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

Naci Saldi, Serdar Yüksel, Tamás Linder

Abstract

For Markov Decision Processes with Borel state and action spaces, the computation of optimal policies are known to be prohibitively hard. In addition, networked control applications require remote controllers to transmit action commands to an actuator with low information rate. These two problems motivate the study of approximating optimal policies by quantized (discretized) policies. In this paper, 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 kernel, which is a significant relaxation of conditions required in the earlier literature. Discretized policies are shown to approximate optimal deterministic stationary policies with arbitrary precision. These results are applied to a fully observed reduction of a partially observed Markov decision process.

Index Terms

Stochastic control, quantization, approximation, partially observed Markov decision processes.

I. INTRODUCTION

In the theory of Markov decision processes (MDPs), the set of stationary policies; that is, control policies realized by measurable mappings from the state space to the action space, is known to be the smallest structured set that contains a globally optimal policy for a large class of infinite horizon discounted cost or average cost optimal control problems. However, when an analytical solution for an optimal policy is missing, computing such a policy even in this

The authors are with the Department of Mathematics and Statistics, Queen’s University, Kingston, ON, Canada, Email: {nsaldi,yuksel,linder}@mast.queensu.ca
This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada. Parts of this work were presented at 53rd IEEE Conference on Decision and Control (CDC), Los Angeles, CA, 2014.
class (e.g., by using value or policy iteration or linear programming algorithms) is in general computationally infeasible for uncountable Borel state and action spaces.

In the networked control literature, the problem of optimal quantization for information transmission from a plant/sensor to a controller for both stabilization and optimization has been studied extensively, see e.g. [1]. However, the problem of transmitting signals from a controller to an actuator has not been studied in detail. In particular, such schemes require a simple encoding/decoding rule since an actuator does not have the computational/intelligence capabilities of a controller to apply the adaptive encoding/decoding rules needed in optimal coding schemes. For this reason, time-invariant uniform quantization is a practically common and a simple encoding rule for controller-actuator communication.

Hence, there is a practical need to approximate optimal stationary policies by policies which are computable and transmittable. With such a motivation, in this paper we investigate the following approximation problem: For uncountable Borel state and action spaces, and weakly continuous transition probability, 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?

There are various methods developed in the literature to tackle the approximation problem: approximate dynamic programming, state aggregation, neuro-dynamic programming (or reinforcement learning), simulation-based techniques, etc. We refer reader to [2]–[7] and references therein for a rather comprehensive survey of these techniques.

The main differences between our paper and the relevant literature can be summarized as follows: (i) We investigate the approximation problem for general state and action spaces, while most of the previous works assume discrete (i.e., finite or countable) state and/or action spaces (see, e.g., [2]–[4]). (ii) Our conditions on the transition probability of the system are weaker than the conditions imposed in prior works that studied general state and action spaces (see, e.g., [5]–[8]). To more explicitly highlight this difference with an example, consider an additive-noise system given by

\[ x_{n+1} = F(x_n, a_n) + v_n, \quad n = 0, 1, 2, \ldots \]

where \( X \) and \( A \) are compact subsets of \( \mathbb{R}^n \) and \( \mathbb{R}^m \), respectively. The noise process \( \{v_n\} \) is a sequence of independent and identically distributed (i.i.d.) random vectors. The conditions required in [6], [8] hold if the noise process has a Lipschitz continuous density and the function \( F \) is also Lipschitz in \((x, a)\). Similar assumptions are needed for the conditions in [5] to hold.
The condition in [7] holds if the noise process admits a continuous density and the function $F$ is continuous in $a$. However, in our case, the continuity of $F$ in $(x, a)$ 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, our conditions do not restrict the noise model and are satisfied by almost all systems studied in the relevant literature, whereas the conditions in prior work hold only if the noise is well behaved in addition to the continuity of $F$. This makes our results uniquely applicable to a large class of partially observed Markov decision processes, as we will discuss later in the paper. (iii) We also study the approximation problem for the challenging average cost case, while those works that have considered general state and action spaces mostly studied the discounted or finite horizon costs. (iv) 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, unlike in [7] where such bounds were derived for transition probabilities that are Lipschitz for the total variation distance.

II. Markov Decision Processes

For a metric space $E$ equipped with its Borel $\sigma$-algebra $\mathcal{B}(E)$, let $C_b(E)$ denote the set of all bounded continuous real functions. For any $u \in C_b(E)$, let $\|u\| := \sup_{x \in E} |u(x)|$ which turns $C_b(E)$ into a Banach space. Let $\mathcal{P}(E)$ denote the set of all probability measures on $E$. If $E$ is a Borel space (i.e., a Borel subset of a complete and separable metric space), then $\mathcal{P}(E)$ is also shown to be a Borel space [9]. A sequence $\{\mu_n\}$ of measures on $E$ is said to converge setwise [10] to a measure $\mu$ if $\mu_n(B) \to \mu(B)$ for all $B \in \mathcal{B}(E)$. 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. For the sake of simplicity, we assume that the set of admissible actions for any $z \in Z$ is $A$ (see Remark 3.2). Let the stochastic kernel $\eta(\cdot | z, a)$ be the transition probability of the next state given that the previous state-action pair is $(z, a)$ [11]. The one-stage cost function $c$ is a measurable function from $Z \times A$ to $[0, \infty)$. The probability measure $\xi$ over $Z$ denotes the initial distribution of the state process.

