Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 53-2 (2020) 7166–7171

Canonical Dynamic Programming Equations subject to Ambiguity ⋆

Ioannis Tzortzis * **Charalambos D. Charalambous** *

* *Department of Electrical Engineering, University of Cyprus, Nicosia, Cyprus (e-mails: tzortzis.ioannis@ucy.ac.cy, chadcha@ucy.ac.cy).*

Abstract: This paper studies the infinite horizon average cost Markov control model subject to ambiguity on the controlled process conditional distribution. The stochastic control problem is formulated as a minimax optimization in which, (i) the existence of optimal policies is established through a pair of canonical dynamic programming equations derived for Borel state and action spaces, and (ii) the controlled process maximizing conditional distribution is characterized through a water-filling solution derived for finite state and action spaces. To obtain average cost optimal policies numerically a policy iteration algorithm is also developed. Finally, as an application of the proposed canonical dynamic programming equations, an example is provided.

Keywords: minimax stochastic control, infinite horizon, average cost, dynamic programming

1. INTRODUCTION

This paper studies the infinite horizon, average cost per unit-time, Markov Control Model (MCM) with deterministic policies, for Borel state and action spaces. The main objective is to address the problem of ambiguous controlled process conditional distributions, and study their effects on the cost-to-go, the dynamic programming recursions, and on the performance of the optimal policies.

MCMs under an average cost criterion have been studied in an anthology of papers (Arapostathis et al., 1993; Borkar, 1984; Sennott, 1995). In such MCMs the existence of optimal policies is established, through the derivation of the dynamic programming recursions, by assuming the controlled process conditional distribution is perfectly known to the control policies. In practice, precise knowledge of the controlled process conditional distribution is rarely available, since it is constructed based on modeling considerations or statistical data. MCMs subject to modeling uncertainties often deal with minimax and risk-sensitive formulations (Bensoussan and Elliot, 1995; Charalambous and Rezaei, 2007). In addition, several robustness approaches have been developed based on different types of uncertainties, i.e., using confidence intervals and moment constraints (Mannor et al., 2016; Yu and Xu, 2016). Distributionally robust approaches utilizing Wasserstein distance can also be found in (Yang, 2019; Xie, 2020). This work differs from existing works, since it is based on a family of controlled process conditional distributions, which are not necessarily absolutely continuous, and contained in a ball with respect to the Total Variation (TV) distance metric centred at a nominal conditional distribution. The emphasis on TV distance metric to model ambiguity is motivated (i) by its generality, since it applies to conditional distributions induced by linear, nonlinear, finite, countable state-space models, etc., and (ii) due to its relation via upper and lower bounds to other distances or distance metrics (Gibbs and Su, 2002).

The current paper extends the results of (Tzortzis et al., 2015; Tzortzis et al., 2019) to canonical dynamic programming equations for Borel spaces. The necessary and sufficient conditions of optimality are established for the per unit-time average cost dynamic programming, based on the concept of canonical triplets (Hernandez-Lerma and Lasserre, 1996). This treatment characterizes optimal policies without imposing any assumptions (i.e., ergodicity assumptions) on the maximizing conditional distribution. The main feature of the canonical dynamic programming equations is that, by utilizing the water-filling properties of the maximizing conditional distribution, they are able to capture the level of ambiguity in distribution, and codify the impact of incorrect distribution models on the performance of the optimal policies. To obtain average cost optimal policies numerically, a policy iteration algorithm is provided. The main feature of the proposed policy iteration algorithm (which is applied for finite state and action spaces), is that the policy evaluation and policy improvement steps are performed using the controlled process maximizing conditional distribution.

The rest of the paper is organized as follows. In Section 2, we formulate the minimax optimization problem subject to ambiguous controlled process conditional distribution. In Section 3, we study the per-unit time infinite horizon average cost Markov control model for Borel spaces, and we derive a pair of canonical dynamic programming equations. In Section 4, we consider finite alphabet spaces, and we provide the solution of the controlled process maximizing conditional distribution along with a policy iteration algorithm. In Section 5, we present an illustrative example, and in Section 6 we conclude with a brief discussion on the main results obtained in this paper.

⋆ This work was partially funded by the European Regional Development Fund and the Republic of Cyprus through the Cyprus Research and Innovation Foundation (Project: POST-DOC/0916/0139).

10.1016/j.ifacol.2020.12.533

2. PROBLEM FORMULATION

2.1 Infinite-Horizon MCM

An infinite horizon MCM with deterministic strategies is a five-tuple

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

consisting of the following:

a) State space. A Borel space $\mathcal{X}$, which models the state space of the controlled random process $\{x_k\in\mathcal{X}:k\in\mathbb{N}\}$, $\mathbb{N}\triangleq 0,1,\dots$.
b) Control (or action) space. A Borel 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 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 $x\in\mathcal{X}$. The feasible state-actions pairs are subsets of $\mathcal{X}\times\mathcal{U}$ defined by $\mathbb{K}\triangleq\{(x,u):x\in\mathcal{X},u\in\mathcal{U}(x)\}$.
d) Controlled Process Distribution. A conditional distribution or stochastic kernel $Q(z|x,u)$ on $\mathcal{X}$ given $(x,u)\in\mathbb{K}$, which corresponds to the controlled process transition probability distribution.
e) One-Stage-Cost. A non-negative function $f:\mathbb{K}\longmapsto[0,\infty]$, called the one-stage-cost.

To ensure the existence of measurable controls we make the following assumption.

Assumption 1. (Hernandez-Lerma and Lasserre, 1996) $\mathbb{K}$ contains the graph of measurable functions from $\mathcal{X}$ to $\mathcal{U}$; that is, there is a measurable function $\varphi:\mathcal{X}\longmapsto\mathcal{U}$ such that $\varphi(x)\in\mathcal{U}(x)$, for all $x\in\mathcal{X}$. The set of all such functions denoted by $\mathbb{F}$ are called selectors of the multifunction $x\longmapsto\mathcal{U}(x)$.

