2016 IEEE 55th Conference on Decision and Control (CDC)
ARIA Resort & Casino
December 12-14, 2016, Las Vegas, USA

Infinite Horizon Discounted Dynamic Programming subject to Total Variation Ambiguity on Conditional Distribution

Ioannis Tzortzis, Charalambos D. Charalambous and Themistoklis Charalambous

***Abstract*— We analyze the infinite horizon minimax discounted cost Markov Control Model (MCM), for a class of controlled process conditional distributions, which belong to a ball, with respect to total variation distance metric, centered at a known nominal controlled conditional distribution with radius $R \in [0, 2]$, in which the minimization is over the control strategies and the maximization is over conditional distributions. Through our analysis (i) we derive a new discounted dynamic programming equation, (ii) we show the associated contraction property, and (iii) we develop a new policy iteration algorithm. Finally, the application of the new dynamic programming and the corresponding policy iteration algorithm are shown via an illustrative example.**

I. Introduction

Consider a discrete-time Markov control model (MCM) with finite state space $\mathcal{X}$, used to represent a stochastic control system observed at times $k = 0, 1, \ldots$. Let x_k denote the state of the system at time k, and assume that there is a finite control (or action) space $\mathcal{U}$ such that for each state $x_k = x \in \mathcal{X}$, $u_k = u \in \mathcal{U}$ denote the control (or action) applied at time k. Let $f(x, u)$ denote the incurred one-stage cost and $Q(\cdot|x, u)$ denote the state transition probabilities. Let $g : \mathcal{X} \longmapsto \mathcal{U}$ denote a deterministic stationary Markov control policy. The infinite horizon discounted cost criterion when policy g is used, and given an initial state $x_0 = x$, is defined by

$$J^o(g) \stackrel{\triangle}{=} \mathbb{E}_x^g \Big\{ \sum_{j=0}^{\infty} \alpha^j f(x_j^g, u_j^g) \Big\} \tag{1}$$

where the discounting factor $\alpha \in (0, 1)$ implies that future costs matter less than similar costs incurred at the present time. The optimal stochastic control problem is to select the stationary control policy, for all initial states x, which minimizes (1), i.e.,

$$J^o(g^*) = \inf_g J^o(g). \tag{2}$$

The dynamic programming equation of the infinite horizon discounted Markov control model (D-MCM) is a function $v_\infty^o : \mathcal{X} \longmapsto \mathbb{R}$ satisfying

$$v_\infty^o(x) = \inf_{u \in \mathcal{U}(x)} \Big\{ f(x, u) + \alpha \int_{\mathcal{X}} v_\infty^o(z) Q(dz|x, u) \Big\}, \quad \text{for all } x \in \mathcal{X}. \tag{3}$$

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.

The infinite horizon discounted cost criterion for a MCM, when the state transition probabilities are known, is addressed by several authors (see, for example, [1]–[3]), to establish existence of optimal decision strategies, to derive necessary and sufficient optimality conditions, and to compute the optimal decision strategies via policy iteration algorithms. However, for situations in which the state transition probabilities (hereafter, called the controlled process conditional distributions) are not known exactly then the optimality and robustness of the optimal decision strategies is not ensured.

Motivated by this implication, in this paper we investigate the effects of the ambiguity in the controlled process conditional distributions, modeled by a ball with respect to total variation distance metric, centered at a known nominal controlled conditional distribution with radius $R \in [0, 2]$ on the dynamic programming equation (3). In particular, we formulate the stochastic control problem using minimax theory, in which the control minimizes the discounted cost criterion while the conditional distribution, from the total variation distance set, maximizes it.

An important difference between the minimax stochastic control problem we investigate in this paper and those investigated by previous authors, i.e., using various robust deterministic and stochastic control approaches, including minimax and risk-sensitive formulations (see [4]–[12] and references therein), is that in our work, the resulting dynamic programming equation is not limited by the assumption that the maximizing controlled process conditional distribution is absolutely continuous with respect to the nominal conditional distribution, and hence defined on the same dimensional state space. This is a nice feature which can be extended for the approximation of systems with high dimensional state spaces by systems of lower dimensions. This remains an interesting subject for further investigation.

In summary, the issues discussed and results obtained in this paper are the following:

(i) formulation of infinite horizon discounted stochastic optimal control subject to conditional distribution ambiguity described by total variation distance via minimax theory;
(ii) characterization of the maximizing conditional distribution and the corresponding new dynamic programming equation;
(iii) the contraction property of the infinite horizon D-MCM dynamic programming and new policy iteration algorithm.

The rest of the paper is organized as follows. In Section II

we provide some background material for the infinite horizon D-MCM, including the maximization of a linear functional subject to total variation distance ambiguity. In Section III we derive a new dynamic programming equation, we show that the dynamic programming operator is contractive, and we develop a new policy iteration algorithm. In Section IV we present an example to illustrate the application of the new dynamic programming equation and the corresponding policy iteration algorithm. Finally, in Section V we draw conclusions.

