Asymptotic Optimality of Quantized Policies in Stochastic Control under Weak Continuity Conditions

Naci Saldi, Tamás Linder, Serdar Yüksel

Abstract— Quantization is an increasingly important operation both because of applications in networked control and the computational benefits of working with finite state spaces. In this paper, we consider quantized approximations of stationary policies for a discrete-time Markov decision process with discounted and average costs and weakly continuous transition probability kernels. We show that deterministic stationary quantizer policies approximate optimal deterministic stationary policies with arbitrary precision under mild technical conditions. We thus extend recent and older results in the literature which consider more stringent continuity conditions for the transition kernels, such as setwise continuity, which limit the applicability of such results. In particular, the weaker continuity requirements allow for the study of partially observable Markov decision processes under practical conditions.

I. INTRODUCTION

In this paper we study the approximation of optimal (deterministic) stationary policies for a Markov decision process (MDP) by stationary policies having finite range. More precisely, our principal objective is to find, for any given \( \varepsilon > 0 \), an \( \varepsilon \)-optimal policy with a finite range.

In the theory of MDPs, it is known that the set of stationary policies is the smallest structured set of control policies in which one can find a globally optimal policy for a large class of infinite horizon discounted cost (see, e.g., [1], [2]) or average cost optimal control problems (see, e.g., [3]–[5]). However, finding an optimal policy even in this class is computationally prohibitive for infinite state and action spaces. Furthermore, considering applications in networked control, transmission of such control actions to an actuator is not realistic when there is an information transmission constraint (physically limited by the presence of a communication channel) between a plant, a controller and an actuator.

Hence, it is of interest to approximate stationary policies, in particular an optimal stationary policy. There are various methods developed in the literature to tackle this problem, most of which assume finite or countable state spaces. Approximate Value Iteration (AVI) and Approximate Policy Iteration (API) (see [6], [7], [8], [9] and references therein) are two well-known methods to approximate an optimal stationary policy in MDPs. As the names imply, the idea is to approximate Value Iteration (VI) and Policy Iteration (PI) algorithms using various techniques (see a review of some techniques in [6]). Another powerful method is state aggregation (see [10]–[12] and references therein). Here, similar states (e.g., with respect to cost functions and transition probabilities) are aggregated to form a reduced MDP whose states are called meta-states. Then, an optimal policy of this MDP is computed, which is then shown to approximate arbitrarily well the optimal policy (in terms of cost function) as the aggregation gets finer. [13]–[17] developed approximation schemes which truncate the state space while calculating the value function in the value iteration algorithm. Such a procedure gives an approximate value function which can be used to compute an approximate optimal policy.

To generalize and then develop new techniques for solving the approximation problem for MDPs having uncountable state and action spaces, [18] investigated the approximation problem from a different point of view. [18] considered stationary policies having finite range (the collection of all such policies is called the set of deterministic stationary quantizer policies), and then posed the following question: Can these policies approximate an optimal stationary policy with arbitrary precision? At first glance the question seems easy when the cost function is continuous, since any function from the state space to the action space can be approximated with arbitrary precision by functions having finite range (i.e., simple functions). However, one should keep in mind that policies which are close to each other in the action space might have completely different evolution in time, which make the problem quite nontrivial. Under some technical assumptions on the components of MDP, the most important of which is the condition that the transition probability is setwise continuous in the action variable, it was shown in [18] that such quantizer policies can approximate any stationary policy with an arbitrary precision. Furthermore, rates of convergence (as a function of the quantization rate) to optimal costs in the absence of quantization were established.

In this paper, we address a version of this problem in which the setwise continuity assumption is replaced with the weak continuity in the state-action variables. This is motivated by the fact that setwise continuity assumption might be too restrictive in certain important cases. In particular, when one formulates a partially observed Markov decision process (POMDP) as a fully observed MDP through state expansion, setwise continuity appears to be unrealistic (see Example 1). Furthermore, although the weak continuity assumption is not directly comparable to the setwise continuity assumption (one is required to hold for state-action variables and the other is imposed only for the action variable) the weak continuity assumption is typically more common in
A. Notation and Basic Facts

