Near Optimality of Quantized Policies in Stochastic Control Under Weak Continuity Conditions[☆]

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

[a]*Department of Mathematics and Statistics, Queen's University, Kingston, ON, Canada.*

Abstract

This paper studies the approximation of optimal control policies by quantized (discretized) policies for a very general class of Markov decision processes (MDPs). The problem is motivated by applications in networked control systems, computational methods for MDPs, and learning algorithms for MDPs. We consider the finite-action approximation of stationary policies for a discrete-time Markov decision process with discounted and average costs under a weak continuity assumption on the transition probability, which is a significant relaxation of conditions required in earlier literature. The discretization is constructive, and quantized policies are shown to approximate optimal deterministic stationary policies with arbitrary precision. The results are applied to the fully observed reduction of a partially observed Markov decision process, where weak continuity is a much more reasonable assumption than more stringent conditions such as strong continuity or continuity in total variation.

Keywords: Stochastic control, quantization, approximation, partially observed Markov decision processes

1. Introduction

In this paper, we study the finite-action approximation of optimal control policies for a discrete time Markov decision process (MDP) with Borel state and action spaces, under discounted and average cost criteria. Various stochastic control problems may benefit from such an investigation.

The optimal information transmission problem in networked control systems is one such example. In many applications to networked control, the perfect transmission of the control actions to an actuator is infeasible when there is a communication channel of finite capacity between a controller and an actuator. Hence, the actions of the controller must be quantized to facilitate reliable transmission to an actuator. Although, the problem of optimal

---

*Corresponding author

Email addresses: `nsaldi@mast.queensu.ca` (Naci Saldi), `yuksel@mast.queensu.ca` (Serdar Yüksel), `linder@mast.queensu.ca` (Tamás Linder)

information transmission from a plant/sensor to a controller has been studied extensively [1], much less is known about the problem of transmitting actions from a controller to an actuator. Such transmission schemes usually require a simple encoding/decoding rule since an actuator does not have the computational/intelligence capability of a controller to use complex algorithms. Therefore, time-invariant uniform quantization is a practically useful encoding rule for controller-actuator communication.

The investigation of the finite-action approximation problem is also useful in computing near optimal policies and learning algorithms for MDPs. In a recent work [2], we consider the development of *finite-state* approximations for obtaining near optimal policies. However, to establish constructive control schemes, one needs to quantize the action space as well. Thus, results on approximate optimality of finite-action models pave the way for practical computation algorithms which are commonly used for finite-state/action MDPs. One other application regarding approximation problems is on learning a controlled Markov chain using simulations. If one can ensure that learning a control model with only finitely many control actions is sufficient for approximate optimality, then it is easier to develop efficient learning algorithms which allow for the approximate computation of finitely many transition probabilities. In particular, results developed in the learning and information theory literature for conditional kernel estimations [3] (with control-free models) can be applied to transition probability estimation for MDPs.

Motivated as above, in this paper we investigate the following approximation problem: For uncountable Borel state and action spaces, under what conditions can the optimal performance (achieved by some optimal stationary policy) be arbitrarily well approximated if the controller action set is restricted to be finite? We show that quantized stationary policies obtained by uniform quantization of the action space can approximate optimal policies with arbitrary precision for an MDP with an unbounded one-stage cost function, under a weak continuity assumption on the transition probability.

Various approximation results, which are somewhat related to our work, have been established for MDPs with Borel state and action spaces in the literature along the theme of computation of near optimal policies. For rather complete surveys of these techniques, we refer the reader to [4, 5, 6, 7, 8, 9, 10, 11, 12] and the references therein. With the exception of [12], these works assume in general restrictive continuity conditions on the transition probability. In [12], the authors considered an approximation problem in which all the components of the original model are *allowed* to vary in the reduced model (varying only the action space corresponds to the setup considered in this paper). Under weak continuity of the transition probability, [12] established the convergence of the reduced models to the original model for the discounted cost when the one-stage cost function is bounded. In this paper we allow the one-stage cost function to be unbounded. In addition, we also study the approximation problem for the challenging average cost case. Hence, our results can be applied to a wider range of stochastic systems. However, analogous with [12], the price we pay for imposing weaker assumptions is that we do not obtain explicit performance bounds in terms of the rate of the quantizer used in the approximations.

In [13] we solved a variant of this problem for the discounted cost under the following assumptions: (i) the action space is compact, (ii) the transition probability is setwise contin-

uous in the action variable, and (iii) the one stage cost function is bounded and continuous in the action variable. The average cost was also considered under some additional restrictions on the ergodicity properties of Markov chains induced by deterministic stationary policies. In this paper we consider a similar problem for systems where the transition probability is weakly continuous in the state-action variables. An important motivation for considering these conditions comes from the fact that for the fully observed reduction of a partially observed MDP (POMDP), the setwise continuity of the transition probability in the action variable is a prohibitive condition even for simple systems such as the one described in Example 2.1 in the next section.

Organization: In Section 2 we give definitions and the problem formulation. The main result for discounted cost is stated and proved in Section 3. In Section 4 an analogous approximation result is obtained for the average cost criterion. In Section 5 the results for the discounted cost are applied to the fully observed reduction of POMDPs via appealing to results by Feinberg *et al.* [14]. Section 6 concludes the paper.

2. Markov Decision Processes

For a metric space E, the Borel σ-algebra is denoted by $\mathcal{B}(\mathsf{E})$. Given any Borel measurable function $w : \mathsf{E} \to [1, \infty)$ and any real valued Borel measurable function u on E, we define w-norm of u as

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

Let $C_w(\mathsf{E})$ and $B_w(\mathsf{E})$ denote the space of all real valued continuous and Borel measurable functions on E with finite w-norm, respectively [15]. Let $\mathcal{P}(\mathsf{E})$ denote the set of all probability measures on E. A sequence $\{\mu_n\}$ of measures on E is said to converge weakly (resp., setwise) [16] to a measure μ if $\int_{\mathsf{E}} g(e)\mu_n(de) \to \int_{\mathsf{E}} g(e)\mu(de)$ for all bounded and continuous real functions g on E (resp., for all bounded and Borel measurable real functions g on E). Unless otherwise specified, the term "measurable" will refer to Borel measurability.

We consider a discrete-time Markov decision process (MDP) with *state space* Z and *action space* A, where Z and A are Borel spaces. In this paper, we assume that the *set of admissible actions* for any $z \in \mathsf{Z}$ is A. Let the *stochastic kernel* $\eta(\,\cdot\,|z, a)$ be the *transition probability* of the next state given that the previous state-action pair is (z, a) [17]. The *one-stage cost* function c is a measurable function from $\mathsf{Z} \times \mathsf{A}$ to $[0, \infty)$. The probability measure ξ over Z denotes the initial distribution of the state process.

Define the history spaces $\mathsf{H}_0 = \mathsf{Z}$ and $\mathsf{H}_t = (\mathsf{Z} \times \mathsf{A})^t \times \mathsf{Z}$, $t = 1, 2, \ldots$ endowed with their product Borel σ-algebras generated by $\mathcal{B}(\mathsf{Z})$ and $\mathcal{B}(\mathsf{A})$. A *policy* is a sequence $\varphi = \{\varphi_t\}$ of stochastic kernels on A given H_t. The set of all policies is denoted by Φ. A *deterministic* policy is a sequence $\varphi = \{\varphi_t\}$ of stochastic kernels on A given H_t which are realized by a sequence of measurable functions $\{f_t\}$ from H_t to A, i.e., $\varphi_t(\,\cdot\,|h_t) = \delta_{f_t(h_t)}(\,\cdot\,)$, where δ_z is the point mass at z. Let $\mathbb{F}$ denote the set of all measurable functions f from Z to A. A *deterministic stationary* policy is a constant sequence of stochastic kernels $\varphi = \{\varphi_t\}$ on A

3

given Z such that $\varphi_t(\cdot|z) = \delta_{f(z)}(\cdot)$ for all t for some $f \in \mathbb{F}$. The set of deterministic stationary policies is identified with the set $\mathbb{F}$.