II. Preliminaries

An infinite horizon discounted-Markov control model (D-MCM) with deterministic strategies is a sextuple

$$\Big(\mathcal{X},\mathcal{U},\{\mathcal{U}(x){:}x{\in}\mathcal{X}\},\{Q(dz|x,u){:}(x,u){\in}\mathcal{X}{\times}\mathcal{U}\},f,\alpha\Big) \quad (4)$$

consisting of the following.

a) State Space. A complete separable metric space (called a Polish space) $\mathcal{X}$, which models the state space of the controlled random process $\{x_k \in \mathcal{X} : k \in \mathbb{N}\}$, $\mathbb{N} \triangleq 0, 1, \ldots$.

b) Control or Action Space. A Polish space $\mathcal{U}$, which models the control or action set of the control random process $\{u_k \in \mathcal{U} : k \in \mathbb{N}\}$.

c) Feasible Controls or Actions. A family $\{\mathcal{U}(x) : x \in \mathcal{X}\}$ of non-empty measurable subsets $\mathcal{U}(x)$ of $\mathcal{U}$, where $\mathcal{U}(x)$ denotes the set of feasible controls or actions, when the controlled process is in state $x \in \mathcal{X}$, and the feasible state-actions pairs are measurable subsets of $\mathcal{X} \times \mathcal{U}$, defined by

$$\mathbb{K} \triangleq \Big\{(x,u) : x \in \mathcal{X}, u \in \mathcal{U}(x)\Big\}. \quad (5)$$

d) Controlled Process Distribution. A conditional distribution or stochastic kernel $Q(dz|x,u)$ on $\mathcal{X}$ given $(x,u) \in \mathbb{K} \subseteq \mathcal{X} \times \mathcal{U}$, which corresponds to the controlled process transition probability distribution.

e) One-Stage-Cost. A non-negative measurable function $f : \mathbb{K} \longmapsto [0,\infty]$, called the one-stage-cost, such that $f(x,\cdot)$ does not take the value $+\infty$ for each $x \in \mathcal{X}$.

(f) Discounting Factor. A real number $\alpha \in (0,1)$ called the discounting factor.

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.

Next, the total variation distance ambiguity class is introduced.

Total Variation Distance Ambiguity. 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. Note that the distance metric induced by the total variation norm does not require absolute continuity of measures when defining the ambiguity ball (i.e., singular measures are admissible).

In this paper, we will derive the analogue of (3), for the class of conditional distributions of the controlled process $Q(dz|x,u)$, $(x,u) \in \mathbb{K}$ which are stationary, and belong to a ball with respect to total variation distance metric, centered at a nominal controlled process distribution $Q^o(dz|x,u)$, $(x,u) \in \mathbb{K}$, having radius $R \in [0,2]$ (specifically, $\{Q(dz|x,u) : ||Q(\cdot|x,u) - Q^o(\cdot|x,u)||_{TV} \le R\}$). The precise definition is the following.

Definition 2.2: For each $g \in G_{SM}$, the nominal controlled process $\{x_t^g : t = 0,1,\ldots\}$ has a time-invariant conditional distribution, conditioned on $x_{t-1}^g = x_{t-1}$, $u_{t-1}^g = u_{t-1}$, defined for every $A \in \mathcal{B}(\mathcal{X})$ by

$$\begin{aligned}&\text{Prob}(x_t \in A|x^{t-1},u^{t-1}) = Q^o(A|x_{t-1},u_{t-1}),\\ &\quad\text{where } Q^o(A|x_{t-1},u_{t-1}) \in \mathcal{Q}(\mathcal{X}|\mathbb{K}),\ t = 0,1,\ldots.\end{aligned} \quad (6)$$

Given the nominal controlled process and $R \in [0,2]$, the true controlled process conditional distributions are stationary, and belong to the total variation distance ball defined by

$$\begin{aligned}\mathbf{B}_R(Q^o)(x,u) \triangleq \Big\{&Q(\cdot|x,u) \in \mathcal{M}_1(\mathcal{X}) :\\ &||Q(\cdot|x,u) - Q^o(\cdot|x,u)||_{TV} \le R\Big\},\ (x,u) \in \mathbb{K}.\end{aligned} \quad (7)$$

Utilizing the above formulation, next we define the minimax stochastic control problem under investigation (which is the analogue of (2)).

Problem 2.3: Given a nominal controlled process distribution and an ambiguity class of Definition 2.2, determine a policy $g^* \in G_{SM}$ and a true controlled process distribution $Q^*(dz|x,u) \in \mathbf{B}_R(Q^o)(x,u)$, which solve the following minimax optimization problem

$$\begin{aligned}&J(g^*,Q^*,x)\\ &= \inf_{g\in G} \sup_{Q(\cdot|x,u)\in\mathbf{B}_R(Q^o)(x,u)} \mathbb{E}_x^g\Big\{\sum_{j=0}^{\infty}\alpha^j f(x_j^g,u_j^g)\Big\}\\ &\equiv J^*(x), \quad \forall x \in \mathcal{X}\end{aligned} \quad (8)$$

where $\mathbb{E}_x^g\{\cdot\}$ indicates the dependence of the expectation operation on the policy g and $x_0 = x$.

