Infinite Horizon Average Cost Dynamic Programming subject to Ambiguity on Conditional Distribution

Ioannis Tzortzis, Charalambos D. Charalambous and Themistoklis Charalambous

***Abstract*— This paper addresses the optimality of stochastic control strategies based on the infinite horizon average cost criterion, subject to total variation distance ambiguity on the conditional distribution of the controlled process. This stochastic optimal control problem is formulated using minimax theory, in which the minimization is over the control strategies and the maximization is over the conditional distributions. Under the assumption that, for every stationary Markov control law the maximizing conditional distribution of the controlled process is irreducible, we derive a new dynamic programming recursion which minimizes the future ambiguity, and we propose a new policy iteration algorithm. The new dynamic programming recursion includes, in addition to the standard terms, the oscillator semi-norm of the cost-to-go. The maximizing conditional distribution is found via a water-filling algorithm. The implications of our results are demonstrated through an example.**

I. Introduction

The stochastic optimal control problem we consider presupposes (a) a discrete-time nominal Markov control model with deterministic strategies, and (b) an ambiguity set of the class of true controlled process conditional distributions described by a ball, with respect to the total variational distance metric centered at the nominal controlled process distribution, and (c) an infinite-horizon average cost criterion. The infinite horizon average cost criterion for a Markov control model, when the controlled process conditional distribution is known, is extensively studied in the literature (see, for example, [1]–[4] and references therein), to establish existence of optimal strategies and to derive necessary and sufficient optimality conditions. However, the resulting cost-to-go and the dynamic programming recursions depend on the conditional distribution of the underlying controlled process (see, e.g., [5]). Thus, any ambiguity in the controlled process conditional distribution will affect the optimality and robustness of the optimal decision strategies.

In this paper, we investigate the effects of the ambiguity in the controlled process conditional distribution, modeled by a ball with respect to the total variation distance metric, centered at a known nominal controlled conditional distribution with radius $R \in [0, 2]$, on the cost-to-go and dynamic programming. In summary, the issues discussed and results obtained in this paper are the following:

I. Tzortzis and C. D. Charalambous are with the Department of Electrical Engineering, University of Cyprus, Nicosia, Cyprus. E-mails: {tzortzis.ioannis,chadcha}@ucy.ac.cy.

T. Charalambous is with the Department of Signals and Systems, Chalmers University of Technology, Gothenburg, Sweden. E-mail: themistoklis.charalambous@chalmers.se.

(i) formulation of the infinite-horizon average cost criterion for a Markov control model and derivation of the dynamic programming recursion, under conditional distribution ambiguity on the controlled process distribution described by total variation distance, via minimax theory;

(ii) characterization of the maximizing conditional distribution of the controlled process belonging to the total variation distance set and the derivation of a new dynamic programming recursion that yields a new optimal policy iteration algorithm.

The paper is organized as follows. In Section II, we introduce the minimax optimization problem of interest, together with the maximization of a linear functional subject to total variation distance ambiguity, and other useful information for the development of the main results of this paper. In Section III, we derive a new dynamic programming recursion and the corresponding policy iteration algorithm. In Section IV, we present an example to illustrate the applications of our results. Section V concludes the paper with a discussion on the results obtained in this paper.

II. Preliminaries

An Markov Control Model (MCM) with deterministic strategies is a quintuple (5-tuple) given by

$$\Big(\mathcal{X}, \mathcal{U}, \{\mathcal{U}(x) : x \in \mathcal{X}\}, \{Q(dz|x,u) : (x,u) \in \mathcal{X} \times \mathcal{U}\}, f\Big) \quad (1)$$

that includes the *state space* $\mathcal{X}$, which models the state space of the controlled random process $\{x_k \in \mathcal{X} : k \in \mathbb{N}\}$, and the *control or action space* $\mathcal{U}$, which models the control or action set of the control random process $\{u_k \in \mathcal{U} : k \in \mathbb{N}\}$. The state space $\mathcal{X}$ and the control or action space $\mathcal{U}$ are both Polish spaces (complete separable metric spaces). For every $x \in \mathcal{X}$, $\mathcal{U}(x)$ is a family of non-empty measurable subsets of $\mathcal{U}$, denoting the *set of feasible controls or actions*, when the controlled process is in state $x \in \mathcal{X}$. The feasible state-actions pairs are defined by

$$\mathbb{K} \triangleq \{(x,u) : x \in \mathcal{X}, u \in \mathcal{U}(x)\} \quad (2)$$

and are measurable subsets of $\mathcal{X} \times \mathcal{U}$. Finally, the *controlled process distribution* $Q(dz|x,u)$ is a conditional distribution on $\mathcal{X}$ given $(x,u) \in \mathbb{K} \subseteq \mathcal{X} \times \mathcal{U}$, and the *one-stage-cost* $f : \mathbb{K} \mapsto [0, \infty]$ is a non-negative measurable function, such that $f(x, \cdot)$ does not take the value $+\infty$ for each $x \in \mathcal{X}$.

The spaces $\mathcal{X}$ and $\mathcal{U}$ are equipped with the natural σ-algebra $\mathcal{B}(\mathcal{X})$ and $\mathcal{B}(\mathcal{U})$, respectively. We use the following notation. Probability distributions on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ are denoted

by $\mathcal{M}_1(\mathcal{X})$, while the family of probability distributions $P(\cdot|y)$ on $(\mathcal{X},\mathcal{B}(\mathcal{X}))$ parametrized by $y \in \mathcal{Y}$, in which $(\mathcal{Y},\mathcal{B}(\mathcal{Y}))$ is another measurable space is described by the set of stochastic Kernels on $(\mathcal{X},\mathcal{B}(\mathcal{X}))$ conditioned on $(\mathcal{Y},\mathcal{B}(\mathcal{Y}))$, denoted by $\mathcal{Q}(\mathcal{X}|\mathcal{Y})$. Next, we give the definition of deterministic stationary Markov control policies.