According to the Ionescu Tulcea theorem [17], an initial distribution ξ on Z and a policy φ define a unique probability measure P_ξ^φ on $\mathsf{H}_\infty = (\mathsf{Z} \times \mathsf{A})^\infty$. The expectation with respect to P_ξ^φ is denoted by $\mathbb{E}_\xi^\varphi$. If $\xi = \delta_z$, we write P_z^φ and $\mathbb{E}_z^\varphi$ instead of $P_{\delta_z}^\varphi$ and $\mathbb{E}_{\delta_z}^\varphi$. The cost functions to be minimized in this paper are the discounted cost with a discount factor $\beta \in (0,1)$ and the average cost, respectively:

$$J(\varphi, \xi) = \mathbb{E}_\xi^\varphi\left[\sum_{t=0}^\infty \beta^t c(z_t, a_t)\right],$$

$$V(\varphi, \xi) = \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}_\xi^\varphi\left[\sum_{t=0}^{T-1} c(z_t, a_t)\right].$$

A policy φ^* is called optimal if $J(\varphi^*, z) = \inf_{\varphi\in\Phi} J(\varphi, z)$ (or $V(\varphi^*, z) = \inf_{\varphi\in\Phi} V(\varphi, z)$ for the average cost) for all $z \in \mathsf{Z}$. It is well known that the set $\mathbb{F}$ of deterministic stationary policies contains an optimal policy for a large class of infinite horizon discounted cost (see, e.g., [17, 18]) and average cost optimal control problems (see, e.g., [19, 18]). In this case we say that $\mathbb{F}$ is an optimal class.

2.1. Problem Formulation

To give a precise definition of the problem we study in this paper, we first give the definition of a quantizer from the state to the action space.

Definition 2.1. A measurable function $q : \mathsf{Z} \to \mathsf{A}$ is called a *quantizer* from Z to A if the range of q, i.e., $q(\mathsf{Z}) = \{q(z) \in \mathsf{A} : z \in \mathsf{Z}\}$, is finite.

The elements of $q(\mathsf{Z})$ (the possible values of q) are called the *levels* of q. The rate $R = \log_2|q(\mathsf{Z})|$ of a quantizer q (approximately) represents the number of bits needed to losslessly encode the output levels of q using binary codewords of equal length. Let $\mathcal{Q}$ denote the set of all quantizers from Z to A. A *deterministic stationary quantizer* policy is a constant sequence $\varphi = \{\varphi_t\}$ of stochastic kernels on A given Z such that $\varphi_t(\cdot|z) = \delta_{q(z)}(\cdot)$ for all t for some $q \in \mathcal{Q}$. For any finite set $\Lambda \subset \mathsf{A}$, let $\mathcal{Q}(\Lambda)$ denote the set of all elements in $\mathcal{Q}$ having range Λ. Analogous with $\mathbb{F}$, the set of all deterministic stationary quantizer policies induced by $\mathcal{Q}(\Lambda)$ will be identified with the set $\mathcal{Q}(\Lambda)$.

Our main objective is to find conditions on the components of the MDP under which there exists a sequence of finite subsets $\{\Lambda_n\}_{n\geq 1}$ of A for which the following holds:

(P) For any initial point z we have $\lim_{n\to\infty} \min_{q\in\mathcal{Q}(\Lambda_n)} J(q, z) = \min_{f\in\mathbb{F}} J(f, z)$ (or $\lim_{n\to\infty} \min_{q\in\mathcal{Q}(\Lambda_n)} V(q, z) = \min_{f\in\mathbb{F}} V(f, z)$ for the average cost), provided that the set $\mathbb{F}$ of deterministic stationary policies is an optimal class for the MDP.

Letting MDP_n be defined as the Markov decision process having the components $\{\mathsf{Z}, \Lambda_n, \eta, c\}$, **(P)** is equivalent to stating that optimal cost of MDP_n converges to the optimal cost of the original MDP.

4

2.2. Setwise Continuity versus Weak Continuity

Requiring that the transition probability of the system be weakly continuous in state-action variables is indeed a fairly mild assumption compared to the setwise continuity in the action variable. Indeed, the latter condition is even prohibitive for certain stochastic systems. In this section we consider two examples, the fully observed reduction of a POMDP and the additive noise model, in order to more explicitly highlight this. We refer the reader to [20, Chapter 4] and Section 5 of this paper for the basics of POMDPs.

Example 2.1. Consider the system dynamics

$$x_{t+1} = x_t + a_t,$$
$$y_t = x_t + v_t,$$

where $x_t \in \mathsf{X}$, $y_t \in \mathsf{Y}$ and $a_t \in \mathsf{A}$, and where X, Y and A are the state, observation and action spaces, respectively. Here, we assume that $\mathsf{X} = \mathsf{Y} = \mathsf{A} = \mathbb{R}_+$ and the 'noise' $\{v_t\}$ is a sequence of i.i.d. random variables uniformly distributed on $[0, 1]$. It is easy to see that the transition probability is weakly continuous (with respect to state-action variables) and the observation channel that gives the conditional distribution of the current observation given the current state is continuous in total variation (with respect to state variable) for this POMDP. Hence, by [14, Theorem 3.7] the transition probability η of the fully observed reduction of the POMDP is weakly continuous in the state-action variables. However, the same conclusion cannot be drawn for the setwise continuity of η with respect to the action variable as shown below.

Let z denote the generic state variable for the fully observed reduced MDP, where the state variables are elements of $\mathcal{P}(\mathsf{X})$ which is equipped with the Borel σ-algebra generated by the topology of weak convergence. If we define the function $F(z, a, y) := \Pr\{x_{t+1} \in \cdot \,|z_t = z, a_t = a, y_{t+1} = y\}$ from $\mathcal{P}(\mathsf{X}) \times \mathsf{A} \times \mathsf{Y}$ to $\mathcal{P}(\mathsf{X})$ and the stochastic kernel $H(\,\cdot\,|z, a) := \Pr\{y_{t+1} \in \cdot \,|z_t = z, a_t = a\}$ on Y given $\mathcal{P}(\mathsf{X}) \times \mathsf{A}$, then η can be written as

$$\eta(\,\cdot\,|z, a) = \int_\mathsf{Y} 1_{\{F(z,a,y) \in \cdot\,\}} H(dy|z, a),$$

where 1_D denotes the indicator function of an event D and z_t denotes the posterior distribution of the state x_t given the past observations, i.e.,

$$z_t(\,\cdot\,) := \Pr\{x_t \in \cdot \,|y_0, \ldots, y_t, a_0, \ldots, a_{t-1}\}.$$

Let us set $z = \delta_0$ (point mass at $0 \in \mathsf{X}$), $\{a_k\} = \{\frac{1}{k}\}$, and $a = 0$. Hence, $a_k \to a$. We show that $\eta(\,\cdot\,|z, a_k) \nrightarrow \eta(\,\cdot\,|z, a)$ setwise.