Define the history spaces $H_0 = Z$ and $H_m = (Z \times A)^m \times Z$, $m = 0, 1, 2, \ldots$ endowed with their product Borel $\sigma$-algebras generated by $\mathcal{B}(Z)$ and $\mathcal{B}(A)$. A policy is a sequence $\varphi = \{\varphi_m\}$ of stochastic kernels on $A$ given $H_m$. The set of all policies is denoted by $\Phi$. A deterministic policy is a sequence $\varphi = \{\varphi_m\}$ of stochastic kernels on $A$ given $H_m$ which are realized by a sequence of measurable functions $\{f_m\}$ from $H_m$ to $A$, i.e., $\varphi_m(\cdot | h_m) = \delta_{f_m(h_m)}(\cdot)$ where $\delta_z$
is the point mass at \( z \). A deterministic stationary policy is a constant sequence of stochastic kernels \( \varphi = \{ \varphi_m \} \) on \( Z \) such that \( \varphi_m(\cdot | z) = \delta_{f(z)}(\cdot) \) for all \( m \) for some measurable function \( f : Z \to A \). We denote by \( S \) the set of deterministic stationary policies.

According to the Ionescu Tulcea theorem [11], an initial distribution \( \xi \) on \( Z \) and a policy \( \varphi \) define a unique probability measure \( P_\xi^\varphi \) on \( H_\infty = (Z \times A)^\infty \). The expectation with respect to \( P_\xi^\varphi \) is denoted by \( E_\xi^\varphi \). If \( \xi = \delta_z \), we write \( P_\varphi^z \) and \( E_\varphi^z \) instead of \( P_\delta_z^\varphi \) and \( E_\delta_z^\varphi \). 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:

\[
J_\beta(\mu, \varphi) = E_\xi^\varphi \left[ \sum_{m=0}^{\infty} \beta^m c(z_m, a_m) \right],
\]

\[
J(\mu, \varphi) = \limsup_{M \to \infty} \frac{1}{M} E_\xi^\varphi \left[ \sum_{m=0}^{M-1} c(z_m, a_m) \right].
\]

A policy \( \varphi^* \) is called optimal if \( J_\beta(\varphi^*, \xi) \) (or \( J(\varphi^*, \xi) \)) for all \( \xi \in \mathcal{P}(Z) \). It is well known that the set \( S \) of deterministic stationary policies contains an optimal policy for a large class of infinite horizon discounted cost (see, e.g., [11], [12]) and average cost optimal control problems (see, e.g., [12], [13]). In this case we say that \( S \) is an optimal class.

A. 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 \( q : Z \to A \) is called a quantizer from \( Z \) to \( A \) if the range of \( q \), i.e., \( q(Z) = \{ q(z) \in A : z \in Z \} \), is finite.

The elements of \( q(Z) \) (the possible values of \( q \)) are called the levels of \( q \). The rate \( R = \log_2 |q(Z)| \) 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 \( \varphi = \{ \varphi_m \} \) of stochastic kernels on \( A \) given \( Z \) such that \( \varphi_m(\cdot | z) = \delta_{q(z)}(\cdot) \) for all \( m \) for some \( q \in \mathcal{Q} \). For any finite set \( \Lambda \subset A \), let \( \mathcal{Q}(\Lambda) \) denote the set of all elements in \( \mathcal{Q} \) having range \( \Lambda \) and let \( S\mathcal{Q}(\Lambda) \) denote the set of all deterministic stationary quantizer policies induced by \( \mathcal{Q}(\Lambda) \).

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:
For any initial distribution $\xi$ we have $\lim_{n \to \infty} \inf_{\phi \in S(Q(\Lambda_n))} J_\beta(\phi, \xi) = \inf_{\phi \in S} J_\beta(\phi, \xi)$ (or $\lim_{n \to \infty} \inf_{\phi \in S(Q(\Lambda_n))} J(\phi, \xi) = \inf_{\phi \in S} J(\phi, \xi)$ for the average cost), provided that the set $S$ of deterministic stationary policies is an optimal class for the MDP.

Note that (P) implies that there exists a near optimal stationary policy making finitely many decisions.

Remark 2.1. In [7] 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 continuous in the action variable, and (iii) the one stage cost function is 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 (P) for problems where the transition probability $\eta$ is weakly continuous in 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, the setwise continuity of the transition probability in the action variable is a prohibitive condition even for a very simple system as shown below. We refer the reader to Section IV of this paper and [14, Chapter 4] for the basics of POMDPs.

