Approximation of Stationary Control Policies by Quantized Control in Markov Decision Processes

Naci Saldi, Tamás Linder, Serdar Yüksel
Department of Mathematics and Statistics, Queen's University, Kingston, ON, Canada
Email: {nsaldi,linder,yuksel}@mast.queensu.ca

***Abstract*— We consider the problem of approximating optimal stationary control policies by quantized control. Stationary quantizer policies are introduced and it is shown that such policies are ε-optimal among stationary policies under mild technical conditions. Quantitative bounds on the approximation error in terms of the rate of the approximating quantizers are also derived. Thus, one can search for ε-optimal policies within quantized control policies. These pave the way for applications in optimal design of networked control systems where controller actions need to be quantized, as well as for a new computational method for the generation of approximately optimal Markov decision policies in general (Borel) state and action spaces for both discounted cost and average cost infinite horizon optimal control problems.**

I. Introduction

In the theory of Markov decision processes (MDP), the set of control policies induced by measurable mappings from the state space to the action space is an important class since it is the smallest structured set in which one can find a globally optimal policy for a large class of MDPs [1], [2]. Such policies are usually called non-randomized (or deterministic) stationary policies, or stationary policies for short, in the literature. Although this set is the smallest structured optimal class for MDPs, computing an optimal policy in this class is in general computationally prohibitive for non-finite Polish (that is, complete, separable and metric) state and action spaces. Hence, it is of interest to approximate optimal policies even in this class.

From the computation point of view, Approximate Value Iteration (AVI) and Approximate Policy Iteration (API) are two powerful methods to approximate an optimal (stationary) policy for an MDP (see [3], [4], [5], [6] and references therein). In AVI, the idea is to compute approximately the value iteration function in each step of the value iteration algorithm. This way one can both approximately find an optimal value function and construct an approximately optimal (stationary) policy. Although, the main purpose of the API is the same as AVI (i.e., to approximate the optimal value function) the algorithm works differently. In each step, first an approximate value function for a given policy is computed. Then, an improved policy is generated using the approximate value function. The main drawback of these algorithms is the accumulation of the approximation error in each step. Another well-known method for approximating an optimal (stationary) policy is *state aggregation*. In this method, similar states (e.g., with respect to cost and transition probabilities) are aggregated to form meta-states, and an optimal policy can then be calculated according to the reduced MDP (see [7], [8], [9] and references therein). The basic issue with this method is how to efficiently aggregate states and construct a reduced MDP from the original one.

For denumerable MDPs, several approaches have been developed to approximate an optimal (stationary) policy. References [10]–[14] used the technique of truncating the state space when evaluating the value function in the value iteration algorithm. In these schemes, at each step the state space is truncated and the corresponding value function calculated; this latter is proved to converge to the true value function. Then, using the truncated value function approximately optimal policies are constructed. In [15] the idea of *embedding* is used to approximate an optimal (stationary) policy. Here, a finite state MDP is constructed, which has the same optimal cost as the original MDP and has an optimal policy which agrees with an optimal policy of the original MDP in the approximating set. This finite state MDP is said to be embedded in the original one. Reference [16] considers the approximation problem for denumerable continuous time MDPs. Here a convergence notion for control models is defined and is then used to show the convergence of optimal policies for the truncated MDPs to an optimal policy for the original MDP.

In this paper we introduce a new method for obtaining approximately optimal control policies with guaranteed performance bounds for a class of cost functions. We first introduce the set of non-randomized stationary quantizer policies, which is a proper subset of the set of non-randomized stationary policies. Policies in this set are induced by quantizers from the state space to the action space. These policies are then used to approximate non-randomized stationary policies. We show that there exists an ε-optimal non-randomized stationary quantizer policy in the set of non-randomized stationary policies. We also demonstrate that the difference between the cost of an optimal nonrandomized stationary policy and the cost of the approximating non-randomized stationary quantizer policy can be upper bounded by a term depending on the rate of the quantizer. Although our method is somewhat similar to the state aggregation approach (i.e. aggregate states which are close to each other in state space), unlike in state aggregation, we can obtain

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

explicit approximation results and error bounds for various cost functions.

The proofs of the results in this paper are presented in the full paper [17] which also deals with the approximation of *randomized* stationary policies.

II. Definition of Markov Decision Process

We consider a discrete time Markov decision process (MDP) with the components as follows:

(i) The state space X is a complete, separable metric (Polish) space equipped with its Borel σ-algebra $\mathcal{B}(\mathsf{X})$.
(ii) The action space A is also a Polish space equipped with its Borel σ-algebra $\mathcal{B}(\mathsf{A})$.
(iii) The transition probability p is a stochastic kernel on X given $\mathsf{X}\times\mathsf{A}$; i.e., $p(\,\cdot\,|x,a)$ is a probability measure on X for all $x\in\mathsf{X}$ and $a\in\mathsf{A}$, and $p(B|\,\cdot\,,\,\cdot\,)$ is a measurable function from $\mathsf{X}\times\mathsf{A}$ to $[0,1]$ for each $B\in\mathcal{B}(\mathsf{X})$.
(iv) The cost function w will be specified later.