We equip the spaces $\mathcal{X}$ and $\mathcal{U}$ with the natural σ-algebra $\mathcal{B}(\mathcal{X})$ and $\mathcal{B}(\mathcal{U})$, respectively. For measurable spaces $(\mathcal{X},\mathcal{B}(\mathcal{X})),(\mathcal{U},\mathcal{B}(\mathcal{U}))$, we denote the set of stochastic Kernels on $(\mathcal{X},\mathcal{B}(\mathcal{X}))$ conditioned on $\mathbb{K}$ by $\mathcal{Q}(\mathcal{X}|\mathbb{K})$, and we denote the set of probability distributions on $(\mathcal{X},\mathcal{B}(\mathcal{X}))$ by $\mathcal{M}_1(\mathcal{X})$. Next, we give the definition of deterministic stationary Markov control policies.

Definition 2. A deterministic stationary Markov control policy is a function $g:\mathcal{X}\longmapsto\mathcal{U}$ such that $g(x_t)\in\mathcal{U}(x_t)$, $\forall x_t\in\mathcal{X}$, $t=0,1,\dots$. The set of such deterministic stationary Markov policies is denoted by G_{SM}, and the set of all deterministic control policies (i.e., non-stationary, possibly non-Markov) is denoted by G.

2.2 Total Variation Distance Ambiguity Class

The Total Variation (TV) distance between two probability measures $||\cdot||_{TV}:\mathcal{M}_1(\mathcal{X})\times\mathcal{M}_1(\mathcal{X})\longmapsto[0,\infty]$, is defined by

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

where $\mathcal{P}(\mathcal{X})$ denotes the collection of all finite partitions of $\mathcal{X}$. In this paper, we derive new dynamic programming equations, for the class of conditional distributions $Q(z|x,u)$, $(x,u)\in\mathbb{K}$ which are stationary, and belong to a ball, with respect to TV distance metric, centered at a nominal controlled process distribution $Q^o(z|x,u)$, $(x,u)\in\mathbb{K}$, with radius $R(x)\in[0,2]$, $x\in\mathcal{X}$.

The precise definition is the following.

Definition 3. For each $g\in G_{SM}$, let $\{x_t^g:t=0,1,\dots\}$ denote the nominal controlled process, with stationary conditional distribution defined by

$$Prob(x_t\in A|x^{t-1},u^{t-1})\triangleq Q^o(A|x_{t-1},u_{t-1}),\ \forall A\in\mathcal{B}(\mathcal{X})$$

where $Q^o(\cdot|\cdot,\cdot)\in\mathcal{Q}(\mathcal{X}|\mathbb{K})$. Given the nominal controlled process and $R(x)\in[0,2]$, $x\in\mathcal{X}$, the true stationary controlled process conditional distribution belongs to the TV distance ambiguity class defined by

$$\mathbf{B}_R(Q^o)(x,u)\triangleq\{Q(\cdot|x,u)\in\mathcal{M}_1(\mathcal{X}):\ ||Q(\cdot|x,u)-Q^o(\cdot|x,u)||_{TV}\le R(x)\},\ (x,u)\in\mathbb{K}. \quad (2)$$

2.3 Minmax Formulation

For any $g\in G$, and $Q(\cdot|x,u)\in\mathbf{B}_R(Q^o)(x,u)$, define the n-stage expected cost by

$$J_n(g,Q,x)\triangleq\mathbb{E}_x^g\Big\{\sum_{k=0}^{n-1}f(x_k,u_k)\Big\} \quad (3)$$

and the corresponding average cost per unit-time by

$$J(g,Q,x)\triangleq\limsup_{n\to\infty}\frac{1}{n}J_n(g,Q,x). \quad (4)$$

Then, the average cost per unit-time subject to ambiguity class (2) is defined by

$$J(g,Q^*,x)\triangleq\sup_{Q(\cdot|x,u)\in\mathbf{B}_R(Q^o)(x,u)}J(g,Q,x) \quad (5)$$

where Q^* denotes the maximizing element. The minimax MCP is to choose a control policy $g^*\in G$ such that

$$J(g^*,Q^*,x)\triangleq\inf_{g\in G}J(g,Q^*,x)=J^*(x),\quad\forall x\in\mathcal{X}. \quad (6)$$

A conditional distribution Q^* that satisfies (5) is called a maximizing conditional distribution, a policy g^* that satisfies (6) is called an average cost optimal policy, and the corresponding $J^*(\cdot)$ is the minimum cost or value function of the minimax MCP.

Next, we introduce an assumption for the minimax MCP defined by (6).

Assumption 4. (a) The map $f:\mathcal{X}\times\mathcal{U}\longmapsto\mathbb{R}$ is bounded, continuous and non-negative.
(b) The set $\mathcal{U}(x)$ is compact for all $x\in\mathcal{X}$.
(c) The map $Q^o(A|\cdot,\cdot)$ is continuous on $\mathbb{K}$ for every Borel set.

Note that it is possible to relax Assumption 4. For example, $f(x,\cdot)$ can be replaced by a lower semi-continuous function on $\mathcal{U}(x)$ for every $x\in\mathcal{X}$, which is non-negative (see Hernandez-Lerma and Lasserre (1996) for several relaxations).

In the next section, we derive the canonical dynamic programming equations which solves the minimax MCP.

3. MINIMAX STOCHASTIC CONTROL

Throughout this section it is assumed that Assumption 4 holds. The characterization of optimal policies for the

minimax MCP defined by (6) will be based on the concept of a canonical triplet adopted to the current formulation.