Example 2.1. Consider the system dynamics

$$x_{m+1} = x_m + a_m,$$
$$y_m = x_m + v_m,$$

where $x_m \in X$, $y_m \in Y$ and $a_m \in A$, and where $X$, $Y$ and $A$ are the state, observation and action spaces, respectively. Here, we assume that $X = Y = A = \mathbb{R}_+$ and $\{v_m\}$ 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 [15, Theorem 3.7] the transition probability $\eta$ 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 $\eta$ with respect to the action variable as shown below.

Let $z$ denote the generic state variable for the fully observed MDP, where the state variables are probability measures over $X$. If we define the function $F(z, a, y) := \Pr\{x_{m+1} \in \cdot | z_m = z, a_m = a, y_{m+1} = y\}$ from $\mathcal{P}(X) \times A \times Y$ to $\mathcal{P}(X)$ and the stochastic kernel $H(\cdot | z, a) :=$
\( \Pr \{ y_{m+1} \in \cdot | z_m = z, a_m = a \} \) on \( Y \) given \( \mathcal{P}(X) \times A \), then \( \eta \) can be written as

\[
\eta(\cdot | z, a) = \int_Y 1_{\{ F(z,a,y) \in \cdot \}} H(dy | z, a),
\]

where \( 1_D \) denotes the indicator function of an event \( D \). Let us set \( z = \delta_0 \) (point mass at \( 0 \in X \)), \( \{ a_k \} = \{ \frac{1}{k} \} \), and \( a = 0 \). Hence, \( a_k \rightarrow a \). We show that \( \eta(\cdot | z, a_k) \not\rightarrow \eta(\cdot | z, a) \) setwise.

Observe that for all \( k \) and \( y \in Y \), 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 \( \mathcal{P}(X) \) as

\[
O := \{ z \in \mathcal{P}(X) : \left| \int_X f(x) \delta_{\frac{1}{k}}(dx) - \int_X f(x) z(dx) \right| < 1 \},
\]

where \( f \) is the symmetric triangular function between \([-1, 1]\) with \( f(0) = 1 \). Observe that we have \( F(z, a_k, y) \in O \) for all \( k \) and \( y \), but \( F(z, a, y) \not\in O \) for all \( y \). Hence,

\[
\eta(O|z, a_k) := \int_Y 1_{\{ F(z,a_k,y) \in O \}} H(dy | z, a_k) = 1,
\]

but

\[
\eta(O|z, a) := \int_Y 1_{\{ F(z,a,y) \in O \}} H(dy | z, a) = 0.
\]

This means that \( \eta(\cdot | z, a_k) \not\rightarrow \eta(\cdot | z, a) \) setwise. Hence, \( \eta \) does not satisfy the setwise continuity assumption.

### III. Near Optimality of Quantized Policies in MDP

In this section the following assumptions will be imposed for both the discounted cost and the average cost. Later, in Section III-B we will make additional assumptions for the average cost case.

**Assumption 1.**

(a) The one stage cost function \( c \) is in \( C_b(Z \times A) \).

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

(c) \( Z \) and \( A \) are compact.

Let \( d_A \) denote the metric on \( A \). Since \( A \) is compact and thus totally bounded, one can find a nested sequence of finite sets \( \{ \{ a_{i,j} \}_{j=1}^{k_n} \}_{n \geq 1} \) such that for all \( n \),

\[
\min_{i \in \{1, \ldots, k_n\}} d_A(a_i, a) < 1/n \text{ for all } a \in A. \tag{1}
\]
In other words, \( \{a_i\}_{i=1}^{k_n} \) is a \( 1/n \)-net in \( A \) and \( \{a_i\}_{i=1}^{k_n} \subset \{a_i\}_{i=1}^{k_{n+1}} \) for all \( n \). Let \( \Lambda_n := \{a_1, \ldots, a_{k_n}\} \) for all \( n \). In the rest of this paper, we assume that the sequence \( \{\Lambda_n\}_{n \geq 1} \) is fixed.

**Remark 3.2.** If for each \( z \) the set of available actions, denoted by \( A(z) \), is restricted to be a compact subset of a Borel space \( A \) (not necessarily compact), then one can recover all the results in this section by finding a nested sequence of finite sets \( \{\{a^z_i\}_{i=1}^{k_n}\}_{n \geq 1} \) such that for all \( n \) and \( z \), \( \{a^z_i\}_{i=1}^{k_n} \) is a \( 1/n \)-net in \( A(z) \).

### A. Discounted Cost

In this section we consider the problem \( (P) \) for the discounted cost with a discount factor \( \beta \in (0, 1) \). Since \( \beta \) is fixed here, we will write \( J \) instead of \( J_\beta \). Define the operator \( T \) on \( C_b(Z) \) by

\[
Tu(z) := \min_{a \in A} \left[ c(z, a) + \beta \int_Z u(y)\eta(dy|z, a) \right],
\]

(2)

In the literature \( T \) is called the **Bellman optimality operator**. It can be proved that \( T \) is a contraction operator with modulus \( \beta \) (see [14, Lemma 2.5]); that is,

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

Define the function \( J^* \) as

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

We call \( J^* \) the **value function** of the MDP for the discounted cost. The following theorem is a widely known result in the theory of Markov decision processes (see e.g., [14, Theorem 2.8, p.23]) which holds even without the compactness assumption on the state space. Here and in the rest of this paper \( \varphi = \{f\} \) means that \( \varphi \) is a deterministic stationary policy induced by the function \( f \), so we might only write \( \{f\} \) to denote this policy.

