W_1(\zeta,\xi) := \sup\left\{\left|\int_{\mathsf{Z}} g\,d\zeta - \int_{\mathsf{Z}} g\,d\xi\right| \colon g \in \mathrm{Lip}(\mathsf{Z},1)\right\}.$$

W_1 is also called the *Kantorovich-Rubinstein distance*. It is known that if Z is compact, then $W_1(\zeta,\xi) \le \mathrm{diam}(\mathsf{Z}) \cdot \|\zeta-\xi\|_{\mathrm{TV}}$; see Villani [39, Theorem 6.15, p. 103]. For compact Z, the Wasserstein distance of order 1 is weaker than total variation distance. Furthermore, for compact Z, the Wasserstein distance of order 1 metrizes the weak topology on the set of probability measures $\mathscr{P}(\mathsf{Z})$ (see Villani [39, Corollary 6.13, p. 97]), which also implies that convergence in this sense is weaker than is setwise convergence.

In this section we impose the following supplementary assumptions in addition to Assumptions 2.1 and 2.2.

Assumption 5.1. (g) *The one-stage cost function c satisfies $c(\cdot,a) \in \mathrm{Lip}(\mathsf{Z}, K_1)$ for all $a \in \mathsf{A}$ for some K_1.*
(h) *The stochastic kernel p satisfies $W_1(p(\cdot \mid z,a), p(\cdot \mid y,a)) \le K_2 d_{\mathsf{Z}}(z,y)$ for all $a \in \mathsf{A}$ for some K_2.*
(j) *Z is an infinite compact subset of $\mathbb{R}^d$ for some $d \ge 1$, equipped with the Euclidean norm.*

We note that Assumption 5.1(j) implies the existence of a constant $\alpha > 0$ and finite subsets $\mathsf{Z}_n \subset \mathsf{Z}$ with cardinality n such that

$$\max_{z\in\mathsf{Z}} \min_{y\in\mathsf{Z}_n} d_{\mathsf{Z}}(z,y) \le \alpha(1/n)^{1/d} \tag{40}$$

for all n, where d_{Z} is the Euclidean distance on Z. In the remainder of this section, we replace Z_n defined in Section 2 with Z_n satisfying (40) to derive *explicit* bounds on the approximation error in terms of the cardinality of Z_n.

5.1. Discounted Cost

Assumptions 2.1 and 5.1 are imposed throughout this section. Additionally, we assume that $K_2\beta < 1$. The last assumption is the key to prove the next result, which states that the value function J^* of the original MDP for the discounted cost is in $\mathrm{Lip}(\mathsf{Z})$. Although this result is known in the literature (see Hinderer [24]), we give a short proof for the sake of completeness using a simple application of the value iteration algorithm.

Theorem 5.1. *The value function J^* for the discounted cost is in $\mathrm{Lip}(\mathsf{Z}, K)$, where $K = K_1(1/(1-\beta K_2))$.*

Proof. Let $u \in \mathrm{Lip}(\mathsf{Z}, K)$ for some $K > 1$; then $g = u/K \in \mathrm{Lip}(\mathsf{Z},1)$, and therefore for all $a \in \mathsf{A}$ and $z, y \in \mathsf{Z}$, we have

$$\begin{aligned}\left|\int_{\mathsf{Z}} u(x)p(dx \mid z,a) - \int_{\mathsf{Z}} u(x)p(dx \mid y,a)\right| &= K\left|\int_{\mathsf{Z}} g(x)p(dx \mid z,a) - \int_{\mathsf{Z}} g(x)p(dx \mid y,a)\right| \\ &\le K W_1(p(\cdot \mid z,a), p(\cdot \mid y,a)) \le K K_2 d_{\mathsf{Z}}(z,y),\end{aligned}$$

by Assumption 5.1(h). Hence, the contraction operator T defined in (2) maps $u \in \mathrm{Lip}(\mathsf{Z}, K)$ to $Tu \in \mathrm{Lip}(\mathsf{Z}, K_1 + \beta K K_2)$ since for all $z, y \in \mathsf{Z}$

$$\begin{aligned}|Tu(z) - Tu(y)| &\le \max_{a\in\mathsf{A}}\left\{|c(z,a) - c(y,a)| + \beta\left|\int_{\mathsf{Z}} u(x)p(dx \mid z,a) - \int_{\mathsf{Z}} u(x)p(dx \mid y,a)\right|\right\}\\ &\le K_1 d_{\mathsf{Z}}(z,y) + \beta K K_2 d_{\mathsf{Z}}(z,y) = (K_1 + \beta K K_2) d_{\mathsf{Z}}(z,y).\end{aligned}$$

Now we apply T recursively to obtain the sequence $\{T^n u\}$ by letting $T^n u = T(T^{n-1}u)$, which converges to the value function J^* by the Banach fixed point theorem. Clearly, by induction we have for all $n \ge 1$

$$T^n u \in \mathrm{Lip}(\mathsf{Z}, K_n),$$

where $K_n = K_1 \sum_{i=0}^{n-1} (\beta K_2)^i + K(\beta K_2)^n$. If we choose $K < K_1$, then $K_n \le K_{n+1}$ for all n, and therefore $K_n \uparrow K_1(1/(1-\beta K_2))$ since $K_2\beta < 1$. Hence, $T^n u \in \mathrm{Lip}(\mathsf{Z}, K_1(1/(1-\beta K_2)))$ for all n, and therefore $J^* \in \mathrm{Lip}(\mathsf{Z}, K_1(1/(1-\beta K_2)))$ since $\mathrm{Lip}(\mathsf{Z}, K_1(1/(1-\beta K_2)))$ is closed with respect to the sup-norm $\|\cdot\|$. □

The following theorem is the main result of this section. Recall that the policy $\hat{f}_n \in \mathbb{F}$ is obtained by extending the optimal policy f_n^* of MDP_n to Z.

