SIAM J. CONTROL OPTIM.
Vol. 57, No. 4, pp. 2843–2872

INFINITE HORIZON AVERAGE COST DYNAMIC PROGRAMMING SUBJECT TO TOTAL VARIATION DISTANCE AMBIGUITY*

IOANNIS TZORTZIS†, CHARALAMBOS D. CHARALAMBOUS†, AND THEMISTOKLIS CHARALAMBOUS‡

Abstract. We analyze the per unit-time infinite horizon average cost Markov control model, subject to a total variation distance ambiguity on the controlled process conditional distribution. This stochastic optimal control problem is formulated as a minimax optimization problem in which the minimization is over the admissible set of control strategies, while the maximization is over the set of conditional distributions which are in a ball, with respect to the total variation distance, centered at a nominal distribution. We derive two new equivalent dynamic programming equations, and a new policy iteration algorithm. The main feature of the new dynamic programming equations is that the optimal control strategies are insensitive to inaccuracies or ambiguities in the controlled process conditional distribution. The main feature of the new policy iteration algorithm is that the policy evaluation and policy improvement steps are performed using the maximizing conditional distribution, which is obtained via a water filling solution of aggregating states together to form new states. Throughout the paper, we illustrate the new dynamic programming equations and the corresponding policy iteration algorithm to various examples.

Key words. stochastic control, Markov control models, minimax, dynamic programming, average cost, infinite horizon, total variation distance, policy iteration

AMS subject classifications. 93E20, 90C39, 90C47

DOI. 10.1137/18M1210514

1. Introduction.

The infinite horizon average cost per unit-time Markov control model (MCM) with deterministic strategies is analyzed in an anthology of papers [1, 6, 7, 23]. In such MCMs, the corresponding cost-to-go and the dynamic programming recursions depend on the conditional distribution of the underlying controlled process [8]. A classical assumption in MCM applications is that the controlled process conditional distribution is perfectly known to the control strategies. In practice, precise knowledge of the controlled process conditional distribution is rarely available, because it is often constructed based on modelling considerations or statistical data. For the decision maker/planner, an issue of central importance is that of ambiguity of the MCM controlled process conditional distribution, and its impact on the optimality of the optimal decision strategies. In this paper, the term "ambiguity" is used to describe conditional distributions of the controlled process, that are described by a set of conditional distributions, which are not absolutely continuous with respect to the nominal conditional distribution, that is used to determine the optimal control strategies. To account for such ambiguity, we model the set of controlled process conditional distributions, by a ball with respect to the total variation

*Received by the editors August 29, 2018; accepted for publication (in revised form) May 28, 2019; published electronically August 21, 2019.
https://doi.org/10.1137/18M1210514
Funding: This work was co-funded by the European Regional Development Fund and the Republic of Cyprus through the Research and Innovation Foundation (Project: POST-DOC/0916/0139). The work of the third author was partly supported by Academy of Finland grant 317726.
†Electrical and Computer Engineering, University of Cyprus, 1678 Nicosia, Cyprus (tzortzis.ioannis@ucy.ac.cy, chadcha@ucy.ac.cy).
‡Department of Electrical Engineering and Automation, Aalto University, 02150 Espoo, Finland (themistoklis.charalambous@aalto.fi).

(TV) distance metric, centered at a nominal controlled process conditional distribution. The emphasis on the TV distance metric to model ambiguity in conditional distributions is motivated by its generality, since it applies to conditional distributions induced by linear, nonlinear, finite, countable state-space models, etc. Further, TV distance is related via upper and lower bounds, to many distances, such as the widely used Kullback–Leibler divergence, L_1 distance, Hellinger distance, Levy–Prohorov distance, and others [15, 16].