**Theorem 3.1.** Suppose Assumption [1] holds. Then, the value function \( J^* \) is the unique fixed point in \( C_b(Z) \) of the contraction operator \( T \), i.e.,

\[
J^* = T J^*.
\]

(3)

Furthermore, a deterministic stationary policy \( \varphi^* = \{f^*\} \) is optimal if and only if

\[
J^*(z) = c(z, f^*(z)) + \beta \int_X J^*(y)\eta(dy|z, f^*(z)).
\]

(4)
Finally, there exists a deterministic stationary policy \( \varphi^* = \{f^*_n\} \) which is optimal, so it satisfies (4).

Define, for all \( n \geq 1 \), the operator \( T_n \) (which will be used to approximate \( T \)) by
\[
    T_n u(z) := \min_{a_i \in \Lambda_n} [c(z, a_i) + \beta \int_Z u(y) \eta(dy|z, a_i)].
\]
(5)

It is straightforward to show that \( T_n \) is a contraction operator with modulus \( \beta \) mapping \( C_b(Z) \) into itself.

We also have \( T_n u \geq T_{n+1} u \) for any \( u \in C_b(Z) \) by the nestedness of the sets \( \{\Lambda_n\}_{n \geq 1} \). Let \( J_n^* \in C_b(Z) \) denote the fixed point of \( T_n \). Using the measurable selection theorem [14, Proposition D.3] it can be proved that there exists a measurable \( f_n^* : Z \to \Lambda_n \) such that
\[
    T_n J_n^*(z) = c(z, f_n^*(z)) + \beta \int_Z J_n^*(y) \eta(dy|z, f_n^*(z)).
\]
(6)

Furthermore, one can also prove that (see [14, Lemma 2.6])
\[
    J(\{f_n^*\}, z) = J_n^*(z) \text{ for all } z \in Z.
\]
(7)

Notice that \( f_n^* \in Q(\Lambda_n) \), so the deterministic stationary policy \( \{f_n^*\} \) is in \( SQ(\Lambda_n) \). The following is the main theorem in this section.

**Theorem 3.2.** Let \( \xi \) be any initial distribution. Under Assumption [1] for any given \( \varepsilon > 0 \) there exists \( n \geq 1 \) and \( \varphi^* \in SQ(\Lambda_n) \) such that
\[
    J(\varphi^*, \xi) < \min_{\rho \in \Phi} J(\rho, \xi) + \varepsilon.
\]
(8)

The proof of Theorem 3.2 requires two lemmas. The first one is about the approximation of \( T \) by \( \{T_n\}_{n \geq 1} \).

**Lemma 3.1.** For any \( u \in C_b(Z) \) we have \( T_n u \to Tu \) monotonically as \( n \to \infty \). Since \( Z \) is compact, we also have \( \|T_n u - Tu\| \to 0 \) by Dini’s theorem [16].

**Proof:** Fix any \( u \in C_b(Z) \). By the measurable selection theorem [14, Proposition D.3] there exists \( f : Z \to A \) such that
\[
    Tu(z) = c(z, f(z)) + \beta \int_Z u(y) \eta(dy|z, f(z)).
\]
Since the action space \( A \) is compact we can approximate \( f \) by a sequence \( \{f_n\}_{n \geq 1} \) of quantizers \( f_n \in Q(\Lambda_n) \) for all \( n \), in the sense that \( f_n \to f \) uniformly. By the definition of \( T_n \) we must have
\[
    T_n u(z) \leq c(z, f_n(z)) + \beta \int_Z u(y) \eta(dy|z, f_n(z)).
\]
Since \( Tu \leq T_n u \) for all \( n \), we have

\[
Tu \leq \liminf_n T_n u \leq \limsup_n T_n u
\]

\[
\leq \limsup_n \left[ c(z, f_n(z)) + \beta \int_Z u(y) \eta(dy|z, f_n(z)) \right] = Tu,
\]

where the last equality follows from the continuity of \( c \) and \( \eta \). This completes the proof.

The next step is to show that the fixed point \( J_n^* \) of \( T_n \) uniformly converges to the fixed point \( J^* \) of \( T \).

Lemma 3.2. \( J(\{f_n^*\}, \cdot) = J_n^*(\cdot) \to J^*(\cdot) = J(\{f^*\}, \cdot) \) uniformly.

Proof: We have

\[
\|J_n^* - J^*\| = \|T_n J_n^* - T J^*\|
\]

\[
\leq \|T_n J_n^* - T_n J^*\| + \|T_n J^* - T J^*\|
\]

\[
\leq \beta \|J_n^* - J^*\| + \|T_n J^* - T J^*\|
\]

and, hence \( \|J_n^* - J^*\| \leq \frac{1}{1-\beta} \|T_n J^* - T J^*\| \). Thus, letting \( n \to \infty \), the result follows by Lemma 3.1.