Let $B(E)$ denote the set of all bounded measurable real functions on a measurable space $(E,\mathcal{E})$ and let $C_b(E)$ denote the set of all bounded continuous real valued functions on a topological space $E$ equipped with its Borel $\sigma$-algebra $B(E)$. For any $u \in C_b(E)$, let $\|u\| := \sup_x |u(x)|$. It is known that $(C_b(E),\| \cdot \|)$ is a Banach space. Let $\mathcal{P}(E)$ denote the set of all probability measures on $E$ and let $\mathcal{M}(E)$ denote the Borel $\sigma$-algebra generated by the weak topology on $\mathcal{P}(E)$ [20]. If $E$ is a Borel space (i.e., Borel subset of a complete, separable metric space), then $\mathcal{P}(E)$ is metrizable with the Prokhorov metric, which makes $\mathcal{P}(E)$ a Borel space [21]. A sequence $\{\mu_n\}$ of measures on a measurable space $(E,\mathcal{E})$ is said to converge setwise [22] to a measure $\mu$ if $\mu_n(B) \to \mu(B)$ for all $B \in \mathcal{E}$, or equivalently, $\int g d\mu_n \to \int g d\mu$ for all $g \in B(E)$. Unless otherwise specified, the term "measurable" will refer to Borel measurability.

II. MARKOV DECISION PROCESSES

We consider a discrete-time Markov decision process (MDP) with components as follows.

(i) The state space $Z$ is a Borel space equipped with its Borel $\sigma$-algebra $B(Z)$.

(ii) The action space $A$ is also a Borel space equipped with its Borel $\sigma$-algebra $B(A)$.

(iii) The transition probability $\eta = \{\eta(z,a)\}$ is a probability measure on $Z \times A$, i.e., $\eta(\cdot | z, a)$ is a probability measure on $Z$ for all $z \in Z$ and $a \in A$, and $\eta(D | \cdot , \cdot )$ is a measurable function from $Z \times A$ to $[0,1]$ for each $D \in B(Z)$.

(iv) The one stage cost function $c$ is a measurable function from $Z \times A$ to $[0,\infty)$.

Define the history spaces $H_0 = Z$ and $H_n = (Z \times A)^n \times Z$, $n = 1,2,\ldots$ endowed with their product Borel $\sigma$-algebras generated by $B(Z)$ and $B(A)$. A randomized policy $\varphi = \{\varphi_n\}$ is a sequence of stochastic kernels on $A$ given $H_n$. A deterministic policy $\varphi = \{\varphi_n\}$ is a sequence of stochastic kernels on $A$ given $H_n$ which are realized by a sequence of measurable functions $\{f_n\}$ from $H_n$ to $A$, i.e., $\varphi_n(\cdot|h_n) = \delta_{f_n(h_n)}(\cdot)$, where $f_n : H_n \to A$ measurable. A randomized Markov policy is a sequence of stochastic kernels $\varphi = \{\varphi_n\}$ on $A$ given $Z$. A deterministic Markov policy is a sequence of stochastic kernels $\varphi = \{\varphi_n\}$ on $A$ given $Z$ which are realized by a sequence of measurable functions $\{f_n\}$ from $Z$ to $A$, i.e., $\varphi_n(\cdot|z) = \delta_{f_n(z)}(\cdot)$, where $f_n : Z \to A$ is measurable. A randomized stationary policy is a sequence of stochastic kernels $\varphi = \{\varphi_n\}$ on $A$ given $Z$ such that $\varphi_n = \varphi_m$ for $m, n = 0,1,2,\ldots$. A deterministic stationary policy is a constant sequence of stochastic kernels $\varphi = \{\varphi_n\}$ on $A$ given $Z$ such that $\varphi_n(\cdot|z) = \delta_{f(z)}(\cdot)$ for all $n$ for some measurable function $f : Z \to A$.

We denote by $R\Delta$, $\Delta$, $RM$, $M$, $RS$ and $S$ the set of all randomized, deterministic, randomized Markov, deterministic Markov, randomized stationary and deterministic stationary policies, respectively. We have the following inclusions: $R\Delta \supset RM \supset RS$, $\Delta \supset M \supset S$, $R\Delta \supset \Delta$, $RM \supset M$ and $RS \supset S$.