A conditional distribution Q^* that satisfies (8) is called *α-discount distribution*, while a policy g^* is called *α-discount optimal*. The corresponding $J^*(\cdot)$ is called the *α-discount cost or value function* of the minimax D-MCM.

A. Total Variation Distance Ambiguity

In this section, we recall certain results from [13], concerning the characterization of the α-*discount distribution* in (8).

Let $(\mathcal{X}, d_{\mathcal{X}})$ denote a complete, separable metric space (a Polish space), and $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ the corresponding measurable space, in which $\mathcal{B}(\mathcal{X})$ is the σ-algebra generated by open sets in $\mathcal{X}$. Define the spaces

$$BC(\mathcal{X}) \triangleq \big\{\text{Bounded continuous functions}, \\ \ell : \mathcal{X} \longmapsto \mathbb{R} : ||\ell|| \triangleq \sup_{x\in\mathcal{X}} |\ell(x)| < \infty\big\}$$

and $BC^+(\mathcal{X}) \triangleq \{\ell \in BC(\mathcal{X}) : \ell \geq 0\}$. For $\ell \in BC^+(\mathcal{X})$, $\mu \in \mathcal{M}_1(\mathcal{X})$ fixed and $R \in [0, 2]$, then we have

$$L(\nu^*) \triangleq \sup_{||\nu-\mu||_{TV}\leq R} \int_{\mathcal{X}} \ell(x)\nu(dx) \tag{9a}$$

$$= \frac{R}{2}\Big\{ \sup_{x\in\mathcal{X}} \ell(x) - \inf_{x\in\mathcal{X}} \ell(x)\Big\} + \int_{\mathcal{X}} \ell(x)\mu(dx) \tag{9b}$$

where the optimal ν^* in (9a) satisfies the constraint $||\xi^*||_{TV} = ||\nu^* - \mu||_{TV} = R$, it is normalized $\nu^*(\mathcal{X}) = 1$, and $\nu^*(A) \in [0, 1]$ on any $A \in \mathcal{B}(\mathcal{X})$. Moreover, by [13] the equality in (9b) is obtained. If $\mathcal{X}$ is a compact set, since $\ell(\cdot) \in BC^+(\mathcal{X})$ then both the supremum and infimum are attained and they are finite. Define[1]

$$x^0 \in \mathcal{X}^0 \triangleq \Big\{x \in \overline{\mathcal{X}} : \ell(x) = \sup\{\ell(x) : x \in \mathcal{X}\} \equiv \ell_{\max}\Big\}$$
$$x_0 \in \mathcal{X}_0 \triangleq \Big\{x \in \overline{\mathcal{X}} : \ell(x) = \inf\{\ell(x) : x \in \mathcal{X}\} \equiv \ell_{\min}\Big\}$$

where $\overline{\mathcal{X}}$ denotes the closure[2] of $\mathcal{X}$. Then, the pay-off $L(\nu^*)$ can be written as

$$L(\nu^*) = \int_{\mathcal{X}^0} \ell_{\max}\nu^*(dx) + \int_{\mathcal{X}_0} \ell_{\min}\nu^*(dx) + \int_{\mathcal{X}\setminus\mathcal{X}^0\cup\mathcal{X}_0} \ell(x)\mu(dx)$$

and the optimal distribution $\nu^* \in \mathcal{M}_1(\mathcal{X})$, which satisfies the total variation constraint, is given by

$$\int_{\mathcal{X}^0} \nu^*(dx) = \min\Big(\mu(\mathcal{X}^0) + \frac{R}{2}\Big) \in [0, 1] \tag{10a}$$

$$\int_{\mathcal{X}_0} \nu^*(dx) = \max\Big(\mu(\mathcal{X}_0) - \frac{R}{2}\Big) \in [0, 1] \tag{10b}$$

$$\nu^*(A) = \mu(A), \quad \forall A \subseteq \mathcal{X} \setminus \mathcal{X}^0 \cup \mathcal{X}_0. \tag{10c}$$

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

III. Minimax Dynamic Programming

In this section, we apply dynamic programming to characterize the solution of Problem 2.3. In addition, we show that the operator associated with the dynamic programming equation is contractive, and we introduce a new policy iteration algorithm.

[1]We adopt the standard definitions; infimum (supremum) of an empty set to be $+\infty$ ($-\infty$).

[2]Closure of a set $\mathcal{X}$ consists of all points in $\mathcal{X}$ plus the limit points of $\mathcal{X}$.