The following notation is from [18]. Define the history spaces $\mathsf{H}_n = (\mathsf{X}\times\mathsf{A})^n\times\mathsf{X}$, $n=0,1,2,\ldots$ with their product Borel σ-algebras generated by $\mathcal{B}(\mathsf{X})$ and $\mathcal{B}(\mathsf{A})$. A *randomized policy* $\pi=\{\pi_n\}$ is a sequence of stochastic kernels on A given H_n. A *non-randomized policy* $\pi=\{\pi_n\}$ 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. $\pi_n(B|h_n)=\delta_{f_n(h_n)}(B)$ where $f_n:\mathsf{H}_n\to\mathsf{A}$ measurable. A *randomized Markov policy* is a sequence of stochastic kernels $\pi=\{\pi_n\}$ on A given X. A *non-randomized Markov policy* is defined as sequence of stochastic kernels $\pi=\{\pi_n\}$ on A given X which are realized by a sequence of measurable functions $\{f_n\}$ from X to A; i.e., $\pi_n(B|x)=\delta_{f_n(x)}(B)$, where $f_n:\mathsf{X}\to\mathsf{A}$ is measurable. A *randomized stationary policy* is a sequence of stochastic kernels $\pi=\{\pi_n\}$ on A given X such that $\pi_n=\pi_m=\eta$ for $m,n=0,1,2,\ldots$. A *non-randomized stationary policy* is a sequence of stochastic kernels $\pi=\{\pi_n\}$ on A given X such that $\pi_n=\pi_m$ for $m,n=0,1,2,\ldots$ and $\pi_n(B|x)=\delta_{f(x)}(B)$ for some measurable function f from X to A.

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

Let $B(\mathsf{E})$ denote the set of all bounded measurable real functions on a measurable space $(\mathsf{E},\mathcal{E})$ and let $C_b(\mathsf{E})$ denote the set of all bounded continuous real valued functions on a topological space E equipped with its Borel σ-algebra $\mathcal{B}(\mathsf{E})$. Also let $\mathcal{P}(\mathsf{E})$ denote the set of all probability measures on E and let $\mathcal{M}(\mathsf{E})$ denote the Borel σ-algebra generated by the weak topology on $\mathcal{P}(\mathsf{E})$. If E is a Polish space, then $\mathcal{P}(\mathsf{E})$ is metrizable with the Prokhorov metric which makes $\mathcal{P}(\mathsf{E})$ into a Polish space. We always equip the set of probability measures with the Borel σ-algebra generated by weak topology. Unless otherwise specified, the term "measurable" will refer to Borel measurability.

According to the Ionescu Tulcea theorem [19], an initial distribution μ on X and a policy π define a unique probability measure P_μ^π on $\mathsf{H}_\infty=(\mathsf{X}\times\mathsf{A})^\infty$, which is called a *strategic measure* [18]. If $\mu=\delta_x$ for some $x\in\mathsf{X}$, we write P_x^π instead of $P_{\delta_x}^\mu$. For $\Delta\subset R\Pi$ define $L_\Delta:=\{P_\mu^\pi:\mu\in\mathcal{P}(\mathsf{X}),\pi\in\Delta\}$. Then $L_{R\Pi}$ is the set of all strategic measures. Clearly $L_{R\Pi}\subset\mathcal{P}(\mathsf{H}_\infty)$. It is known that $L_{R\Pi}$, L_Π, L_{RM}, L_M, L_{RS} and L_S are all in $\mathcal{M}(\mathsf{H}_\infty)$ [18, Theorem 3.2]. Hence, the restriction of $\mathcal{M}(\mathsf{H}_\infty)$ to L_Δ, for each of $\Delta=R\Pi,\Pi,RM,M,RS,S$, coincides with the Borel σ-algebra on L_Δ generated by the weak topology. The cost function w is defined to be a measurable function from $L_{R\Pi}$ to $[0,\infty]$, i.e.,

$$w: L_{R\Pi}\to[0,\infty]. \tag{1}$$

Let c and c_n, $n=0,1,2,\ldots$, be measurable functions from $\mathsf{X}\times\mathsf{A}$ to $[0,\infty]$. The following are examples for the type of cost functions defined in (1) (see [20]). Here the expectations are taken with respect to strategic measures induced by the policies and initial distributions.

i) Expected Finite Horizon Cost: $E[\sum_{n=0}^N c_n(x_n,a_n)]$ for some $N<\infty$.
ii) Expected Total Cost: $E[\sum_{n=0}^\infty c_n(x_n,a_n)]$.
iii) Expected Discounted Cost: $E[\sum_{n=0}^\infty \beta^n c(x_n,a_n)]$ for some $\beta\in(0,1)$.
iv) Expected Average Cost: $\limsup_{N\to\infty}\frac{1}{N}E[\sum_{n=0}^N c(x_n,a_n)]$.

Note that both the expected finite horizon cost and the expected discounted cost are special cases of the expected total cost.

A measurable function $q:\mathsf{X}\to\mathsf{A}$ is called a *quantizer* from X to A if the range of q, i.e., $q(\mathsf{X})=\{q(x):x\in\mathsf{X}\}$, is finite. The rate R of an any quantizer q is defined as the logarithm of the cardinality of its range, i.e., $R:=\log(q(\mathsf{X}))$. Let $\mathcal{Q}$ denote the set of all quantizers from X to A. In this paper we introduce a new type of policy called a *non-randomized stationary quantizer policy*. Such a policy is a sequence $\pi=\{\pi_n\}$ of stochastic kernels on A given X such that $\pi_n=\pi_m$, $m,n=0,1,2,\ldots$, and $\pi_n(B|x)=\delta_{q(x)}(B)$ for some $q\in\mathcal{Q}$. Let $S\mathcal{Q}$ denote the set of all non-randomized stationary quantizer policies and define the the set of strategic measures generated by $S\mathcal{Q}$ as $L_{S\mathcal{Q}}=\{P_\mu^\pi:\mu\in\mathcal{P}(\mathsf{X}),\pi\in S\mathcal{Q}\}$.