According to the Ionescu Tulcea theorem [2], an initial distribution $\xi$ on $Z$ and a policy $\varphi$ define a unique probability measure $P^\xi_\varphi$ on $H_\infty = (Z \times A)^\infty$ which is called a strategic measure [23]. Thus $P^\xi_\varphi$ is symbolically given by

$$P^\xi_\varphi(dz_0 da_0 dz_1 da_1 \ldots) := \prod_{n=0}^{\infty} \eta(dz_n|z_{n-1},a_{n-1})\varphi_n(da_n|h_n),$$

where $h_n = (z_0, a_0, \ldots, z_{n-1}, a_{n-1}, z_n)$ and $\eta(dz_0|z_{-1},a_{-1}) := \xi(dz_0)$. If $\xi = \delta_z$ (point mass at $z \in Z$), we write $P^z_\varphi$ instead of $P^\xi_\varphi$. 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:

$$E\left[ \sum_{n=0}^{\infty} \beta^n c(z_n,a_n) \right],$$

$$\limsup_{N \to \infty} \frac{1}{N} E\left[ \sum_{n=0}^{N-1} c(z_n,a_n) \right].$$

Here the expectations are taken with respect to strategic measures induced by the policies and initial distributions.

A. Problem Formulation

A measurable function $q : Z \to A$ is called a quantizer from $Z$ to $A$ if the range of $q$, i.e., $q(Z) = \{q(z) : z \in Z\}$, is finite. The elements of $q(Z)$ (the possible values of $q$) are called the levels of $q$. Let $Q$ denote the set of all quantizers from $Z$ to $A$. A deterministic stationary quantizer policy (see [18]) is a constant sequence $\varphi = \{\varphi_n\}$ of stochastic kernels on $A$ given $Z$ such that $\varphi_n(\cdot|z) = \delta_{q(z)}(\cdot)$ for all $n$ for some quantizer $q \in Q$. Let $SQ$ denote the set of all deterministic stationary quantizer policies.

Let $J_\beta(\varphi, \xi)$ ( $J(\varphi, \xi)$ ) denote the discounted (average) cost function of policy $\varphi \in R\Delta$ and initial distribution $\xi$. The principal objective of this paper is to find conditions on the MDP such that the following holds:

(P) For any given $\varepsilon > 0$ there exists a $\varphi^* \in SQ$ satisfying $J_\beta(\varphi^*, \xi) < \inf_{\varphi \in SQ} J_\beta(\varphi, \xi) + \varepsilon$ (or $J(\varphi^*, \xi) < \inf_{\varphi \in SQ} J(\varphi, \xi) + \varepsilon$), provided that the set $S$ of deterministic stationary policies is an optimal class for the MDP.

Remark 1: [18] investigated and solved this problem for the discounted cost under the assumptions that the action space is compact, the transition probability is setwise continuous in the action variable, the one stage cost function is also continuous in the action variable. The average cost was also considered under some additional restrictions on the ergodicity properties and on the rates of convergence to the invariant measures of Markov chains induced by deterministic stationary policies. The objective of this paper
is to solve problem (P) when the transition probability $\eta$ is weakly continuous in state-action variables. The motivation for this setup 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 and [24, Chapter 4] for the basics of POMDPs.

**Example 1:** Consider the system

$$x_{n+1} = x_n + a_n,$$

$$y_n = x_n + v_n$$

where $X = Y = A = \mathbb{R}_+$ and $\{v_n\}$ is a sequence of i.i.d. random variables uniformly distributed on $[0, 1]$. It is easy to see that the transition probability $p$ is weakly continuous and the observation channel $q$ is continuous in total variation for this POMDP (see Section IV). Hence, by [19, Theorem 4.4] the transition probability $\eta$ of the fully observed MDP is weakly continuous in the state-action variables. Let us set $z = \delta_0$, $\{a_k\}$ and $a = 0$. Hence, $a_k \to a$. We will show that as $a_k \to a$, $\eta(\cdot|z, a_k) \to \eta(\cdot|z, a)$ setwise. Observe that for all $k$ and $y \in Y$, the nonlinear filtering equation gives

$$F(z, a, y) := \Pr\{x_{n+1} \in \cdot | z_n = \delta_0, a_n = 1/k, y_{n+1} = y\}$$

and similarly

$$F(z, a, y) = \delta_0.$$  

Define the open set $O$ in $\mathcal{P}(X)$ with respect to weak topology as

$$O := \{z \in \mathcal{P}(X) : \| f(x) \delta_1(dx) - \int f(x) z(dx) \| < 1 \},$$

where $f$ is the symmetric triangle function between $[-1, 1]$ with $f(0) = 1$. Observe that for this open set $O$, 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,

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

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

where $H(\cdot|z, a) := \Pr\{y_{n+1} \in \cdot | z_n = z, a_n = a\}$. This implies that $\eta(\cdot|z, a_k) \to \eta(\cdot|z, a)$ setwise.