Consider the finite horizon version of Problem 2.3, with cost criterion

$$\sup_{Q(\cdot|x,u)\in\mathbf{B}_R(Q^o)(x,u)} \mathbb{E}_x^g\Big\{\sum_{j=0}^{n-1} \alpha^j f(x_j^g, u_j^g)\Big\} \tag{11}$$

In [14] the value function of (11), is determined by the dynamic programming equations

$$V_n(x) = 0$$
$$V_j(x) = \inf_{u\in\mathcal{U}(x)} \Big\{\alpha^j f(x, u) + \int_{\mathcal{X}} V_{j+1}(z)Q^o(dz|x, u) \\ + \frac{R}{2}\big(\sup_{z\in\mathcal{X}} V_{j+1}(z) - \inf_{z\in\mathcal{X}} V_{j+1}(z)\big)\Big\}, \ x \in \mathcal{X}.$$

Define $v_i(x) = \alpha^{i-n}V_{n-i}(x)$, where $0 \leq i \leq n$ is the time to go. Then,

$$v_0(x) = 0 \tag{12}$$

$$v_i(x) = \inf_{u\in\mathcal{U}(x)} \Big\{f(x, u) + \alpha \int_{\mathcal{X}} v_{i-1}(z)Q^o(dz|x, u) \\ + \alpha\frac{R}{2}\big(\sup_{z\in\mathcal{X}} v_{i-1}(z) - \inf_{z\in\mathcal{X}} v_{i-1}(z)\big)\Big\} \tag{13}$$

which is obtained as follows:

$$\begin{aligned} v_0(x) &= \alpha^{-n}V_n(x) = 0, \\ v_i(x) &= \alpha^{i-n}V_{n-i}(x) \\ &= \inf_{u\in\mathcal{U}(x)} \Big\{\alpha^{i-n}\alpha^{n-i} f(x, u) \\ &\quad + \alpha^{i-n} \int_{\mathcal{X}} V_{n-i+1}(z)Q^o(dz|x, u) \\ &\quad + \alpha^{i-n}\frac{R}{2}\Big(\sup_{z\in\mathcal{X}} V_{n-i+1}(z) - \inf_{z\in\mathcal{X}} V_{n-i+1}(z)\Big)\Big\} \\ &= \inf_{u\in\mathcal{U}(x)} \Big\{f(x, u) + \alpha \int_{\mathcal{X}} v_{i-1}(z)Q^o(dz|x, u) \\ &\quad + \alpha\frac{R}{2}\Big(\sup_{z\in\mathcal{X}} v_{i-1}(z) - \inf_{z\in\mathcal{X}} v_{i-1}(z)\Big)\Big\}. \end{aligned}$$

In contrast with finite horizon case the one given by (12)-(13) proceeds from lower to higher values of indices i. The dynamic programming for the discounted cost

$$\sup_{Q(\cdot|x,u)\in\mathbf{B}_R(Q^o)(x,u)} \mathbb{E}_x^g\Big\{\sum_{j=0}^{\infty} \alpha^j f(x_j^g, u_j^g)\Big\} \tag{14}$$

is given by

$$v_\infty(x) = \inf_{u\in\mathcal{U}(x)} \Big\{f(x, u) + \alpha \int_{\mathcal{X}} v_\infty(z)Q^o(dz|x, u) \\ + \alpha\frac{R}{2}\Big(\sup_{z\in\mathcal{X}} v_\infty(z) - \inf_{z\in\mathcal{X}} v_\infty(z)\Big)\Big\}. \tag{15}$$

The maximizing conditional distribution is

$$Q^*\left(\mathcal{X}^+|x, u\right) = Q^o(\mathcal{X}^+|x, u) + \frac{R}{2} \in [0, 1], \quad (x, u)\in\mathbb{K} \tag{16}$$

$$Q^*\left(\mathcal{X}^-|x, u\right) = Q^o(\mathcal{X}^-|x, u) - \frac{R}{2} \in [0, 1], \quad (x, u)\in\mathbb{K} \tag{17}$$

$$Q^*\left(A|x, u\right) = Q^o(A|x, u), \ \forall A\subseteq\mathcal{X}\setminus\mathcal{X}^+\cup\mathcal{X}^-, (x, u)\in\mathbb{K} \tag{18}$$

where

$$\mathcal{X}^+ \triangleq \Big\{x \in \mathcal{X} : V(x) = \sup\{V(x) : x \in \mathcal{X}\}\Big\} \quad (19)$$

$$\mathcal{X}^- \triangleq \Big\{x \in \mathcal{X} : V(x) = \inf\{V(x) : x \in \mathcal{X}\}\Big\}. \quad (20)$$

Next, we recall the following theorem from [15], which we invoke to show that the operator in the right hand side of (15) is contractive.

Theorem 3.1: Let $(L, ||\cdot||)$ be a complete normed space and let $T : L \longrightarrow L$ satisfy the following inequality for some $0 < \alpha < 1$,

$$||TV_1 - TV_2|| \leq \alpha||V_1 - V_2||, \text{ for all } V_1, V_2 \in L. \quad (21)$$

A mapping T satisfying (21) is called a contraction mapping. Then, the following hold.

1) There exists a unique $w \in L$ satisfying $Tw = w$, called the fixed point of T.