Theorem 5.2. *We have*

$$\|J(\hat{f}_n, \cdot) - J^*\| \le \frac{\tau(\beta, K_2) K_1 (1/(1-\beta K_2)) + 2K_1/(1-\beta)}{1-\beta} 2\alpha (1/n)^{1/d},$$

where $\tau(\beta, K_2) = (2+\beta)\beta K_2 + (\beta^2 + 4\beta + 2)/(1-\beta)^2$ *and* α *is the coefficient in* (40).

Proof. To prove the theorem, we obtain upper bounds on the expressions derived in Section 2.1 in terms of the cardinality n of Z_n. The proof of Theorem 2.2 gives

$$\|J(\hat{f}_n, \cdot) - J^*\| \le \frac{\|T_{\hat{f}_n} J^* - \hat{T}_{\hat{f}_n} J^*\| + (1+\beta)\|\hat{J}_n^* - J^*\|}{1-\beta}.$$

To prove the theorem we upper bound $\|T_{\hat{f}_n} J^* - \hat{T}_{\hat{f}_n} J^*\|$ and $\|\hat{J}_n^* - J^*\|$ in terms n. For the first term we have

$$\begin{aligned}\|T_{\hat{f}_n} J^* - \hat{T}_{\hat{f}_n} J^*\| &= \sup_{z\in\mathsf{Z}} |T_{\hat{f}_n} J^*(z) - \hat{T}_{\hat{f}_n} J^*(z)|\\ &\le \sup_{z\in\mathsf{Z}} \int \left| c(z, \hat{f}_n(z)) + \beta \int_{\mathsf{Z}} J^*(y) p(dy \mid z, \hat{f}_n(z)) - c(x, \hat{f}_n(x)) - \beta \int_{\mathsf{Z}} J^*(y) p(dy \mid x, \hat{f}_n(x)) \right| \nu_{n, i_n(z)}(dx)\\ &\le \sup_{z\in\mathsf{Z}} \int \left(K_1 d_{\mathsf{Z}}(x,z) + \beta \left| \int_{\mathsf{Z}} J^*(y) p(dy \mid z, \hat{f}_n(z)) - \int_{\mathsf{Z}} J^*(y) p(dy \mid x, \hat{f}_n(z)) \right| \right) \nu_{n, i_n(z)}(dx)\\ &\qquad\qquad \text{(since } \hat{f}_n(x) = \hat{f}_n(z) \text{ for all } x \in \mathcal{S}_{n, i_n(z)}\text{)}\\ &\le \sup_{z\in\mathsf{Z}} \int (K_1 + \beta \|J^*\|_{\mathrm{Lip}} K_2) d_{\mathsf{Z}}(x,z) \nu_{n, i_n(z)}(dx)\\ &\le (K_1 + \beta \|J^*\|_{\mathrm{Lip}} K_2) \max_{i\in\{1,\ldots,n\}} \mathrm{diam}(\mathcal{S}_{n,i})\\ &\le (K_1 + \beta \|J^*\|_{\mathrm{Lip}} K_2) 2\alpha (1/n)^{1/d}. \end{aligned} \tag{41}$$

For the second term, the proof of Theorem 2.4 gives

$$\|\hat{J}_n^* - J^*\| \le \frac{\|\hat{T}_n J^* - F_n J^*\| + (1+\beta)\|J^* - u_n^*\|}{1-\beta}.$$

First consider $\|\hat{T}_n J^* - F_n J^*\|$. Define

$$l(z,a) := c(z,a) + \beta \int_{\mathsf{X}} J^*(y) p(dy \mid z, a),$$

so that

$$J^*(z) = \min_{a\in\mathsf{A}} l(z,a).$$

It is straightforward to show that $l(\cdot,a)\in \mathrm{Lip}(\mathsf{Z},K_l)$ for all $a\in\mathsf{A}$, where $K_l=K_1+\beta\|J^*\|_{\mathrm{Lip}}K_2$. By adapting the proof of Lemma 2.3 to the value function J^*, we obtain

$$\begin{aligned}\|\hat{T}_nJ^*-F_nJ^*\| &= \sup_{z\in\mathsf{Z}}\left|\min_{a\in\mathsf{A}}\int l(x,a)\nu_{n,i_n(z)}(dx)-\int \min_{a\in\mathsf{A}} l(x,a)\nu_{n,i_n(z)}(dx)\right|\\ &\le \sup_{z\in\mathsf{Z}}\int \sup_{y\in\mathscr{S}_{n,i_n(z)}}|l(y,a_i)-J^*(y)|\nu_{n,i_n(z)}(dy)\\ &\le \max_{i\in\{1,\dots,n\}}\int \sup_{y\in\mathscr{S}_{n,i}}\{|l(y,a_i)-l(z_i,a_i)|+|J^*(z_i)-J^*(y)|\}\nu_{n,i}(dy)\\ &\le \max_{i\in\{1,\dots,n\}}\int \sup_{y\in\mathscr{S}_{n,i_n}}\{K_l d_{\mathsf{Z}}(y,z_i)+\|J^*\|_{\mathrm{Lip}}d_{\mathsf{Z}}(z_i,y)\}\nu_{n,i}(dy)\\ &\le (K_l+\|J^*\|_{\mathrm{Lip}})\max_{i\in\{1,\dots,n\}}\mathrm{diam}(\mathscr{S}_{n,i})\\ &\le (K_l+\|J^*\|_{\mathrm{Lip}})2\alpha(1/n)^{1/d}. \end{aligned}\tag{42}$$

For the expression $\|J^*-u_n^*\|$, by Lemma 2.2 we have

$$\|u_n^*-J^*\|\le\frac{2}{1-\beta}\inf_{r\in\mathsf{Z}^{k_n}}\|J^*-\Phi_r\|,\qquad \text{where}\quad \Phi_r(z)=\sum_{i=1}^{k_n}r_i1_{S_{n,i}}(z),\quad r=(r_1,\dots,r_{k_n}).$$