### III. APPROXIMATION OF DETERMINISTIC STATIONARY POLICIES IN MDP

We impose the following assumptions on the components of the MDP.

**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 $(z_n, a_n) \to (z, a)$, then $\eta(\cdot|z_n, a_n) \to \eta(\cdot|z, a)$ weakly.

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

### 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

$$Tu(z) := \min_{a \in A} \left[ c(z, a) + \beta \int Z u(z') \eta(\text{d}z'|z, a) \right].$$

It can be proved that $T$ is a contraction operator with modulus $\beta$ from $C_b(Z)$ to $C_b(Z)$ (see [24, Lemma 2.5]), i.e.,

$$\|Tu - Tv\| \leq \beta \|u - v\|$$

for all $u, v \in C_b(Z)$.

Define the value function $J^*$ of the MDP as

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

The following theorem is a widely known result in the theory of Markov decision processes (see e.g., [24, Theorem 2.8, pg.23]) without a compactness assumption on $Z$.

**Theorem 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.,

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

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

$$J^*(z) = c(z, f^*(z)) + \beta \int X J^*(z') \eta(\text{d}z'|z, f^*(z)).$$

Finally, there exists a deterministic stationary policy $\varphi^* = \{f^*\}$ which is optimal, so it satisfies (3).

Since the action space $A$ is compact and thus totally bounded, one can find a nested sequence of finite sets $\{\{a_i\}_{i=1}^{k_n}\}_{n \geq 1}$ such that for all $n$

$$\min_{i \in \{1, \ldots, k_n\}} d_a(a, a_i) < 1/n$$

for all $a \in A$. (4)

In other words, $\{a_i\}_{i=1}^{k_n}$ is an $1/n$-net of $A$ and $\{a_i\}_{i=1}^{k_n} \subset \{a_i\}_{i=1}^{k_{n+1}}$ for all $n$.

**Remark 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 Section III by finding a nested sequence of finite sets $\{\{a_i\}_{i=1}^{k_n}\}_{n \geq 1}$ such that for all $n$,  

$$\{a_i\}_{i=1}^{k_n} \subset \{a_i\}_{i=1}^{k_{n+1}}$$

is an $1/n$-net of $A(z)$.

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

$$T_n u(z) := \min_{i \in \{1, \ldots, k_n\}} \left[ c(z, a_i) + \beta \int Z u(z') \eta(\text{d}z'|z, a_i) \right].$$

(5)

$T_n$ can be shown to be a contraction operator with modulus $\beta$ mapping $C_b(Z)$ into itself (see [24, Lemma 2.5]). We also have $T_n u \geq T_{n+1} u$ for any $u \in C_b(Z)$ by the nestedness of the sets $\{\{a_i\}_{i=1}^{k_n}\}_{n \geq 1}$. Let $J^*_n \subset C_b(Z)$ denote the fixed point of $T_n$. Using the measurable selection theorem
[24, Proposition D.3] it can be proved that there exists a measurable $f^* : Z \to A$ such that
\[ T_n J^*(z) = c(z, f^*_n(z)) + \beta \int_Z J^*(z') \eta(dz'|z, f^*_n(z)). \]
(6)
Furthermore, one can also prove that (see [24, Lemma 2.6])
\[ J(\{f^*_n\}, z) = J^*(z) \quad \text{for all } z \in Z. \]
(7)
Notice that $f^*_n \in Q$, so the deterministic stationary policy $\{f^*_n\}$ is in $SQ$. We now state our main theorem in this section.