2) For $V \in L$, define $\{T^nV : n \in \mathbb{Z}_+\}$ by $TV = V$, $T^{n+1}V = T^n(TV)$ then

$$\lim_{n\longrightarrow\infty} ||T^nV - w|| = 0, \text{ for all } V \in L, \quad (22)$$

where w is the fixpoint defined in 1).

Lemma 3.2: Let L be the class of all measurable functions $V : \mathcal{X} \longrightarrow \mathbb{R}$, with finite norm $||V|| \triangleq \max_{x\in\mathcal{X}} |V(x)|$, and $T : L \longmapsto L$ defined by

$$(TV)(x) = \inf_{u\in\mathcal{U}(x)} \Big\{f(x,u) + \alpha \int_{\mathcal{X}} V(z)Q^o(dz|x,u) + \alpha\frac{R}{2}\Big(\sup_{z\in\mathcal{X}} V(z) - \inf_{z\in\mathcal{X}} V(z)\Big)\Big\}. \quad (23)$$

If $V \in BC^+(\mathcal{X})$ and $\sup_{z\in\mathcal{X}} V(z)$, $\inf_{z\in\mathcal{X}} V(z)$ are finite, then T is a contraction.

Proof: For $V_1, V_2 \in L$,

$$\begin{aligned}(TV_1)(x) - (TV_2)(x) = &\inf_{u\in\mathcal{U}(x)} \Big\{f(x,u) + \alpha \int_{\mathcal{X}} V_1(z)Q^o(dz|x,u) + \alpha\frac{R}{2}\Big(\sup_{z\in\mathcal{X}} V_1(z) - \inf_{z\in\mathcal{X}} V_1(z)\Big)\Big\} \\ &- \inf_{u\in\mathcal{U}(x)} \Big\{f(x,u) + \alpha \int_{\mathcal{X}} V_2(z)Q^o(dz|x,u) + \alpha\frac{R}{2}\Big(\sup_{z\in\mathcal{X}} V_2(z) - \inf_{z\in\mathcal{X}} V_2(z)\Big)\Big\}.\end{aligned}$$

Let

$$v \triangleq \arg\inf_{u\in\mathcal{U}(x)} \Big\{f(x,u) + \alpha \int_{\mathcal{X}} V_2(z)Q^o(dz|x,u) + \alpha\frac{R}{2}\Big(\sup_{z\in\mathcal{X}} V_2(z) - \inf_{z\in\mathcal{X}} V_2(z)\Big)\Big\}.$$

Then,

$$\begin{aligned}&(TV_1)(x) - (TV_2)(x) \\ &= \inf_{u\in\mathcal{U}(x)} \Big\{f(x,u) + \alpha \int_{\mathcal{X}} V_1(z)Q^o(dz|x,u) + \alpha\frac{R}{2}\Big(\sup_{z\in\mathcal{X}} V_1(z) - \inf_{z\in\mathcal{X}} V_1(z)\Big)\Big\} \\ &\quad - \Big\{f(x,v) + \alpha \int_{\mathcal{X}} V_2(z)Q^o(dz|x,v) + \alpha\frac{R}{2}\Big(\sup_{z\in\mathcal{X}} V_2(z) - \inf_{z\in\mathcal{X}} V_2(z)\Big)\Big\} \\ &\leq \Big\{f(x,v) + \alpha \int_{\mathcal{X}} V_1(z)Q^o(dz|x,v) + \alpha\frac{R}{2}\Big(\sup_{z\in\mathcal{X}} V_1(z) - \inf_{z\in\mathcal{X}} V_1(z)\Big)\Big\} \\ &\quad - \Big\{f(x,v) + \alpha \int_{\mathcal{X}} V_2(z)Q^o(dz|x,v) + \alpha\frac{R}{2}\Big(\sup_{z\in\mathcal{X}} V_2(z) - \inf_{z\in\mathcal{X}} V_2(z)\Big)\Big\} \\ &\overset{(a)}{=} \Big\{\alpha \int_{\mathcal{X}} V_1(z)Q^{V_1}(dz|x,v)\Big\} - \Big\{\alpha \int_{\mathcal{X}} V_2(z)Q^o(dz|x,v) + \alpha\frac{R}{2}\Big(\sup_{z\in\mathcal{X}} V_2(z) - \inf_{z\in\mathcal{X}} V_2(z)\Big)\Big\} \\ &\overset{(b)}{\leq} \Big\{\alpha \int_{\mathcal{X}} V_1(z)Q^{V_1}(dz|x,v)\Big\} - \Big\{\alpha \int_{\mathcal{X}} V_2(z)Q^{V_1}(dz|x,v)\Big\} \\ &= \alpha \int_{\mathcal{X}} (V_1(z) - V_2(z))\, Q^{V_1}(dz|x|v) \\ &\leq \alpha \sup_{z\in\mathcal{X}} |V_1(z) - V_2(z)| = \alpha||V_1 - V_2||,\end{aligned}$$