Proof of Theorem 3.2: Recall that \( J^*(z) = J(\{f^*\}, z) \) for all \( z \) by Theorem 3.1. Hence, \( \min_{\varphi \in \Phi} J(\varphi, \xi) = J(\{f^*\}, \xi) \). By Lemma 3.2 there exists \( f_n^* \in \mathcal{SQ}(A_n) \) for all \( n \) such that \( J(\{f_n^*\}, \cdot) \to J(\{f^*\}, \cdot) \) uniformly. Thus, \( J(\{f_n^*\}, \xi) \to J(\{f^*\}, \xi) \) which implies the result.

B. Average Cost

In this section we consider the problem \( (P) \) for the average cost. We will also need the discounted cost which we now denote by \( J_\beta^* \) since the dependence on the discount factor \( \beta \) is needed. Note that since the one stage cost function \( c \) is bounded by some \( L \geq 0 \), we must have

(d) \((1 - \beta)J_\beta^*(z) \leq L \) for all \( \beta \in (0, 1) \) and \( z \in Z \).

In addition to Assumption 1, we impose the following assumption in this section.

Assumption 2.

There exists \( \alpha \in (0, 1) \) and \( N \geq 0 \), a nonnegative function \( b \) on \( Z \) and a state \( z_0 \in Z \) such that,

(e) \(-N \leq h_\beta(z) \leq b(z) \) for all \( z \in Z \) and \( \beta \in [\alpha, 1] \), where \( h_\beta(z) := J_\beta^*(z) - J_\beta^*(z_0) \).

(f) The sequence \( \{h_\beta(k)\} \) is equicontinuous, where \( \{\beta(k)\} \) is a sequence of discount factors converging to 1 which satisfies \( \lim_{k \to \infty} (1 - \beta(k)) J_{\beta(k)}^*(z) = \rho^* \) for all \( z \in Z \) for some \( \rho^* \in [0, L] \).
Notice that Assumption 2-(e) is the usual condition imposed in the literature (see, e.g., [11, Section 5.4]) which guarantees the existence of an optimal stationary policy via the vanishing discount approach.

The following theorem states the existence of a solution for the Average Cost Optimality Equation (ACOE) and an optimal stationary policy under Assumptions 1 and 2. It can be proved using the same steps as in [11, Theorem 5.4.3] with the difference being the condition of the weak continuity of the transition kernel and the continuity of the function $h$; therefore, we only give a sketch of the proof.

**Theorem 3.3.** Under Assumptions 1 and 2, there exist a constant $\rho^* \geq 0$, a continuous and bounded $h$ from $Z$ to $\mathbb{R}$ with $-N \leq h(\cdot) \leq b(\cdot)$, and $\{f^*\} \in S$ such that $(\rho^*, h, f^*)$ satisfies the ACOE; that is,

$$\rho^* + h(z) = \min_{a \in A} \left[ c(z, a) + \int_Z h(y)\eta(dy|z, a) \right]$$

$$= c(z, f^*(z)) + \int_Z h(y)\eta(dy|z, f^*(z)),$$

for all $z \in Z$. Moreover, $\{f^*\}$ is optimal and $\rho^*$ is the value function, i.e.,

$$\inf_{\varphi} J(\varphi, z) =: J^*(z) = J(\{f^*\}, z) = \rho^*,$$

for all $z \in Z$.

**Sketch of the proof:** By Assumption 2-(f) and the Arzela-Ascoli theorem, there exists a subsequence $\{h_{\beta(k_l)}\}$ of $\{h_{\beta(k)}\}$ which converges uniformly to a continuous and bounded function $h$. Take the limit in equation [11, (5.4.10)] along this subsequence, i.e., consider

$$\rho^* + h(z) = \lim_{l} \min_{a \in A} \left[ c(z, a) + \int_Z h_{\beta(k_l)}(y)\eta(dy|z, a) \right]$$

$$= \min_{a \in A} \lim_{l} \left[ c(z, a) + \int_Z h_{\beta(k_l)}(y)\eta(dy|z, a) \right]$$

$$= \min_{a \in A} \left[ c(z, a) + \int_Z h(y)\eta(dy|z, a) \right],$$

where the exchange of limit and minimum follows from the compactness of $A$ and since

$$c(z, a) + \beta(k_l) \int_Z h_{\beta(k_l)}(y)\eta(dy|z, a) \to c(z, a) + \int_Z h(y)\eta(dy|z, a)$$

uniformly as $l \to \infty$. The rest of the proof coincides with the proof of [11, Theorem 5.4.3].