Consider the MCM (1), where $(\mathcal{X},\mathcal{U})$ are Borel spaces, and let $h : \mathcal{X} \longmapsto \mathbb{R}$ be a bounded, continuous and non-negative function. Denote the expected n-stage cost, with a terminal cost h, policy g, and $x_0 = x$, by $J_0(g, Q, x, h) = h(x)$, and for $n \geq 1$, by

$$J_n(g, Q, x, h) = J_n(g, Q, x) + \mathbb{E}_x^g\{h(x_n)\}$$

with $J_n(g, Q, x) = J_n(g, Q, x, 0)$. The corresponding maximizing expected n-stage cost is given by

$$\begin{aligned} J_n(g,x,h) &= \sup_{Q(\cdot|x,u)\in\mathbf{B}_R(Q^o)(x,u)} \mathbb{E}_x^g\Big\{\sum_{k=0}^{n-1} f(x_k,u_k)+h(x_n)\Big\} \\ &= \mathbb{E}_x^{g,Q^*}\Big\{\sum_{k=0}^{n-1} f(x_k,u_k)+h(x_n)\Big\} \\ &= J_n(g,x)+\mathbb{E}_x^{g,Q^*}\{h(x_n)\} \end{aligned} \tag{7}$$

with $J_n(g, x) = J_n(g, x, 0)$, and with $Q^*(\cdot|x, u)$ denoting the maximizing conditional distribution. Then,

$$J_n^*(x,h) = \inf_{g\in G} J_n(g,x,h); \tag{8a}$$

$$J_n^*(x) = \inf_{g\in G} J_n(g,x,h), \quad \text{if } h(\cdot)=0. \tag{8b}$$

Throughout this section it is assumed that there exists a policy $g \in G$ and an initial state $x \in \mathcal{X}$ such that $J(g, x) < \infty$ (i.e., see (5)). The definition of a canonical triplet is introduced next, following (Hernandez-Lerma and Lasserre, 1996) with a slight variation, to account for the maximizing conditional distribution over which the dynamic programming equation is expressed.

Definition 5. Let ρ and h be real-valued, bounded, continuous, non-negative, measurable functions on $\mathcal{X}$ and $\varphi \in \mathbb{F}$ a given selector. Then (ρ, h, φ) is said to be a canonical triplet if for all $x \in \mathcal{X}$ and $n = 0, 1, \ldots,$

$$J_n(g^\infty, x, h) = J_n^*(x, h) = n\rho(x) + h(x). \tag{9}$$

A selector $\varphi \in \mathbb{F}$ (of a stationary policy $g^\infty \in G_{SM}$) is called canonical if it is an element of some canonical triplet.

Note that with the appropriate choice of h as the terminal cost the policy g^∞ is optimal for the n-stage problem for all $n = 0, 1, \ldots$. The following Theorem characterizes the canonical triplets for the minimax MCP problem, with respect to the new dynamic programming equation.

Theorem 6. (ρ, h, φ) is a canonical triplet if and only if, for every $x \in \mathcal{X}$, the following holds.

$$\rho(x) = \inf_{u\in\mathcal{U}(x)} \int_{\mathcal{X}} \rho(z) Q^*(dz|x,u) \tag{a}$$

$$\rho(x)+h(x) = \inf_{u\in\mathcal{U}(x)} \Big\{ f(x,u) + \int_{\mathcal{X}} h(z) Q^*(dz|x,u)\Big\} \tag{b}$$

$\varphi(x)\in\mathcal{U}(x)$ attains the minimum in (a)-(b), that is, (c)

$$\rho(x) = \int_{\mathcal{X}} \rho(z) Q^*(dz|x,\varphi) \tag{10}$$

$$\rho(x)+h(x) = \Big\{ f(x,\varphi) + \int_{\mathcal{X}} h(z) Q^*(dz|x,\varphi)\Big\} \tag{11}$$

Proof. (Necessity). Suppose that (ρ, h, φ) is a canonical triplet, i.e., (9) holds $\forall x \in \mathcal{X}$ and $n \geq 0$. The dynamic programming equation corresponding to minimax MCP (6) is given by

$$\begin{aligned} V_j(x) = \inf_{u\in\mathcal{U}(x)} \sup_{Q(\cdot|x,u)\in\mathbf{B}_R(Q^o)(x,u)} \Big\{ f(x,u) \\ + \int_{\mathcal{X}} V_{j+1}(z) Q(dz|x,u)\Big\}. \end{aligned} \tag{12}$$

Letting $Q^*(\cdot|\cdot,\cdot)$ denoting the maximizing conditional distribution, then (12) may be written as follows.

$$V_j(x) = \inf_{u\in\mathcal{U}(x)} \Big\{ f(x,u) + \int_{\mathcal{X}} V_{j+1}(z) Q^*(dz|x,u)\Big\}. \tag{13}$$

Let us define $\overline{V}_j(x) = V_{n-j}(x)$, $(j = 0, 1, \ldots, n)$. Then, (13) may be expressed in the "forward" form

$$\overline{V}_{j+1}(x) = \inf_{u\in\mathcal{U}(x)} \Big\{ f(x,u) + \int_{\mathcal{X}} \overline{V}_j(z) Q^*(dz|x,u)\Big\}. \tag{14}$$

Substituting (14) into (7)-(8), we have

$$J_{n+1}^*(x,h) = \inf_{u\in\mathcal{U}(x)} \Big\{ f(x,u) + \int_{\mathcal{X}} J_n^*(z,h) Q^*(dz|x,u)\Big\}. \tag{15}$$

Thus, from (9) we have

$$\begin{aligned} (n+1)\rho(x)+h(x) = \inf_{u\in\mathcal{U}(x)} \Big\{ f(x,u) \\ + \int_{\mathcal{X}} \big(n\rho(z)+h(z)\big) Q^*(dz|x,u)\Big\}. \end{aligned} \tag{16}$$

Evaluating (16) at $n = 0$ we obtain (a). Furthermore, since $\rho(\cdot)$, $h(\cdot)$ and $f(\cdot,\cdot)$ are bounded, then multiplying both sides of (16) by $1/n$ and letting $n \longrightarrow \infty$ yields (b). Finally, for any deterministic stationary policy $g^\infty \in G_{SM}$, we have that

$$J_{n+1}(g^\infty, x, h) = f(x,\varphi) + \int_{\mathcal{X}} J_n(g^\infty, z, h) Q^*(dz|x,\varphi). \tag{17}$$

Thus, if $\varphi \in \mathbb{F}$ satisfies (9), then by (15)-(17) we have that

$$(n+1)\rho(x)+h(x) = f(x,\varphi)+\int_{\mathcal{X}} \big(n\rho(z)+h(z)\big) Q^*(dz|x,\varphi)$$

which, as before, gives (10) and (11).