Definition 2.1: A deterministic stationary Markov control policy is a measurable function $g : \mathcal{X} \mapsto \mathcal{U}$ such that $g(x_t) \in \mathcal{U}(x_t)$, $\forall x_t \in \mathcal{X}$. The set of all deterministic stationary Markov control policies is denoted by $G_{SM} \subset G$, where G denotes the set of all non-necessarily stationary control policies.

The infinite horizon average cost criterion when policy $g \in G$ is used, and given an initial state $x_0 = x$, is defined by

$$J(g) \triangleq \limsup_{t\to\infty} \frac{1}{t}\mathbb{E}_x^g\Big\{\sum_{k=0}^{t-1} f(x_k,u_k)\Big\} \tag{3}$$

The infinite horizon average cost stochastic optimal control problem is to choose a control policy $g^* \in G$ such that

$$J(g^*) = \inf_{g\in G} J(g).$$

In [5], it is shown, under appropriate conditions, that for all stationary Markov control laws $g \in G_{SM}$, there exists a solution $V : \mathcal{X} \mapsto \mathbb{R}$ and $J^* \in \mathbb{R}$ to the dynamic programming of the infinite horizon average cost stochastic optimal control problem satisfying

$$J^* + V(x) = \inf_{u\in\mathcal{U}(x)} \Big\{f(x,u) + \int_{\mathcal{X}} Q(dz|x,u)V(z)\Big\}. \tag{4}$$

In [6], [7], it is shown that if the state space $\mathcal{X}$ and the action space $\mathcal{U}$ are of finite cardinality, and for all stationary Markov control laws $g \in G_{SM}$ the transition probability matrix of the controlled process $Q(z|x,u)$ is irreducible (that is, all stationary policies have at most one recurrent class), then there exists a solution $V : \mathcal{X} \mapsto \mathbb{R}$ and $J^* \in \mathbb{R}$ to the dynamic programming of the infinite-horizon average cost stochastic optimal control problem satisfying

$$J^* + V(x) = \inf_{u\in\mathcal{U}(x)} \Big\{f(x,u) + \sum_{z\in\mathcal{X}} Q(z|x,u)V(z)\Big\}. \tag{5}$$

The optimality criterion (3) and the dynamic programming equations (5) and (4) are functionals of the conditional distribution of the controlled process, and hence the optimality and robustness of the optimal decision strategies depend on the complete knowledge of $Q(z|x,u)$, $Q(dz|x,u)$, respectively.

A. Variation Distance Ambiguity

Motivated by the above discussion, our objective is to investigate dynamic programming under ambiguity of the conditional distribution of the controlled process $\{Q(dz|x,u) : (x,u) \in \mathbb{K}\}$. The ambiguity of the conditional distribution of the controlled process is modeled by total variation distance, as follows.

Recall the total variation distance between two probability measures, $||\cdot||_{TV} : \mathcal{M}_1(\Sigma)\times\mathcal{M}_1(\Sigma) \longmapsto [0,\infty]$ defined by

$$||\alpha-\beta||_{TV} \triangleq \sup_{P\in\mathcal{P}(\Sigma)} \sum_{F_i\in P} |\alpha(F_i)-\beta(F_i)|, \quad \alpha,\beta \in \mathcal{M}_1(\Sigma)$$

where $\mathcal{M}_1(\Sigma)$ denotes the set of probability measures on the σ-algebra $\mathcal{B}(\Sigma)$ and $\mathcal{P}(\Sigma)$ denotes the collection of all finite partitions of Σ.

Next, we give the definitions of nominal controlled process distribution and the corresponding set of ambiguous controlled process distributions.

Definition 2.2: (Nominal Controlled Process Distribution) A nominal controlled state process $\{x_t^g : t = 0,1,\ldots, g \in G_{SM}\}$ corresponds to a sequence of stationary stochastic kernels as follows: for every $A \in \mathcal{B}(\mathcal{X})$

$$\text{Prob}(x_t \in A|x^{t-1},u^{t-1}) = Q^o(A|x_{t-1},u_{t-1})$$

where $Q^o(A|x_{t-1},u_{t-1}) \in \mathcal{Q}(\mathcal{X}|\mathbb{K})$, $t = 0,1,\ldots$.

The class of controlled processes is described by the sequence of stochastic kernels, $\{Q(dx_t|x_{t-1},u_{t-1}) \in \mathcal{Q}(\mathcal{X}|\mathbb{K}) : t = 0,1,\ldots\}$ belonging to a total variation distance set as follows.

Definition 2.3: (Class of Controlled Process Distributions) Given the nominal controlled process stochastic kernel of Definition 2.2, and $R \in [0,2]$, the class of stationary controlled process distributions is defined by

$$\mathbf{B}_R(Q^o)(x,u) \triangleq \Big\{Q(\cdot|x,u) \in \mathcal{M}_1(\mathcal{X}) : ||Q(\cdot|x,u) - Q^o(\cdot|x,u)||_{TV} \le R\Big\}, (x,u) \in \mathbb{K}.$$

The analogue of (3) is defined as follows.