Since the action space $A$ is compact, there exists $f_n \in Q(\Lambda_n)$ for all $n$ such that $f_n \to f^*$ uniformly. Then, by the weak continuity of $\eta$ and the continuity of $c$, we have

$$c(z, f_n(z)) + \int_Z h(y)\eta(dy|z, f_n(z)) \to c(z, f^*(z)) + \int_Z h(y)\eta(dy|z, f^*(z)).$$

\(\text{October 28, 2014 DRAFT}\)
pointwise. For each \( n \) let \( f^*_n \in \mathcal{Q}(\Lambda_n) \) satisfy
\[
c(z, f^*_n(z)) + \int_Z h(y)\eta(dy|z, f^*_n(z)) = \min_{a_i \in \Lambda_n} [c(z, a_i) + \int_Z h(y)\eta(dy|z, a_i)].
\] (10)
The existence of such \( f_n \) again follows from the measurable selection theorem. Since (10) is upper bounded by the first term in (9) for all \( n \), we must have
\[
c(z, f^*_n(z)) + \int_Z h(y)\eta(dy|z, f^*_n(z)) \to c(z, f^*(z)) + \int_Z h(y)\eta(dy|z, f^*(z))
\] monotonically since \( \Lambda_n \subset \Lambda_{n+1} \), and hence uniformly by Dini’s theorem. We now state and prove the main theorem for this section.

**Theorem 3.4.** Let \( \xi \) be any initial distribution. Under Assumptions 1 and 2, for any given \( \varepsilon > 0 \) there exists \( n \geq 1 \) and \( \varphi^* \in \mathcal{S}(\Lambda_n) \) such that
\[
J(\varphi^*, \xi) < \min_{\varphi \in \Phi} J(\varphi, \xi) + \varepsilon.
\] (12)
**Proof:** Fix any \( \varepsilon > 0 \). Choose \( n \) large enough so that the maximum difference between the terms in (11) is less than \( \varepsilon \). Hence, for this \( n \), the ACOE gives
\[
c(z, f^*_n(z)) + \int_Z h(y)\eta(dy|z, f^*_n(z)) \leq \rho^* + \varepsilon + h(z),
\] for all \( z \in Z \). Iteration of the above equation and Assumption 2(e) yield (see [11, (5.2.23)])
\[
l(\rho^* + \varepsilon) + h(z) \geq J^l(\{f^*_n\}, z) + E_Z[f^*_n h(z_i)]
\geq J^l(\{f^*_n\}, z) - N,
\] where \( J^l(\varphi, z) := E_Z[\sum_{m=0}^{l-1} c(z_m, a_m)] \) for any policy \( \varphi \). Thus, multiplying both sides by \( \frac{1}{l} \) and taking the limsup, we obtain \( \rho^* + \varepsilon \geq J(\{f^*_n\}, z) \). Since \( \rho^* = \inf_{\varphi} J(\varphi, z) \) for all \( z \in Z \) by Theorem 3.3, this completes the proof.

The assumption that \( \{h_{\beta(k)}\} \) is equicontinuous is the key in the proof of Theorem 3.4, however, verifying this assumption is difficult in general. Therefore, to find MDPs for which Theorem 3.4 is applicable, it is necessary to find more tractable conditions. One such condition is the geometric ergodicity of the Markov chains induced by deterministic stationary policies; that is, the existence of a continuous function \( K(z) \) and a constant \( \kappa \in (0,1) \) such that for all \( \varphi \in S \), there is a (necessarily unique) probability measure \( \nu_\varphi \in \mathcal{P}(Z) \) satisfying \( \|\lambda^{\varphi,z}_m - \nu_\varphi\|_{TV} \leq K(z)\kappa^m \) for all \( z \in Z \) and \( m \geq 1 \), where \( \lambda^{\varphi,z}_m \) denotes the law of \( x_m \) under the policy \( \varphi \) and the initial point \( z \). In this situation, it is straightforward to prove that \( \|h_{\beta(k)}\| \leq \|c\|/(1 - \kappa)(K(z) + K(z_0)) \) which implies the equicontinuity of \( \{h_\beta\} \). Note that geometric ergodicity holds under any of
the conditions $R_i, i \in \{0, 1, 1(a), 1(b), 2, \ldots, 5\}$ in [17]. Another condition is the convexity of the value functions for discounted costs. To be more precise, an MDP is said to be convex [18] if the state space $Z$ is an open convex subset of a separable Banach space (e.g., $\mathbb{R}^n$) and the value functions $J^*_\beta(k)$ is convex for each $\beta(k)$ in Assumption [2](f). For this MDP, if the function $b$ in Assumption [2](e) is upper semi-continuous and satisfies $\int_Z b(y)\eta(dy|z, a) < \infty$ for all $(z, a) \in Z \times A$, then it is proved in [18] that functions $\{h_\beta(k)\}$ satisfy the local Lipschitz property which implies the equicontinuity. To show the convexity of any MDP a basic method is described in [18, p. 181] which uses the value iteration algorithm.

IV. APPLICATION TO PARTIALLY OBSERVED MDPs

In this section we apply the result obtained in Section III-A 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 $X \times A$ to $[0, \infty)$.