(Sufficiency). Conversely, suppose (ρ, h, φ) satisfy (a)-(c). Proceeding by induction equation (9) is trivially satisfied when $n = 0$. Suppose that is true for some $n \geq 0$. Then, the following is obtained

$$\begin{aligned} J_{n+1}^*(x,h) &= \inf_{u\in\mathcal{U}(x)} \Big\{ f(x,u) + \int_{\mathcal{X}} \big(n\rho(z)+h(z)\big) Q^*(dz|x,u)\Big\} \\ &\geq \inf_{u\in\mathcal{U}(x)} \Big\{ f(x,u) + \int_{\mathcal{X}} h(z) Q^*(dz|x,u)\Big\} \\ &\quad + n \inf_{u\in\mathcal{U}(x)} \Big\{ \int_{\mathcal{X}} \rho(z) Q^*(dz|x,u)\Big\} = (n+1)\rho(x)+h(x). \end{aligned}$$

On the other hand,

$$\begin{aligned} &J_{n+1}^*(x,h) \leq J_{n+1}(g^\infty,x,h) \\ &= f(x,\varphi) + \int_{\mathcal{X}} \big(n\rho(z)+h(z)\big) Q^*(dz|x,\varphi) \\ &= f(x,\varphi) + \int_{\mathcal{X}} h(z) Q^*(dz|x,\varphi)\Big\} + n\int_{\mathcal{X}} \rho(z) Q^*(dz|x,\varphi) \\ &= (n+1)\rho(x)+h(x) \end{aligned}$$

This implies, $J_{n+1}^*(x,h) = J_{n+1}(g^\infty,x,h) = (n+1)\rho(x) + h(x)$. ■

Due to the fact that the average cost as an optimality criterion is underselective, (i.e., with limitations in distinguishing optimal policies with different costs), next we

introduce a more selective criterion. For other underselective and overselective optimality criteria see (Flynn, 1976, 1980).

Definition 7. A policy $g^\dagger$ is said to be

(a) (Dynkin and Yushkevich, 1979) strong average cost optimal if

$$J(g^\dagger, x) \leq \liminf_{n\to\infty} \frac{1}{n} J_n(g, x), \quad \forall g \in G,\ x \in \mathcal{X}. \tag{18}$$

(b) (Flynn, 1980) F-strong average cost optimal if

$$\lim_{n\to\infty} \frac{1}{n}\Big(J_n(g^\dagger, x) - J_n^*(x)\Big) = 0, \quad \forall x \in \mathcal{X} \tag{19}$$

where $J_n^*(x) = \inf_{g\in G} J_n(g, x)$.

Based on Definition 7, next we derive stronger results.

Theorem 8. (Hernandez-Lerma and Lasserre, 1996) Suppose the cost function f satisfies Assumption 4, and let (ρ, h, φ) be a canonical triplet (with h not necessarily bounded).

(a) If for every $g \in G$ and $x \in \mathcal{X}$

$$\lim_{n\to\infty} \mathbb{E}_x^{g,Q^*}\Big\{\frac{h(x_n)}{n}\Big\} = 0 \tag{20}$$

then g^∞ is an average cost optimal policy and ρ is the average cost value function

$$J^*(x) = \rho(x) = J(g^\infty, x) = \lim_{n\to\infty} \frac{1}{n} J_n(g^\infty, x), \quad \forall x. \tag{21}$$

(b) If for every $x \in \mathcal{X}$

$$\lim_{n\to\infty} \sup_{g\in G} \mathbb{E}_x^{g,Q^*}\Big\{\frac{h(x_n)}{n}\Big\} = 0 \tag{22}$$

then g^∞ is strong average cost optimal and F-strong average cost optimal and

$$J^*(x) = \lim_{n\to\infty} \frac{1}{n} J_n^*(x). \tag{23}$$

Proof. See (Hernandez-Lerma and Lasserre, 1996). ■

Note that, in the case in which $\rho(\cdot)$ is constant (i.e., ρ does not vary with x), then the optimality equation (a) of Theorem 6 is redundant and hence (a)-(c) reduce to

$$\rho^* + h(x) = \inf_{u\in\mathcal{U}(x)} \Big\{f(x,u) + \int_{\mathcal{X}} h(z) Q^*(dz|x,u)\Big\}$$
$$\rho^* + h(x) = f(x,\varphi) + \int_{\mathcal{X}} h(z) Q^*(dz|x,\varphi).$$

In the next section, we provide the solution of the inner optimization in minimax MCP applied for finite state and control spaces. Toward this end, we address the extremum problem of maximizing a linear functional subject to TV distance ambiguity.

4. MAXIMIZATION OVER TOTAL VARIATION DISTANCE AMBIGUITY

Let $(\mathcal{X}, \mathcal{U})$ be finite sets of cardinality $|\mathcal{X}|$ and $|\mathcal{U}|$, respectively. Define the set of conditional probability vectors on $|\mathcal{X}|$ conditioned on $x \in \mathcal{X}$, $u \in \mathcal{U}$, by

$$\mathbb{P}_{x,u}(\mathcal{X}) \triangleq \big\{P(\cdot|x,u) : P(z|x,u) \geq 0, z = 1, \ldots, |\mathcal{X}|, \sum_{z\in\mathcal{X}} P(z|x,u) = 1\big\}, \quad x \in \mathcal{X},\ u \in \mathcal{U}. \tag{24}$$

Let $\ell \triangleq \{\ell(x) : x \in \mathcal{X}\} \in \mathbb{R}_+^{|\mathcal{X}|}$ (i.e., the set of non-negative vectors of dimension $\mathcal{X}$). The precise optimization problem is the following.