**Theorem 2:** Under Assumption 1, for any given $\epsilon > 0$ there exists $\varphi^* \in SQ$ such that
\[ J(\varphi^*, \xi) < \min J(\varphi, \xi) + \epsilon. \]
(8)
**Sketch of proof:** For any $\varphi \in X$, we can prove that $T_n u \to Tu$ pointwise as $n \to \infty$. Since, $T_{n+1} u \leq T_n u$ for all $n$, we have $\|T_n u - Tu\| \to 0$ by Dini’s theorem [25]. Using this result, we can show that the fixed points of the contraction operators $\{T_n f\}_{n \geq 1}$ converge uniformly to the fixed point of the contraction operator $T$, that is, $J(\{f^*_n\}, \cdot) = J^*_n(\cdot) \to J^*(\cdot) = J(\{f^*\}, \cdot)$ uniformly.

From this, the result follows.

**B. Average Cost**

In this section we consider the problem (P) for the average cost. Here, for the discounted cost we write $J_\beta$ since the dependence on the discount factor $\beta$ is needed here. Throughout this section we are imposing Assumption 1. Note that since the one stage cost function $c$ is bounded, say, by $L \geq 0$, we must have $\|J^*\| \leq L$ where $J^*(z) := \inf_{\varphi} J(\varphi, z)$. This is because for any policy $\varphi$ and initial state $z$ one can write
\[ J(\varphi, z) = \limsup_{N \to \infty} \int_Z c(z, \mu_N) d\mu_N, \]
where $\{\mu_N\}$ is the set of probability measures (i.e. occupation measures) defined in [2, 5.7.18 pg. 118]. One can further prove using Tauberian theorem [2, Lemma 5.3.1] that $(1 - \beta) J^* \leq L$ uniformly by $L$. Hence, $(d) (1 - \beta) J^*_\beta(z) \leq L$ for all $\beta \in (0, 1)$ and $z \in Z$.

We will impose the following assumption in addition to Assumption 1 in this section.

**Assumption 2:**
There exists numbers $\alpha \in (0, 1)$ and $N \geq 0$, a nonnegative function $h$ on $Z$ and a state $\tilde{z} \in Z$ such that,
\[ (e) -N \leq h(\beta z) \leq b(z) \quad \text{for all } z \in Z \quad \text{and } \beta \in [\alpha, 1), \]
where $h(\beta z) := J^*_\beta(z) - J^*_\beta(\tilde{z})$.
\[ (f) \] The sequence $\{h(\beta_n)\}$ is equicontinuous where $\{\beta_n\}$ is the sequence of discount factors converging up to 1 and satisfies [2, Lemma pg. 88].

The following theorem is the same as [2, Theorem 5.4.3] except that the function $h$ is continuous and bounded here which results from Assumption 2-(f) and the Arzela-Ascoli theorem, and that instead of setwise continuity in the action variable, we only require weak continuity in state-action variables of the transition probability.

**Theorem 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 $\{f^*_n\} \in S$ such that $(\rho^*, b, f^*)$ satisfies the Average Cost Optimality Inequality (ACOII), i.e.,

\[ \rho^* + h(z) \geq \min_{a \in A} [c(z, a) + \int_Z h(z') \eta(dz'|z, a)] \]
\[ = c(z, f^*(z)) + \int_Z h(z') \eta(dz' | z, f^*(z)) \quad \forall z \in Z. \]

Moreover, $\{f^*_n\}$ is optimal and $\rho^*$ is the value function, i.e.,
\[ \inf_{\varphi} J(\varphi, z) = J^*(z) = J(\{f^*_n\}, z) = \rho^* \quad \forall z \in Z. \]

Let $\{f_n\} \subset Q$ such that $f_n \to f^*$ uniformly and $|f_n(X)| \leq k_n$. Then, by the weak continuity of $\eta$ and the continuity of $c$, we have
\[ c(z, f^*_n(z)) + \int_Z h(z') \eta(dz' | z, f^*_n(z)) \]
\[ \to c(z, f^*(z)) + \int_Z h(z') \eta(dz' | z, f^*(z)) \quad (9) \]
pointwise. Let $\{f^*_n\}$ be a sequence of measurable functions which satisfy for all $n$.

\[ c(z, f^*_n(z)) + \int_Z h(z') \eta(dz' | z, f^*_n(z)) \]
\[ = \min_{a_i \in \{1,...,k_n\}} [c(z, a_i) + \int_Z h(z') \eta(dz' | z, a_i)], \quad n \geq 1. \]

The existence of these functions again follows from the measurable selection theorem. Since (10) is upper bounded by the first term in (9) for all $n$, we must have
\[ c(z, f^*_n(z)) + \int_Z h(z') \eta(dz' | z, f^*_n(z)) \]
\[ \to c(z, f^*(z)) + \int_Z h(z') \eta(dz' | z, f^*(z)) \quad (11) \]
monotonically, and hence uniformly by Dini’s theorem. We now state the main theorem for this section, where it has a simple proof based on the iteration of the ACOII.

