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

Ioannis Tzortzis, Charalampos 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 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) 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

\[
(\mathcal{X}, \mathcal{U}, \{\mathcal{U}(x) : x \in \mathcal{X}\}, \{Q(dz|x,u):(x,u)\in\mathcal{X}\times\mathcal{U}\}, f) \tag{1}
\]

that includes the state space \( \mathcal{X} \), which models the state space of the controlled random process \( \{x_k : x_k \in \mathcal{X} : k \in \mathbb{N}\} \), 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

\[
\mathcal{K} \triangleq \{(x,u) : x \in \mathcal{X}, u \in \mathcal{U}(x)\} \tag{2}
\]

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 \( (x,u) \in \mathcal{K} \subseteq \mathcal{X} \times \mathcal{U} \), and the one-stage-cost function \( f : \mathcal{K} \to [0, \infty] \) 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 \( \sigma \)-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 $\mathcal{P}(\cdot|y)$ 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 $g : \mathcal{X} \rightarrow \mathcal{U}$ 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

$$J(g) = \limsup_{t \rightarrow \infty} \frac{1}{t} \mathbb{E}_x \left\{ \sum_{k=0}^{t-1} f(x_k, u_k) \right\}$$

(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 $V : \mathcal{X} \rightarrow \mathbb{R}$ and $J^* \in \mathbb{R}$ to the dynamic programming of the infinite horizon average cost stochastic optimal control problem satisfying

$$J^* + V(x) = \inf_{u \in \mathcal{U}(x)} \left\{ f(x,u) + \int \mathcal{X} Q(dz|x,u)V(z) \right\}.$$  

(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 $Q(\cdot|x,u)$ is irreducible (that is, all stationary policies have at most one recurrent class), then there exists a solution $V : \mathcal{X} \rightarrow \mathbb{R}$ and $J^* \in \mathbb{R}$ to the dynamic programming of the infinite-horizon average cost stochastic optimal control problem satisfying

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

(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 $Q(\cdot|x,u)$, $Q(dz|x,u)$, 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, $||| \cdot |||_{TV} : \mathcal{M}_1(\Sigma) \times \mathcal{M}_1(\Sigma) \rightarrow [0, \infty]$ defined by

$$||| \alpha - \beta |||_{TV} = \sup_{P \in \mathcal{P}(\Sigma)} \sum_{F_i \in P} |\alpha(F_i) - \beta(F_i)|$$

for $\alpha, \beta \in \mathcal{M}_1(\Sigma)$ where $\mathcal{M}_1(\Sigma)$ denotes the set of probability measures on the $\sigma$-algebra $\mathcal{B}(\Sigma)$ and $\mathcal{P}(\Sigma)$ denotes the collection of all finite partitions of $\Sigma$.

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 process is $\{x_t^0 : t = 0, 1, \ldots \in G_{SM}\}$ corresponds to a sequence of stationary stochastic kernels as follows: for every $A \in \mathcal{B}(\mathcal{X})$

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

where $Q^o(\cdot|x_t, u_t) \in \mathcal{Q}(\mathcal{X}|\mathcal{K})$, $t = 0, 1, \ldots$.

The class of controlled processes is described by the sequence of stochastic kernels, $\{Q(dx_t|x_t, u_t) : t = 0, 1, \ldots \}$ 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

$$\mathcal{B}_R(Q^o)(x, u) = \left\{ \{Q(\cdot|x, u) \in \mathcal{M}_1(\mathcal{X}) : \|Q(\cdot|x, u) - Q^o(\cdot|x, u)\|_{TV} \leq R \} \right\}, (x,u) \in \mathbb{K}.$$

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

$$J(g) = \limsup_{t \rightarrow \infty} \sup_{Q(\cdot|x,u) \in \mathcal{B}_R(Q^o)(x,u)} \frac{1}{t} \mathbb{E}_x \left\{ \sum_{k=0}^{t-1} f(x_k, u_k) \right\}$$

where $\mathbb{E}_x \{ \cdot \}$ denotes the expectation with respect to any conditional distribution $Q(\cdot|x_t, u_t) \in \mathcal{B}_R(Q^o)(x,u)$, $t = 0, 1, \ldots$ (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 metric spaces, is more involved 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 $\Sigma$ 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, $\{v(x) : x \in \Sigma \} \in \mathcal{M}_1(\Sigma)$, $\{\mu(x) : x \in \Sigma \} \in \mathcal{M}_1(\Sigma)$. Also, let $\ell = \{\ell(x) : x \in \Sigma \} \in \mathcal{M}_1(\Sigma)$, define the maximum and minimum values of $\{\ell(x) : x \in \Sigma \}$ by