Definition 2.4: (Pay-Off Functional) For every $g \in G$, the infinite-horizon average cost criterion for the class of controlled conditional distributions is defined by

$$J(g) = \limsup_{t\to\infty} \sup_{Q(\cdot|x,u)\in\mathbf{B}_R(Q^o)(x,u)} \frac{1}{t}\mathbb{E}_Q\Big\{\sum_{k=0}^{t-1} f(x_k,u_k)\Big\}$$

where $\mathbb{E}_Q\{\cdot\}$ denotes the expectation with respect to any conditional distribution $Q(\cdot|x_{t-1},u_{t-1}) \in \mathbf{B}_R(Q^o)(x,u)$, $t = 0,1,\ldots$ (e.g., it belongs to the total variation distance ball of Definition 2.3).

For the rest of the paper, the analysis is carried out under the assumption that $(\mathcal{X},\mathcal{U})$ are spaces with finite cardinality; the analysis when these are complete separable metrix spaces, is more invlolved and it is found in [8].

B. Maximization with Total Variation Distance Ambiguity

In this section, we recall results from [9] to aid the derivation of the maximizing probability distribution subject to total variation distance ambiguity of the controlled process.

Let Σ be a non-empty set of finite cardinality $|\Sigma|$ endowed with the discrete topology, with $\mathcal{M}_1(\Sigma)$ identified with the standard probability simplex in $\mathbb{R}^{|\Sigma|}$, i.e., the set of all $|\Sigma|$-dimensional vectors which are probability vectors, $\{\nu(x) : x \in \Sigma\} \in \mathcal{M}_1(\Sigma)$, $\{\mu(x) : x \in \Sigma\} \in \mathcal{M}_1(\Sigma)$. Also, let $\ell \triangleq \{\ell(x) : x \in \Sigma\} \in \mathbb{R}_+^{|\Sigma|}$. Define the maximum and minimum values of $\{\ell(x) : x \in \Sigma\}$ by

$$\ell_{\max} \triangleq \max_{x\in\Sigma} \ell(x), \quad \ell_{\min} \triangleq \min_{x\in\Sigma} \ell(x)$$

and their corresponding support sets by

$$\Sigma^0 \triangleq \big\{x \in \Sigma : \ell(x)=\ell_{\max}\big\}, \ \Sigma_0 \triangleq \big\{x \in \Sigma : \ell(x)=\ell_{\min}\big\}.$$

For all remaining sequence, $\{\ell(x) : x \in \Sigma \setminus \Sigma^0 \cup \Sigma_0\}$, define recursively the set of indices for which the sequence achieves its $(k+1)$th smallest value by

$$\Sigma_k \triangleq \Big\{x \in \Sigma : \ell(x) = \min\big\{\ell(\alpha) : \alpha \in \Sigma \setminus \Sigma^0 \cup (\bigcup_{j=1}^{k} \Sigma_{j-1})\big\}\Big\}$$

and the corresponding values of the sequence on Σ_k sets by

$$\ell(\Sigma_k) \triangleq \min_{x \in \Sigma \setminus \Sigma^0 \cup (\bigcup_{j=1}^{k} \Sigma_{j-1})} \ell(x),$$

where $k \in \{1, 2, \ldots, r\}$, and r is the number of Σ_k sets which is at most $|\Sigma \setminus \Sigma^0 \cup \Sigma_0|$, i.e., $1 \le r \le |\Sigma \setminus \Sigma^0 \cup \Sigma_0|$. In [9] it is shown that the maximum pay-off subject to total variation constraint is given by

$$L(\nu^*) = \ell_{\max}\nu^*(\Sigma^0) + \ell_{\min}\nu^*(\Sigma_0) + \sum_{k=1}^{r} \ell(\Sigma_k)\nu^*(\Sigma_k), \quad (6)$$

and that the optimal probabilities are given by the following equations (involving water-filling).

$$\nu^*(\Sigma^0) = \mu(\Sigma^0) + \frac{\alpha}{2} \quad (7a)$$

$$\nu^*(\Sigma_0) = \Big(\mu(\Sigma_0) - \frac{\alpha}{2}\Big)^+ \quad (7b)$$

$$\nu^*(\Sigma_k) = \Big(\mu(\Sigma_k) - \Big(\frac{\alpha}{2} - \sum_{j=1}^{k} \mu(\Sigma_{j-1})\Big)^+\Big)^+ \quad (7c)$$

$$\alpha = \min\Big(R, 2(1 - \mu(\Sigma^0))\Big) \quad (7d)$$

where $R \in [0, 2]$ and $k \in \{1, 2, \ldots, r\}$.

In the next section, we apply the above solution of to the dynamic programming recursion under ambiguity on the conditional distribution.

III. Minimax Stochastic Control

In this section, we study the infinite horizon average cost Markov Control Model and we derive the new dynamic programming recursion, with respect to total variation distance ambiguity, under an irreducibility assumption on the conditional distribution of the controlled process. In addition, we propose a new policy iteration algorithm.