**Theorem 4:** Under Assumptions 1 and 2, for any given $\epsilon > 0$ there exists $\varphi^* \in SQ$ such that
\[ J(\varphi^*, \xi) < \min_{\varphi \in R^\Delta} J(\varphi, \xi) + \epsilon. \]
(12)

A MDP is said to be convex if the state space $Z$ is an open convex subset of a separable Banach space, e.g., $\mathbb{R}^n$, and the value functions $J^*_\beta(n)$ is convex for each $\beta(n)$ in Assumption 2-(f). We refer the reader to papers [26], [27] for sufficient conditions that results in convex $J^*_\beta$ for each $\beta$. For this MDP, if the function $b$ in Assumption 2-(e) is upper semi-continuous and satisfies $\int_Z b(z') \eta(dz' | z, a) < \infty$ for all $(z, a) \in Z \times A$, then under Assumption 2-(e) it is proved in [28] that functions $\{h(\beta_n)\}$ satisfy the local Lipschitz property, and hence satisfy Assumption 2-(f). This means that Theorem 4 is applicable to any convex MDP satisfying the above condition on the function $b$. 

1082
IV. APPLICATION TO PARTIALLY OBSERVED MDP

In this section we apply the result obtained in Section III-A to the partially observed Markov decision processes (POMDPs) defined as follows:

(i) The state space $X$ is a Borel space equipped with its Borel $\sigma$-algebra $B(X)$.
(ii) The observation space $Y$ is also a Borel space equipped with its Borel $\sigma$-algebra $B(Y)$.
(iii) The action space $A$ is also a Borel space equipped with its Borel $\sigma$-algebra $B(A)$.
(iv) The transition probability $p$ is a stochastic kernel on $X$ given $X \times A$.
(v) The observation channel $q$ is a stochastic kernel on $Y$ given $X$.
(vi) The one stage cost function $c$ is a measurable function from $X \times A$ to $[0, \infty)$.

Once again, define the history spaces $\tilde{H}_n = (Y \times A)^n \times Y$, $n = 0, 1, 2, \ldots$ endowed with their product Borel $\sigma$-algebras generated by $B(Y)$ and $B(A)$. A randomized policy $\pi = \{\pi_n\}$ is a sequence of stochastic kernels on $A$ given $H_n$. A deterministic policy $\pi = \{\pi_n\}$ is a sequence of stochastic kernels on $A$ given $H_n$, which are realized by a sequence of measurable functions $\{g_n\}$ from $H_n$ to $A$, i.e., $\pi_n(\cdot | \tilde{h}_n) = \delta_{g_n(\tilde{h}_n)}(\cdot)$ where $g_n : H_n \rightarrow A$ measurable. We denote by $\mathcal{R}$, $\mathcal{H}$ the set of all randomized and deterministic policies, respectively.

A. Reduction of POMDPs to MDPs

It is known that any POMDP can be reduced to a (completely observable) MDP [29], [30], whose states are the posterior state distributions or beliefs of the observer, i.e., the state at time $n + 1$ is

$$\Pr\{x_{n+1} \in \cdot | y_0, \ldots, y_n, a_0, \ldots, a_{n-1}\}.$$

Let us name this equivalent MDP as belief-MDP. The belief-MDP has the following components. Here, $Z := \mathcal{P}(X)$.