where (a) is obtained by applying (9b), with $\ell \equiv \alpha V_1$, $\nu^*(\cdot) \equiv Q^{V_1}(\cdot|\cdot)$, $\mu(\cdot) \equiv Q^o(\cdot|\cdot)$, and (b) is obtained by first applying (9b) as in (a) with Q^{V_2} and then replacing Q^{V_2} by Q^{V_1} which is suboptimal hence, the upper bound. By reversing the roles of V_1 and V_2 we get $(TV_2)(x) - (TV_1)(x) \leq \alpha||V_2 - V_1||$. Hence, $|(TV_1)(x) - (TV_2)(x)| \leq \alpha||V_1 - V_2||$ for all $x \in \mathcal{X}$, and

$$||TV_1 - TV_2|| \triangleq \max_{x\in\mathcal{X}} |(TV_1)(x) - (TV_2)(x)| \leq \alpha||V_1 - V_2||$$

which implies that $T : L \longmapsto L$ is a contraction. ■

Utilizing Lemma 3.2 we obtain the following theorem.

Theorem 3.3: Assume that $v_\infty \in BC^+(\mathcal{X})$ and $\sup_{z\in\mathcal{X}} v_\infty(z)$, $\inf_{z\in\mathcal{X}} v_\infty(z)$ are finite.

a) The dynamic programming equation

$$v_\infty(x) = \inf_{u\in\mathcal{U}(x)} \Big\{f(x,u) + \alpha \int_{\mathcal{X}} v_\infty(z)Q^o(dz|x,u) + \alpha\frac{R}{2}\Big(\sup_{z\in\mathcal{X}} v_\infty(z) - \inf_{z\in\mathcal{X}} v_\infty(z)\Big)\Big\}$$

has a unique solution.

b) Moreover,

$$v_\infty(x) = \inf_{g\in\mathcal{U}(x)} \mathbb{E}_{Q^*}\Big\{\sum_{j=0}^{\infty} \alpha^j f(x_j, u_j)|x_0 = x\Big\}.$$

c) The mapping T defined by

$$(TV)(x) = \inf_{u\in\mathcal{U}(x)} \Big\{ f(x,u) + \alpha \int_{\mathcal{X}} V(z) Q^o(dz|x,u) + \alpha \frac{R}{2} \Big(\sup_{z\in\mathcal{X}} V(z) - \inf_{z\in\mathcal{X}} V(z) \Big) \Big\}$$

is a contraction mapping with respect to the norm $||V|| = \max_{x\in\mathcal{X}} |V(x)|$.

d) For any V, $\lim_{n\to\infty} ||T^n V - v_\infty|| = 0$ and so

$$\lim_{n\longrightarrow\infty} (T^n V)(x) = v_\infty(x), \quad \text{for all} \quad x \in \mathcal{X}.$$

Proof: a) Follows from [15] (Theorem 6.3.6, part (a)).

b) We need to show that $v_\infty(x)$ is the minimum value of $\mathbb{E}_{Q^*}\Big\{\sum_{j=0}^{\infty} \alpha^j f(x_j,u_j)\Big\}$ starting in state $x_0 = x$. Recall that $0 \le f(x,u) \le M$ for all $x \in \mathcal{X}$, $u \in \mathcal{U}(x)$. Clearly, with $x_0 = x$ and for all n,

$$\inf_{g\in\mathcal{U}(x)} \mathbb{E}_{Q^*}\Big\{\sum_{j=0}^{\infty} \alpha^j f(x_j,u_j)\Big\} \ge \inf_{g\in\mathcal{U}(x)} \mathbb{E}_{Q^*}\Big\{\sum_{j=0}^{n-1} \alpha^j f(x_j,u_j)\Big\} = v_n(x).$$

Hence, $\inf_{g\in\mathcal{U}(x)} \mathbb{E}_{Q^*}\Big\{\sum_{j=0}^{\infty} \alpha^j f(x_j,u_j)\Big\} \ge \lim_{n\to\infty} v_n(x) = v_\infty(x)$. Conversely, for all n

$$\begin{aligned} &\inf_{g\in\mathcal{U}(x)} \mathbb{E}_{Q^*}\Big\{\sum_{j=0}^{\infty} \alpha^j f(x_j,u_j)\Big\} \\ &\le \inf_{g\in\mathcal{U}(x)} \mathbb{E}_{Q^*}\Big\{\sum_{j=0}^{n-1} \alpha^j f(x_j,u_j)\Big\} + \sum_{j=n}^{\infty} \alpha^j M \\ &= v_n(x) + \frac{\alpha^n M}{1-\alpha} \end{aligned}$$

and so

$$\inf_{g\in\mathcal{U}(x)} \mathbb{E}_{Q^*}\Big\{\sum_{j=0}^{\infty} \alpha^j f(x_j,u_j)\Big\} \le \lim_{n\to\infty} \Big[v_n(x) + \frac{\alpha^n M}{1-\alpha}\Big] = v_\infty(x).$$

Hence, $\inf_{g\in\mathcal{U}(x)} \mathbb{E}_{Q^*}\Big\{\sum_{j=0}^{\infty} \alpha^j f(x_j,u_j)\Big\} = v_\infty(x)$.