Problem 9. For $\ell \in \mathbb{R}_+^{|\mathcal{X}|}$ and $Q^o(\cdot|x,u) \in \mathbb{P}_{x,u}(\mathcal{X})$, $(x,u) \in \mathcal{X}\times\mathcal{U}$, define the average pay-off by

$$\mathbb{L}_1(Q) \triangleq \sum_{z\in\mathcal{X}} \ell(z) Q(z|x,u). \tag{25}$$

The objective is to find the solution of the extremum problem

$$L(R) \triangleq \max_{Q(\cdot|x,u)\in\mathbf{B}_R(Q^o)(x,u)} \mathbb{L}_1(Q) \tag{26}$$

where

$$\mathbf{B}_R(Q^o)(x,u) \triangleq \{Q(\cdot|x,u) \in \mathbb{P}_{x,u}(\mathcal{X}) : ||Q(\cdot|x,u) - Q^o(\cdot|x,u)||_{TV} = \sum_{z\in\mathcal{X}} |Q(z|x,u) - Q^o(z|x,u)| \leq R(x)\},\ (x,u) \in \mathcal{X}\times\mathcal{U}. \tag{27}$$

Problem 9 is a convex optimization problem on the space of probability measures with the property that, $L(R)$ is a non-decreasing concave function of $R(x)$ and

$$L(R) = \sup_{Q(\cdot|x,u)\in\mathbf{B}_R(Q^o)(x,u)} \mathbb{L}_1(Q) \tag{28}$$

for values of $R(x) \leq r_{\max}(x,u)$, where $r_{\max}(x,u)$ is the smallest non-negative number belonging to $[0,2]$ such that $L(R)$ is constant in $[r_{\max}(x,u), 2]$, $(x,u) \in \mathbb{K}$. The proof of the above statement can be found in (Charalambous et al., 2014, Lemma 3.1).

Next, we recall results from (Charalambous et al., 2014), adopted to conditional distributions, concerning the characterization of the solution of Problem 9 for finite alphabet spaces. In particular, the solution of Problem 9 is obtained by first identifying the partition of $\mathcal{X}$ into disjoint sets $(\mathcal{X}^0, \mathcal{X}\setminus\mathcal{X}^0)$, and then by finding upper and lower bounds on the probabilities of $\mathcal{X}^0$ and $\mathcal{X}\setminus\mathcal{X}^0$, which are achievable.

Toward this end, let us define the maximum and minimum values of $\{\ell(x) : x \in \mathcal{X}\}$ by $\ell_{\max} \triangleq \max_{x\in\mathcal{X}} \ell(x)$ and $\ell_{\min} \triangleq \min_{x\in\mathcal{X}} \ell(x)$, respectively, and their corresponding support sets by

$$\mathcal{X}^0 \triangleq \{x \in \mathcal{X} : \ell(x) = \ell_{\max}\} \tag{29}$$
$$\mathcal{X}_0 \triangleq \{x \in \mathcal{X} : \ell(x) = \ell_{\min}\}. \tag{30}$$

For all remaining elements, $\{\ell(x) : x \in \mathcal{X}\setminus\{\mathcal{X}^0 \cup \mathcal{X}_0\}\}$, such that $\mathcal{X}^0 \cup \mathcal{X}_0 \subset \mathcal{X}$, and for $1 \leq r \leq |\mathcal{X}\setminus\{\mathcal{X}^0\cup\mathcal{X}_0\}|$, we define recursively the set of indices for which the sequence achieves its $(k+1)^{th}$ smallest value by

$$\mathcal{X}_k \triangleq \big\{x\in\mathcal{X} : \ell(x) = \min\big\{\ell(z) : z\in\mathcal{X}\setminus\mathcal{X}^0\cup\big(\bigcup_{j=1}^{k}\mathcal{X}_{j-1}\big)\big\}\big\} \tag{31}$$

for $k \in \{1, 2, \ldots, r\}$, until all the elements of $\mathcal{X}$ are exhausted. Further, we define the corresponding values of the sequence on sets $\mathcal{X}_k$ by

$$\ell(\mathcal{X}_k) \triangleq \min_{x\in\mathcal{X}\setminus\mathcal{X}^0\cup(\bigcup_{j=1}^{k}\mathcal{X}_{j-1})} \ell(x), \quad k \in \{1, 2, \ldots, r\}$$

where r is the number of $\mathcal{X}_k$ sets which is at most $|\mathcal{X}\setminus\mathcal{X}^0 \cup \mathcal{X}_0|$. Next, the solution of Problem 9 is given.

Theorem 10. The maximum pay-off (26) subject to the TV distance ambiguity is given by

$$L(R) = \ell_{\max} Q^*(\mathcal{X}^0|x,u) + \ell_{\min} Q^*(\mathcal{X}_0|x,u) + \sum_{k=1}^{r} \ell(\mathcal{X}_k) Q^*(\mathcal{X}_k|x,u) \quad (32)$$

and the maximizing conditional distribution is given by the following water-filling equations

$$Q^*(\mathcal{X}^0|x,u) = Q^o(\mathcal{X}^0|x,u) + \frac{\alpha(x,u)}{2} \quad (33a)$$

$$Q^*(\mathcal{X}_0|x,u) = \Big(Q^o(\mathcal{X}_0|x,u) - \frac{\alpha(x,u)}{2}\Big)^+ \quad (33b)$$

$$Q^*(\mathcal{X}_k|x,u) = \Big(Q^o(\mathcal{X}_k|x,u) - \Big(\frac{\alpha(x,u)}{2} - \sum_{j=1}^{k} Q^o(\mathcal{X}_{j-1}|x,u)\Big)^+\Big)^+ \quad (33c)$$

$$\alpha(x,u) = \min\Big(R(x), r_{\max}(x,u)\Big), \quad (33d)$$

where $r_{\max}(x,u) = 2(1 - Q^o(\mathcal{X}^0|x,u))$, $R(x) \in [0,2]$, $k \in \{1,2,\dots,r\}$, r is the number of $\mathcal{X}_k$ sets, and $(x)^+ \triangleq \max(0,x)$.