Observe that for all k and $y \in \mathsf{Y}$, we have $F(z, a_k, y) = \delta_{\frac{1}{k}}$ and $F(z, a, y) = \delta_0$. Define the open set O with respect to the weak topology in $\mathcal{P}(\mathsf{X})$ as

$$O := \Big\{z \in \mathcal{P}(\mathsf{X}) : \Big|\int_\mathsf{X} g(x)\delta_1(dx) - \int_\mathsf{X} g(x)z(dx)\Big| < 1\Big\},$$

where g is the symmetric triangular function on $[-1,1]$ with $g(0) = 1$ and $g(-1) = g(1) = 0$. Observe that we have $F(z, a_k, y) \in O$ for all k and y, but $F(z, a, y) \notin O$ for all y. Hence,

$$\eta(O|z, a_k) := \int_{\mathsf{Y}} 1_{\{F(z,a_k,y)\in O\}} H(dy|z, a_k) = 1,$$

but

$$\eta(O|z, a) := \int_{\mathsf{Y}} 1_{\{F(z,a,y)\in O\}} H(dy|z, a) = 0$$

implying that $\eta(\cdot|z, a_k)$ does not converge to $\eta(\cdot|z, a)$ setwise. Hence, η does not satisfy the setwise continuity assumption.

Example 2.2. In this example we consider an additive-noise system given by

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

where $\mathsf{Z} = \mathbb{R}^n$ and $\mathsf{A} = \mathbb{R}^m$ for some $n, m \geq 1$. The noise process $\{v_t\}$ is a sequence of independent and identically distributed (i.i.d.) random vectors. In such a system, the continuity of F in (z, a) (which holds for most practical systems) is sufficient to imply the weak continuity of the transition probability, and no assumptions are needed on the noise process (not even the existence of a density is required). Hence, weak continuity does not restrict the noise model and is satisfied by almost all systems in the applications, whereas other conditions, such as strong continuity or continuity in total variation distance, hold only if the noise is well behaved in addition to the continuity of F. For example, for setwise continuity, it is usually required that F is continuous in a for every z, the noise admits a density, and this density is continuous [17, Example C.8].

3. Near Optimality of Quantized Policies with Discounted Cost

In this section we consider the problem **(P)** for the discounted cost. The following assumptions will be imposed for both the discounted cost and the average cost. We note that these assumptions are often used in the literature for studying discounted Markov decision processes with unbounded one-stage cost and weakly continuous transition probability.

Assumption 3.1.

(a) The one stage cost function c is nonnegative and continuous.

(b) The stochastic kernel $\eta(\cdot|z, a)$ is weakly continuous in $(z, a) \in \mathsf{Z} \times \mathsf{A}$, i.e., if $(z_k, a_k) \to (z, a)$, then $\eta(\cdot|z_k, a_k) \to \eta(\cdot|z, a)$ weakly.

(c) A is compact.

(d) There exist nonnegative real numbers M and $\alpha \in [1, \frac{1}{\beta})$, and a continuous weight function $w : \mathsf{Z} \to [1, \infty)$ such that for each $z \in \mathsf{Z}$, we have

$$\sup_{a \in \mathsf{A}} c(z, a) \leq Mw(z), \tag{1}$$

6

$$\sup_{a \in \mathsf{A}} \int_{\mathsf{Z}} w(y)\eta(dy|z,a) \leq \alpha w(z), \tag{2}$$

and $\int_{\mathsf{Z}} w(y)\eta(dy|z,a)$ is continuous in (z,a).

Let d_{A} denote the metric on A. Since A is compact and thus totally bounded, one can find a sequence of finite sets $\Lambda_n = \{a_{n,1}, \ldots, a_{n,k_n}\} \subset \mathsf{A}$ such that for all n,

$$\min_{i \in \{1,\ldots,k_n\}} d_{\mathsf{A}}(a, a_{n,i}) < 1/n \text{ for all } a \in \mathsf{A}. \tag{3}$$

In other words, Λ_n is a $1/n$-net in A. In the rest of this paper, we assume that the sequence $\{\Lambda_n\}_{n \geq 1}$ is fixed. To ease the notation in the sequel, let us define the mapping $\Upsilon_n : \mathbb{F} \to \mathcal{Q}(\Lambda_n)$ by

$$\Upsilon_n(f)(z) = \arg\min_{a \in \Lambda_n} d_{\mathsf{A}}(f(z), a). \tag{4}$$

Hence, for all $f \in \mathbb{F}$, we have

$$\sup_{z \in \mathsf{Z}} d_{\mathsf{A}}\big(\Upsilon_n(f)(z), f(z)\big) < 1/n. \tag{5}$$

Define the operator T on the set of real-valued measurable functions on Z by

$$Tu(z) := \min_{a \in \mathsf{A}} \left[c(z,a) + \beta \int_{\mathsf{Z}} u(y)\eta(dy|z,a) \right]. \tag{6}$$

In the literature T is called the *Bellman optimality operator*.

Lemma 3.1. *For any $u \in C_w(\mathsf{Z})$ the function $l_u(z,a) := \int_{\mathsf{Z}} u(y)\eta(dy|z,a)$ is continuous in (z,a).*

Proof. For any nonnegative continuous function g on Z, the function $l_g(z,a) = \int_{\mathsf{Z}} g(y)\eta(dy|z,a)$ is lower semi-continuous in (z,a), if η is weakly continuous (see, e.g., [17, Proposition E.2]). Define the nonnegative continuous function g by letting $g = bw + u$, where $b = \|u\|_w$. Then l_g is lower semi-continuous. Since $l_u = l_g - bl_w$ and l_w is continuous by Assumption 3.1-(d), l_u is lower semi-continuous. Analogously, define the nonnegative continuous function v by letting $v = -u + bw$. Then l_v is lower semi-continuous. Since $l_u = bl_w - l_v$ and l_w is continuous by Assumption 3.1-(d), l_u is also upper semi-continuous. Therefore, l_u is continuous. $\square$

Lemma 3.2. *Let Y be any of the compact sets A or Λ_n. Define the operator T_{Y} on $B_w(\mathsf{Z})$ by letting*

$$T_{\mathsf{Y}}u(z) := \min_{a \in \mathsf{Y}} \left[c(z,a) + \beta \int_{\mathsf{Z}} u(y)\eta(dy|z,a) \right].$$

Then T_{Y} maps $C_w(\mathsf{Z})$ into itself. Moreover, $C_w(\mathsf{Z})$ is closed with respect to the w-norm.

Proof. Note that $T_Y u(z) = \min_{a \in Y} \big(c(z,a) + \beta l_u(z,a) \big)$. The function l_u is continuous by Lemma 3.1, and therefore, $T_Y u$ is also continuous by [21, Proposition 7.32]. Since T_Y maps $B_w(Z)$ into itself, $T_Y u \in C_w(Z)$.

For the second statement, let u_n converge to u in w-norm in $C_w(Z)$. It is sufficient to prove that u is continuous. Let $z_k \to z_0$. Since $B := \{z_0, z_1, z_2, \dots\}$ is compact, w is bounded on B. Therefore, $u_n \to u$ uniformly on B which implies that $\lim_{k \to \infty} u(z_k) = u(z_0)$. This completes the proof. $\qquad\square$

Lemma 3.2 implies that T maps $C_w(Z)$ into itself. It can also be proved that T is a contraction operator with modulus $\sigma := \beta\alpha$ (see [15, Lemma 8.5.5]); that is,

$$\|Tu - Tv\|_w \le \sigma\|u - v\|_w \text{ for all } u, v \in C_w(Z).$$

Define the discounted value function J^* by

$$J^*(z) := \inf_{\varphi \in \Phi} J(\varphi, z).$$

The following theorem is a known result in the theory of Markov decision processes (see e.g., [15, Section 8.5, p. 65]).

Theorem 3.1. *Suppose Assumption 3.1 holds. Then, the value function J^* is the unique fixed point in $C_w(Z)$ of the contraction operator T, i.e.,*

$$J^* = TJ^*. \tag{7}$$

Furthermore, a deterministic stationary policy f^* is optimal if and only if

$$J^*(z) = c(z, f^*(z)) + \beta \int_Z J^*(y)\eta(dy|z, f^*(z)). \tag{8}$$

Finally, there exists a deterministic stationary policy f^* which is optimal, so it satisfies (8).