The primary goal of this paper is to find conditions on the spaces X and A, initial distribution μ, the stochastic kernel p, and the cost function w such that the following statements hold:

(P1) For any given $\varepsilon>0$ there exists a non-randomized stationary quantizer policy π^* satisfying $w(P_\mu^{\pi^*})<\inf_{\pi\in S}w(P_\mu^\pi)+\varepsilon$.
(P2) For any $\pi\in S$ there exists an approximating sequence $\{\pi^k\}\in S\mathcal{Q}$ such that the difference $|w(P_\mu^\pi)-w(P_\mu^{\pi^k})|$ can be upper bounded by a term depending on the rates of quantizers inducing $\{\pi^k\}$.

III. Approximation of Non-randomized Stationary Policies

A sequence $\{\mu_n\}$ of measures on measurable space $(\mathsf{E}, \mathcal{E})$ is said to converge setwise [21] to a measure μ 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 $g \in B(\mathsf{E})$. In this section, we will use the following assumptions:

(a) The stochastic kernel $p(\cdot\,|x,a)$ is setwise continuous in $a \in \mathsf{A}$, i.e., if $a_n \to a$, then $p(\cdot\,|x,a_n) \to p(\cdot\,|x,a)$ setwise for all $x \in \mathsf{X}$.
(b) A is compact.

We now define the ws^∞ topology on $\mathcal{P}(\mathsf{H}_\infty)$ which was first introduced by M. Schäl in [22]. Let $\mathcal{C}(\mathsf{H}_0) = B(\mathsf{X})$ and let $\mathcal{C}(\mathsf{H}_n)$ $(n \geq 1)$ be the set of real valued functions g on H_n such that $g \in B(\mathsf{H}_n)$ and $g(x_0, \cdot, x_1, \cdot, \ldots, x_{n-1}, \cdot, x_n) \in C_b(\mathsf{A}^n)$ for all $(x_0, \ldots, x_n) \in \mathsf{X}^{n+1}$. The ws^∞ topology on $\mathcal{P}(\mathsf{H}_\infty)$ is defined as the smallest topology which makes the mappings $P \mapsto \int_{\mathsf{H}_\infty} g dP$, $g \in \bigcup_{n=0}^\infty \mathcal{C}(\mathsf{H}_n)$, continuous. Similarly, the weak topology on $\mathcal{P}(\mathsf{H}_\infty)$ can also be defined as the smallest topology which makes the mappings $P \mapsto \int_{\mathsf{H}_\infty} g dP$, $g \in \bigcup_{n=0}^\infty C_b(\mathsf{H}_n)$, continuous [22, Lemma 4.1]. A theorem due to E.J. Balder [23, page 149] and A.S. Nowak [24] states the weak topology and the ws^∞ topology on $L_{R\Pi}$ are equivalent. Hence, the ws^∞ topology is metrizable with the Prokhorov metric on $L_{R\Pi}$.

The following theorem is a Corollary of [25, Theorem 2.4] which will be used in this paper frequently. It is the generalization of the dominated convergence theorem.

Theorem 1. *Let $(\mathsf{E}, \mathcal{E})$ be a measurable space and let μ, μ_n, $n = 1, 2, \ldots$ be measures with the same finite total mass. Suppose $\mu_n \to \mu$ setwise, $\lim_{n\to\infty} h_n(x) = h(x)$ for all $x \in \mathsf{X}$, and h, h_n $(n \geq 1)$ are uniformly bounded. Then, $\lim_{n\to\infty} \int h_n d\mu_n = \int h d\mu$.*

Since the action space A is compact, we can uniformly approximate any measurable function $f : \mathsf{X} \to \mathsf{A}$ by a sequence of simple functions $\{q_k\} \in \mathcal{Q}$ (quantizers in our context), i.e., such that $q_k(x)$ converges uniformly to $f(x)$ as $k \to \infty$. The following proposition will be proved using Theorem 1.

Proposition 1. *Assume (a) and (b) hold. Let $\pi \in S$ be induced by $f : \mathsf{X} \to \mathsf{A}$ and let $\{q_k\}$ be the sequence of quantizers which converge uniformly to f. Let $\{\pi^k\} \in S\mathcal{Q}$ be induced by $\{q_k\}$. Then, $P_\mu^{\pi^k} \to P_\mu^\pi$ in ws^∞ topology for an arbitrary initial distribution μ.*

A. Expected Total and Discounted Costs

In this section, we consider the first approximation problem **(P1)** for the expected total cost criterion $E[\sum_{n=0}^\infty c_n(x_n, a_n)]$ and its special case, the expected discounted cost criterion $E[\sum_{n=0}^\infty \beta^n c(x_n, a_n)]$. Let E_μ^π denote the expectation with respect to P_μ^π on H_∞. Define

$$w_t(P_\mu^\pi) := E_\mu^\pi \left[\sum_{n=0}^\infty c_n(x_n, a_n) \right].$$

In this section we impose the following assumptions in addition to assumptions (a) and (b):

(c) c and c_n $(n \geq 1)$ are non-negative, bounded measurable functions satisfying $c(x, \cdot)$, $c_n(x, \cdot) \in C_b(\mathsf{A})$ for all $x \in \mathsf{X}$.
(d) $\sup_{\pi\in S} \sum_{n=N+1}^\infty \int_{\mathsf{H}_\infty} c_n(x_n, a_n) P_\mu^\pi \to 0$ as $N \to \infty$.

Since the per stage cost functions c_n are non-negative, assumption (d) is equivalent to Condition (C) in Schäl's paper [22, page 349]. Clearly, the expected discounted cost and expected finite horizon cost satisfy the assumption (d) under assumption (c).