Define the history spaces $\tilde{H}_m = (Y \times A)^m \times Y, m = 0, 1, 2, \ldots$ endowed with their product Borel $\sigma$-algebras generated by $B(Y)$ and $B(A)$. A policy $\pi = \{\pi_m\}$ is a sequence of stochastic kernels on $A$ given $\tilde{H}_m$. We denote by $\Pi$ the set of all policies. Let $\tilde{J}_\beta(\pi, \mu)$ denote the discounted cost function of the policy $\pi \in \Pi$ with initial distribution $\mu$ of the POMDP.

It is known that any POMDP can be reduced to a (completely observable) MDP [19], [20], whose states are the posterior state distributions or beliefs of the observer; that is, the state at time $m$ is

$$\Pr\{x_m \in \cdot|y_0, \ldots, y_m, a_0, \ldots, a_{m-1}\} \in \mathcal{P}(X).$$

We call this equivalent MDP the belief-MDP. The belief-MDP has the state space $Z = \mathcal{P}(X)$ and action space $A$. The transition probability $\eta$ of the belief-MDP can be constructed as in Example [2](i) (see also [14]). The one-stage cost function $c$ of the belief-MDP is given by

$$c(z, a) := \int_X \tilde{c}(x, a)z(dx). \quad (13)$$

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

For the belief-MDP define the history spaces $H_m = (Z \times A)^m \times Z, m = 0, 1, 2, \ldots$ as in Section II. Again, $\Phi$ denotes the set of all policies for the belief-MDP, where the policies are
defined in an usual manner. Let \( J_\beta(\varphi, \xi) \) denote the discounted cost function of policy \( \varphi \in \Phi \) for initial distribution \( \xi \) of the belief-MDP.

Notice that any history vector \( h_m = (z_0, \ldots, z_m, a_0, \ldots, a_{m-1}) \) of the belief-MDP is a function of the history vector \( \tilde{h}_m = (y_0, \ldots, y_m, a_0, \ldots, a_{m-1}) \) of the POMDP. Let us write this relation as \( i(\tilde{h}_m) = h_m \). Hence, for a policy \( \varphi = \{ \varphi_m \} \in \Phi \), we can define a policy \( \pi \varphi = \{ \pi \varphi_m \} \in \Pi \) as
\[
\pi \varphi_m(\cdot | \tilde{h}_m) := \varphi_m(\cdot | i(\tilde{h}_m)).
\]
Let us write this as a mapping from \( \Phi \) to \( \Pi \):
\[
\Phi \ni \varphi \mapsto i(\varphi) = \pi \varphi \in \Pi.
\]
It is straightforward to show that the cost function \( J_\beta(\varphi, \xi) \) and the cost function \( \tilde{J}_\beta(\pi \varphi, \mu) \) are the same. One can also prove that (see \cite{19}, \cite{20}) for any \( \beta \in (0, 1) \)
\[
\inf_{\varphi \in \Phi} J_\beta(\varphi, \xi) = \inf_{\pi \in \Pi} \tilde{J}_\beta(\pi, \mu). \tag{14}
\]
and furthermore, that if \( \varphi \) is an optimal policy for belief-MDP, then \( \pi \varphi \) 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 3.**

(a) The one stage cost function is \( \tilde{c} \in C_b(X \times A) \).

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

(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) \( X \) and \( A \) are compact.

It is known that the set of probability measures is a compact metric space for the weak topology if the space itself is compact. Hence, \( Z \) is compact under the Assumption 3-(d). The one stage cost function \( c \), which is defined in (13), is in \( C_b(Z \times A) \) under Assumption 3-(a)(d).

Indeed, let \( (z_k, a_k) \to (z, a) \) in \( Z \times A \). Define \( v_k(x) := \tilde{c}(x, a_k) \) and \( v(x) := \tilde{c}(x, a) \). Note that \( v_k \) converges continuously to \( v \) \cite{21}; that is, \( v_k(x_k) \to v(x) \) for any \( x_k \to x \). By \cite{21} Theorem 3.3, we have \( c(z_k, a_k) \to c(z, a) \). Hence, the belief-MDP satisfies Assumption 1 if \( \eta \) is weakly continuous. The following theorem is a consequence of \cite{15} Theorem 3.7, Example 4.1] and Example 2.1

**Theorem 4.1.** Under Assumption 3-(b(c), the stochastic kernel \( \eta \) for belief-MDP is weakly continuous in \( (z, a) \). If we relax the continuity of the observation channel in total variation to
setwise or weak continuity, then \( \eta \) may not be weakly continuous even if the transition probability \( p \) of POMDP is continuous in total variation. Finally, \( \eta \) may not be setwise continuous in \( a \), even if the observation channel is continuous in total variation.

Theorem 4.1 implies that belief-MDP satisfies Assumption \( \Pi \) and so Theorem 3.2 is applicable. However, note that continuity of the observation channel in total variation in Assumption \( \Pi \) 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 \( \eta \). Hence, results from prior work reviewed in Section \( \Pi \) cannot be applied to the POMDP we consider even though we have a fairly strong condition on the observation channel.