The main objective of this paper is to investigate the effects of the ambiguity in the controlled process conditional distribution on the cost-to-go and dynamic programming for discrete-time MCMs when the criterion is the per unit-time infinite horizon average cost. We formulate the optimal control problem using minimax optimization techniques, in which admissible control strategies are chosen to minimize the cost, while the controlled process conditional distributions are chosen to maximize the cost, from a set, described by a ball with respect to the TV distance, centered at a nominal distribution. The usefulness of such a minimax formulation lies in the characterization of the maximizing controlled process conditional distribution, which states the following: as the radius of the TV distance increases, then the maximizing conditional distribution is constructed from the nominal conditional distribution, by means of a water-filling of aggregating states together, such that the maximizing distribution is not absolutely continuous with respect to the nominal controlled process conditional distribution. This property is particularly attractive in finite state MCMs with a large number of states because the optimal control strategy is then computed based on the reduced maximizing conditional distribution. Moreover, it allows the designer to impose different levels of ambiguity on the transition probabilities of the controlled process, depending on the current state.

The literature on MCMs, subject to modelling uncertainties, is quite extensive, and includes both deterministic and stochastic control models. Such approaches are often dealt with using minimax and risk-sensitive formulations. Representative papers, although not exhaustive, are [2, 3, 4, 9, 10, 18, 21, 26] and references therein. In addition, several robustness approaches have been developed based on different types of uncertainty sets. Good overviews of the developed approaches using confidence intervals and moment constraints can be found in [28, 30]. Additional work which deals with modelling uncertainty, without incorporating any a priori conditional distribution information, using rectangular uncertainty sets, and applied for finite horizon problems, can be found in [20]. A related formulation, which deals with uncertainties of the noise in MCMs through the use of Wasserstein distance, is developed in [29], where the author made extensive use of convex optimization techniques. The fundamental difference of our work is that we employ a methodology which is not limited by the assumption that the maximizing conditional distribution of the controlled process is absolutely continuous with respect to the nominal conditional distribution of the controlled process. The current paper complements previous work [24], where the TV distance ambiguity model is investigated in the context of MCMs, with discounted pay-off. The reader may find in [24] a quantified analysis of the performance of optimal strategies compared to the optimal strategies of risk-sensitive formulations. However, as is well known, the analysis of per unit-time infinite horizon average cost dynamic programming is fundamentally different from the analysis of the discounted average cost dynamic programming. The method of deriving the second equivalent dynamic programming equation of this paper can also be used with some modifications to derive an analogous equation for the discounted pay-off of [24]. The second equivalent dynamic programming equation is rather informative, because it includes terms which

are related to the level of ambiguity in distribution, and codify the impact of incorrect distribution models on the performance of the optimal decisions.

The paper is structured as follows. We begin our analysis by considering MCMs defined on finite state and control spaces. Then, we employ certain results from [12] to characterize the maximizing conditional distribution and the corresponding dynamic programming equations. The main feature of the maximizing conditional distribution is its explicit characterization via a water-filling solution, which is similar in spirit to extremum problems encountered in information theory, such as channel capacity and lossy data compression [13]. Subsequently, we derive two new dynamic programming equations and show these are equivalent. Under the assumption that the nominal controlled process distribution is irreducible, for every stationary Markov control law the maximizing conditional distribution of the controlled process is also irreducible, the optimal control law exists, and a new policy iteration algorithm is derived. The main feature of the corresponding policy iteration algorithm is that the policy evaluation and the policy improvement steps are performed using the maximizing conditional distribution. We include examples to illustrate the water-filling properties of the maximizing distribution, its impact on the dynamic programming equation, and the choice of optimal strategies.

The remainder of the paper is organized as follows. In section 1.1, we introduce the classical infinite horizon dynamic programming equation of MCM with an average cost per unit-time optimality criterion, and we briefly discuss the main results derived in the paper. In section 2, we give some preliminary results concerning the maximization of a linear functional subject to TV distance that we apply in subsequent sections in the context of MCMs. In section 3, we study the infinite horizon average cost Markov decision problem (MDP) for finite state and control spaces, and we derive new dynamic programming recursions and the corresponding policy iteration algorithm. In section 4, we present an example which illustrates the implications of the new dynamic programming recursions on the corresponding policy iteration algorithm.

1.1. Discussion on the main results. In this section, we briefly describe the two dynamic programming derived in this paper, by first recalling the classical dynamic programming equations.

1.1.1. Formulation of infinite-horizon MCM. An infinite horizon MCM with deterministic strategies is a five-tuple

$$\Big(\mathcal{X},\mathcal{U},\{\mathcal{U}(x):x\in\mathcal{X}\},\{Q(z|x,u):(x,u)\in\mathcal{X}\times\mathcal{U}\},f\Big) \tag{1.1}$$

consisting of the following.