We have the following proposition about the continuity of the expected total cost w_t.

Proposition 2. *Under assumptions (c) and (d), $\pi \mapsto w_t(P_\mu^\pi)$ is sequentially continuous on $L_{R\Pi}$ under the ws^∞ topology.*

Theorem 2. *Under assumptions (a), (b), (c) and (d) for any $\varepsilon > 0$ there exists $\pi^* \in S\mathcal{Q}$ such that $w_t(P_\mu^{\pi^*}) < \inf_{\pi\in S} w_t(P_\mu^\pi) + \varepsilon$. Hence, **(P1)** is true under assumptions (a), (b), (c) and (d) for the expected total cost criterion.*

Proof: This follows from Proposition 1 and 2.

In the rest of this section we consider the expected discounted cost criterion, i.e., $w_\beta(P_\mu^\pi) := \sum_{n=0}^\infty \beta^n \int_{\mathsf{H}_\infty} c(x_n, a_n) dP_\mu^\pi$. Recall that w_β satisfies (d) under the assumption (c), so Theorem 2 holds for w_β under assumptions (a), (b), (c). However, we can also obtain Theorem 2 for w_β by considering occupation measures. For any initial distribution μ and any policy π the occupation measure defined is as follows:

$$\nu_\mu^\pi(B) := (1-\beta) \sum_{n=0}^\infty \beta^n P_\mu^\pi\big((x_n, a_n) \in B\big), \qquad (2)$$

where $B \in \mathcal{B}(\mathsf{X} \times \mathsf{A})$. It is clear that ν_μ^π is a well defined probability measure on $\mathsf{X} \times \mathsf{A}$. It is also immediate that $w_\beta(P_\mu^\pi)$ can be written as an integral of $\frac{1}{(1-\beta)}c$ with respect to the occupation measure ν_μ^π, i.e.,

$$w_\beta(P_\mu^\pi) = \frac{1}{(1-\beta)} \int_{\mathsf{X}\times\mathsf{A}} c(x,a) d\nu_\mu^\pi. \qquad (3)$$

Similar to the ws^∞ topology, we now define the ws topology on $\mathcal{P}(\mathsf{X} \times \mathsf{A})$ which was also introduced by M. Schäl in [22]. A sequence of probability measures $\{\nu_k\}$ on $\mathsf{X}\times\mathsf{A}$ converges in the ws topology to a probability measure ν on $\mathsf{X} \times \mathsf{A}$ if and only if $\int g d\nu_k \to \int g d\nu$ for all bounded measurable function g satisfying $g(x, \cdot) \in C_b(\mathsf{A})$ for all $x \in \mathsf{X}$.

Proposition 3. *Let $P_\mu^\pi, \{P_\mu^{\pi^k}\} \in L_{R\Pi}$ $(k \geq 1)$. If $P_\mu^{\pi^k} \to P_\mu^\pi$ in the ws^∞ topology, then $\nu_\mu^{\pi^k} \to \nu_\mu^\pi$ in the ws topology.*

Let $w_{o,\beta}$ denote the expected discounted cost function when it is written with respect to the occupation measure, i.e.,

$$w_{o,\beta}(\nu_\mu^\pi) := \frac{1}{(1-\beta)} \int_{\mathsf{X}\times\mathsf{A}} c(x,a) d\nu_\mu^\pi,$$

and note that $w_{o,\beta}(\nu_\mu^\pi) = w_\beta(P_\mu^\pi)$.

Proposition 4. *Under assumption (c) if a sequence of occupation measures $\{\nu_\mu^{\pi^k}\}$ converges to an occupation measure ν_μ^π in the ws topology, then $w_{o,\beta}(\nu_\mu^{\pi^k}) \to w_{o,\beta}(\nu_\mu^\pi)$.*

Proof: This follows from the definition of the ws topology.

Theorem 3. *Under assumptions (a), (b) and (c) for any given $\varepsilon > 0$ there exists $\pi^* \in SQ$ such that $w_{o,\beta}(\nu_\mu^{\pi^*}) < \inf_{\pi\in S} w_{o,\beta}(\nu_\mu^\pi) + \varepsilon$.*

Proof: This follows from Propositions 1, 3 and 4.

B. Expected Average Cost

In this section we consider the first approximation problem **(P1)** for the expected average cost function $\limsup_{N\to\infty} \frac{1}{N} E[\sum_{n=0}^{N-1} c(x_n, a_n)]$. We are still assuming (b) and (c). Recall that the goal is to obtain an ε-optimal non-randomized stationary quantizer policy in the set of non-randomized stationary policies for any given $\varepsilon > 0$. In contrast to the expected total cost and discounted cost cases, the expected average cost cannot be expected to be sequentially continuous on the set of strategic measures $L_{R\Pi}$ for the ws^∞ topology under reasonable assumptions. Hence, in this case it is not convenient working with strategic measures.