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

and their corresponding support sets by

$$\Sigma_{\ell_{max}} = \{ x \in \Sigma : \ell(x) = \ell_{max} \}, \quad \Sigma_{\ell_{min}} = \{ x \in \Sigma : \ell(x) = \ell_{min} \}.$$
For all remaining sequences, \( \ell(x) : x \in \Sigma \setminus \Sigma_0 \cup \Sigma_0 \), define recursively the set of indices for which the sequence achieves its \((k+1)\)th smallest value by

\[
\Sigma_k \triangleq \left\{ x \in \Sigma : \ell(x) = \min \{ \ell(\alpha) : \alpha \in \Sigma \setminus \Sigma_0 \cup \bigcup_{j=1}^{k} \Sigma_{j-1} \} \right\}
\]

and the corresponding values of the sequence on \( \Sigma_k \) 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 \( \Sigma_k \) sets which is at most \(|\Sigma \setminus \Sigma_0 \cup \Sigma_0|\), i.e., \( 1 \leq r \leq |\Sigma \setminus \Sigma_0 \cup \Sigma_0| \).

In [9] it is shown that the maximum payoff subject to total variation constraint is given by

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

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

\[
\begin{align*}
\nu^*(\Sigma^0) &= \mu(\Sigma^0) + \frac{\alpha}{2} \\
\nu^*(\Sigma_0) &= \left( \mu(\Sigma_0) - \frac{\alpha}{2} \right) \\
\nu^*(\Sigma_k) &= \left( \mu(\Sigma_k) - \left( \frac{\alpha}{2} - \sum_{j=1}^{k} \mu(\Sigma_{j-1}) \right) + \right) \\
\alpha &= \min \left\{ R, 2(1 - \mu(\Sigma^0)) \right\}
\end{align*}
\]

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

In the next section, we apply the above solution 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

\[
J(g) = \lim_{j \to \infty} \max_{Q(\cdot|x,u) \in B_R(Q^*)(x,u)} \frac{1}{j} E_Q \left\{ \sum_{k=0}^{j-1} f(x_k, u_k) \right\}.
\]

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

\[
\max_{Q(\cdot|x,u) \in B_R(Q^*)(x,u)} E_Q \left\{ \sum_{k=0}^{n-1} f(x_k, u_k) \right\}
\]

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

\[
V_j(x) = \min_{u \in U(x)} \max_{Q(\cdot|x,u) \in B_R(Q^*)(x,u)} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} V_{j+1}(z) Q(z|x,u) \right\}
\]

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

\[
V_j(x) = \min_{u \in U(x)} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} V_{j+1}(z) Q^*(z|x,u) \right\}
\]

\[
= \min_{u \in U(x)} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} V_{j+1}(z) Q^*(z|x,u) + R \left( \max_{z \in \mathcal{X}} V_{j+1}(z) - \min_{z \in \mathcal{X}} V_{j+1}(z) \right) \right\}.
\]

Define \( \nabla_j(x) = V_{n-j}(x) \). Then \( \nabla_j \) satisfies the equation

\[
\nabla_j(x) + \frac{1}{j} \nabla_j(x) = \min_{u \in U(x)} \max_{Q(\cdot|x,u) \in B_R(Q^*)(x,u)} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} \nabla_{j+1}(z) + \frac{1}{j} \nabla_j(x) \right\}.
\]

which can be rewritten as follows

\[
\nabla_j(x) + \frac{1}{j} \nabla_j(x) = \min_{u \in U(x)} \max_{Q(\cdot|x,u) \in B_R(Q^*)(x,u)} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} \left( \nabla_{j-1}(z) + \frac{1}{j} \nabla_j(x) \right) Q(z|x,u) \right\}.
\]

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

\[
\lim_{j \to \infty} (\nabla_j(x) - j J^*) = V(x), \quad \forall x \in \mathcal{X}.
\]

Under Assumption 3.1, then

\[
\lim_{j \to \infty} \frac{1}{j} \nabla_j(x) = J^*, \quad \forall x \in \mathcal{X}
\]

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

\[
\lim_{j \to \infty} \max_{x \in \mathcal{X}} (\nabla_j(x) - j J^*) = \max_{x \in \mathcal{X}} V(x).
\]

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

\[ J^* + V(x) = \lim_{j \to \infty} \left( \frac{1}{j} \mathcal{V}_j(x) + (\mathcal{V}_j(x) - jJ^*) \right) \]

(a) \[ \lim_{j \to \infty} \min_{u \in \mathcal{U}(x)} \max_{Q(x,u) \in \mathcal{B}_R(Q^*)} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} (\mathcal{V}_{j-1}(z) + \frac{1}{j} \mathcal{V}_j(x)) Q(z|x, u) - jJ^* \right\} \]