Proof. See (Charalambous et al., 2014, Theorem 4.1). ■

To obtain average cost optimal policies numerically next a policy iteration algorithm is proposed. In the proposed Algorithm 1 the policy evaluation and the policy improvement steps utilize the maximizing conditional distribution.

Algorithm 1 Policy Iteration

Data: $f : \mathcal{X} \times \mathcal{U} \longmapsto \mathbb{R}$ cost function, $Q^o : \mathcal{U} \longmapsto \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ nominal transition probability matrix, $R(x) \in [0,2]$, $\forall x \in \mathcal{X}$ TV distance parameter, $\mathcal{P}_{\mathcal{X}}$ family of partitions of $\mathcal{X}$ in the sense of (29) - (31), n number of partitions, $m = 0$ iteration index, $g_0 : \mathcal{X} \longmapsto \mathcal{U}$ arbitrary stationary Markov control policy.

1. (policy evaluation) For all $P(i) \in \mathcal{P}_{\mathcal{X}}$, $i = 1,2,\dots,n$, calculate $Q^{P(i)}(g_m)$ using (33), and solve
$$J_{Q^{P(i)}}(g_m) = Q^{P(i)}(g_m) J_{Q^{P(i)}}(g_m) \quad (34a)$$
$$J_{Q^{P(i)}}(g_m) + h_{Q^{P(i)}}(g_m) = f(g_m) + Q^{P(i)}(g_m) h_{Q^{P(i)}}(g_m) \quad (34b)$$
for $J_{Q^{P(i)}}(g_m) \in \mathbb{R}^{|\mathcal{X}|}$ and $h_{Q^{P(i)}}(g_m) \in \mathbb{R}^{|\mathcal{X}|}$. Identify the support sets of (34b) using (29) - (31) (with h replacing ℓ) and let $S^{P(i)}$ denote the grouping of these sets. For all $i = 1,2,\dots,n$, calculate
$$\mathcal{L}^{P(i)}(g_m) = Q^{P(i)}(g_m) h_{Q^{P(i)}}(g_m). \quad (35)$$
If
$$P(i) = \arg\max_{P \in \mathcal{P}_{\mathcal{X}}} \mathcal{L}^P(g_m), \text{ and} \quad (36a)$$
$$P(i) \text{ is consistent with } S^{P(i)} \quad (36b)$$
let $P^*(g_m) = P(i)$, $Q^*(g_m) = Q^{P^*}(g_m)$, $h_{Q^*}(g_m) = h_{Q^{P^*}}(g_m)$, $J_{Q^*}(g_m) = J_{Q^{P^*}}(g_m)$, and proceed to step 2.
2. (policy improvement) Let
$$g_{m+1} = \arg\min_{g \in \mathbb{R}^{|\mathcal{X}|}} \big\{f(g) + Q^*(g) h_{Q^*}(g_m)\big\}. \quad (37)$$
3. If $g_{m+1} = g_m$, let $g^* = g_m$; else let $m = m+1$ and return to step 1.

In the next section, an example is provided as an application of the canonical dynamic programming equations.

5. EXAMPLE – INFINITE HORIZON MINIMAX MDP

Consider a stochastic control system with $\mathcal{X} = \{0,1\}$ and control set $\mathcal{U} = \{u_1, u_2\}$. The nominal transition probabilities under controls u_1 and u_2 are given by

$$Q^o(u_1) = \frac{1}{9}\begin{pmatrix} 0 & 9 \\ 0 & 9 \end{pmatrix}, \quad Q^o(u_2) = \frac{1}{9}\begin{pmatrix} 2 & 7 \\ 3 & 6 \end{pmatrix}. \quad (38)$$

The TV distance radius is set equal to $R(x) = [6/9 \;\; 12/9]$. The cost function under each state and control is given by $f(0,u_1) = 2$, $f(0,u_2) = 0.5$, $f(1,u_1) = 1$, and $f(1,u_2) = 3$. Since $|\mathcal{X}| = 2$, then the family of partitions $P(i) \in \mathcal{P}_{\mathcal{X}}$, $i = 1,2$, in the sense of (29) - (31) is given by $P(1) = \{\mathcal{X}^0 = \{0\}, \mathcal{X}_0 = \{1\}\}$, and $P(2) = \{\mathcal{X}^0 = \{1\}, \mathcal{X}_0 = \{0\}\}$. Select (arbitrarily) the initial policies $g_0 : \mathcal{X} \longmapsto \mathcal{U}$ by $g_0(0) = u_1$, $g_0(1) = u_2$, and apply Algorithm 1.