(a) Finite state space. The state space $\mathcal{X}$ of the controlled random process $\{x_k \in \mathcal{X} : k \in \mathbb{N}\}$, $\mathbb{N} \triangleq 0, 1, \ldots$.

(b) Finite control (or action) space. The control or action space $\mathcal{U}$ 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 nonempty 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 nonnegative function $f : \mathbb{K} \longmapsto [0, \infty]$, called the one-stage-cost.

We denote the set of stochastic kernels on $\mathcal{X}$ conditioned on $\mathbb{K}$ by $\mathcal{Q}(\mathcal{X}|\mathbb{K})$, and the set of probability distributions on $\mathcal{X}$ by $\mathcal{M}_1(\mathcal{X})$. Next, we give the definition of deterministic stationary Markov control policies.

DEFINITION 1.1. *A deterministic stationary Markov control policy is a function $g : \mathcal{X} \longmapsto \mathcal{U}$ such that $g(x_t) \in \mathcal{U}(x_t)$ $\forall x_t \in \mathcal{X}$, $t = 0, 1, \ldots$. The set of such deterministic stationary Markov policies is denoted by G_{SM}, and the set of all deterministic control policies (i.e., nonstationary, possibly non-Markov) is denoted by G.*

Define the n-stage expected cost, for a fixed $x_0 = x$, by

$$J_n^o(g, x) \triangleq \mathbb{E}_x^g \left\{ \sum_{k=0}^{n-1} f(x_k, u_k) \right\}, \tag{1.2}$$

where $\mathbb{E}_x^g\{\cdot\}$ indicates the dependence of the expectation operation on the policy $g \in G$ and $x_0 = x$. Then, the average cost per unit-time, when policy $g \in G$ is used and given $x_0 = x$, is defined by

$$J^o(g, x) \triangleq \limsup_{n \to \infty} \frac{1}{n} J_n^o(g, x). \tag{1.3}$$

The Markov control problem (MCP) is to find a control policy $g^* \in G$ such that

$$J^o(g^*, x) \triangleq \inf_{g \in G} J^o(g, x) = J^{o,*}(x) \quad \forall x \in \mathcal{X}. \tag{1.4}$$

1.1.2. The classical dynamic programming equation. For finite cardinality spaces $(\mathcal{X}, \mathcal{U})$, it is known [17, 19, 22, 27] that if for all Markov control policies $g \in G_{SM}$ the transition probability matrix $Q(z|x, u)$ is irreducible (that is, all stationary policies have at most one recurrent class), then there exists a solution $V^o : \mathcal{X} \longmapsto \mathbb{R}$ and $J^{o,*} \in \mathbb{R}$ (independent of $x \in \mathcal{X}$) such that the pair $(J^{o,*}, V^o(x))$ is the solution of the dynamic programming equation (of the infinite-horizon MCP (1.4))

$$J^{o,*} + V^o(x) = \inf_{u \in \mathcal{U}(x)} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} Q(z|x, u) V^o(z) \right\} \tag{1.5}$$

from which existence of optimal policy $g^* \in G_{SM}$ is obtained. Note that if the irreducibility condition is not satisfied (i.e., there is more than one recurrent class), then the dynamic programming equation (1.5) may not be sufficient to give the optimal policy and the minimum cost. In this case, (1.5) is replaced by the following equations [17, 22]:

$$J^{o,*}(x) = \inf_{u \in \mathcal{U}(x)} \left\{ \sum_{z \in \mathcal{X}} Q(z|x, u) J^{o,*}(z) \right\}, \tag{1.6a}$$

$$J^{o,*}(x) + V^o(x) = \inf_{u \in \mathcal{U}(x)} \left\{ f(x, u) + \sum_{z \in \mathcal{X}} Q(z|x, u) V^o(z) \right\}. \tag{1.6b}$$

We refer to (1.6a) as the first general dynamic programming equation and to (1.6b) as the second general dynamic programming equation (some authors use the term multichain, instead). Note that the pair of generalized dynamic programming equations

(1.6a)–(1.6b) solves the MCP (1.4) without imposing irreducibility of the conditional distribution of the controlled process. Since the MCP (1.4) and the dynamic programming equation (1.5) are functionals of the conditional distribution of the controlled process, then the optimal strategies $g \in G$ are obtained based on the assumption of having accurate knowledge of the conditional distribution $Q(z|x,u)$. Hence, any ambiguity or mismatch of $Q(z|x,u)$ from the true conditional distribution will affect the optimality of the strategies. Motivated by this implication, we consider the problem formulation discussed in the next section.