Let $J(\pi, \mu)$ denote the expected average cost associated with the initial distribution μ and policy π, $J(\pi,\mu) = \limsup_{N\to\infty} \frac{1}{N} E_\mu^\pi[\sum_{n=0}^{N-1} c(x_n, a_n)]$. If $\mu = \delta_x$, we write $J(\pi, x)$ instead of $J(\pi, \delta_x)$. Observe that any non-randomized stationary policy π, induced by f, defines a stochastic kernel on X given X:

$$Q_\pi(\,\cdot\,|x) := \int_{\mathsf{A}} p(\,\cdot\,|x,a)\delta_{f(x)}(da) = p(\,\cdot\,|x, f(x)). \quad (4)$$

Define the function c_π on X corresponding to policy π as follows: $c_\pi(x) := \int_{\mathsf{A}} c(x,a)\delta_{f(x)}(da) = c(x, f(x))$. Clearly, c_π is a bounded measurable function. Let Q_π^n denote the n-step transition probability for Q_π. Let us write $Q_\pi^n g(x) := \int_{\mathsf{X}} g(\tilde{x}) Q_\pi^n(d\tilde{x}|x)$ for any measurable function g on X. If Q_π admits an ergodic invariant probability measure ν_π, then by Theorem 2.3.4 and Proposition 2.4.2 in [21], there exists an invariant set with full ν_π measure such that for all x in that set we have

$$\begin{aligned} J(\pi, x) &= \limsup_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \int_{\mathsf{H}_\infty} c(x_n, a_n) dP_x^\pi \\ &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} Q_\pi^n c_\pi(x) = \int_{\mathsf{X}} c_\pi(x)\nu_\pi(dx). \quad (5) \end{aligned}$$

Let $\mathsf{M}_\pi \subset \mathsf{X}$ be the set of all $x \in \mathsf{X}$ such that convergence in (5) holds. Hence, $\nu_\pi(\mathsf{M}_\pi) = 1$ if ν_π exists. By working with invariant probability measures for the induced stochastic kernels Q_π instead of strategic measures, we can derive an approximation result similar to Theorem 2 by invoking Theorem 1. However, to do that we need setwise convergence of invariant probability measures ν_π. The following results give sufficient conditions for the setwise convergence of invariant probability measure ν_π corresponding to stochastic kernels Q_π.

Proposition 5. *Let $\pi \in S$ be induced by f. Let $\{q_k\}$ and $\{\pi^k\}$ be as before and let Q_π and Q_{π^k} ($k \geq 1$) be the corresponding stochastic kernels as defined in (4). Assume that*

(i) *Q_{π^k} has an invariant probability measure ν_{π^k} for each k;*
(ii) *$\{\nu_{\pi^k}\}$ is sequentially relatively compact for the setwise topology.*

Then every setwise limit of the sequence $\{\nu_{\pi^k}\}$ is an invariant probability measure for Q_π. In particular, if Q_π has an unique invariant probability measure, then every convergent subsequence of $\{\nu_{\pi^k}\}$ converges to this invariant measure.

Proposition 6. *Let π, f, Q_π and $\{\pi^k\}$, $\{q_k\}$, $\{Q_{\pi^k}\}$ ($k \geq 1$) be as in the Proposition 5. Under the following assumptions*

(i) *Q_π, Q_{π^k} ($k \geq 1$) have an invariant probability measures ν_π, ν_{π^k} ($k \geq 1$),*
(ii) *For all $B \in \mathcal{B}(\mathsf{X})$, $Q_{\tilde{\pi}}^n(B|x) \to \nu_{\tilde{\pi}}(B)$ uniformly in $\tilde{\pi} \in \{\pi, \pi^1, \pi^2, \cdots\}$ for some x,*

we have $\nu_{\pi^k} \to \nu_\pi$ setwise.

Recall that M_π is the set of all initial points x such that the convergence in (5) holds. The following assumptions will be imposed in the next theorem.

(e) For any $\pi \in S$, Q_π has an unique invariant probability measure ν_π.
(f1) The set $\Gamma_S := \{\nu \in \mathcal{P}(\mathsf{X}) : \nu Q_\pi = \nu$ for some $\pi \in S\}$ is relatively sequentially compact in the setwise topology.
(f2) There exits $x \in \mathsf{X}$ such that for all $B \in \mathcal{B}(\mathsf{X})$, $Q_\pi^n(B|x) \to \nu_\pi(B)$ uniformly in $\pi \in S$.
(g) $\mathsf{M} := \bigcap_{\pi\in S} \mathsf{M}_\pi \neq \emptyset$.

Theorem 4. *Let the initial distribution μ be concentrated on some $x \in \mathsf{M}$. Then, for any $\varepsilon > 0$ there exists $\pi^* \in SQ$ such that $J(\pi^*, x) < \inf_{\tilde{\pi}\in S} J(\tilde{\pi}, x) + \varepsilon$ under assumptions (e), (f1) (or (f2)), and (g).*

Proof: This follows from (5), Propositions 5 and 6.

In the rest of this section we will derive conditions under which assumptions (e), (f1), (f2), (g) hold. In particular, we will consider an additive-noise system with Gaussian noise to find sufficient conditions under which assumptions (e), (f1), (f2) and (g) hold.

To begin with, assumptions (e), (f2) and (g) are satisfied under any of the conditions Ri, $i \in \{0, 1, 1(a), 1(b), 2, \ldots, 6\}$ in [26]. Moreover, $\mathsf{M} = \mathsf{X}$ in (g) at least one of the above conditions is hold. The next step is to find sufficient conditions for assumptions (e), (f1) and (g) to hold. The following gives a sufficient condition for sequential relative compactness in setwise topology.

Lemma 1. *If the set of probability measures Γ on X is majorized by a finite measure γ, then Γ sequentially relatively compact in the setwise topology.*

Proof: This follows from Prokhorov's theorem and [21, Theorem 1.5.5].

The following proposition gives a sufficient condition for the existence of an invariant probability measure for a stochastic kernel which is not necessarily Feller. It can be proved by modifying the proof of [27, Theorem 4.17].