Since $\|J^*\|_{\mathrm{Lip}}<\infty$, we have $\inf_{r\in\mathsf{Z}^{k_n}}\|J^*-\Phi_r\|\le\|J^*\|_{\mathrm{Lip}}\max_{i\in\{1,\dots,n\}}\mathrm{diam}(\mathscr{S}_{n,i})\le\|J^*\|_{\mathrm{Lip}}2\alpha(1/n)^{1/d}$. Hence,

$$\|u_n^*-J^*\|\le\frac{2}{1-\beta}\|J^*\|_{\mathrm{Lip}}2\alpha(1/n)^{1/d}.\tag{43}$$

Hence, by (42) and (43) we obtain

$$\|\hat{J}_n^*-J^*\|\le\left(\left(\beta K_2+\frac{\beta+3}{(1-\beta)^2}\right)\|J^*\|_{\mathrm{Lip}}+\frac{K_1}{1-\beta}\right)2\alpha(1/n)^{1/d}.\tag{44}$$

Thus, the result follows from (41) and (44), and the fact $\|J^*\|_{\mathrm{Lip}}\le K_1(1/(1-\beta K_2))$. □

Remark 5.1. It is important to point out that if we replace Assumption 5.1(h) with the uniform Lipschitz continuity of $p(\cdot\,|\,z,a)$ in z with respect to total variation distance, then Theorem 5.2 remains valid (with possibly different constants in front of the term $(1/n)^{1/d}$). However, in this case, we do not need the assumption $K_2\beta<1$.

Remark 5.2. For the average cost case, instead of assuming from the outset the uniform Lipschitz continuity of c and p in the z variable, we first derive a rate of convergence result in terms of the moduli of continuity of the functions ω_c and ω_p in the z variable of $c(z,a)$ and $p(\cdot\,|\,z,a)$, where the total variation distance is used to define ω_p. We next state that explicit rate of convergence result can be given if we impose some structural assumptions on ω_c and ω_p such as linearity, which corresponds to the uniform Lipschitz continuity of $c(z,a)$ and $p(\cdot\,|\,z,a)$ in z. However, this is not the right approach for the discounted cost case as the modulus of continuity function ω_p is calculated using the Wasserstein distance of order 1. Indeed, to obtain a similar result as in the average cost case, we must relate ω_c and ω_p to the modulus of continuity ω_{J^*} of the value function J^*. This can be established if ω_c and ω_p are affine functions (i.e., $\omega_c(r)=K_1r+L_1$ and $\omega_p(r)=K_2r+L_2$) using the dual formulation of the Wasserstein distance of order 1 (Villani [39, Theorem 5.10]):

$$W_1(\mu,\nu)=\sup_{\substack{(\psi,\varphi)\in C_b(\mathsf{Z})\times C_b(\mathsf{Z})\\ \psi(x)-\varphi(y)\le d_{\mathsf{Z}}(x,y)}}\left|\int_{\mathsf{Z}}\psi(z)\mu(dz)-\int_{\mathsf{Z}}\varphi(z)\nu(dz)\right|.$$

However, in this situation we can explicitly compute the convergence rate only if $L_1=L_2=0$, which is the uniform Lipschitz continuity case.

5.2. Average Cost

In this section, we suppose that Assumptions 2.2 and 5.1(j) hold. We define the modulus of continuity functions in the z variable of $c(z,a)$ and $p(\cdot \mid z,a)$ as follows

$$\omega_c(r) := \sup_{a\in\mathsf{A}} \sup_{z,y\in\mathsf{Z}:\, d_{\mathsf{Z}}(z,y)\le r} |c(z,a)-c(y,a)|, \qquad \omega_p(r) := \sup_{a\in\mathsf{A}} \sup_{z,y\in\mathsf{Z}:\, d_{\mathsf{Z}}(z,y)\le r} \|p(\cdot \mid z,a)-p(\cdot \mid y,a)\|_{\mathrm{TV}}.$$

Since $c(z,a)$ and $p(\cdot \mid z,a)$ are uniformly continuous, we have $\lim_{r\to 0}\omega_c(r)=0$ and $\lim_{r\to 0}\omega_p(r)=0$. Note that when ω_c and ω_p are linear, $c(z,a)$ and $p(\cdot \mid z,a)$ are uniformly Lipschitz in z. In the remainder of this section, we first derive a rate of convergence result in terms of ω_c and ω_p. Next, we explicitly compute the convergence rate for the Lipschitz case as a corollary of this result.

To obtain convergence rates for the average cost, we first prove a rate of convergence result for Lemma 2.6. To this end, for each $n\ge 1$, let $d_n := 2\alpha(1/n)^{1/d}$, where α is the coefficient in (40).

Lemma 5.1. *For all $t\ge 1$, we have*

$$\sup_{(y,f)\in\mathsf{Z}\times\mathbb{F}} \|p^t(\cdot \mid y,f(y)) - q_n^t(\cdot \mid y,f(y))\|_{\mathrm{TV}} \le t\omega_p(d_n).$$

Proof. Similar to the proof of Lemma 2.6, we use induction. For $t=1$, recalling the proof of Lemma 2.6, the claim holds by the following argument:

$$\sup_{(y,f)\in\mathsf{Z}\times\mathbb{F}} \|p(\cdot \mid y,f(y)) - q_n(\cdot \mid y,f(y))\|_{\mathrm{TV}} \le \sup_{y\in\mathsf{Z}} \sup_{(x,a)\in\mathcal{S}_{n,i_n(y)}\times\mathsf{A}} \|p(\cdot \mid y,a)-p(\cdot \mid x,a)\|_{\mathrm{TV}} \le \omega_p(d_n).$$

Now, assume the claim is true for $t\ge 1$. Again recalling the proof of Lemma 2.6, we have

$$\begin{aligned}\sup_{(y,f)\in\mathsf{Z}\times\mathbb{F}} \|p^{t+1}(\cdot \mid y,f(y)) - q_n^{t+1}(\cdot \mid y,f(y))\|_{\mathrm{TV}} &\le \sup_{(y,f)\in\mathsf{Z}\times\mathbb{F}} \|p^t(\cdot \mid y,f(y)) - q_n^t(\cdot \mid y,f(y))\|_{\mathrm{TV}} \\ &\quad + \sup_{(z,f)\in\mathsf{Z}\times\mathbb{F}} \|p(\cdot \mid z,f(z)) - q_n(\cdot \mid z,f(z))\|_{\mathrm{TV}} \\ &\le t\omega_p(d_n) + \omega_p(d_n) = (t+1)\omega_p(d_n).\end{aligned}$$