Define, for all $n \ge 1$, the operator T_n (which will be used to approximate T) by

$$T_n u(z) := \min_{a \in \Lambda_n} \left[c(z, a) + \beta \int_Z u(y)\eta(dy|z, a) \right]. \tag{9}$$

Note that T_n is the Bellman optimality operator for MDP_n having components $\{Z, \Lambda_n, \eta, c\}$. Analogous with T, it can be shown that T_n is a contraction operator with modulus $\sigma = \alpha\beta$ mapping $C_w(Z)$ into itself. Let $J_n^* \in C_w(Z)$ (discounted value function of MDP_n) denote the fixed point of T_n.

The following theorem is the main result of this section which states that the discounted value function of MDP_n converges to the discounted value function of the original MDP.

Theorem 3.2. *For any compact set $K \subset Z$ we have*

$$\lim_{n \to \infty} \sup_{z \in K} |J_n^*(z) - J^*(z)| = 0. \tag{10}$$

Therefore,

$$\lim_{n \to \infty} |J_n^*(z) - J^*(z)| = 0 \quad \text{for all } z \in Z.$$

8

To prove Theorem 3.2, we need following results.

Lemma 3.3. *For any compact subset K of Z and for any $\varepsilon > 0$, there exists a compact subset K_ε of Z such that*

$$\sup_{(z,a)\in K\times\mathsf{A}} \int_{K_\varepsilon^c} w(y)\eta(dy|z,a) < \varepsilon. \tag{11}$$

Proof. Let us define the set of measures Ξ on Z as

$$\Xi := \left\{ Q(\,\cdot\,|z,a) : Q(D|z,a) = \int_D w(y)\eta(dy|z,a), \ (z,a) \in K \times \mathsf{A} \right\}.$$

Note that Ξ is uniformly bounded since

$$\sup_{(z,a)\in K\times\mathsf{A}} \int_{\mathsf{Z}} w(y)\eta(dy|z,a) \le \alpha \sup_{z\in K} w(z) < \infty.$$

If the mapping $Q : K \times \mathsf{A} \ni (z,a) \mapsto Q(\,\cdot\,|z,a) \in \Xi$ is continuous with respect to the weak topology on Ξ, then Ξ (being a continuous image of the compact set $K \times \mathsf{A}$) is compact with respect to the weak topology. Then, by an extension of Prohorov's theorem to non-probability measures [22, Theorem 8.6.2], Ξ is tight, completing the proof. Hence, we only need to prove the continuity of the mapping Q.

By Lemma 3.1, for any $u \in C_w(\mathsf{Z})$, $\int_{\mathsf{Z}} u(y)\eta(dy|z,a)$ is continuous in (z,a). Let $(z_k, a_k) \to (z,a)$ in $K \times \mathsf{A}$. Note that for any $g \in C_b(\mathsf{Z})$, $gw \in C_w(\mathsf{Z})$. Therefore, we have

$$\lim_{k\to\infty} \int_{\mathsf{Z}} g(y)Q(dy|z_k,a_k) = \lim_{k\to\infty} \int_{\mathsf{Z}} g(y)w(y)\eta(dy|z_k,a_k)$$
$$= \int_{\mathsf{Z}} g(y)w(y)\eta(dy|z,a) = \int_{\mathsf{Z}} g(y)Q(dy|z,a)$$

proving that $Q(\,\cdot\,|z_k,a_k) \to Q(\,\cdot\,|z,a)$ weakly. $\qquad\square$

Lemma 3.4. *Let $\{u_n\}$ be a sequence in $C_w(\mathsf{Z})$ with $\sup_n \|u_n\|_w := L < \infty$. If u_n converges to $u \in C_w(\mathsf{Z})$ uniformly on each compact subset of Z, then for any $f \in \mathbb{F}$ and compact subset K of Z we have*

$$\lim_{n\to\infty} \sup_{z\in K} \left| \int_{\mathsf{Z}} u_n(y)\eta(dy|z, f_n(z)) - \int_{\mathsf{Z}} u(y)\eta(dy|z, f(z)) \right| = 0,$$

where $f_n = \Upsilon_n(f)$ (see (4)).

Proof. Fix a compact subset K of Z. Then for K_ε as in Lemma 3.3,

$$\sup_{z\in K} \left| \int_{\mathsf{Z}} u_n(y)\eta(dy|z, f_n(z)) - \int_{\mathsf{Z}} u(y)\eta(dy|z, f(z)) \right|$$

9

$$\leq \sup_{z \in K} \left| \int_{\mathsf{Z}} u_n(y)\eta(dy|z, f_n(z)) - \int_{\mathsf{Z}} u(y)\eta(dy|z, f_n(z)) \right|$$

$$+ \sup_{z \in K} \left| \int_{\mathsf{Z}} u(y)\eta(dy|z, f_n(z)) - \int_{\mathsf{Z}} u(y)\eta(dy|z, f(z)) \right|$$

$$\leq \sup_{z \in K} \left| \int_{K_\varepsilon} u_n(y)\eta(dy|z, f_n(z)) - \int_{K_\varepsilon} u(y)\eta(dy|z, f_n(z)) \right|$$

$$+ \sup_{z \in K} \left| \int_{K_\varepsilon^c} u_n(y)\eta(dy|z, f_n(z)) - \int_{K_\varepsilon^c} u(y)\eta(dy|z, f_n(z)) \right|$$

$$+ \sup_{z \in K} \left| \int_{\mathsf{Z}} u(y)\eta(dy|z, f_n(z)) - \int_{\mathsf{Z}} u(y)\eta(dy|z, f(z)) \right|$$

$$\leq \sup_{y \in K_\varepsilon} |u_n(y) - u(y)| + 2L\varepsilon + \sup_{z \in K} \left| \int_{\mathsf{Z}} u(y)\eta(dy|z, f_n(z)) - \int_{\mathsf{Z}} u(y)\eta(dy|z, f(z)) \right|.$$

Let us define $l(z, a) := \int_{\mathsf{Z}} u(y)\eta(dy|z, a)$. Since $u \in C_w(\mathsf{Z})$, by Lemma 3.1 l is continuous, and therefore, uniformly continuous on $K \times \mathsf{A}$. Note that in the last expression as $n \to \infty$: (i) the first term goes to zero since $u_n \to u$ uniformly on K_ε and (ii) the last term goes to zero since l is uniformly continuous on $K \times \mathsf{A}$ and $f_n \to f$ uniformly. Then the result follows by observing that ε is arbitrary. $\qquad\square$

Let us define $v^0 = v_n^0 = 0$, and $v^{t+1} = Tv^t$ and $v_n^{t+1} = T_n v_n^t$ for $t \geq 1$; that is, $\{v^t\}_{t \geq 1}$ and $\{v_n^t\}_{t \geq 1}$ are successive approximations to the discounted value functions of the original MDP and MDP$_n$, respectively. Lemma 3.2 implies that v^t and v_n^t are in $C_w(\mathsf{Z})$ for all t and n. By [15, Theorem 8.3.6, p. 47], [15, (8.3.34), p. 52] and [15, Section 8.5, p. 65] we have

$$v^t(z) \leq J^*(z) \leq M\frac{w(z)}{1 - \sigma}, \tag{12}$$

$$\|v^t - J^*\|_w \leq M\frac{\sigma^t}{1 - \sigma}, \tag{13}$$

and

$$v_n^t(z) \leq J_n^*(z) \leq M\frac{w(z)}{1 - \sigma}, \tag{14}$$

$$\|v_n^t - J_n^*\|_w \leq M\frac{\sigma^t}{1 - \sigma}. \tag{15}$$

Since for each n and u, $Tu \leq T_n u$, we also have $v^t \leq v_n^t$ for all $t \geq 1$ and $J^* \leq J_n^*$.