A. Optimality of Control Policies for Finite State and Control Spaces

Consider the problem of minimizing the average cost criterion

$$J(g) = \limsup_{j\to\infty} \max_{Q(\cdot|x,u) \in \mathbf{B}_R(Q^o)(x,u)} \frac{1}{j}\mathbb{E}_Q\Big\{\sum_{k=0}^{j-1} f(x_k, u_k)\Big\}.$$

For the finite horizon optimal stochastic control problem with pay-off

$$\max_{Q(\cdot|x,u) \in \mathbf{B}_R(Q^o)(x,u)} \mathbb{E}_Q\Big\{\sum_{k=0}^{n-1} f(x_k, u_k)\Big\} \quad (8)$$

we show in [10] that the value function satisfies the dynamic programming equation

$$V_j(x) = \min_{u\in\mathcal{U}(x)} \max_{Q(\cdot|x,u)\in\mathbf{B}_R(Q^o)(x,u)} \Big\{f(x,u) + \sum_{z\in\mathcal{X}} V_{j+1}(z)Q(z|x,u)\Big\} \quad (9)$$

which is equivalent to (since Q^* exists and is given by a variation of (7))

$$\begin{aligned} V_j(x) &= \min_{u\in\mathcal{U}(x)} \Big\{f(x,u) + \sum_{z\in\mathcal{X}} V_{j+1}(z)Q^*(z|x,u)\Big\} \\ &= \min_{u\in\mathcal{U}(x)} \Big\{f(x,u) + \sum_{z\in\mathcal{X}} V_{j+1}(z)Q^o(z|x,u) \\ &\quad + \frac{R}{2}\Big(\max_{z\in\mathcal{X}} V_{j+1}(z) - \min_{z\in\mathcal{X}} V_{j+1}(z)\Big)\Big\}. \end{aligned} \quad (10)$$

Define $\overline{V}_j(x) = V_{n-j}(x)$. Then $\overline{V}_j$ satisfies the equation

$$\overline{V}_j(x) = \min_{u\in\mathcal{U}(x)} \max_{Q(\cdot|x,u)\in\mathbf{B}_R(Q^o)(x,u)} \Big\{f(x,u) + \sum_{z\in\mathcal{X}} \overline{V}_{j-1}(z)Q(z|x,u)\Big\}.$$

which can be rewritten as follows

$$\overline{V}_j(x) + \frac{1}{j}\overline{V}_j(x) = \min_{u\in\mathcal{U}(x)} \max_{Q(\cdot|x,u)\in\mathbf{B}_R(Q^o)(x,u)} \Big\{f(x,u) + \sum_{z\in\mathcal{X}} \big(\overline{V}_{j-1}(z) + \frac{1}{j}\overline{V}_j(x)\big)Q(z|x,u)\Big\}. \quad (11)$$

Assumption 3.1: Assume that there exists a V and a $J^* \in \mathbb{R}$ such that

$$\lim_{j\to\infty} \big(\overline{V}_j(x) - jJ^*\big) = V(x), \quad \forall x \in \mathcal{X}. \quad (12)$$

Under Assumption 3.1, then

$$\lim_{j\to\infty} \frac{1}{j}\overline{V}_j(x) = J^*, \quad \forall x \in \mathcal{X} \quad (13)$$

and the limit does not depend on $x \in \mathcal{X}$. In addition, taking maximization with respect to $x \in \mathcal{X}$ on both sides of (12) (using the finite cardinality of $\mathcal{X}$ which enables us to exchange the limit and max) then

$$\lim_{j\to\infty} \max_{x\in\mathcal{X}} \big(\overline{V}_j(x) - jJ^*\big) = \max_{x\in\mathcal{X}} V(x). \quad (14)$$

By Assumption 3.1 and (13), we have the following identi-

ties.

$$
\begin{aligned}
&J^* + V(x) \\
&= \lim_{j\to\infty} \Big(\frac{1}{j}\overline{V}_j(x) + (\overline{V}_j(x) - jJ^*) \Big) \\
&\overset{(a)}{=} \lim_{j\to\infty} \min_{u\in\mathcal{U}(x)} \max_{Q(\cdot|x,u)\in\mathbf{B}_R(Q^o)(x,u)} \\
&\quad \Big\{ f(x,u) + \sum_{z\in\mathcal{X}} \big(\overline{V}_{j-1}(z) + \frac{1}{j}\overline{V}_j(x)\big) Q(z|x,u) - jJ^* \Big\} \\
&\overset{(b)}{=} \lim_{j\to\infty} \min_{u\in\mathcal{U}(x)} \\
&\quad \Big\{ f(x,u) - jJ^* + \sum_{z\in\mathcal{X}} \big(\overline{V}_{j-1}(z) + \frac{1}{j}\overline{V}_j(x)\big) Q^o(z|x,u) \\
&\quad + \frac{R}{2}\Big(\max_{z\in\mathcal{X}} \big(\overline{V}_{j-1}(z) + \frac{1}{j}\overline{V}_j(x)\big) - \min_{z\in\mathcal{X}} \big(\overline{V}_{j-1}(z) + \frac{1}{j}\overline{V}_j(x)\big) \Big) \Big\} \\
&\overset{(c)}{=} \lim_{j\to\infty} \min_{u\in\mathcal{U}(x)} \Big\{ f(x,u) \\
&\quad + \sum_{z\in\mathcal{X}} \big(\overline{V}_{j-1}(z) - (j-1)J^* + \frac{1}{j}\overline{V}_j(x) - J^*\big) Q^o(z|x,u) \\
&\quad + \frac{R}{2}\Big(\max_{z\in\mathcal{X}} \big(\overline{V}_{j-1}(z) - jJ^*\big) - \min_{z\in\mathcal{X}} \big(\overline{V}_{j-1}(z) - jJ^*\big) \Big) \Big\}
\end{aligned}
$$

where
(a) is obtained by (11),
(b) is obtained by the equivalent formulation (10), and
(c) by adding and subtracting $J^*(1 + j\frac{R}{2})$.
By the finite cardinality of $\mathcal{X}$ and $\mathcal{U}$, then