c) This follows from Lemma 3.2.

d) Follows from [15] (Theorem 6.3.6, part (b)). ■

1) Policy Iteration Algorithm: Next, we present a policy iteration algorithm in which the policy improvement and policy evaluation steps must be performed using the maximizing conditional distribution obtained under total variation distance ambiguity constraint. Hence, in addition to the classical case (see, for example, [1]), in which the policy improvement and evaluation steps are performed using the nominal conditional distribution, here, under the assumption that $f(\cdot)$ is bounded and non-negative, we propose a modified algorithm which is expected to converge to a stationary policy in a finite number of iterations, since both state space $\mathcal{X}$ and control space $\mathcal{U}$ are finite sets, and that at each iteration a better stationary policy will be obtained.

First, we introduce some notation. Since the state space $\mathcal{X}$ is a finite set, with say, n elements, any function $V : \mathcal{X} \longrightarrow \mathbb{R}^n$ may be represented by vector in $\mathbb{R}^n$ defined by

$$V(x) \triangleq \begin{pmatrix} V(x_1) & \cdots & V(x_n) \end{pmatrix}^T \in \mathbb{R}^n.$$

Write $z \le y$, if $z(i) \le y(i)$, for $\forall i \in \mathbb{Z}^n \triangleq \{1,2,\ldots,n\}$; and $z < y$ if $z \le y$ and $z \ne y$. For a stationary control law g, let $f(g) = (f(x_1,g(x_1)) \quad \cdots \quad f(x_n,g(x_n))^T$, and define each entry of the transition matrix $Q^o(g) \in \mathbb{R}^{n\times n}$ by $Q^o_{ij}(g) = Q^o(x_j|x_i,g(x_i)) \equiv Q^{g,o}(x_i|x_j)$. Rewrite (23) (with $\sup_{z\in\mathcal{X}} V(z)$ denoting componentwise supremum, and similarly for the infimum) as

$$TV = \min_{g\in\mathbb{R}^n} \Big\{ f(g) + \alpha Q^o(g)V + \alpha \frac{R}{2} \{ \sup_{z\in\mathcal{X}} V(z) - \inf_{z\in\mathcal{X}} V(z) \} \Big\}$$

which is equivalent to

$$TV = \min_{g\in\mathbb{R}^n} \Big\{ f(g) + \alpha Q^*(g)V \Big\}$$

where $Q^*(g) \in \mathbb{R}^{n\times n}$ is given by (16)-(18). Note that, the minimization is taken componentwise, i.e., $g(x_1)$ is the minimum of the first component of $f(g) + \alpha Q^*(g)V$ and so on. For each stationary policy g, define $T(g) : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ by

$$T(g)V = f(g) + \alpha Q^*(g)V.$$

Then, $T(g)$ is a contraction mapping on the space of bounded continuous functions to itself, and from Theorem 3.3 it follows that

$$V(g) = T(g)V = f(g) + \alpha Q^*(g)V$$

has a unique solution $V(g) \in \mathbb{R}^n$. Next, we give the policy iteration algorithm.

Algorithm 3.4 (Policy Iteration): Consider the notation above.

Initialization. Let $m = 0$ and select $g_0 : \mathcal{X} \longmapsto \mathcal{U}$ be an arbitrary stationary control law. Solve the equation

$$f(g_0) + \alpha Q^o(g_0) V_{Q^o}(g_0) = V_{Q^o}(g_0) \quad \text{for} \quad V_{Q^o}(g_0) \in \mathbb{R}^n.$$

Identify the support sets using (19)-(20), and construct the matrix $Q^*(g_0)$ using (16)-(18). Solve the equation

$$f(g_0) + \alpha Q^*(g_0) V_{Q^*}(g_0) = V_{Q^*}(g_0), \text{ for } V_{Q^*}(g_0) \in \mathbb{R}^n.$$

1. For $m = m+1$ while $\min_{g\in\mathbb{R}^n} \Big\{ f(g) + \alpha Q^*(g) V_{Q^*}(g_{m-1}) \Big\} < V_{Q^*}(g_{m-1})$ do:

(a) (Policy Improvement) Let $g_m \in \mathbb{R}^n$ be such that

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

(b) (Policy Evaluation) Solve the following equation for $V_{Q^o}(g_m) \in \mathbb{R}^n$

$$f(g_m) + \alpha Q^o(g_m) V_{Q^o}(g_m) = V_{Q^o}(g_m).$$

Identify the support sets using (19)-(20), and construct the matrix $Q^*(g_m)$ using (16)-(18). Solve the equation

$$f(g_m) + \alpha Q^*(g_m) V_{Q^*}(g_m) = V_{Q^*}(g_m), \quad \text{for } V_{Q^*}(g_m) \in \mathbb{R}^n.$$

2. Set $g^* = g_m$.

In the next section, we illustrate through an example how the theoretical results obtained in preceding sections are applied.