Lemma 3.5. *For any compact set $K \subset \mathsf{Z}$ and $t \geq 1$, we have*

$$\lim_{n \to \infty} \sup_{z \in K} |v_n^t(z) - v^t(z)| = 0. \tag{16}$$

Proof. We prove (16) by induction. For $t = 1$, the claim holds since $v^0 = v_n^0 = 0$, and c is uniformly continuous on $K \times A$ for any compact subset K of Z. Assume the claim is true for $t \geq 1$. We fix any compact set K. Let f_t^* denote the selector of $Tv^t = v^{t+1}$; that is,

$$v^{t+1}(z) = Tv^t(z) = c(z, f_t^*(z)) + \beta \int_Z v^t(y)\eta(dy|z, f_t^*(z)),$$

and let $f_{t,n}^* := \Upsilon_n(f_t^*)$ (see (4)). By (12) and (14) we have

$$v^t(z) \leq M \frac{w(z)}{1 - \sigma} \tag{17}$$

$$v_n^t(z) \leq M \frac{w(z)}{1 - \sigma}, \tag{18}$$

for all t and n. For each $n \geq 1$, we have

$$\sup_{z \in K} |v_n^{t+1}(z) - v^{t+1}(z)|$$
$$= \sup_{z \in K} (v_n^{t+1}(z) - v^{t+1}(z)) \quad (\text{as } v^{t+1} \leq v_n^{t+1})$$
$$= \sup_{z \in K} \left(\min_{\Lambda_n} \left[c(z, a) + \beta \int_Z v_n^t(y)\eta(dy|z, a) \right] - \min_A \left[c(z, a) + \beta \int_Z v^t(y)\eta(dy|z, a) \right] \right)$$
$$\leq \sup_{z \in K} \left(\left[c(z, f_{t,n}^*(z)) + \beta \int_Z v_n^t(y)\eta(dy|z, f_{t,n}^*(z)) \right] - \left[c(z, f_t^*(z)) + \beta \int_Z v^t(y)\eta(dy|z, f_t^*(z)) \right] \right)$$
$$\leq \sup_{z \in K} |c(z, f_{t,n}^*(z)) - c(z, f_t^*(z))| + \beta \sup_{z \in K} \left| \int_Z v_n^t(y)\eta(dy|z, f_{t,n}^*(z)) - \int_Z v^t(y)\eta(dy|z, f_t^*(z)) \right|.$$

Note that in the last expression, as $n \to \infty$ the first term goes to zero since c is uniformly continuous on $K \times A$ and $f_{t,n}^* \to f_t^*$ uniformly, and the second term goes to zero by Lemma 3.4, (17), and (18). $\square$

Now, using Lemma 3.5 we prove Theorem 3.2.

Proof of Theorem 3.2. Let us fix any compact set $K \subset Z$. Since w is bounded on K, it is enough to prove $\lim_{n \to \infty} \sup_{z \in K} \frac{|J_n^*(z) - J^*(z)|}{w(z)} = 0$. We have

$$\sup_{z \in K} \frac{|J_n^*(z) - J^*(z)|}{w(z)} \leq \sup_{z \in K} \frac{|J_n^*(z) - v_n^t(z)|}{w(z)} + \sup_{z \in K} \frac{|v_n^t(z) - v^t(z)|}{w(z)} + \sup_{z \in K} \frac{|v^t(z) - J^*(z)|}{w(z)}$$
$$\leq 2M \frac{\sigma^t}{1 - \sigma} + \sup_{z \in K} \frac{|v_n^t(z) - v^t(z)|}{w(z)} \quad (\text{by (13) and (15)}).$$

Since $w \geq 1$, $\sup_{z \in K} \frac{|v_n^t(z) - v^t(z)|}{w(z)} \to 0$ as $n \to \infty$ for all t by Lemma 3.5. Hence, the last expression can be made arbitrarily small since $t \geq 1$ is arbitrary and $\sigma \in (0, 1)$, this completes the proof. $\square$

4. Near Optimality of Quantized Policies with Average Cost

In this section we consider the problem **(P)** for the average cost. We prove an approximation result analogous to Theorem 3.2. To do this, some new assumptions are needed on the components of the original MDP in addition to the conditions in Assumption 3.1. A version of these assumptions were used in [23] and [24] to study the existence of the solution to the Average Cost Optimality Equality (ACOE) and Inequality (ACOI). For any probability measure ϑ and measurable function h on Z, let $\vartheta(h) := \int_{\mathsf{Z}} h(z)\vartheta(dz)$.

Assumption 4.1. Suppose Assumption 3.1 holds with (2) replaced by condition (e) below. Moreover, suppose there exist a probability measure λ on Z and a continuous function $\phi : \mathsf{Z} \times \mathsf{A} \to [0, \infty)$ such that

(e) $\int_{\mathsf{Z}} w(y)\eta(dy|z, a) \leq \alpha w(z) + \lambda(w)\phi(z, a)$ for all $(z, a) \in \mathsf{Z} \times \mathsf{A}$, where $\alpha \in (0, 1)$.

(f) $\eta(D|z, a) \geq \lambda(D)\phi(z, a)$ for all $(z, a) \in \mathsf{Z} \times \mathsf{A}$ and $D \in \mathcal{B}(\mathsf{Z})$.

(g) The weight function w is λ-integrable.

(h) $\int_{\mathsf{Z}} \phi(z, f(z))\lambda(dz) > 0$ for all $f \in \mathbb{F}$.

Any $f \in \mathbb{F}$ gives rise to a time-homogeneous Markov chain $\{z_t\}_{t=1}^{\infty}$ (state process) with the transition probability $\eta(\,\cdot\,|z, f(z))$ on Z given Z. For any $t \geq 1$, let $\eta^t(\,\cdot\,|z, f(z))$ denote the t-step transition probability of this Markov chain given the initial point z. Hence, $\eta^t(\,\cdot\,|z, f(z))$ is recursively given by

$$\eta^{t+1}(\,\cdot\,|z, f(z)) = \int_{\mathsf{Z}} \eta(\,\cdot\,|y, f(y))\eta^t(dy|z, f(z)).$$

For any $z \in \mathsf{Z}$, let

$$V^*(z) := \inf_{\varphi \in \Phi} V(\varphi, z).$$

V^* is called the average cost value function of the MDP. The following theorem is a consequence of [23, Theorems 3.3 and 3.6].

Theorem 4.1. *Under Assumption 4.1 the following holds.*

(i) *For each $f \in \mathbb{F}$, the stochastic kernel $\eta(\,\cdot\,|z, f(z))$ is positive Harris recurrent with unique invariant probability measure ν_f. Furthermore, w is ν_f-integrable, and therefore, $\rho_f := \int_{\mathsf{Z}} c(z, f(z))\nu_f(dz) < \infty$.*

(ii) *There exist $f^* \in \mathbb{F}$ and $h^* \in C_w(\mathsf{Z})$ such that the triple (h^*, f^*, ρ_{f^*}) satisfies the average cost optimality equality (ACOE), i.e.,*

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

12

$$= c(z, f^*(z)) + \int_{\mathsf{Z}} h^*(y)\eta(dy|z, f^*(z)),$$

and therefore,

$$V^*(z) = \rho_{f^*},$$

for all $z \in \mathsf{Z}$.

Proof. The only statement that does not directly follow from [23, Theorems 3.3 and 3.6] is the fact: $h^* \in C_w(\mathsf{Z})$. Hence, we only prove this.