$$
\begin{aligned}
J^* + V(x) = \min_{u\in\mathcal{U}(x)} \Big\{ & f(x,u) \\
& + \sum_{z\in\mathcal{X}} V(z) Q^o(z|x,u) + \frac{R}{2}\big(\max_{z\in\mathcal{X}} V(z) - \min_{z\in\mathcal{X}} V(z) \big) \Big\}
\end{aligned} \tag{15}
$$

which is equivalent to

$$
\begin{aligned}
J^* + V(x) = \min_{u\in\mathcal{U}(x)} \max_{Q(\cdot|x,u)\in\mathbf{B}_R(Q^o)(x,u)} \\
\Big\{ f(x,u) + \sum_{z\in\mathcal{X}} V(z) Q(z|x,u) \Big\}.
\end{aligned} \tag{16}
$$

Equation (15) (or, equivalently, (16)) is the dynamic programming equation for average cost minimax problem, under the assumption that (12) and (14) are satisfied. It can be shown that, J^* satisfying (16) is the minimal average payoff, and further that if g^* is a stationary policy such that $g^*(x)$ achieves the minimum on the right-hand side of (16) for every $x \in \mathcal{X}$, then g^* is optimal. Next, we derive more general conditions under which (15) is valid.

B. Existence

Dynamic programming equation (15) is valid under the assumption that (12) and (14) are satisfied. Here, we derive more general conditions under which (15) is valid. First, we introduce some notation.

Let the state space $\mathcal{X}$ be a finite set, with say, n elements. Then, any function $V : \mathcal{X} \longrightarrow \mathbb{R}^n$ may be represented by a vector in $\mathbb{R}^n$. Any stationary control law $g \in G_{SM}$, $g : \mathcal{X} \longmapsto \mathbb{R}$, may also be identified with a $g \in \mathbb{R}^n$. For any g, let $Q(g) \in \mathbb{R}^{n\times n}$ and

$$
f(g) = \begin{pmatrix} f(x_1, g(x_1)) & \cdots & f(x_n, g(x_n)) \end{pmatrix}^T \in \mathbb{R}^n.
$$

Let $q_0 \in \mathbb{R}^n$ be defined by $q_0(i) \triangleq Q(\{x_0 = i\})$ and $e \triangleq (1, \cdots, 1)^T \in \mathbb{R}^n$. The maximum expected cost on a finite horizon is then

$$
\max_{Q(\cdot|x,u)\in\mathbf{B}_R(Q^o)(x,u)} \mathbb{E}_Q \Big\{ \sum_{k=0}^{j-1} f(x_k, u_k) \Big\} \tag{17a}
$$

$$
= \max_{Q(\cdot|x,u)\in\mathbf{B}_R(Q^o)(x,u)} q_0^T \Big\{ \sum_{k=0}^{j-1} Q(g)^k \Big\} f(g) \tag{17b}
$$

$$
= q_0^T \Big\{ \sum_{k=0}^{j-1} Q^*(g)^k \Big\} f(g) \tag{17c}
$$

where $Q^*(\cdot|x,u)$ denotes the maximizing conditional distribution of (17b). The maximum average cost criterion is then

$$
J(g) = \limsup_{j\to\infty} \frac{1}{j} q_0^T \Big\{ \sum_{k=0}^{j-1} Q^*(g)^k \Big\} f(g).
$$

Since $q_0 \in \mathbb{R}^n$ and $f(g) \in \mathbb{R}^n$ are independent of j, we only need to investigate the conditions under which the limit of the term inside the brackets exists.

Proposition 3.2: [6], [7] Let g be a stationary Markov control law which defines $g : \mathcal{X} \longmapsto \mathcal{U}$. Assume that the maximizing conditional distribution $Q^*(g) \in \mathbb{R}_+^{n\times n}$ in (17) is irreducible. Then,

(a) there exists a unique $q(g) \in \mathbb{R}_+^n$ such that

$$
Q^*(g)q(g) = q(g), \;\; e^T g = 1, \;\; e = (1, \ldots, 1)^T; \tag{18}
$$

(b) the average cost associated with the control law $g \in G_{SM}$ is

$$
J(g) = q(g)^T f(g); \tag{19}
$$

(c) there exists a $V(g) \in \mathbb{R}^n$ such that

$$
J(g)e + V(g) = f(g) + Q^*(g)V(g). \tag{20}
$$

Consequently, it can be shown that if for any stationary control law $g \in G_{SM}$, $Q^*(g) \in \mathbb{R}_+^{n\times n}$ is irreducible, and if there exists a stationary Markov control law $g \in G_{SM}$ such that

$$
J(g^*) = \min_{g\in G_{SM}} J(g)
$$

then there exists a pair $(V(g^*,\cdot), J(g^*))$, $V(g^*,\cdot) : \mathcal{X} \longmapsto \mathbb{R}$ and $J(g) \in \mathbb{R}$ that is a solution to the dynamic programming equation

$$
\begin{aligned}
&J(g^*) + V(g^*, x) \\
&= \min_{u\in\mathcal{U}} \max_{Q(\cdot|x,u)\in\mathbf{B}_R(Q^o)(x,u)} \Big\{ f(x,u) + \sum_{z\in\mathcal{X}} V(g^*, z) Q(z|x,u) \Big\} \\
&= \min_{u\in\mathcal{U}} \Big\{ f(x,u) + \sum_{z\in\mathcal{X}} V(g^*, z) Q^*(z|x,u) \Big\}.
\end{aligned}
$$

Then we obtain the following theorem.