A. Let $m = 0$ (iteration index).

1. (policy evaluation) *Under partition $P(1) \in \mathcal{P}_{\mathcal{X}}$:* From (33d),

$$\alpha(0,u_1) = \min(R(x=0), r_{\max}(0,u_1)) = \min(R(x=0), 2(1-Q^o(0|0,u_1))) = \min(\frac{6}{9}, 2)$$
$$\alpha(1,u_1) = \min(R(x=1), r_{\max}(1,u_1)) = \min(R(x=1), 2(1-Q^o(0|1,u_1))) = \min(\frac{12}{9}, 2).$$

Following a similar procedure $\alpha(0,u_2) = 6/9$, and $\alpha(1,u_2) = 12/9$. From (33a)-(33c),

$$Q^{P(1)}(g_0) = \begin{pmatrix} q^o_{00}(u_1) + \frac{\alpha(0,u_1)}{2} & \Big(q^o_{01}(u_1) - \frac{\alpha(0,u_1)}{2}\Big)^+ \\ q^o_{10}(u_2) + \frac{\alpha(1,u_2)}{2} & \Big(q^o_{11}(u_2) - \frac{\alpha(1,u_2)}{2}\Big)^+ \end{pmatrix} = \frac{1}{9}\begin{pmatrix} 3 & 6 \\ 9 & 0 \end{pmatrix}.$$

Next, we proceed to solve (34). The optimality equations are given by

$$J_{Q^{P(1)}}(g_0,0) = \frac{3}{9} J_{Q^{P(1)}}(g_0,0) + \frac{6}{9} J_{Q^{P(1)}}(g_0,1),$$
$$J_{Q^{P(1)}}(g_0,1) = J_{Q^{P(1)}}(g_0,0),$$
$$J_{Q^{P(1)}}(g_0,0) + h_{Q^{P(1)}}(g_0,0) = 2 + \frac{3}{9} h_{Q^{P(1)}}(g_0,0) + \frac{6}{9} h_{Q^{P(1)}}(g_0,1),$$
$$J_{Q^{P(1)}}(g_0,1) + h_{Q^{P(1)}}(g_0,1) = 3 + h_{Q^{P(1)}}(g_0,0).$$

Since $h_{Q^{P(1)}}(g_0)$ is uniquely determined up to an additive constant, we let $h_{Q^{P(1)}}(g_0,0) = 0$. The solution is given by

$$h_{Q^{P(1)}}(g_0) = [0 \;\; \frac{3}{5}], \quad J_{Q^{P(1)}}(g_0,0) = J_{Q^{P(1)}}(g_0,1) = \frac{12}{5}.$$

Using (29) - (31), we identify the support sets and we let $S^{P(1)}$ denote the grouping of these sets, i.e., $S^{P(1)} : \{\mathcal{X}^0 = \{1\}, \mathcal{X}_0 = \{0\}\}$. Calculating (35), we obtain $\mathcal{L}^{P(1)}(g_0) = [2/5 \;\; 0]^T$.

Under partition $P(2) \in \mathcal{P}_{\mathcal{X}}$: From (33d), $\alpha(x,u_1) = [0 \;\; 0]$, and $\alpha(x,u_2) = [4/9 \;\; 6/9]$. Applying (33a)-(33c), we obtain

$$Q^{P(2)}(g_0) = \frac{1}{9}\begin{pmatrix} 0 & 9 \\ 0 & 9 \end{pmatrix}.$$

	$Q^{P(i)}$	$J_{Q^{P(i)}}$	$h_{Q^{P(i)}}$	$S^{P(i)}$	$\mathcal{L}^{P(i)}$
i=1	$\frac{1}{9}\begin{pmatrix}5 & 4\\ 6 & 3\end{pmatrix}$	$\begin{pmatrix}7/10\\ 7/10\end{pmatrix}$	$\begin{pmatrix}0\\ 9/20\end{pmatrix}$	$\mathcal{X}^0=\{1\}$ $\mathcal{X}_0=\{0\}$	$\begin{pmatrix}4/20\\ 3/20\end{pmatrix}$
i=2	$\frac{1}{9}\begin{pmatrix}0 & 9\\ 0 & 9\end{pmatrix}$	$\begin{pmatrix}1\\ 1\end{pmatrix}$	$\begin{pmatrix}0\\ 1/2\end{pmatrix}$	$\mathcal{X}^0=\{1\}$ $\mathcal{X}_0=\{0\}$	$\begin{pmatrix}1/2\\ 1/2\end{pmatrix}$

Table 1. Policy evaluation step for $m=1$.

Next, we solve (34). The optimality equations are given by

$$J_{Q^{P(2)}}(g_0,0)=J_{Q^{P(2)}}(g_0,1),\quad J_{Q^oP(2)}(g_0,1)=J_{Q^{P(2)}}(g_0,1),$$
$$J_{Q^{P(2)}}(g_0,0)+h_{Q^{P(2)}}(g_0,0)=2+h_{Q^{P(2)}}(g_0,1),$$
$$J_{Q^{P(2)}}(g_0,1)+h_{Q^{P(2)}}(g_0,1)=3+h_{Q^{P(2)}}(g_0,1).$$

Since $h_{Q^{P(1)}}(g_0)$ is uniquely determined up to an additive constant, let $h_{Q^{P(12)}}(g_0,0)=0$. The solution is given by

$$h_{Q^{P(2)}}(g_0)=[0\ \ 1],\quad J_{Q^{P(2)}}(g_0,0)=J_{Q^{P(2)}}(g_0,1)=3.$$

Using (29) - (31), the grouping of the support sets is given by, $S^{P(2)}:\{\mathcal{X}^0=\{1\},\mathcal{X}_0=\{0\}\}$. Calculating (35), we obtain $\mathcal{L}^{P(2)}(g_0)=[1\ \ 1]^T$.

Since partition $P(2)$ is the one which satisfies (36) we let $P^*(g_0)=P(2)$, $Q^*(g_0)=Q^{P(2)}(g_0)$, $h_{Q^*}(g_0)=h_{Q^{P(2)}}(g_0)$, and $J_{Q^*}(g_0)=J_{Q^{P(2)}}(g_0)$.