This completes the proof. □

The following theorem is the main result of this section. A somewhat similar result was obtained in Hernández-Lerma [20, Section 3.5], where identical assumptions are imposed on both the original model and the approximating model (see Hernández-Lerma [20, Assumption 5.1]). Moreover, the approximating transition probability and one-stage cost function are assumed to converge to the original transition probability and one-stage cost function with respect to some rate; that is, $\rho(n) := \sup_{(x,a)\in\mathsf{X}\times\mathsf{A}} |b_n(x,a)-c(x,a)|$ and $\pi(n) := \sup_{(x,a)\in\mathsf{X}\times\mathsf{A}} \|q_n(\cdot \mid x,a)-p(\cdot \mid x,a)\|_{\mathrm{TV}}$ with $\rho(n),\pi(n)\to 0$ as $n\to\infty$. Although our result may appear to be a special case of the results in Hernández-Lerma [20, Section 3.5], there are several differences: (i) our assumptions are only imposed for the original model and (ii) in Hernández-Lerma [20, Section 3.5], the approximating models do not have finite state space, whereas our approximating models are obtained by extending finite state models to the original state space, thereby allowing for a constructive numerical method to calculate near optimal policies.

Recall that the optimal policy $\tilde{f}_n^*$ for $\widetilde{\mathrm{MDP}}_n$ is obtained by extending the optimal policy f_n^* for MDP_n to Z, and R and κ are the constants in Theorem 2.5.

Theorem 5.3. *For all $t\ge 1$, we have*

$$|\rho_{\tilde{f}_n^*} - \rho_{f^*}| \le 4\|c\|R\kappa^t + 2\omega_c(d_n) + 2\|c\|t\omega_p(d_n).$$

Proof. The proof of Theorem 2.6 gives

$$|\rho_{\tilde{f}_n^*} - \rho_{f^*}| \le |\rho_{\tilde{f}_n^*} - \hat{\rho}^n_{\tilde{f}_n^*}| + |\hat{\rho}^n_{\tilde{f}_n^*} - \hat{\rho}^n_{\hat{f}_n^*}| + |\hat{\rho}^n_{\hat{f}_n^*} - \rho_{f^*}|.$$

Hence, to prove the theorem we obtain an upper bounds on the three terms in the sum. Consider the first term (recall the proof of Lemma 2.7)

$$\begin{aligned}|\rho_{\tilde{f}_n^*} - \hat{\rho}^n_{\tilde{f}_n^*}| \le \sup_{f\in\mathbb{F}} |\hat{\rho}^n_f - \rho_f| &\le 2R\kappa^t\|c\| + \|c\| \sup_{(y,f)\in\mathsf{Z}\times\mathbb{F}} \|q_n^t(\cdot \mid y,f(y)) - p^t(\cdot \mid y,f(y))\|_{\mathrm{TV}} \\ &\le 2R\kappa^t\|c\| + \|c\|t\omega_p(d_n) \quad \text{(by Lemma 5.1).} \end{aligned} \tag{45}$$

For the second term, the proof of Lemma 2.11 gives

$$\begin{aligned}|\hat{\rho}^n_{\tilde{f}^*_n} - \hat{\rho}^n_{\hat{f}^*_n}| &\le |\hat{\rho}^n_{\tilde{f}^*_n} - \tilde{\rho}^n_{\tilde{f}^*_n}| + |\tilde{\rho}^n_{\tilde{f}^*_n} - \hat{\rho}^n_{\hat{f}^*_n}| \\ &\le \sup_{f\in\mathbb{F}} |\hat{\rho}^n_f - \tilde{\rho}^n_f| + |\inf_{f\in\mathbb{F}} \tilde{\rho}^n_f - \inf_{f\in\mathbb{F}} \hat{\rho}^n_f| \\ &\le 2\sup_{f\in\mathbb{F}} |\hat{\rho}^n_f - \tilde{\rho}^n_f| \\ &\le 2\|b_n - c\| \quad \text{(see the proof of Lemma 2.9)} \\ &\le 2 \sup_{(z,a)\in \mathsf{Z}\times\mathsf{A}} \int |c(x,a) - c(z,a)| \nu_{n,i_n(z)}(dx) \\ &\le 2\omega_c(d_n). \end{aligned} \tag{46}$$

For the last term, we have

$$|\hat{\rho}^n_{\hat{f}^*_n} - \rho_{f^*}| = \left|\inf_{f\in\mathbb{F}} \hat{\rho}^n_f - \inf_{f\in\mathbb{F}} \rho_f\right| \le \sup_{f\in\mathbb{F}} |\hat{\rho}^n_f - \rho_f| \le 2R\kappa^t\|c\| + \|c\| t\omega_p(d_n) \quad \text{(by (45))}. \tag{47}$$

Combining (45)–(47) implies the result. □

To explicitly calculate a convergence rate, we need to impose some structural assumptions on ω_c and ω_p. One such assumption is linearity, which corresponds to the uniform Lipschitz continuity of $c(z,a)$ and $p(\cdot\,|\,z,a)$ in z. This means that $\omega_c(r) = K_1 r$ and $\omega_p(r) = K_2 r$ or, equivalently, $|c(z,a) - c(y,a)| \le K_1 d_{\mathsf{Z}}(z,y)$ and $\|p(\cdot\,|\,z,a) - p(\cdot\,|\,y,a)\| \le K_2 d_{\mathsf{Z}}(z,y)$ for all $z, y \in \mathsf{Z}$ and $a \in \mathsf{Z}$. In this case, by Theorem 5.3, for all $t \ge 1$ we have

$$|\rho_{\tilde{f}^*_n} - \rho_{f^*}| \le 4\|c\|R\kappa^t + 4K_1\alpha(1/n)^{1/d} + 4\|c\|K_2\alpha(1/n)^{1/d}t. \tag{48}$$