By [23, Theorem 3.5], h^* is the unique fixed point of the following contraction operator with modulus α

$$Fu(z) := \min_{a \in \mathsf{A}} \left[c(z, a) + \int_{\mathsf{Z}} u(y)\eta(dy|z, a) - \lambda(u)\phi(z, a) \right].$$

Since ϕ is continuous, by Lemma 3.1 the function inside the minimization is continuous in (z, a) if $u \in C_w(\mathsf{Z})$. Then by Lemma 3.2, F maps $C_w(\mathsf{Z})$ into itself. Therefore, $h^* \in C_w(\mathsf{Z})$. $\qquad\square$

This theorem implies that for each $f \in \mathbb{F}$, the average cost is given by $V(f, z) = \int_{\mathsf{Z}} c(y, f(y))\nu_f(dy)$ for all $z \in \mathsf{Z}$ (instead of ν_f-a.e.).

Remark 4.1. If the state space Z is compact and the transition probability $\eta(\cdot|z, a)$ has a strictly positive density $g(y|z, a)$ with respect to some probability measure ϑ which is continuous in (y, z, a), then Assumption 4.1 holds for $w = 1$, $\lambda = \vartheta$, and $\phi(z, a) = \min_{y \in \mathsf{Z}} g(y|z, a)$.

Note that all the statements in Theorem 4.1 are also valid for MDP_n with an optimal policy f_n^* and a canonical triplet $(h_n^*, f_n^*, \rho_{f_n^*})$. Analogous with F, define the contraction operator F_n (with modulus α) corresponding to MDP_n as

$$F_n u(z) := \min_{a \in \Lambda_n} \left[c(z, a) + \int_{\mathsf{Z}} u(y)\eta(dy|z, a) - \lambda(u)\phi(z, a) \right],$$

and therefore, $h_n^* \in C_w(\mathsf{Z})$ is its fixed point.

The next theorem is the main result of this section which states that the average cost value function, denoted as V_n^*, of MDP_n converges to the average cost value function V^* of the original MDP.

Theorem 4.2. *We have*

$$\lim_{n \to \infty} |V_n^* - V^*| = 0,$$

where V_n^* and V^* are both constants.

Let us define $u^0 = u_n^0 = 0$, and $u^{t+1} = Fu^t$ and $u_n^{t+1} = F_n u_n^t$ for $t \geq 1$; that is, $\{u^t\}_{t \geq 1}$ and $\{u_n^t\}_{t \geq 1}$ are successive approximations to h^* and h_n^*, respectively. Lemma 3.2 implies that u^t and u_n^t are in $C_w(\mathsf{Z})$ for all t and n.

13

Lemma 4.1. *For all $u, v \in C_w(\mathsf{Z})$ and $n \geq 1$, the following results hold: (i) if $u \leq v$, then $Fu \leq Fv$ and $F_n u \leq F_n v$; (ii) $Fu \leq F_n u$.*

Proof. Define the sub-stochastic kernel $\hat{\eta}$ by letting

$$\hat{\eta}(\cdot|z, a) := \eta(\cdot|z, a) - \lambda(\cdot)\phi(z, a).$$

Using $\hat{\eta}$, F and F_n can be written as

$$Fu(z) := \min_{a \in \mathsf{A}} \left[c(z, a) + \int_{\mathsf{Z}} u(y)\hat{\eta}(dy|z, a) \right],$$

$$F_n u(z) := \min_{a \in \Lambda_n} \left[c(z, a) + \int_{\mathsf{Z}} u(y)\hat{\eta}(dy|z, a) \right].$$

Then the results follow from the fact that $\hat{\eta}(\cdot|z, a) \geq 0$ by Assumption 4.1-(f). $\qquad \square$

Lemma 4.1 implies that $u^0 \leq u^1 \leq u^2 \leq \ldots \leq h^*$ and $u_n^0 \leq u_n^1 \leq u_n^2 \leq \ldots \leq h_n^*$. Note that $\|u^1\|_w, \|u_n^1\|_w \leq M$ by Assumption 3.1-(d). Since

$$\|h^*\|_w \leq \|h^* - u^1\|_w + \|u^1\|_w = \|Fh^* - Fu^0\|_w + \|u^1\|_w \leq \alpha\|h^*\|_w + \|u^1\|_w$$
$$\|h_n^*\|_w \leq \|h_n^* - u_n^1\|_w + \|u_n^1\|_w = \|F_n h_n^* - F_n u_n^0\|_w + \|u_n^1\|_w \leq \alpha\|h_n^*\|_w + \|u_n^1\|_w,$$

we have

$$u^t(z) \leq h^*(z) \leq M\frac{w(z)}{1 - \alpha},$$

and

$$u_n^t(z) \leq h_n^*(z) \leq M\frac{w(z)}{1 - \alpha}.$$

By inequalities above and the facts $\|u^t - h^*\|_w \leq \alpha^t\|h\|_w$ and $\|u_n^t - h_n^*\|_w \leq \alpha^t\|h_n\|_w$, we also have

$$\|u^t - h^*\|_w \leq M\frac{\alpha^t}{1 - \alpha},$$

and

$$\|u_n^t - h_n^*\|_w \leq M\frac{\alpha^t}{1 - \alpha}.$$

By Lemma 4.1, for each n and $v \in C_w(\mathsf{Z})$, we have $Fv \leq F_n v$. Therefore, by the monotonicity of F and the fact $u^0 = u_n^0 = 0$, we have

$$\begin{aligned} u^t &\leq u_n^t \\ h^* &\leq h_n^*, \end{aligned} \tag{19}$$

for all t and n.

Lemma 4.2. *For any compact set $K \subset \mathsf{Z}$ and $t \geq 1$, we have*

$$\lim_{n \to \infty} \sup_{z \in K} |u_n^t(z) - u^t(z)| = 0. \tag{20}$$

Proof. Note that for each $t \geq 1$, by the dominated convergence theorem and $\lambda(w) < \infty$, we have $\lambda(u_n^t) \to \lambda(u^t)$ if $u_n^t \to u^t$ pointwise. The proof can be finished using the same arguments as in the proof of Lemma 3.5 and so we omit the details. $\quad\square$

Lemma 4.3. *For any compact set $K \subset \mathsf{Z}$, we have*

$$\lim_{n \to \infty} \sup_{z \in K} |h_n^*(z) - h^*(z)| = 0.$$

Proof. The lemma can be proved using the same arguments as in the proof of Theorem 3.2. $\quad\square$

Now, using Lemma 4.3 we prove Theorem 4.2

Proof of Theorem 4.2. Recall that $V^* = \rho_{f*}$ and $V_n^* = \rho_{f_n^*}$, and they satisfy the following ACOEs:

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

$$h_n^*(z) + \rho_{f_n^*} = \min_{a \in \Lambda_n} \left[c(z, a) + \int_{\mathsf{Z}} h_n^*(y) \eta(dy|z, a) \right] = c(z, f_n^*(z)) + \int_{\mathsf{Z}} h_n^*(y) \eta(dy|z, f_n^*(z)).$$