(i) The state space $Z$ is the set of all probability measures over $X$ equipped with the Borel $\sigma$-algebra $B(Z)$ generated by the topology of weak convergence.
(ii) The action space $A$ is the same as the action space of POMDP.
(iii) The transition probability $\eta$ is a stochastic kernel on $Z$ given $Z \times A$.
(iv) The one stage cost function $\tilde{c}$ is a measurable function from $Z \times A$ to $[0, \infty)$ defined as:

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

For the explicit construction of the transition probability $\eta$ of the belief-MDP see e.g. [24]. Hence, belief-MDP is a Markov decision process with the components $(Z, A, \eta, c)$

Let $J_{\beta}(\pi, \mu)$ denote the discounted cost function of the policy $\pi \in \mathcal{R}$ and initial distribution $\mu$ of the POMDP. Define the history spaces $H_n = (Z \times A)^n \times Z$, $n = 0, 1, 2, \ldots$ endowed with their product Borel $\sigma$-algebras generated by $B(Z)$ and $B(A)$ for the belief-MDP. Notice that any history vector $\tilde{h}_n = (z_0, \ldots, z_n, a_0, \ldots, a_{n-1})$ of the belief-MDP is indeed a function of the history vector $h_n = (y_0, \ldots, y_n, a_0, \ldots, a_{n-1})$ of the POMDP. Let us write this relation as

$$i(\tilde{h}_n) = h_n.$$

Hence, for a policy $\varphi = \{\varphi_n\} \in R\Delta$, we can define a policy $\pi^\varphi = \{\pi_n^\varphi\} \in R\Pi$ as

$$\pi_n^\varphi(\cdot | \tilde{h}_n) := \varphi_n(\cdot | i(\tilde{h}_n)).$$

Let us write this as a mapping from $R\Delta$ to $R\Pi$: $R\Delta \ni \varphi \mapsto i(\varphi) = \pi^\varphi \in R\Pi$. Hence, we have $i(R\Delta) \subset R\Pi$. Furthermore, 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. These two observations imply that for any $\beta \in (0, 1)$

$$\inf_{\varphi \in R\Delta} J_{\beta}(\varphi, \xi) \geq \inf_{\pi \in R\Pi} \tilde{J}_{\beta}(\pi, \mu).$$

(14) is simply a consequence of the construction we made, but one can also prove that (see [29], [30], [31]) for any $\beta \in (0, 1)$

$$\inf_{\varphi \in R\Delta} J_{\beta}(\varphi, \xi) = \inf_{\pi \in R\Pi} \tilde{J}_{\beta}(\pi, \mu).$$

(15) and furthermore, 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.

We will impose the following assumptions on the components of the original POMDP.

Assumption 3:

(a) The one stage cost function $\tilde{c} \in C_b(X \times A)$.
(b) The stochastic kernel $p(\cdot | x, a)$ is weakly continuous in $(x, a) \in X \times A$.
(c) The stochastic kernel $q(\cdot | x)$ is continuous in total variation, i.e., if $x_n \rightarrow x$, then $q(\cdot | x_n) \rightarrow q(\cdot | x)$ in total variation.
(d) $X$ and $A$ are compact.

It is known that the set of probability measures is compact in 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 $(z_n, a_n) \rightarrow (z, a)$ in $Z \times A$. Define $v_n(x) := \tilde{c}(x, a_n)$ and $v(x) := \tilde{c}(x, a)$. Note that $v_n$ converges continuously to $v$ [32], that is, $v_n(x_n) \rightarrow v(x)$ for any $x_n \rightarrow x$ where $x \in X$. Hence, by [32, Theorem 3.3], we have

$$\lim_{n \rightarrow \infty} \int_X v_n(x) z_n(dx) = \int_X v(x) z(dx).$$

Since $c$ is bounded, $c \in C_b(Z \times A)$. Furthermore, it is proved in [19, Theorem 4.4] that under Assumption 3-(b)(c), the stochastic kernel $\eta$ for belief-MDP is weakly continuous in $(z, a) \in Z \times A$. Hence, the belief-MDP satisfies Assumption 1, and so Theorem 2 is applicable.

Theorem 5: Under Assumption 3 on the POMDP, for any given $\varepsilon > 0$ there exists $\varphi^* \in SQ$ such that

$$J(\varphi^*, \xi) \leq \min_{\varphi \in R\Delta} J(\varphi, \xi) + \varepsilon$$

for the belief-MDP.
V. CONCLUSION

In this paper, approximation of optimal stationary policies with stationary quantizer policies in MDPs is considered with discounted and average costs. It is assumed that the transition probability is weakly continuous in state-action variables. It was shown that one can always find a quantized policies which \( \varepsilon \)-optimal. These results were applied to the fully observed reduction of the partially observed MDP (belief-MDP).

One future 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. One 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