Proposition 7. *Let Q be a stochastic kernel on X given X. If there exists $\tilde{x}$ in X such that the sequence $\{Q^n(\cdot|\tilde{x})\}$ is majorized by a finite measure γ, then Q has an invariant probability measure.*

Observe that the stochastic kernel p on X given $\mathsf{X}\times\mathsf{A}$ can be written as a measurable mapping from $\mathsf{X}\times\mathsf{A}$ to $\mathcal{P}(\mathsf{X})$ if $\mathcal{P}(\mathsf{X})$ is equipped with its Borel σ-algebra $\mathcal{M}(\mathsf{X})$, i.e.,

$$p(\cdot|x,a) : \mathsf{X}\times\mathsf{A} \to \mathcal{P}(\mathsf{X}).$$

We impose the following assumption

(e1) $p(\cdot|x,a) \le \gamma(\cdot)$ for all $x \in \mathsf{X}$, $a \in \mathsf{A}$ for some finite measure γ on X.

Fact 1. *Let $\pi \in S$ be induced by f and let Q_π be the corresponding stochastic kernel defined as $Q_\pi(\cdot|x) = p(\cdot|x, f(x))$. Under assumption (e1), $\{Q_\pi^n(\cdot|x)\}$ is majorized by γ for all x.*

Proposition 8. *Suppose (e1) holds. Then, for any $\pi \in S$, Q_π has an invariant probability measure ν_π which is majorized by γ. Hence, (e1) implies assumption (f1) by Lemma 1. In addition, if these invariant measures are unique, then assumptions (e) and (g) also hold with $\mathsf{M} = \mathsf{X}$ in (g).*

Example 1. Let us consider an additive-noise system with a Gaussian noise given by

$$x_{n+1} = F(x_n, a_n) + v_n, \;\; n = 0, 1, 2, \ldots$$

where $\mathsf{X} = \mathbb{R}^n$ and the v_n's are i.i.d. random vectors having a non-degenerate Gaussian distribution. For any $\pi \in S$, if Q_π has an invariant probability measure, then it has to be unique since $\{x_n\}$ is irreducible because the Gaussian noise has a density which is positive all X. Hence if this system satisfies assumption (e1), then assumptions (e), (f1) and (g) with $\mathsf{M} = \mathsf{X}$ hold by Proposition 8. It is not difficult to see that assumption (e1) holds if F has a bounded range. On the other hand, the boundedness of F also implies $R1(a)$ in [26] which further implies assumptions (e), (f2) and (g) with $\mathsf{M} = \mathsf{X}$ [26, Theorem 3.2]. Hence, if F is bounded, then (e), (f1), (f2) and (g) hold with $\mathsf{M} = \mathsf{X}$. This means that if F is bounded, then Theorem 4 holds for this system.

IV. Upper Bounds on the Approximation Error Based on the Rate of Quantization

In this section our aim is to find an upper bound in terms of the rates of quantizers used on how well stationary quantizer policies can approximate general non-randomized stationary policies. Recall that the rate of a quantizer q is defined as the logarithm of the cardinality of its range, i.e., $R := \log|q(\mathsf{X})|$. Let $\|\cdot\|_{TV}$ [21] denote the total variation distance between measures and let d_{A} denote the metric of the space A. We will use the following assumptions in this section:

(h) A is infinite compact subset of $\mathbb{R}^d$ for some $d \ge 1$.

(j) $|c(x,\tilde{a}) - c(x,a)| \le K_1 d_{\mathsf{A}}(\tilde{a}, a)$ for all x, and some $K_1 \ge 0$.

(k) $\|p(\cdot|x,\tilde{a}) - p(\cdot|x,a)\|_{TV} \le K_2 d_{\mathsf{A}}(\tilde{a}, a)$ for all x, and some $K_2 \ge 0$.

(l) For each non-randomized stationary policy π, the stochastic kernel $Q_\pi(dy|x)$ has a density $g_\pi(y|x)$ with respect to a σ-finite measure m on X, and there exists $\varepsilon > 0$ and $C \in \mathcal{B}(\mathsf{X})$ such that $m(C) > 0$ and

$$g_\pi(y|x) \ge \varepsilon \text{ for all } y \in C, x \in \mathsf{X}, \pi \in S.$$

Indeed, assumption (l) is the same as condition $R1(a)$ in [26] which was mentioned in Section III-B. However, we define it as a new assumption for the sake of completeness and clarity. In the following two section we will obtain bounds for the expected discounted cost and expected average cost criteria. Assumptions (h), (j) and (k) will be imposed for both cases, but (l) will only be assumed for the expected average cost case.

The following result holds since A is a compact (thus bounded) subset of a d-dimensional Euclidean space and so there exists a finite subset $C \subset \mathsf{A}$ with cardinality $|C| \le k$ such that $\max_{x\in\mathsf{A}} \min_{y\in C} d_{\mathsf{A}}(x,y) \le \lambda(1/k)^{1/d}$ for some $\lambda > 0$, where d_{A} is the Euclidean distance on $\mathbb{R}^d$.

Lemma 2. *Let $\mathsf{A} \subset \mathbb{R}^d$ be compact . Then for any measurable function $f : \mathsf{X} \to \mathsf{A}$ we can construct a sequence of quantizers $\{q_k\}$ from X to A such that $|q_k(\mathsf{X})| = k$ and $\sup_{x\in\mathsf{X}} d_{\mathsf{A}}(q_k(x), f(x)) \le \lambda(1/k)^{1/d}$ for some constant λ.*

In the rest of this section we are assuming that any non-randomized stationary policy π induced by f is approximated by a sequence $\{\pi^k\}$ of non-randomized stationary quantizer policies which are induced by a sequence $\{q_k\}$ of quantizers as in Lemma 2. By an abuse of notation, let $P_\mu^\pi(dx_n)$ denote the marginal distribution of the state x_n. The following proposition is the key result in this section.