**Theorem 4.2.** Let \( \xi \) be any initial distribution on \( Z \). Under Assumption \( \Pi \) on the POMDP, for any given \( \varepsilon > 0 \) there exists \( n \geq 1 \) and \( \varphi^* \in SQ(\Lambda_n) \) such that

\[
J(\varphi^*, \xi) < \min_{\varphi \in \Phi} J(\varphi, \xi) + \varepsilon
\]

for the belief-MDP, where \( SQ(\Lambda_n) \) is defined in a similar way as in Section \( \Pi \).

The significance of Theorem 4.2 is reinforced by the following observation. If we define \( D\Pi Q(\Lambda_n) \) as the set of deterministic policies in \( \Pi \) taking values in \( \Lambda_n \), then the above theorem implies that for any given \( \varepsilon > 0 \) there exists \( n \geq 1 \) and \( \pi^* \in D\Pi Q(\Lambda_n) \) such that

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

where \( \pi^* = \pi \varphi^* \). This means that if there is an information transmission constraint from the controller to the plant, then one might get \( \varepsilon \)-close to the value function for any small \( \varepsilon \) by quantizing the controller’s actions and sending the encoded levels.

**V. DISCUSSION**

In this paper, we considered the finite action approximation of stationary policies for a discrete-time Markov decision process with discounted and average costs under the mild weak continuity assumptions.

One direction for future work is to establish similar converge results when the state space is also discretized, which may give rise to a computationally feasible algorithm that gives a near optimal policy. Another direction is to solve the problem (P) under milder conditions for the average cost, so that it can be applicable to a wider range of 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 and T. Başar, Stochastic Networked Control Systems: Stabilization and Optimization under Information Constraints. Boston, MA, Birkhauser, 2013.
[2] D. Bertsekas and J. Tsitsiklis, Neuro-Dynamic Programming. Athena Scientific, 1996.
[3] Z. Ren and B. Krogh, “State aggregation in Markov decision processes,” in IEEE Conf. Decision Control, Las Vegas, Dec 2002, pp. 3819 – 3824.
[4] R. Cavazos-Cadena, “Finite-state approximations for denumerable state discounted Markov decision processes,” Appl. Math. Optim., vol. 14, pp. 1–26, 1986.
[5] D. Bertsekas, “Convergence of discretization procedures in dynamic programming,” IEEE Trans. Autom. Control, vol. 20, no. 3, pp. 415–419, Jun. 1975.
[6] F. Dufour and T. Prieto-Rumeau, “Approximation of Markov decision processes with general state space,” J. Math. Anal. Appl., vol. 388, pp. 1254–1267, 2012.
[7] N. Saldi, T. Linder, and S. Yüksel, “Asymptotic optimality and rates of convergence of quantized stationary policies in stochastic control,” to appear in IEEE Trans. Autom. Control, 2015.
[8] F. Dufour and T. Prieto-Rumeau, “Finite linear programming approximations of constrained discounted Markov decision processes,” SIAM J. Control Optim., vol. 51, no. 2, pp. 1298–1324, 2013.
[9] K. Parthasarathy, Probability Measures on Metric Spaces. AMS Bookstore, 1967.
[10] O. Hernández-Lerma and J. Lasserre, Markov Chains and Invariant Probabilities. Birkhauser, 2003.
[11] ——., Discrete-Time Markov Control Processes: Basic Optimality Criteria. Springer, 1996.
[12] E. Feinberg, P. Kasyanov, and N. Zadoianchuk, “Average cost Markov decision processes with weakly continuous transition probabilities,” Math. Oper. Res., vol. 37, no. 4, pp. 591–607, Nov. 2012.
[13] V. Borkar, “Convex analytic methods in Markov decision processes,” in Handbook of Markov Decision Processes, E. Feinberg and A. Shwartz, Eds. Kluwer Academic Publisher, 2002.
[14] O. Hernández-Lerma, Adaptive Markov Control Processes. Springer-Verlag, 1989.
[15] E. Feinberg, P. Kasyanov, and M. Zgurovy, “Partially observable total-cost Markov decision process with weakly continuous transition probabilities,” arXiv:1401.2168 2014.
[16] R. M. Dudley, Real Analysis and Probability. Cambridge University Press, 2004.
[17] O. Hernández-Lerma, R. Montes-De-Oca, and R. Cavazos-Cadena, “Recurrence conditions for Markov decision processes with Borel state space: a survey,” Ann. Oper. Res., vol. 28, no. 1, pp. 29–46, 1991.
[18] E. Fernández-Gaucherand, S. Marcus, and A. Arapostathis, “Structured solutions for stochastic control problems,” in Stochastic Theory and Adaptive Control, ser. Lecture Notes in Control and Information Sciences. Springer Berlin Heidelberg, 1992, vol. 184, pp. 172–184.
[19] 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., vol. 21, pp. 153–158, 1976.
[20] D. Rhenius, “Incomplete information in Markovian decision models,” Ann. Statist., vol. 2, pp. 1327–1334, 1974.
[21] R. Serfozo, “Convergence of Lebesgue integrals with varying measures,” Sankhya Ser.A, pp. 380–402, 1982.