To obtain a proper rate of convergence result (i.e., an upper bound that only depends on n), the dependence of the upper bound on t has to be written as a function of n. This can be done by (approximately) minimizing the upper bound in (48) with respect to t for each n. Let us define the constants $I_1 := 4\|c\|R$, $I_2 := 4K_1\alpha$, and $I_3 := 4\|c\|K_2\alpha$. The upper bound in (48) then becomes

$$I_1\kappa^t + I_2(1/n)^{1/d} + I_3(1/n)^{1/d}t. \tag{49}$$

For each n, it is straightforward to compute that

$$t'(n) := \ln\left(\frac{n^{1/d}}{I_4}\right)\frac{1}{\ln(1/\kappa)}$$

is the zero of the derivative of the convex term in (49), where $I_4 := I_3/(I_1\ln(1/\kappa))$. Letting $t = \lceil t'(n)\rceil$ in (49), we obtain the following result.

Corollary 1. *Suppose that $c(z,a)$ and $p(\cdot\,|\,z,a)$ are uniformly Lipschitz continuous in z in addition to the assumptions imposed at the beginning of this section. Thus, we have*

$$|\rho_{\tilde{f}^*_n} - \rho_{f^*}| \le (I_1I_4 + I_2)(1/n)^{1/d} + \frac{I_3}{\ln(1/\kappa)}(1/n)^{1/d}\ln\left(\frac{n^{1/d}}{I_4}\right).$$

6. Order Optimality for Approximation Errors in the Rate of Quantization

The following example demonstrates that the order of the performance losses in Theorem 5.2 and Corollary 1 cannot be better than $O((1/n)^{1/d})$. More precisely, we exhibit a simple standard example where we can lower bound the performance loss by $L(1/n)^{1/d}$ for some positive constant L. A similar result was obtained in Saldi et al. [33, Section IV] for the case of quantization of action space, where the action space was a compact subset of $\mathbb{R}^m$ for some $m \ge 1$. Therefore, when both state and action spaces are quantized, the resulting construction is order optimal in the above sense as the approximation error, in this case, is bounded by the sum of the approximation errors in quantization of state space and quantization of action space.

In what follows $h(\cdot)$ and $h(\cdot\,|\,\cdot)$ denote differential and conditional differential entropies, respectively; see Cover and Thomas [12, Chapter 8].

Consider the additive-noise system:

$$z_{t+1} = F(z_t, a_t) + v_t, \quad t = 0, 1, 2, \ldots,$$

where $z_t, a_t, v_t \in \mathbb{R}^d$. We assume that $\sup_{(z,a)\in\mathbb{R}^d\times\mathbb{R}^d}(\|F(z,a)\|/(\|z\| + \|a\|)) < 1/2$. The noise process $\{v_t\}$ is a sequence of i.i.d. random vectors whose common distribution has density g supported on some compact subset V of $\mathbb{R}^d$. We choose V such that $\mathsf{Z} = \mathsf{A}$ can be taken to be compact subsets of $\mathbb{R}^d$. For simplicity suppose that the initial distribution μ has the same density g. It is assumed that the differential entropy $h(g) := -\int_{\mathsf{Z}} g(z)\log g(z)dz$ is finite. Let the one-stage cost function be $c(z,a) := \|z - a\|$. Clearly, the optimal stationary policy f^* is induced by the identity $f^*(z) = z$, having the optimal cost $J(f^*, \mu) = 0$ and $V(f^*, \mu) = 0$. Let $\hat{f}_n$ be the piecewise constant extension of the optimal policy f_n^* of the MDP$_n$ to the set Z. Fix $n \geq 1$ and define $D_t := E_\mu^{\hat{f}_n}[c(z_t, a_t)]$ for all t. Since $a_t = \hat{f}_n(z_t)$ can take at most n values in A, by the Shannon lower bound (SLB) (see Yamada et al. [44, p. 12]) we have for $t \geq 1$

$$\begin{aligned}\log n \geq R(D_t) \geq h(z_t) + \theta(D_t) &= h(F(z_{t-1}, a_{t-1}) + v_{t-1}) + \theta(D_t) \\ &\geq h(F(z_{t-1}, a_{t-1}) + v_{t-1} \mid z_{t-1}, a_{t-1}) + \theta(D_t) \qquad (50)\\ &= h(v_{t-1}) + \theta(D_t), \qquad (51)\end{aligned}$$

where $\theta(D_t) = -d + \log((1/(dV_d\Gamma(d)))(d/D_t)^d)$, $R(D_t)$ is the rate-distortion function of z_t, V_d is the volume of the unit sphere $S_d = \{z\colon \|z\| \leq 1\}$, and Γ is the gamma function. Here, (50) follows from the fact that conditioning reduces the entropy (see Cover and Thomas [12, Theorem 2.6.5, p. 29]) and (51) follows from the independence of v_{t-1} and the pair (z_{t-1}, a_{t-1}). Note that $h(v_{t-1}) = h(g)$ for all t. Thus, $D_t \geq L(1/n)^{1/d}$, where $L := (d/2)(2^{h(g)}/(dV_d\Gamma(d)))^{1/d}$. Since we have obtained stagewise error bounds, these give $|J(f^*, \mu) - J(\hat{f}_n, \mu)| \geq (L/(1-\beta))(1/n)^{1/d}$ and $|V(f^*, \mu) - V(\hat{f}_n, \mu)| \geq L(1/n)^{1/d}$.

Remark 6.1. We note that if $h(x_{t+1} \mid x_t, a_t)$ can be lower bounded by some constant k for all $t \geq 1$, the above analysis still holds by replacing $h(g)$ with k. For instance, this is the case if the transition probability $p(\cdot \mid x, a)$ admits a density that is bounded from above uniformly in (x, a).

7. Numerical Examples

In this section, we consider two examples, the additive noise model and fisheries management problem, in order to illustrate our results numerically. Since computing true costs of the policies obtained from the finite models is intractable, we only compute the value functions of the finite models and illustrate their converge to the value function of the original MDP as $n \to \infty$.