Theorem 3.3: Assume that for all stationary Markov control laws $g \in G_{SM}$, and for a given total variation parameter

R, the maximizing transition matrix $Q^*(g)$ of (17c) is irreducible.

(a) There exists a solution $V : \mathcal{X} \longmapsto \mathbb{R}$ and $J^* \in \mathbb{R}$ to the dynamic programming equation

$$J^* + V(x) = \min_{u\in\mathcal{U}} \Big\{ f(x,u) + \sum_{z\in\mathcal{X}} V(z) Q^*(z|x,u) \Big\}. \quad (21)$$

The maximizing conditional distribution is

$$Q^*(\mathcal{X}^+|x,u) = Q^o(\mathcal{X}^+|x,u) + \frac{R}{2} \in [0,1] \quad (22)$$

$$Q^*(\mathcal{X}^-|x,u) = Q^o(\mathcal{X}^-|x,u) - \frac{R}{2} \in [0,1] \quad (23)$$

$$Q^*(A|x,u) = Q^o(A|x,u), \quad \forall A \subseteq \mathcal{X} \backslash \mathcal{X}^+ \cup \mathcal{X}^- \quad (24)$$

where

$$\mathcal{X}^+ \triangleq \{x \in \mathcal{X} : V(x) = \max\{V(x) : x \in \mathcal{X}\}\} \quad (25)$$

$$\mathcal{X}^- \triangleq \{x \in \mathcal{X} : V(x) = \min\{V(x) : x \in \mathcal{X}\}\}. \quad (26)$$

Moreover,

$$J^* + V(x) = \min_{u\in\mathcal{U}} \Big\{ f(x,u) \quad (27)$$

$$+ \sum_{z\in\mathcal{X}} V(z) Q^o(z|x,u) + \frac{R}{2} \big(\max_{z\in\mathcal{X}} V(z) - \min_{z\in\mathcal{X}} V(z) \big) \Big\}.$$

(b) If $g^*(x)$ attains the minimum in (21) for every x, then g^* is an optimal policy.

1) The minimum cost is J^*.

Proof: Theorem 3.3 is obtained by combining the results of Sections II-B, III-A and III-B. ■

The main observation is that in specific applications one may employ either dynamic programming equation (21) or (27). Part (a) of Theorem 3.3, shows that for a stationary Markov control policy and for an irreducible stochastic matrix Q^* there exists a solution to the dynamic programming equation (21). Specifically, the maximizing stochastic matrix Q^* which is given by (22)-(24), is calculated based on the support sets (25)-(26), the nominal stochastic matrix Q^o, and the value of the total variation parameter R. However, due to the water-filling behavior of the solution it turns out that, as the value of total variation parameter increases, the maximizing stochastic matrix Q^*, eventually, will be transformed into a reducible stochastic matrix (see (22)-(24)). Hence, our proposed method for solving minimax stochastic control problem with average cost is valid for values of total variation parameter $R \in [0, R_{\max}] \subseteq [0,2]$, and $R_{\max}$ is strictly less than 2.

Remark 3.4: For values of $R \in [R_{\max}, 2]$, for which irreducibility condition of the maximizing conditional distribution of the controlled process is violated, dynamic programming equation (21) may not be sufficient to give the optimal policy and the minimum cost. In particular, if irreducibility condition is not satisfied then (21) need not have a unique solution. To overcome this limitation and to characterize the average optimality criterion for any ball of radius $R \in [0,2]$, one may employ a pair of dynamic programming equations, also known as multichain dynamic programming equations [5], [11]. This generalization is addressed in [8].

1) Policy Iteration Algorithm: In this section, we provide a modified version of the classical policy iteration algorithm for average cost dynamic programming (i.e., see [6], [7]), in which policy evaluation and policy improvement steps must be performed using the maximizing conditional distribution obtained under total variation distance ambiguity constraint.[1]

Algorithm 1 Policy iteration algorithm

1) Let $m = 0$ and select an arbitrary stationary Markov control law $g_0 : \mathcal{X} \longmapsto \mathcal{U}$.
2) (Policy Evaluation) Solve the equation

$$J_{Q^o}(g_m)e + V_{Q^o}(g_m) = f(g_m) + Q^o(g_m)V_{Q^o}(g_m) \quad (28)$$

for $J_{Q^o}(g_m) \in \mathbb{R}$ and $V_{Q^o}(g_m) \in \mathbb{R}^n$. Identify the support sets of (28) using (25)-(26), and construct the matrix $Q^*(g_m)$ using (22)-(24). Solve the equation

$$J_{Q^*}(g_m)e + V_{Q^*}(g_m) = f(g_m) + Q^*(g_m)V_{Q^*}(g_m) \quad (29)$$

for $J_{Q^*}(g_m) \in \mathbb{R}$ and $V_{Q^*}(g_m) \in \mathbb{R}^n$.
3) (Policy Improvement) Let

$$g_{m+1} = \arg\min_{g\in\mathbb{R}^n} \Big\{ f(g) + Q^*(g)V_{Q^*}(g_m) \Big\}. \quad (30)$$

4) If $g_{m+1} = g_m$, let $g^* = g_m$; else let $m = m+1$ and return to step 2.

In the next section, we illustrate the new dynamic programming equation and the corresponding policy iteration algorithm through an example.