2. (policy improvement) Let $g_1=\operatorname{argmin}_{g\in\mathbb{R}^2}\{f(g)+Q^*(g)h_{Q^*}(g_0)\}$, where

$$Q^*(u_1)=Q^{P(2)}(u_1)=\frac{1}{9}\begin{pmatrix}0 & 9\\ 0 & 9\end{pmatrix}=Q^*(u_2)=Q^{P(2)}(u_2). \quad (39)$$

Then,

$$\begin{aligned}&g_1(0)\\ &=\operatorname{argmin}\{f(0,u_1)+q^*_{00}(u_1)h_{Q^*}(g_0,0)+q^*_{01}(u_1)h_{Q^*}(g_0,1),\\ &\qquad f(0,u_2)+q^*_{00}(u_2)h_{Q^*}(g_0,0)+q^*_{01}(u_2)h_{Q^*}(g_0,1)\}\\ &=\operatorname{argmin}\{3,1.5\}\end{aligned}$$

and similarly, $g_1(1)=\operatorname{argmin}\{2,4\}$. Thus, $g_1(0)=u_2$, and $g_1(1)=u_1$.

3. Since $g_1\neq g_0$, let $m=1$ and return to step 1.

B. Let $m=1$ (iteration index).

1. (policy evaluation) Following similar calculations as in $m=0$, and using $g_1:\mathcal{X}\longmapsto\mathcal{U}$, the results of the policy evaluation step for all $P(i)\in\mathcal{P}_\mathcal{X}$, $i=1,2$, are summarized in Table 1. Since $P(2)$ is the partition which satisfies (36), we let $P^*(g_1)=P(2)$, $Q^*(g_1)=Q^{P(2)}(g_1)$, $h_{Q^*}(g_1)=h_{Q^{P(2)}}(g_1)$, and $J_{Q^*}(g_1)=J_{Q^{P(2)}}(g_1)$.

2. (policy improvement) Let $g_2=\operatorname{argmin}_{g\in\mathbb{R}^2}\{f(g)+Q^*(g)h_{Q^*}(g_1)\}$. Since $P^*(g_1)=P^*(g_0)=P(2)$, the maximizing transition probability matrix $Q^*(\cdot)$, under each possible control, is given by (39). Then, $g_2(0)=\operatorname{argmin}\{2.5,1\}$, and $g_2(1)=\operatorname{argmin}\{1.5,3.5\}$. Thus, $g_2(0)=u_2$, and $g_2(1)=u_1$.

3. Since, $g_2=g_1$, then $g^*=g_1$ is an optimal control policy with $J_{Q^*}=[1\ \ 1]$, and $h_{Q^*}=[0\ \ 1/2]$.

6. CONCLUSION

In this paper, we studied the optimality of the minimax Markov decision problem via dynamic programming on an infinite horizon, when the ambiguity class is described by the TV distance metric. As an optimality criterion we considered the per unit-time average cost. By working on Borel spaces, we established the existence of optimal policies through a pair of canonical dynamic programming equations, and a policy iteration algorithm was provided and applied for finite alphabet spaces.

REFERENCES

Arapostathis, A., Borkar, V.S., Fernandez-Gaucherand, E., Ghosh, M.K., and Marcus, S.I. (1993). Discrete-time controlled Markov processes with average cost criterion: a survey. *SIAM J. Control Optim.*, 31(2), 282–344.

Bensoussan, A. and Elliot, R. (1995). A finite dimensional risk-sensitive control problem. *SIAM J. Control Optim.*, 33(6), 1834–1846.

Borkar, V.S. (1984). On minimum cost per unit time control of Markov chains. *SIAM J. Control Optim.*, 22(6), 965–978.

Charalambous, C. and Rezaei, F. (2007). Stochastic uncertain systems subject to relative entropy constraints: Induced norms and monotonicity properties of minimax games. *IEEE Trans. Autom. Control*, 52(4), 647–663.

Charalambous, C.D., Tzortzis, I., Loyka, S., and Charalambous, T. (2014). Extremum problems with total variation distance and their applications. *IEEE Trans. Autom. Control*, 59(9), 2353–2368.

Dynkin, E.B. and Yushkevich, A.A. (1979). *Controlled Markov processes*. Springer-Verlag, New York.

Flynn, J. (1976). Conditions for the equivalence of optimality criteria in dynamic programming. *Ann. Statist.*, 4, 936–953.

Flynn, J. (1980). On optimality criteria for dynamic programs with long finite horizons. *J. Math. Anal. Appl.*, 76, 202–208.

Gibbs, A. and Su, F. (2002). On choosing and bounding probability metrics. *Internat. Statist. Rev*, 70(3), 419–435.

Hernandez-Lerma, O. and Lasserre, J.B. (1996). *Discrete-time Markov control processes: Basic optimality criteria*. Number v. 1 in Applications of Mathematics Stochastic Modelling and Applied Probability. Springer Verlag.

Mannor, S., Mebel, O., and Xu, H. (2016). Robust MDPs with k-rectangular uncertainty. *Math. Oper. Res.*, 41(4), 1484–1509.

Sennott, L.I. (1995). Another set of conditions for average optimality in Markov control processes. *Systems and Control Letters*, 24(2), 147–151.

Tzortzis, I., Charalambous, C.D., and Charalambous, T. (2015). Infinite horizon average cost dynamic programming subject to ambiguity on conditional distribution. In *2015 54th IEEE Conference on Decision and Control (CDC)*, 7171–7176.

Tzortzis, I., Charalambous, C.D., and Charalambous, T. (2019). Infinite Horizon Average Cost Dynamic Programming subject to Total Variation Distance Ambiguity. *SIAM J. Control Optim.*, 57(4), 2843–2872.

Xie, W. (2020). On distributionally robust chance constrained programs with Wasserstein distance. *arXiv: Optimization and Control*.

Yang, I. (2019). Wasserstein distributionally robust stochastic control: A data-driven approach. *arXiv: Optimization and Control*.

Yu, P. and Xu, H. (2016). Distributionally robust counterpart in Markov decision processes. *IEEE Trans. Autom. Control*, 61, 2538–2543.