Before proceeding to the examples, we note that all results in this paper apply with straightforward modifications for the case of maximizing reward instead of minimizing cost.

7.1. Additive Noise System

In this example, the additive noise system is given by

$$x_{t+1} = F(x_t, a_t) + v_t, \quad t = 0, 1, 2, \ldots$$

where $x_t, a_t, v_t \in \mathbb{R}$ and $\mathsf{X} = \mathbb{R}$. The noise process $\{v_t\}$ is a sequence of $\mathbb{R}$-valued i.i.d. random variables with common density g. Hence, the transition probability $p(\cdot \mid x, a)$ is given by

$$p(D \mid x, a) = \int_D g(v - F(x, a))m(dv), \quad \text{for all } D \in \mathscr{B}(\mathbb{R}),$$

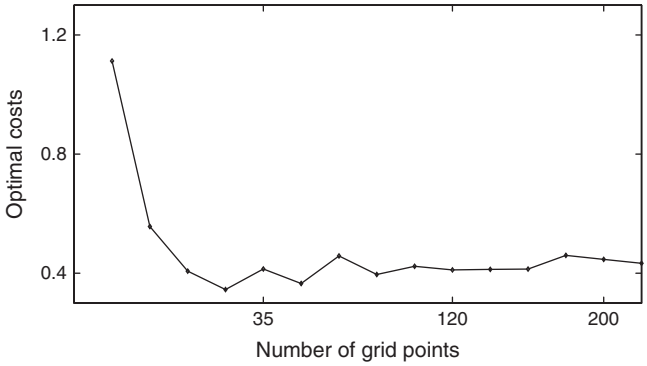
where m is the Lebesgue measure. The one-stage cost function is $c(x, a) = (x - a)^2$, the action space is $\mathsf{A} = [-L, L]$ for some $L > 0$, and the cost function to be minimized is the discounted cost.

We assume that (i) g is a Gaussian probability density function with zero mean and variance σ^2, (ii) $\sup_{a\in\mathsf{A}} |F(x, a)|^2 \leq k_1x^2 + k_2$ for some $k_1, k_2 \in \mathbb{R}_+$, (iii) $\beta < 1/\alpha$ for some $\alpha \geq k_1$, and (iv) F is continuous. Hence, Assumption 3.1 holds for this model with $w(x) = k + x^2$ and $M = 4(L^2/k + x^2)$ for some $k \in \mathbb{R}_+$.

For the numerical results, we use the following parameters: $F(x, a) = x + a$, $\beta = 0.3$, $L = 0.5$, and $\sigma = 0.1$.

We selected a sequence $\{[-l_n, l_n]\}_{n=1}^{15}$ of nested closed intervals, where $l_n = 0.5 + 0.25n$, to approximate $\mathbb{R}$. Each interval is uniformly discretized using $\lceil 2k_{\lceil n/3\rceil}l_n\rceil$ grid points, where $k_m = 5m$ for $m = 1, \ldots, 5$ and $\lceil q\rceil$ denotes the smallest integer greater than or equal to $q \in \mathbb{R}$. Therefore, the discretization is gradually refined.

Figure 1. Optimal Costs of the Finite Models When the Initial State Is $x = 0.7$



For each n, the finite state space is given by $\{x_{n,i}\}_{i=1}^{k_n} \cup \{\Delta_n\}$, where $\{x_{n,i}\}_{i=1}^{k_n}$ are the representation points in the uniform quantization of the closed interval $[-l_n, l_n]$ and Δ_n is a pseudostate. We also uniformly discretize the action space $\mathsf{A} = [-0.5, 0.5]$ by using $2k_{\lceil n/3 \rceil}$ grid points. For each n, the finite state models are constructed as in Section 2 by replacing Z with $[-l_n, l_n]$ and by setting $\nu_n(\cdot) = \frac{1}{2}m_n(\cdot) + \frac{1}{2}\delta_{\Delta_n}(\cdot)$, where m_n is the Lebesgue measure normalized over $[-l_n, l_n]$.

We use the value iteration algorithm to compute the value functions of the finite models. Figure 1 displays the graph of these value functions corresponding to the different values for the number of grid points when the initial state is $x = 0.7$. The figure illustrates that the value functions of the finite models converge to the value function of the original model.

7.2. Fisheries Management Problem

In this example we consider the following population growth model, called a Ricker model; see Hernández-Lerma and Lasserre [21, Section 1.3]:

$$x_{t+1} = \theta_1 a_t \exp\{-\theta_2 a_t + v_t\}, \quad t = 0, 1, 2, \ldots \tag{52}$$

where $\theta_1, \theta_2 \in \mathbb{R}_+$, x_t is the population size in season t, and a_t is the population to be left for spawning for the next season; in other words, $x_t - a_t$ is the amount of fish captured in the season t. The one-stage "reward" function is $u(x_t - a_t)$, where u is some utility function. In this model, the goal is to maximize the average reward.

The state and action spaces are $\mathsf{X} = \mathsf{A} = [\kappa_{\min}, \kappa_{\max}]$ for some $\kappa_{\min}, \kappa_{\max} \in \mathbb{R}_+$. Since the population left for spawning cannot be greater than the total population, for each $x \in \mathsf{X}$, the set of admissible actions is $\mathsf{A}(x) = [\kappa_{\min}, x]$ which is not consistent with our assumptions. However, we can (equivalently) reformulate above problem so that the admissible actions $\mathsf{A}(x)$ will become A for all $x \in \mathsf{X}$. In this case, instead of dynamics in Equation (52) we have

$$x_{t+1} = \theta_1 \min(a_t, x_t) \exp\{-\theta_2 \min(a_t, x_t) + v_t\}, \quad t = 0, 1, 2, \ldots$$

and $\mathsf{A}(x) = [\kappa_{\min}, \kappa_{\max}]$ for all $x \in \mathsf{X}$. The one-stage reward function is $u(x_t - a_t)1_{\{x_t \geq a_t\}}$.