IV. Example

In this section, we illustrate an application of the infinite horizon minimax problem for average cost criterion, by considering a stochastic control system with state space $\mathcal{X} = \{1,2,3\}$ and control set $\mathcal{U} = \{u_1, u_2\}$. Assume that the nominal transition probabilities under controls u_1 and u_2 are given by

$$Q^o(u_1) = \frac{1}{9}\begin{pmatrix} 3 & 1 & 5 \\ 4 & 2 & 3 \\ 1 & 6 & 2 \end{pmatrix}, Q^o(u_2) = \frac{1}{9}\begin{pmatrix} 1 & 2 & 6 \\ 4 & 2 & 3 \\ 4 & 1 & 4 \end{pmatrix} \quad (31)$$

the total variation distance radius is $R = 6/9$, and the cost function under each state and action is $f(1,u_1) = 2$, $f(2,u_1) = 1$, $f(3,u_1) = 3$, $f(1,u_2) = 0.5$, $f(2,u_2) = 3$ and $f(3,u_2) = 0$. To obtain an optimal stationary policy of the infinite horizon minimax problem for average cost, policy iteration Algorithm 1 is applied.

A. Let $m = 0$.

1) Select the initial policies as follows $g_0(1) = u_1$, $g_0(2) = u_2$, $g_0(3) = u_2$.

2) Solve the equation $J_{Q^o}(g_0)e + V_{Q^o}(g_0) = f(g_0) + Q^o(g_0)V_{Q^o}(g_0)$ for $J_{Q^o}(g_0) \in \mathbb{R}$ and $V_{Q^o}(g_0) \in \mathbb{R}^3$. The

[1] This follows by Theorem 3.3, part (a).

$$Q^*(u)=\begin{pmatrix} \left(q^o_{11}(u)-\left(\frac{R}{2}-q^o_{13}(u)\right)^+\right)^+ & \min\left(1,q^o_{12}(u)+\frac{R}{2}\right) & \left(q^o_{13}(u)-\frac{R}{2}\right)^+ \\ \left(q^o_{21}(u)-\left(\frac{R}{2}-q^o_{23}(u)\right)^+\right)^+ & \min\left(1,q^o_{22}(u)+\frac{R}{2}\right) & \left(q^o_{23}(u)-\frac{R}{2}\right)^+ \\ \left(q^o_{31}(u)-\left(\frac{R}{2}-q^o_{33}(u)\right)^+\right)^+ & \min\left(1,q^o_{32}(u)+\frac{R}{2}\right) & \left(q^o_{33}(u)-\frac{R}{2}\right)^+ \end{pmatrix} \tag{32}$$

optimality equations (28) are given by

$$J_{Q^o}(g_0)+\frac{6}{9}V_{Q^o}(g_0,1)=2+\frac{1}{9}V_{Q^o}(g_0,2)+\frac{5}{9}V_{Q^o}(g_0,3) \tag{33a}$$
$$J_{Q^o}(g_0)+\frac{7}{9}V_{Q^o}(g_0,2)=3+\frac{4}{9}V_{Q^o}(g_0,1)+\frac{3}{9}V_{Q^o}(g_0,3) \tag{33b}$$
$$J_{Q^o}(g_0)+\frac{5}{9}V_{Q^o}(g_0,3)=\frac{4}{9}V_{Q^o}(g_0,1)+\frac{1}{9}V_{Q^o}(g_0,2). \tag{33c}$$

Since $V_{Q^o}(g_0)$ is uniquely determined up to an additive constant, let $V_{Q^o}(g_0,3)=0$. The solution is

$$V_{Q^o}(g_0,1)=1.8,\ V_{Q^o}(g_0,2)=3.375,\ V_{Q^o}(g_0,3)=0$$
$$J_{Q^o}(g_0)=1.175.$$

Note that, $V_{Q^o}\triangleq\{V_{Q^o}(1),V_{Q^o}(2),V_{Q^o}(3)\}$, and hence by (25)-(26), the support sets based on the values of V_{Q^o} are $\mathcal{X}^+=\{2\}$, $\mathcal{X}^-=\{3\}$ and $\mathcal{X}_1=\{1\}$. Once the partition is been identified, (22)-(24) are applied to obtain (32). Hence,

$$Q^*(u_1)=\frac{1}{9}\begin{pmatrix}3&4&2\\4&5&0\\0&9&0\end{pmatrix}, Q^*(u_2)=\frac{1}{9}\begin{pmatrix}1&5&3\\4&5&0\\4&4&1\end{pmatrix} \tag{34}$$

Note that, since every state can reach every other state, matrix $Q^*(u)$ remains irreducible under both controls.

Next, solve the equation $J_{Q^*}(g_0)e+V_{Q^*}(g_0)=f(g_0)+Q^*(g_0)V_{Q^*}(g_0)$ for $J_{Q^*}(g_0)\in\mathbb{R}$ and $V_{Q^*}(g_0)\in\mathbb{R}^3$. The optimality equations (29) are given by

$$J_{Q^*}(g_0)+\frac{6}{9}V_{Q^*}(g_0,1)=2+\frac{4}{9}V_{Q^*}(g_0,2)+\frac{2}{9}V_{Q^*}(g_0,3) \tag{35a}$$
$$J_{Q^*}(g_0)+\frac{4}{9}V_{Q^*}(g_0,2)=3+\frac{4}{9}V_{Q^*}(g_0,1) \tag{35b}$$
$$J_{Q^*}(g_0)+\frac{8}{9}V_{Q^*}(g_0,3)=\frac{4}{9}V_{Q^*}(g_0,1)+\frac{4}{9}V_{Q^*}(g_0,2). \tag{35c}$$

Since $V_{Q^*}(g_0)$ is uniquely determined up to an additive constant, let $V_{Q^*}(g_0,3)=0$. The solution is

$$V_{Q^*}(g_0,1)=1.8,\ V_{Q^*}(g_0,2)=3.375,\ V_{Q^*}(g_0,3)=0$$
$$J_{Q^*}(g_0)=2.3.$$

3) Let $g_1=\text{argmin}_{g\in\mathbb{R}^3}\{f(g)+Q^*(g)V_{Q^*}(g_0)\}$, then the resulting optimal control laws are $g_1(1)=u_2$, $g_1(2)=u_1$ and $g_1(3)=u_2$. Since, $g_1\neq g_0$, let $m=1$ and return to step 2.