(b) \[ \lim_{j \to \infty} \min_{u \in \mathcal{U}(x)} \max_{Q(x,u) \in \mathcal{B}_R(Q^*)} \left\{ f(x, u) - jJ^* + \sum_{z \in \mathcal{X}} (\mathcal{V}_{j-1}(z) + \frac{1}{j} \mathcal{V}_j(x)) Q(z|x, u) \right\} \]

(c) \[ \lim_{j \to \infty} \min_{u \in \mathcal{U}(x)} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} (\mathcal{V}_{j-1}(z) - (j - 1)J^* + \frac{1}{j} \mathcal{V}_j(x) - J^*) Q(z|x, u) \right\} \]

where (a) is obtained by (11), (b) is obtained by the equivalent formulation (10), and (c) by adding and subtracting \( J^*(1 + \frac{j}{j}) \).

By the finite cardinality of \( \mathcal{X} \) and \( \mathcal{U} \), then

\[ J^* + V(x) = \min_{u \in \mathcal{U}(x)} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} V(z) Q^*(z|x, u) + \frac{R}{2} \left( \max_{z \in \mathcal{X}} V(z) - \min_{z \in \mathcal{X}} V(z) \right) \right\} \]

which is equivalent to

\[ J^* + V(x) = \min_{u \in \mathcal{U}(x)} \max_{Q(x,u) \in \mathcal{B}_R(Q^*)} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} V(z) Q(z|x, u) \right\}. \]

Equation (15) (or, equivalently, (16)) is the dynamic programming equation for average cost minimization 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 \( V : \mathcal{X} \to \mathbb{R}^n \) may be represented by a vector in \( \mathbb{R}^n \). Any stationary control law \( g \in \mathcal{G}_{SM} \), \( g : \mathcal{X} \to \mathbb{R}^n \), may also be identified with a \( g \in \mathbb{R}^n \). For any \( g \), let \( Q(g) \in \mathbb{R}^{n \times n} \) and

\[ f(g) = (f(x_1, g(x_1)) \cdots f(x_n, g(x_n)))^T \in \mathbb{R}^n. \]

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

\[ \max_{Q(x,u) \in \mathcal{B}_R(Q^*)} \mathbb{E}_Q \left\{ \sum_{k=0}^{j-1} f(x_k, u_k) \right\} \]

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

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

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

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

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 \( g : \mathcal{X} \to \mathcal{U} \). Assume that the maximizing conditional distribution \( Q^*(g) \in \mathbb{R}_{T}^{n \times n} \) in (17) is irreducible. Then,

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

\[ Q^*(g) q(g) = q(g), \quad e^T g = 1, \quad e = (1, \ldots, 1)^T; \]

(b) the average cost associated with the control law \( g \in \mathcal{G}_{SM} \) is

\[ J(g) = q(g)^T f(g); \]

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

\[ J(g) e + V(g) = f(g) + Q^*(g) V(g). \]

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

\[ J(g^*) = \min_{g \in \mathcal{G}_{SM}} J(g) \]

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

\[ J(g^* + V(g^*, x)) = \min_{u \in \mathcal{U}} \max_{Q(x,u) \in \mathcal{B}_R(Q^*)} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} V(g^*, z) Q(z|x, u) \right\} \]

Then we obtain the following theorem.

**Theorem 3.3:** Assume that for all stationary Markov control laws \( g \in \mathcal{G}_{SM} \), and for a given total variation parameter
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 \( g_0 : \mathcal{X} \rightarrow \mathcal{U} \).
2) (Policy Evaluation) Solve the equation
\[
J_{Q^c}(g_m)e + V_{Q^c}(g_m) = f(g_m) + Q^o(g_m)V_{Q^c}(g_m)
\]
for \( J_{Q^c}(g_m) \in \mathbb{R} \) and \( V_{Q^c}(g_m) \in \mathbb{R}^n \). Identify the support sets of (28) using (25)-(26), and construct the matrix \( Q^*(g_m) \) using (22)-(24). Solve the equation
\[
J_{Q^e}(g_m)e + V_{Q^e}(g_m) = f(g_m) + Q^*(g_m)V_{Q^e}(g_m)
\]
for \( J_{Q^e}(g_m) \in \mathbb{R} \) and \( V_{Q^e}(g_m) \in \mathbb{R}^n \).
3) (Policy Improvement) Let
\[
g_{m+1} = \arg \min_{g \in \mathbb{R}^n} \left\{ f(g) + Q^*(g)V_{Q^e}(g) \right\}.
\]
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
\[
Q^o(u_1) = \frac{1}{9} \begin{pmatrix} 3 & 1 & 5 \\ 4 & 2 & 3 \\ 1 & 6 & 2 \end{pmatrix}, \quad Q^o(u_2) = \frac{1}{9} \begin{pmatrix} 1 & 2 & 6 \\ 4 & 2 & 3 \\ 4 & 1 & 4 \end{pmatrix}
\]
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 \( J_{Q^e}(g_0)e + V_{Q^e}(g_0) = f(g_0) + Q^o(g_0)V_{Q^e}(g_0) \) for \( J_{Q^e}(g_0) \in \mathbb{R} \) and \( V_{Q^e}(g_0) \in \mathbb{R}^3 \). The