Note that $h_n^* \geq h^*$ (see (19)) and $\rho_{f_n^*} \geq \rho_{f*}$. For each n, let $f_n := \Upsilon_n(f^*)$. Then for any $z \in \mathsf{Z}$ we have

$$
\begin{aligned}
\limsup_{n \to \infty} \left(h_n^*(z) + \rho_{f_n^*} \right) &= \limsup_{n \to \infty} \left(\min_{a \in \Lambda_n} \left[c(z, a) + \int_{\mathsf{Z}} h_n^*(y) \eta(dy|z, a) \right] \right) \\
&= \limsup_{n \to \infty} \left(c(z, f_n^*(z)) + \int_{\mathsf{Z}} h_n^*(y) \eta(dy|z, f_n^*(z)) \right) \\
&\leq \limsup_{n \to \infty} \left(c(z, f_n(z)) + \int_{\mathsf{Z}} h_n^*(y) \eta(dy|z, f_n(z)) \right) \\
&= c(z, f^*(z)) + \int_{\mathsf{Z}} h^*(y) \eta(dy|z, f^*(z)) \tag{21} \\
&= h^*(z) + \rho_{f*} \\
&\leq \liminf_{n \to \infty} \left(h_n^*(z) + \rho_{f_n^*} \right),
\end{aligned}
$$

where (21) follows from Lemma 3.4 and the fact that h_n^* converges to h^* uniformly on any compact subset K of Z and $\sup_n \|h_n^*\|_w \leq \frac{M}{1-\alpha}$. Since $\lim_{n \to \infty} h_n^*(z) = h^*(z)$ by Lemma 4.3, we have $\lim_{n \to \infty} \rho_{f_n^*} = \rho_{f*}$. This completes the proof. $\quad\square$

15

5. Application to Partially Observed MDPs

In this section we apply the result obtained in Section 3 to partially observed Markov decision processes (POMDPs). Consider a discrete time POMDP with state space X, action space A, and observation space Y, all Borel spaces. Let $p(\,\cdot\,|x,a)$ denote the transition probability of the next state given the current state-action pair is (x,a), and let $r(\,\cdot\,|x)$ denote the transition probability of the current observation given the current state variable x. The one-stage cost function, denoted by $\tilde{c}$, is again a measurable function from $\mathsf{X} \times \mathsf{A}$ to $[0,\infty)$.

Define the history spaces $\tilde{\mathsf{H}}_t = (\mathsf{Y} \times \mathsf{A})^t \times \mathsf{Y}$, $t = 0, 1, 2, \dots$ endowed with their product Borel σ-algebras generated by $\mathcal{B}(\mathsf{Y})$ and $\mathcal{B}(\mathsf{A})$. A *policy* $\pi = \{\pi_t\}$ is a sequence of stochastic kernels on A given $\tilde{\mathsf{H}}_t$. We denote by Π the set of all policies. Hence, for any initial distribution μ and policy π we can think of POMDP as a stochastic process $\{x_t, y_t, a_t\}_{t \geq 0}$ defined on a probability space $\left(\Omega, \mathcal{B}(\Omega), P_\mu^\pi\right)$ where $\Omega = \tilde{\mathsf{H}}_\infty \times \mathsf{X}^\infty$, x_t is a X-valued random variable, y_t is a Y-valued random variable, a_t is a A-valued random variable, and P_μ^π-almost surely they satisfy

$$P_\mu^\pi(x_0 \in \,\cdot\,) = \mu(\,\cdot\,)$$
$$P_\mu^\pi(x_t \in \,\cdot\,|x_{[0,t-1]}, y_{[0,t-1]}, a_{[0,t-1]}) = P_\mu^\pi(x_t \in \,\cdot\,|x_{t-1}, a_{t-1}) = p(\,\cdot\,|x_{t-1}, a_{t-1})$$
$$P_\mu^\pi(y_t \in \,\cdot\,|x_{[0,t]}, y_{[0,t-1]}, a_{[0,t-1]}) = P_\mu^\pi(y_t \in \,\cdot\,|x_t) = r(\,\cdot\,|x_t)$$
$$P_\mu^\pi(a_t \in \,\cdot\,|x_{[0,t]}, y_{[0,t]}, a_{[0,t-1]}) = \pi_t(\,\cdot\,|y_{[0,t]}, a_{[0,t-1]})$$

where $x_{[0,t]} = (x_0, \dots, x_t)$, $y_{[0,t]} = (y_0, \dots, y_t)$, and $a_{[0,t]} = (a_0, \dots, a_t)$ $(t \geq 1)$. Let $\tilde{J}(\pi, \mu)$ denote the discounted cost function of the policy $\pi \in \Pi$ with initial distribution μ of the POMDP.

It is known that any POMDP can be reduced to a (completely observable) MDP [25], [26], whose states are the posterior state distributions or beliefs of the observer; that is, the state at time t is

$$\Pr\{x_t \in \,\cdot\,|y_0, \dots, y_t, a_0, \dots, a_{t-1}\} \in \mathcal{P}(\mathsf{X}).$$

We call this equivalent MDP the belief-MDP. The belief-MDP has state space $\mathsf{Z} = \mathcal{P}(\mathsf{X})$ and action space A. The transition probability η of the belief-MDP can be constructed as in Example 2.1 (see also [20])

$$\eta(\,\cdot\,|z, a) = \int_\mathsf{Y} 1_{\{F(z,a,y) \in \,\cdot\,\}} H(dy|z, a),$$

where $F(z, a, y) := \Pr\{x_{t+1} \in \,\cdot\,|z_t = z, a_t = a, y_{t+1} = y\}$, $H(\,\cdot\,|z, a) := \Pr\{y_{t+1} \in \,\cdot\,|z_t = z, a_t = a\}$, and z_t denotes the posterior distribution of the state x_t given the past observations. The one-stage cost function c of the belief-MDP is given by

$$c(z, a) := \int_\mathsf{X} \tilde{c}(x, a) z(dx). \tag{22}$$

Hence, the belief-MDP is a Markov decision process with the components $(\mathsf{Z}, \mathsf{A}, \eta, c)$.

For the belief-MDP define the history spaces $\mathsf{H}_t = (\mathsf{Z} \times \mathsf{A})^t \times \mathsf{Z}$, $t = 0, 1, 2, \ldots$ as in Section 2. Again, Φ denotes the set of all policies for the belief-MDP, where the policies are defined in an usual manner. Let $J(\varphi, \xi)$ denote the discounted cost function of policy $\varphi \in \Phi$ for initial distribution ξ of the belief-MDP.

Notice that any history vector $h_t = (z_0, \ldots, z_t, a_0, \ldots, a_{t-1})$ of the belief-MDP is a function of the history vector $\tilde{h}_t = (y_0, \ldots, y_t, a_0, \ldots, a_{t-1})$ of the POMDP. Let us write this relation as $i(\tilde{h}_t) = h_t$. Hence, for a policy $\varphi = \{\varphi_t\} \in \Phi$, we can define a policy $\pi^\varphi = \{\pi_t^\varphi\} \in \Pi$ as

$$\pi_t^\varphi(\,\cdot\,|\tilde{h}_t) := \varphi_t(\,\cdot\,|i(\tilde{h}_t)).$$

Let us write this as a mapping from Φ to Π: $\Phi \ni \varphi \mapsto i(\varphi) = \pi^\varphi \in \Pi$. It is straightforward to show that the cost functions $J(\varphi, \xi)$ and $\tilde{J}(\pi^\varphi, \mu)$ are the same. One can also prove that (see [25], [26])

$$\inf_{\varphi \in \Phi} J(\varphi, \xi) = \inf_{\pi \in \Pi} \tilde{J}(\pi, \mu) \tag{23}$$

and furthermore, that if φ is an optimal policy for belief-MDP, then π^φ is optimal for the POMDP as well. Hence, the POMDP and the corresponding belief-MDP are equivalent in the sense of cost minimization. We will impose the following assumptions on the components of the original POMDP.

Assumption 5.1.

(a) The one stage cost function $\tilde{c}$ is continuous and bounded.

(b) The stochastic kernel $p(\,\cdot\,|x, a)$ is weakly continuous in $(x, a) \in \mathsf{X} \times \mathsf{A}$.