Proposition 9. *Let $\pi \in S$ be induced by f. Let $\{q_k\}$ be as in the Lemma 2 inducing policies $\{\pi^k\}$. For any initial distribution μ we have*

$$\|P_\mu^\pi(dx_n) - P_\mu^{\pi^k}(dx_n)\|_{TV} \le \lambda K_2(2n-1)(1/k)^{1/d} \quad (6)$$

for all $n \ge 1$ under assumptions (h), (j) and (k).

A. Upper Bound for the Expected Discounted Cost Case

In this case, for any initial distribution μ and any $\pi \in S$, induced by f, the expected discounted cost can be written as

$$w_\beta(P_\mu^\pi) = \sum_{n=0}^{\infty} \beta^n \int_{\mathsf{X}} c(x_n, f(x_n)) P_\mu^\pi(dx_n). \quad (7)$$

We will write $w_\beta(\pi)$ instead of $w_\beta(P_\mu^\pi)$. The following theorem essentially follows from Proposition 9.

Theorem 5. *Let $\pi \in S$ induced by f. Let $\{q_k\}$ as in the Lemma 2 inducing policies $\{\pi^k\}$. For any initial distribution μ, we have*

$$|w_\beta(\pi) - w_\beta(\pi^k)| \leq K(1/k)^{1/d} \tag{8}$$

*where $K = \frac{\lambda}{1-\beta}(K_1 - \beta K_2 M + \frac{2\beta M K_2}{1-\beta})$ with $M := \sup_{(x,a)\in \mathsf{X}\times\mathsf{A}} |c(x,a)|$ under assumptions (h), (j) and (k). Hence, **(P2)** is true under assumptions (h), (j) and (k) for the expected discounted cost criterion.*

B. Upper Bound for the Expected Average Cost Case

For the expected average cost criterion we cannot apply the same method as for the discounted cost since the bound obtained there converges to infinity as β approaches 1. However, as in Section III-B we approach the problem by writing the expected average cost as an integral of the one stage cost function with respect to an invariant probability measure for the induced stochastic kernel. This way we obtain a bound on the difference between the actual and the approximated cost. However, the bound for this case will depend both on the rates of quantizers approximating the actual policy and an extra term which changes with the system parameters. However, as we show this extra term goes to zero as $n \to \infty$.

Lemma 3. *Suppose (c) and (l) hold. Then, for any $\pi \in S$ and $x \in \mathsf{X}$ we have*

$$J(\pi, x) = \int_{\mathsf{X}} c_\pi(x)\nu_\pi(dx) \tag{9}$$

where ν_π is the unique invariant probability measure for the induced stochastic kernel Q_π (see (4)).

Proof: By [26, Theorem 3.2], assumption (l) implies the *Uniform Ergodicity* property in [26, page 33]. The Uniform Ergodicity property implies the existence of a unique invariant probability measure ν_π for the Q_π and it also implies that

$$\frac{1}{N}\sum_{n=0}^{N-1} P_x^\pi(dx_n) \to \nu_\pi(\,\cdot\,) \text{ as } N \to \infty \text{ setwise} \tag{10}$$

for all $x \in \mathsf{X}$ and all $\pi \in S$. Since c is bounded by assumption, (10) implies (9).

Lemma 4. *Suppose assumption (l) holds. Then for any $\pi \in S$ and $x \in \mathsf{X}$, we have*

$$\|Q_\pi^n(\,\cdot\,|x) - \nu_\pi(\,\cdot\,)\|_{TV} \leq 2\left(\frac{(2-\varepsilon m(C))}{2}\right)^n$$

for all n where ν_π is the unique invariant probability measure in Lemma 3.

Proof: By assumption (l), for any $\pi \in S$ and $x \in \mathsf{X}$ we have

$$Q_\pi(\,\cdot\,|x) \geq \gamma(\,\cdot\,)$$

where the measure γ is defined as $\gamma(E) := \int_E \rho(x)m(dx)$ and $\rho(x) := \varepsilon I_C(x)$. Here $I_C(\,\cdot\,)$ is the indicator function of C. Clearly, $\gamma(\mathsf{X}) = \varepsilon m(C) > 0$. By Lemma 3.3 and its proof in [28, page 57] we can obtain the desired result. Lemmas 3 and 4 imply the following theorem.

Theorem 6. *Let $\pi \in S$ and let $\{\pi^k\} \in S\mathcal{Q}$ approximating π. Under assumptions (h), (j), (k) and (l) for any $x \in \mathsf{X}$ we have*

$$|w(\pi, \delta_x) - w(\pi^k, \delta_x)| \leq 4M\left(\frac{2-\varepsilon m(C)}{2}\right)^n + K_n(1/k)^{1/d} \tag{11}$$

for all $n \geq 0$ where $K_n = \big((2n-1)K_2\lambda M + K_1\lambda\big)$ and $M := \sup_{(x,a)\in\mathsf{X}\times\mathsf{A}} |c(x,a)|$.

Observe that depending on the values of ε and $m(C)$, we can first make the first term in (11) small enough by choosing sufficiently large n, and then for this n we can choose k large enough such that the second term in (11) is small. The following is an example of how to find ε and C in assumption (l).