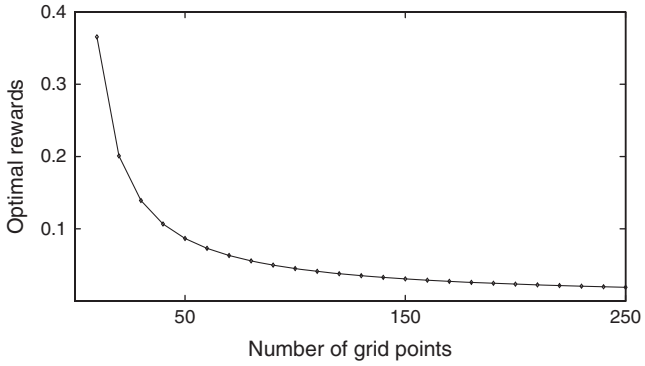
Since X is already compact, it is sufficient to discretize $[\kappa_{\min}, \kappa_{\max}]$. The noise process $\{v_t\}$ is a sequence of i.i.d. random variables that have common density g supported on $[0, \lambda]$. Therefore, the transition probability $p(\cdot \mid x, a)$ is given by

$$\begin{aligned} p(D \mid x, a) &= \Pr\{x_{t+1} \in D \mid x_t = x, a_t = a\} \\ &= \Pr\{\theta_1 \min(a, x) \exp\{-\theta_2 \min(a, x) + v\} \in D\} \\ &= \int_D g(\log(v) - \log(\theta_1 \min(a, x)) + \theta_2 \min(a, x)) \frac{1}{v} m(dv), \end{aligned}$$

for all $D \in \mathcal{B}(\mathbb{R})$. To make the model consistent, we must have $\theta_1 y \exp\{-\theta_2 y + v\} \in [\kappa_{\min}, \kappa_{\max}]$ for all $(y, v) \in [\kappa_{\min}, \kappa_{\max}] \times [0, \lambda]$.

We assume that (i) $g > \epsilon$ for some $\epsilon \in \mathbb{R}_+$ on $[0, \lambda]$, (ii) g is continuous on $[0, \lambda]$, and (iii) the utility function u is continuous. Define $h(v, x, a) := g(\log(v) - \log(\theta_1 \min(a, x)) + \theta_2 \min(a, x))(1/v)$, and for each $(x, a) \in \mathsf{X} \times \mathsf{A}$, let $S_{x,a}$ denote the support of $h(\cdot, x, a)$. Assumption 2.2 then holds for this model with $\theta(x, a) = \inf_{v \in S_a} h(v, x, a)$ (provided that it is measurable), $\zeta = m_\kappa$ (Lebesgue measure restricted on $[\kappa_{\min}, \kappa_{\max}]$), and for some $\lambda \in (0, 1)$.

Figure 2. Optimal Rewards of the Finite Models When the Initial State Is $x = 2$



For the numerical results, we use the following values of the parameters:

$$\theta_1 = 1.1, \qquad \theta_2 = 0.1, \qquad \kappa_{\max} = 7, \qquad \kappa_{\min} = 0.005, \qquad \lambda = 0.5.$$

We assume that the noise process is distributed uniformly over $[0, 0.5]$. Hence, $g \equiv 1$ on $[0, 0.5]$ and otherwise zero. The utility function u is taken to be the shifted isoelastic utility function (see Dufour and Prieto-Rumeau [13, Section 4.1])

$$u(z) = 3((z + 0.5)^{1/3} - (0.5)^{1/3}).$$

We selected 25 different values for the number n of grid points to discretize the state space: $n = 10, 20, 30, \ldots, 250$. The grid points are chosen uniformly over the interval $[\kappa_{\min}, \kappa_{\max}]$. We also uniformly discretize the action space A by using the following number of grid points: $5n = 50, 100, 150, \ldots, 1{,}250$.

We use the relative value iteration algorithm (see Bertsekas [4, Chapter 4.3.1]) to compute the value functions of the finite models. For each n, the finite state models are constructed as in Section 2 by replacing Z with $[\kappa_{\min}, \kappa_{\max}]$ and by setting $v_n(\cdot) = m_\kappa(\cdot)$.

Figure 2 shows the graph of the value functions of the finite models corresponding to the different values of n (number of grid points) when the initial state is $x = 2$. It can be seen that the value functions converge (to the value function of the original model).

8. Conclusion

The approximation of a discrete time MDP by finite-state MDPs was considered for discounted and average costs for both compact and noncompact state spaces. Under usual conditions imposed for studying Markov decision processes, it was shown that if one uses a sufficiently large number of grid points to discretize the state space, then the resulting finite-state MDP yields a near optimal policy. Under the Lipschitz continuity of the transition probability and the one-stage cost function, explicit bounds were derived on the performance loss due to discretization in terms of the number of grid points for the compact state case. These results were then illustrated numerically by considering two different MDP models.

Acknowledgments

Parts of this work were presented at the American Control Conference in July 2015, Chicago, IL.

References

[1] Aliprantis CD, Border KC (2006) *Infinite Dimensional Analysis* (Springer, Berlin).
[2] Bartoszynski R (1961) A characterization of the weak convergence of measures. *Ann. Math. Statist.* 32(2):561–576.
[3] Bertsekas DP (1975) Convergence of discretization procedures in dynamic programming. *IEEE Trans. Autom. Control* 20(3):415–419.
[4] Bertsekas DP (1995) *Dynamic Programming and Optimal Control*, Vol. II (Athena Scientific, Belmont, MA).
[5] Bertsekas DP, Shreve SE (1978) *Stochastic Optimal Control: The Discrete Time Case* (Academic Press, New York).
[6] Bertsekas DP, Tsitsiklis JN (1996) *Neuro-Dynammic Programming* (Athena Scientific, Belmont, MA).
[7] Blackwell D, Freedman D, Orkin M (1974) The optimal reward operator in dynamic programming. *Ann. Probab.* 2(2):926–941.
[8] Borkar V (2002) Convex analytic methods in Markov decision processes. Feinberg E, Shwartz A, eds. *Handbook of Markov Decision Processes* (Kluwer Academic Publishers, Dordrecht, Netherlands), 347–375.
[9] Cavazos-Cadena R (1986) Finite-state approximations for denumerable state discounted Markov decision processes. *Appl. Math. Optim.* 14(1):1–26.
[10] Chang HS, Fu MC, Hu J, Marcus SI (2007) A survey of some simulation-based methods in Markov decision processes. *Comm. Inform. System* 7(1):59–92.