(c) The stochastic kernel $r(\,\cdot\,|x)$ is continuous in total variation, i.e., if $x_k \to x$, then $r(\,\cdot\,|x_k) \to r(\,\cdot\,|x)$ in total variation.

(d) A is compact.

We refer the reader to [14, Section 8] for examples satisfying Assumption 5.1-(c). Note that by [21, Proposition 7.30], the one stage cost function c, which is defined in (22), is in $C_b(\mathsf{Z} \times \mathsf{A})$ under Assumption 5.1-(a),(b). Hence, the belief-MDP satisfies the conditions in Theorem 3.2 for $w = 1$ if η is weakly continuous. The following theorem is a consequence of [14, Theorem 3.7, Example 4.1] and Example 2.1.

Theorem 5.1.

(i) *Under Assumption 5.1-(b),(c), the stochastic kernel η for belief-MDP is weakly continuous in (z, a).*

(ii) *If we relax the continuity of the observation channel in total variation to setwise or weak continuity, then η may not be weakly continuous even if the transition probability p of POMDP is continuous in total variation.*

17

(iii) Finally, η may not be setwise continuous in a, even if the observation channel is continuous in total variation.

Part (i) of Theorem 5.1 implies that belief-MDP satisfies conditions in Theorem 3.2. However, note that continuity of the observation channel in total variation in Assumption 5.1 cannot be relaxed to weak or setwise continuity. On the other hand, the continuity of the observation channel in total variation is not enough for the setwise continuity of η. Hence, results in [13] cannot be applied to the POMDP we consider even though we put a fairly strong condition on the observation channel.

Theorem 5.2. *Suppose Assumption 5.1 holds for the POMDP. Then we have*

$$\lim_{n\to\infty} |J_n^*(z) - J^*(z)| = 0 \quad \text{for all } z \in \mathsf{Z},$$

where J_n^* is the discounted value function of the belief-MDP with the components $\{\mathsf{Z}, \Lambda_n, \eta, c\}$ and J^* is the discounted value function of the belief-MDP with the components $\{\mathsf{Z}, \mathsf{A}, \eta, c\}$.

The significance of Theorem 5.2 is reinforced by the following observation. If we define $D\Pi\mathcal{Q}(\Lambda_n)$ as the set of deterministic policies in Π taking values in Λ_n, then the above theorem implies that for any given $\varepsilon > 0$ there exists $n \geq 1$ and $\pi^* \in D\Pi\mathcal{Q}(\Lambda_n)$ such that

$$\tilde{J}(\pi^*, \mu) < \min_{\pi\in\Pi} \tilde{J}(\pi, \mu) + \varepsilon,$$

where $\pi^* = \pi^{\varphi^*}$. This means that even when is an information transmission constraint from the controller to the plant, one might get ε-close to the value function for any small ε by quantizing the controller's actions and sending the encoded levels.

6. Discussion

In this paper, we considered the finite-action approximation of stationary policies for a discrete-time Markov decision process with either discounted or average costs. Under mild weak continuity assumptions it was shown that if one uses a sufficiently large number of points to discretize the action space, then the resulting finite-action MDP can approximate the original model with arbitrary precision. The results obtained for the discounted cost were also applied to the finite-action approximation problem for POMDPs.

One direction for future work is to further investigate the problem **(P)** for the average cost under specific conditions for POMDPs, so that the results obtained for the average cost can be applicable to the belief-MDPs. In this case, a possible solution methodology is to investigate conditions on the POMDP under which the Markov chain arising from the belief-MDP with a stationary policy is ergodic and hence has a unique invariant measure.

References

[1] S. Yüksel, T. Başar, Stochastic Networked Control Systems: Stabilization and Optimization under Information Constraints, Boston, MA, Birkhäuser, 2013.

[2] N. Saldi, S. Yüksel, T. Linder, Asymptotic optimality of finite approximations to Markov decision processes with Borel spaces, arXiv:1503.02244 (2015).

[3] L. Györfi, M. Kohler, Nonparametric estimation of conditional distributions, IEEE Trans. Inf. Theory 53 (2007) 1872–1878.

[4] D. White, Finite-state approximations for denumerable state infinite horizon discounted Markov decision processes, J. Math. Anal. Appl. 74 (1980) 292–295.

[5] D. Bertsekas, Convergence of discretization procedures in dynamic programming, IEEE Trans. Autom. Control 20 (3) (1975) 415–419.

[6] F. Dufour, T. Prieto-Rumeau, Approximation of Markov decision processes with general state space, J. Math. Anal. Appl. 388 (2012) 1254–1267.

[7] F. Dufour, T. Prieto-Rumeau, Finite linear programming approximations of constrained discounted Markov decision processes, SIAM J. Control Optim. 51 (2) (2013) 1298–1324.

[8] F. Dufour, T. Prieto-Rumeau, Approximation of average cost Markov decision processes using empirical distributions and concentration inequalities, Stochastics (2014) 1–35.

[9] R. Liu, Q. Zhang, G. Yin, Nearly optimal control of nonlinear Markovian systems subject to weak and strong interactions, Stochastic Analysis and Applications 19 (2001) 361–386.

[10] R. Liu, Q. Zhang, G. Yin, Nearly optimal control of singularly perturbed Markov decision processes in discrete time, App. Math. Optim. 44 (2001) 105–129.

[11] R. Jain, P. Varaiya, Simulation-based uniform value function estimates of Markov decision processes, SIAM J. Control Optim. 45 (5) (2006) 1633–1656.

[12] H. Langen, Convergence of dynamic programming models, Math. Oper. Res. 6 (4) (1981) 493–512.

[13] N. Saldi, T. Linder, S. Yüksel, Asymptotic optimality and rates of convergence of quantized stationary policies in stochastic control, IEEE Trans. Autom. Control 60 (2) (2015) 553–558.

[14] E. Feinberg, P. Kasyanov, M. Zgurovsky, Partially observable total-cost Markov decision process with weakly continuous transition probabilities, arXiv:1401.2168 (2014).

[15] O. Hernández-Lerma, J. Lasserre, Further Topics on Discrete-Time Markov Control Processes, Springer, 1999.

[16] O. Hernández-Lerma, J. Lasserre, Markov Chains and Invariant Probabilities, Birkhäuser, 2003.

[17] O. Hernández-Lerma, J. Lasserre, Discrete-Time Markov Control Processes: Basic Optimality Criteria, Springer, 1996.

[18] E. Feinberg, P. Kasyanov, N. Zadioanchuk, Average cost Markov decision processes with weakly continuous transition probabilities, Math. Oper. Res. 37 (4) (2012) 591–607.

[19] V. Borkar, Convex analytic methods in Markov decision processes, in: E. Feinberg, A. Shwartz (Eds.), Handbook of Markov Decision Processes, Kluwer Academic Publisher, 2002.

[20] O. Hernández-Lerma, Adaptive Markov Control Processes, Springer-Verlag, 1989.

[21] D. P. Bertsekas, S. E. Shreve, Stochastic optimal control: The discrete time case, Academic Press New York, 1978.

[22] V. Bogachev, Measure Theory: Volume II, Springer, 2007.

[23] O. Vega-Amaya, The average cost optimality equation: a fixed point approach, Bol. Soc. Mat. Mexicana 9 (3) (2003) 185–195.

[24] E. Gordienko, O. Hernández-Lerma, Average cost Markov control processes with weighted norms: Existence of canonical policies, Appl. Math. 23 (2) (1995) 199–218.

[25] A. Yushkevich, Reduction of a controlled Markov model with incomplete data to a problem with complete information in the case of Borel state and control spaces, Theory Prob. Appl. 21 (1976) 153–158.

[26] D. Rhenius, Incomplete information in Markovian decision models, Ann. Statist. 2 (1974) 1327–1334.