Example 2. Consider the additive-noise system with Gaussian noise given by

$$x_{n+1} = F(x_n, a_n) + v_n, \ n = 0, 1, 2, \cdots$$

where $\mathsf{X} = \mathbb{R}$ and the v_n's are i.i.d. real valued Gaussian random variables with zero mean and variance σ^2. Assume F has a bounded range in $\mathbb{R}$, say, $F(\mathbb{R}) \subset [-L, L]$,where $L > 0$. Let m denote the Lebesgue measure on $\mathbb{R}$. For any $\pi \in S$ induced by f and for any $x \in \mathsf{X}$, $Q_\pi(\,\cdot\,|x)$ is absolutely continuous with respect to m with density $g_\pi(y|x) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-(y-F(x,f(x)))^2/2\sigma^2}$. Hence, assumption (l) holds with $\varepsilon = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-(2L)^2/2\sigma^2}$ and $C = [-L, L]$ for this system.

V. Conclusion

In this paper, we introduced stationary quantizer policies and showed under not too restrictive conditions that one can always find a non-randomized stationary quantizer policy which is ε optimal, in terms of the cost, in the set of all non-randomized stationary policies. We also found an upper bound on the error for approximating optimal policies in terms of the rates of the quantizers. Our continuity results also apply to the approximation of randomized stationary policies. A detailed discussion is given in [17].

References

[1] V. S. Borkar, "Convex analytic methods in markov decision processes," in *Handbook of Markov Decision Processes*, E. Feinberg and A. Shwartz, Eds. Kluwer Academic Publisher, 2002.

[2] O. Hernández-Lerma and J. Lasserre, "Weak conditions for average optimality in Markov control processes," *Systems Control Lett.*, vol. 22, pp. 287–291, 1994.

[3] A. Farahmand, R. Munos, and C. Szepesvari, "Error propagation for approximate policy and value iteration," *Advances in Neural Information Processing Systems*, 2010.

[4] L. Busoniu, D. Ernst, B. Schutter, and R. Babuska, "Approximate dynamic programming with a fuzzy parametrization," *Automatica*, vol. 46, pp. 804–814, 2010.

[5] D. Bertsekas and J. Tsitsiklis, *Neuro-Dynammic Programming*. Athena Scientific, 1996.

[6] J. Tsitsiklis and B. Roy, "Feature-based methods for large scale dynamic programming," *Machine Learning*, vol. 22, pp. 59–94, 1996.
[7] Z. Ren and B. Krogh, "State aggregation in Markov decision processes," in *CDC 2002*, Las Vegas, December 2002.
[8] C. Beck, S. Lall, T. Liang, and M. West, "Model reduction, optimal prediction, and the Mori-Zwanzig representation of Markoc chains," in *CDC 2009*, Shanghai, December 2009.
[9] R. Ortner, "Pseudometrics for state aggregation in average reward Markov decision processes," in *Algorithmic Learning Theory*. Springer-Verlag, 2007.
[10] B. Fox, "Finite-state approximations to denumerable state dynamic programs," *J. Math. Anal. Appl.*, vol. 34, pp. 665–670, 1971.
[11] D. White, "Finite-state approximations for denumerable state infinite horizon discounted Markov decision processes," *J. Math. Anal. Appl.*, vol. 74, pp. 292–295, 1980.
[12] ——, "Finite-state approximations for denumerable state infinite horizon discounted Markov decision processes with unbounded rewards," *J. Math. Anal. Appl.*, vol. 186, pp. 292–306, 1982.
[13] R. Cavazos-Cadena, "Finite-state approximations for denumerable state discounted Markov decision processes," *Appl. Math. Optim.*, vol. 14, pp. 1–26, 1986.
[14] O. Hernández-Lerma, "Finite-state approximations for denumerable multidimensional state discounted Markov decision processes," *J. Math. Anal. Appl.*, vol. 113, pp. 382–388, 1986.
[15] A. Leizarowitz and A. Shwartz, "Exact finite approximations of average-cost countable Markov decision processes," *Automatica*, vol. 44, pp. 1480–1487, 2008.
[16] T. Prieto-Rumeau and J. Lorenzo, "Approximating ergodic average reward continous-time controlled Markov chain," *IEEE Trans. Autom. Control*, vol. 55, no. 1, pp. 201–207, Jan. 2008.
[17] N. Saldi, T. Linder, and S. Yüksel, "Approximation of stationary control policies by quantized control in Markov decision processes," *Technical Report, Queen's University*.
[18] E. Feinberg, "On measurability and representation of strategic measures in Markov decision processes," *Statistics, Probability and Game Theory*, vol. 30, pp. 29–43, 1996.
[19] O. Hernández-Lerma and J. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer, 1996.
[20] E. Feinberg, "Controlled Markov processes with arbitrary numerical criteria," *Theory Prob. Appl.*, vol. 27, pp. 486–502, 1982.
[21] O. Hernández-Lerma and J. Lasserre, *Markov Chains and Invariant Probabilities*. Birkhauser, 2003.
[22] M. Schäl, "On dynamic programming: compactness of the space of policies," *Stochastic Process. Appl.*, vol. 3, no. 4, pp. 345–364, 1975.
[23] E. Balder, "On the compactness of the space of policies in stochastic dynamic programming," *Stochastic Process. Appl.*, vol. 32, no. 1, pp. 141–150, 1989.
[24] A. Nowak, "On the weak topology on a space of probability measures induced by policies," *Bull. Polish Acad. Sci. Math.*, vol. 36, pp. 181–186, 1988.
[25] R. Serfozo, "Convergence of Lebesgue integrals with varying measures," *Sankhya Ser.A*, pp. 380–402, 1982.
[26] 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.
[27] M. Hairer, "Ergodic properties of Markov processes," *Lecture Notes*, 2006.
[28] O. Hernández-Lerma, *Adaptive Markov Control Processes*. Springer-Verlag, 1989.