B. Let $m=1$.

2) Following the same procedure as in step.1 the solution of $J_{Q^o}(g_1)e+V_{Q^o}(g_1)=f(g_1)+Q^o(g_1)V_{Q^o}(g_1)$, for $J_{Q^o}(g_1)\in\mathbb{R}$ and $V_{Q^o}(g_1)\in\mathbb{R}^3$ with $V_{Q^o}(g_1,3)=0$ is

$$V_{Q^o}(g_1,1)=0.468,\ V_{Q^o}(g_1,2)=1.125,\ V_{Q^o}(g_1,3)=0$$
$$J_{Q^o}(g_1)=0.333.$$

Next, we proceed with the identification of the support sets, which are $\mathcal{X}^+=\{2\}$, $\mathcal{X}^-=\{3\}$ and $\mathcal{X}_1=\{1\}$. Since the partition is the same as in $m=0$ then $Q^*(u_1)$ and $Q^*(u_2)$ are given by (34).

Solving equation $J_{Q^*}(g_1)e+V_{Q^*}(g_1)=f(g_1)+Q^*(g_1)V_{Q^*}(g_1)$, for $J_{Q^*}(g_1)\in\mathbb{R}$ and $V_{Q^*}(g_1)\in\mathbb{R}^3$, with $V_{Q^*}(g_1,3)=0$, the solution is

$$V_{Q^*}(g_1,1)=0.468,\ V_{Q^*}(g_1,2)=1.125,\ V_{Q^*}(g_1,3)=0$$
$$J_{Q^*}(g_1)=0.708.$$

3) Let $g_2=\text{argmin}_{g\in\mathbb{R}^3}\{f(g)+Q^*(g)V_{Q^*}(g_1)\}$, then the resulting control laws are $g_2(1)=u_2$, $g_2(2)=u_1$ and $g_2(3)=u_2$.

4) Since, $g_2=g_1$, then $g^*=g_1$ is an optimal control law with $J_{Q^*}=0.708$, $V_{Q^*}(1)=0.468$, $V_{Q^*}(2)=1.125$ and $V_{Q^*}(3)=0$.

V. Conclusions

In this paper, we examined the optimality of stochastic control strategies via dynamic programming on an infinite horizon, when the ambiguity class is described by a ball with respect to the total variation distance. As optimality criterion we considered the average cost criterion, and we introduced a new dynamic programming recursion which minimize the future ambiguity with respect to total variation distance, and we derived a new policy iteration algorithm which is performed using the maximizing conditional distribution of the controlled process.

References

[1] A. Arapostathis, V. S. Borkar, E. Fernandez-Gaucherand, M. K. Ghosh, and S. I. Marcus, "Discrete-time controlled Markov processes with average cost criterion: a survey," *SIAM J. Control Optim.*, vol. 31, no. 2, pp. 282–344, 1993.

[2] V. S. Borkar, "On minimum cost per unit time control of Markov chains," *SIAM J. Control Optim.*, vol. 22, no. 6, pp. 965–978, 1984.

[3] ——, "Control of Markov chains with long-run average cost criterion: the dynamic programming equations," *SIAM J. Control Optim.*, vol. 27, no. 3, pp. 642–657, 1989.

[4] L. I. Sennott, "Another set of conditions for average optimality in Markov control processes," *Systems and Control Letters*, vol. 24, no. 2, pp. 147–151, 1995.

[5] O. Hernandez-Lerma and J. B. Lasserre, *Discrete-time Markov control processes: Basic optimality criteria*, ser. Applications of Mathematics Stochastic Modelling and Applied Probability. Springer Verlag, 1996, no. v. 1.

[6] P. R. Kumar and P. Varaiya, *Stochastic systems: Estimation, identification, and adaptive control*. Prentice Hall, 1986.

[7] J. H. Van Schuppen, *Mathematical control and system theory of discrete-time stochastic systems*. Preprint, 2014.

[8] I. Tzortzis, C. D. Charalambous, and T. Charalambous, "Infinite horizon average cost dynamic programming subject to total variation distance ambiguity," *SIAM J. Control Optim.*, 2015 (submitted).

[9] C. D. Charalambous, I. Tzortzis, S. Loyka, and T. Charalambous, "Extremum problems with total variation distance and their applications," *IEEE Trans. Autom. Control*, vol. 59, no. 9, pp. 2353–2368, Sep. 2014.

[10] I. Tzortzis, C. D. Charalambous, and T. Charalambous, "Dynamic programming subject to total variation distance ambiguity," *SIAM J. Control Optim.*, vol. 53, no. 4, pp. 2040–2075, July 2015.

[11] M. L. Puterman, *Markov decision Processes*. New York: Wiley, 1994.