[11] Chow C-S, Tsitsiklis JN (1991) An optimal one-way multigrid algorithm for discrete-time stochastic control. *IEEE Trans. Automatic Control* 36(8):898–914.
[12] Cover TM, Thomas JA (2006) *Elements of Information Theory*, 2nd ed. (John Wiley & Sons, Hoboken, NJ).
[13] Dufour F, Prieto-Rumeau T (2012) Approximation of Markov decision processes with general state space. *J. Math. Anal. Appl.* 388(2):1254–1267.
[14] Dufour F, Prieto-Rumeau T (2013) Finite linear programming approximations of constrained discounted Markov decision processes. *SIAM J. Control Optim.* 51(2):1298–1324.
[15] Dufour F, Prieto-Rumeau T (2015) Approximation of average cost Markov decision processes using empirical distributions and concentration inequalities. *Stochastics* 87(2):273–307.
[16] Feinberg EA, Kasyanov PO, Zadioanchuk NV (2012) Average cost Markov decision processes with weakly continuous transition probabilities. *Math. Oper. Res.* 37(4):591–607.
[17] Fox BL (1971) Finite-state approximations to denumerable state dynamic programs. *J. Math. Anal. Appl.* 34(3):665–670.
[18] Gordienko E, Hernandez-Lerma O (1995) Average cost Markov control processes with weighted norms: Existence of canonical policies. *Appl. Math.* 23(2):199–218.
[19] Gray GM, Neuhoff DL (1998) Quantization. *IEEE Trans. Inf. Theory* 44(6):2325–2383.
[20] Hernández-Lerma O (1989) *Adaptive Markov Control Processes* (Springer, New York).
[21] Hernández-Lerma O, Lasserre JB (1996) *Discrete-Time Markov Control Processes: Basic Optimality Criteria* (Springer, New York).
[22] Hernández-Lerma O, Lasserre JB (1999) *Further Topics on Discrete-Time Markov Control Processes* (Springer, New York).
[23] Hernández-Lerma O, Lasserre JB (2003) *Markov Chains and Invariant Probabilities* (Birkhäuser, Basel, Switzerland).
[24] Hinderer K (2005) Lipschitz continuity of value functions in Markovian desision processes. *Math. Methods Oper. Res.* 62(1):3–22.
[25] Jain R, Varaiya PP (2006) Simulation-based uniform value function estimates of Markov decision processes. *SIAM J. Control Optim.* 45(5):1633–1656.
[26] Jaśkiewicz A, Nowak AS (2006) On the optimality equation for average cost Markov control processes with Feller transition probabilities. *J. Math. Anal. Appl.* 316(2):495–509.
[27] Kuratowski K (1966) *Topology*, Vol. I (Academic Press, New York).
[28] Langen HJ (1981) Convergence of dynamic programming models. *Math. Oper. Res.* 6(4):493–512.
[29] Meyn SP, Tweedie RL (1993) *Markov Chains and Stochastic Stability* (Springer, New York).
[30] Ortner R (2007) Pseudometrics for state aggregation in average reward Markov decision processes. *Algorithmic Learning Theory* (Springer, Berlin), 373–387.
[31] Puterman ML (2005) *Markov Decision Processes* (John Wiley & Sons, Hoboken, NJ).
[32] Ren Z, Krogh BH (2002) State aggregation in Markov decision processes. *Proc. 41st IEEE Conf. Decision Control* (IEEE, Piscataway, NJ), 3819–3824.
[33] Saldi N, Linder T, Yüksel S (2015) Asymptotic optimality and rates of convergence of quantized stationary policies in stochastic control. *IEEE Trans. Autom. Control* 60(2):553–558.
[34] Saldi N, Yüksel S, Linder T (2016) Near optimality of quantized policies in stochastic control under weak continuity conditions. *J. Math. Anal. Appl.* 435(1):321–337.
[35] Serfozo R (1982) Convergence of Lebesgue integrals with varying measures. *Sankhya Ser. A* 44(3):380–402.
[36] Shreve SE, Bertsekas DP (1979) Universally measurable policies in dynamic programming. *Math. Oper. Res.* 4(1):15–30.
[37] Van Roy B (2006) Performance loss bounds for approximate value iteration with state aggregation. *Math. Oper. Res.* 31(2):234–244.
[38] Vega-Amaya O (2003) The average cost optimality equation: A fixed point approach. *Bol. Soc. Mat. Mexicana* 9(3):185–195.
[39] Villani C (2009) *Optimal Transport: Old and New* (Springer, Berlin).
[40] White DJ (1980) Finite-state approximations for denumerable state infinite horizon discounted Markov decision processes. *J. Math. Anal. Appl.* 74(1):292–295.
[41] White DJ (1982) Finite-state approximations for denumerable state infinite horizon discounted Markov decision processes with unbounded rewards. *J. Math. Anal. Appl.* 86(1):292–306.
[42] Whitt W (1978) Approximations of dynamic programs I. *Math. Oper. Res.* 3(3):231–243.
[43] Whitt W (1979) Approximations of dynamic programs II. *Math. Oper. Res.* 4(2):179–185.
[44] Yamada Y, Tazaki S, Gray RM (1980) Asymptotic performance of block quantizers with difference distortion measures. *IEEE Trans. Inf. Theory* 26(1):6–14.
[45] Yu H, Bertsekas DP (2004) Discretized approximations for POMDP with average cost. *The 20th Conf. UAI* (AUAI Press, Arlington, VA).
[46] Yüksel S, Başar T (2013) *Stochastic Networked Control Systems: Stabilization and Optimization Under Information Constraints* (Birkhäuser, Boston).