¹This follows by Theorem 3.3, part (a).
optimality equations (28) are given by

\[ J_{Q^*}(g_0) + \frac{6}{9} V_{Q^*}(g_0) = 2 + \frac{1}{9} J_{Q^*}(g_0) + \frac{2}{9} V_{Q^*}(g_0, 3) \quad (33a) \]

\[ J_{Q^*}(g_0) + \frac{4}{9} V_{Q^*}(g_0, 2) = 3 + \frac{4}{9} J_{Q^*}(g_0, 1) + \frac{3}{9} V_{Q^*}(g_0, 3) \quad (33b) \]

\[ J_{Q^*}(g_0) + \frac{4}{9} V_{Q^*}(g_0, 3) = \frac{4}{9} V_{Q^*}(g_0, 1) + \frac{1}{9} V_{Q^*}(g_0, 2). \quad (33c) \]

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

\[ V_{Q^*}(g_0, 1) = 1.8, \quad V_{Q^*}(g_0, 2) = 3.375, \quad V_{Q^*}(g_0, 3) = 0 \]

\[ J_{Q^*}(g_0) = 1.175. \]

Note that, \( V_{Q^*} \triangleq \{ V_{Q^*}(1), V_{Q^*}(2), V_{Q^*}(3) \} \), and hence by (25)-(26), the support sets based on the values of \( V_{Q^*} \) are \( X^+ = \{ 2 \}, X^- = \{ 3 \} \) and \( X_1 = \{ 1 \} \). Once the partition is been identified, (22)-(24) are applied to obtain (32). Hence,

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

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

\[ J_{Q^*}(g_0) + \frac{6}{9} V_{Q^*}(g_0, 1) = 2 + \frac{1}{9} J_{Q^*}(g_0, 2) + \frac{2}{9} V_{Q^*}(g_0, 3) \quad (35a) \]

\[ J_{Q^*}(g_0) + \frac{4}{9} V_{Q^*}(g_0, 2) = 3 + \frac{4}{9} J_{Q^*}(g_0, 1) + \frac{3}{9} V_{Q^*}(g_0, 3) \quad (35b) \]

\[ J_{Q^*}(g_0) + \frac{4}{9} V_{Q^*}(g_0, 3) = \frac{4}{9} V_{Q^*}(g_0, 1) + \frac{1}{9} V_{Q^*}(g_0, 2). \quad (35c) \]

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

\[ V_{Q^*}(g_0, 1) = 1.8, \quad V_{Q^*}(g_0, 2) = 3.375, \quad V_{Q^*}(g_0, 3) = 0 \]

\[ J_{Q^*}(g_0) = 2.3. \]

3) Let \( g_1 = \arg \min_{g \in \mathbb{R}^3} \{ f(g) + Q^*(g) V_{Q^*}(g_0) \} \), then the resulting optimal control laws are \( g_1(1) = u_2, \quad g_1(2) = u_1 \) 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 \( J_{Q^*}(g_1) e + V_{Q^*}(g_1) = f(g_1) + Q^*(g_1) V_{Q^*}(g_1) \), for \( J_{Q^*}(g_1) \in \mathbb{R} \) and \( V_{Q^*}(g_1) \in \mathbb{R}^3 \) with \( V_{Q^*}(g_1, 3) = 0 \) is

\[ V_{Q^*}(g_1, 1) = 0.468, \quad V_{Q^*}(g_1, 2) = 1.125, \quad V_{Q^*}(g_1, 3) = 0 \]

\[ J_{Q^*}(g_1) = 0.333. \]

Next, we proceed with the identification of the support sets, which are \( X^+ = \{ 2 \}, X^- = \{ 3 \} \) and \( X_1 = \{ 1 \} \). 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

\[ V_{Q^*}(g_1, 1) = 0.468, \quad V_{Q^*}(g_1, 2) = 1.125, \quad V_{Q^*}(g_1, 3) = 0 \]

\[ J_{Q^*}(g_1) = 0.708. \]

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

4) Since, \( g_2 = g_1 \), then \( g^* = g_1 \) is an optimal control law with \( J_{Q^*} = 0.708, \quad V_Q^*(1) = 0.468, \quad V_Q^*(2) = 1.125 \) 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