IV. Example

Here, we illustrate an application of the infinite horizon minimax problem for discounted cost, by considering the stochastic control system shown in Fig. 1, with state space $\mathcal{X} = \{1, 2, 3\}$ and control set $\mathcal{U} = \{u_1, u_2\}$.

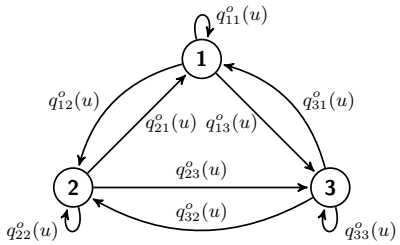

Fig. 1: Stochastic Control System

Assume 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 discount factor is $\alpha = 0.9$, the total variation distance radius is $R = 6/9$, and the cost function under each state and action is

$$f(1,u_1) = 2, \quad f(2,u_1) = 1, \quad f(3,u_1) = 3,$$
$$f(1,u_2) = 0.5, \quad f(2,u_2) = 3, \quad f(3,u_2) = 0.$$

Using policy iteration Algorithm 3.4, with initial policies $g_0(1) = u_1$, $g_0(2) = u_2$, $g_0(3) = u_2$, the algorithm converges to the following optimal policy and value after two iterations.

$$g^* = g_2 \triangleq \begin{pmatrix} g_2(1) & g_2(2) & g_2(3) \end{pmatrix} = \begin{pmatrix} u_2 & u_1 & u_2 \end{pmatrix}$$
$$V_{Q^*}(g^*) = V_{Q^*}(g_2) = \begin{pmatrix} 6.79 & 7.43 & 6.32 \end{pmatrix}.$$

Fig. 2 depicts the optimal value functions for all possible values of R, and shows that, the value functions are non-decreasing and concave functions of total variation parameter.

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 the total variation distance between the conditional distribution of the controlled process and the nominal conditional distribution.

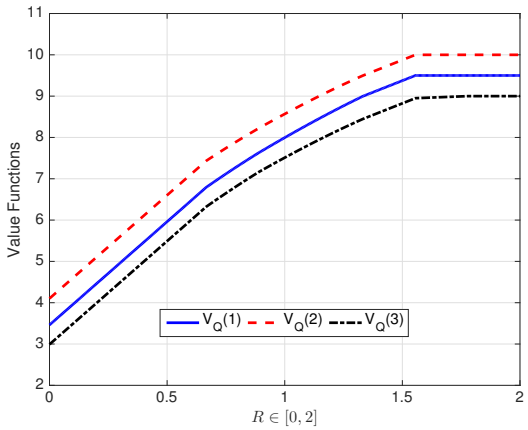

Fig. 2: Optimal Value Function

As an optimality criterion we consider the discounted cost criterion, and we derive a new dynamic programming equation. Moreover, we showed that the dynamic programming operator is contractive, and a new policy iteration algorithm is developed for computing the optimal decision strategies.

References

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

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

[3] 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.

[4] N. Ahmed, *Linear and nonlinear filtering for scientists and engineers*. Singapore, New Jersey, London, Hong Kong: World Scientific Publishers, 1999.

[5] J. S. Baras and M. Rabi, "Maximum entropy models, dynamic games, and robust output feedback control for automata," in *Proceedings of the 44th IEEE Conference on Decision and Control, and the European Control Conference*, Dec. 2005, pp. 1043–1049.

[6] T. Basar and P. Bernhard, *H-infinity optimal control and related minimax design problems: A dynamic game approach*, ser. Collection Systèmes complexes. Birkhuser, 1995.

[7] A. Bensoussan and R. Elliot, "A finite dimensional risk-sensitive control problem," *SIAM J. Control Optim.*, vol. 33, no. 6, pp. 1834–1846, 1995.

[8] C. D. Charalambous and F. Rezaei, "Stochastic uncertain systems subject to relative entropy constraints: Induced norms and monotonicity properties of minimax games," *IEEE Trans. Autom. Control*, vol. 52, no. 4, pp. 647–663, Apr. 2007.

[9] C. D. Charalambous and J. Hibey, "Minimum principle for partially observable nonlinear risk-sensitive control problems using measure-valued decompositions," *Stoch.Stoch.Rep.*, vol. 57, no. 3-4, pp. 247–288, 1996.

[10] I. Petersen, M. James, and P. Dupuis, "Minimax optimal control of stochastic uncertain systems with relative entropy constraints," *IEEE Trans. Autom. Control*, vol. 45, no. 3, pp. 398–412, Mar. 2000.

[11] P. D. Pra, L. Meneghini, and W. J. Runggaldier, "Connections between stochastic control and dynamic games," *Math. Control Signals Systems*, vol. 9, no. 4, pp. 303–326, 1996.

[12] V. Ugrinovskii and I. Petersen, "Finite horizon minimax optimal control of stochastic partially observed time varying uncertain systems," *Math. Control Signals Systems*, vol. 12, no. 1, pp. 1–23, 1999.

[13] 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.

[14] 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.

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