1.1.3. Dynamic programming equation of per unit-time infinite-horizon MCM with total variation distance ambiguity. Let $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ denote an arbitrary measurable space, and let $\mathcal{M}_1(\mathcal{X})$ denote the set of probability measures on $\mathcal{X}$. The TV distance between two probability measures is a function $||\cdot||_{TV} : \mathcal{M}_1(\mathcal{X}) \times \mathcal{M}_1(\mathcal{X}) \longmapsto [0,\infty]$, defined by [14]

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

where $\mathcal{P}(\mathcal{X})$ denotes the collection of all finite partitions of $\mathcal{X}$. Note that $||\alpha - \beta||_{TV} \leq ||\alpha||_{TV} + ||\beta||_{TV}$ and equality holds if $\alpha(\cdot)$ and $\beta(\cdot)$ are defined on nonoverlapping support sets.

In this paper, we will derive the analogues of (1.5) and (1.6a)–(1.6b) for the class of conditional distributions of the controlled process $Q(z|x,u)$, $(x,u) \in \mathbb{K}$ which are stationary, and belong to a ball, with respect to the 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 1.2. *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*

$$Prob(x_t \in A | x^{t-1}, u^{t-1}) \triangleq Q^o(A|x_{t-1}, u_{t-1}) \ \forall A \in \mathcal{B}(\mathcal{X}), \ t = 0, 1, \dots,$$

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 ball defined by

(1.7)
$$\mathbf{B}_R(Q^o)(x,u) \triangleq \{Q(\cdot|x,u) \in \mathcal{M}_1(\mathcal{X}) : ||Q(\cdot|x,u) - Q^o(\cdot|x,u)||_{TV} \leq R(x)\}, \ (x,u) \in \mathbb{K}.$$

Remark 1.3. Regarding the TV distance parameter $R(x) \in [0,2]$ $\forall x \in \mathcal{X}$, we emphasize the following two extreme cases: (1) For $R(x) = 0$ $\forall x \in \mathcal{X}$, the nominal and the true controlled process distributions are identical, i.e., $Q(\cdot|x,u) = Q^o(\cdot|x,u)$ $\forall (x,u) \in \mathbb{K}$. Intuitively, in this case we assume complete and accurate knowledge of the controlled process distribution. Consequently, the minimax MCP reduces to the classical MCP. (2) For $R(x) = 2$ $\forall x \in \mathcal{X}$, the nominal and the true controlled process distributions have nonoverlapping support sets,[1] i.e., $\text{supp}(Q(\cdot|x,u)) \cap \text{supp}(Q^o(\cdot|x,u)) = \emptyset$ $\forall (x,u) \in \mathbb{K}$. This is the largest possible distance we can have between the nominal and the true controlled process distributions.

Next, we consider the analogue of (1.4), based on Definition 1.2. For any $g \in G$

[1] $\text{supp}(f) \triangleq \{x \in \mathcal{X} | f(x) \neq 0\}$.

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

$$J_n(g,Q,x) \triangleq \mathbb{E}_x^g\Big\{\sum_{k=0}^{n-1} f(x_k,u_k)\Big\} \tag{1.8}$$

and the corresponding average cost per unit-time by

$$J(g,Q,x) \triangleq \limsup_{n\to\infty} \frac{1}{n} J_n(g,Q,x). \tag{1.9}$$

Then, the average cost per unit-time subject to ambiguity class (1.7) is defined by

$$J(g,Q^*,x) \triangleq \sup_{Q(\cdot|x,u)\in\mathbf{B}_R(Q^o)(x,u)} J(g,Q,x). \tag{1.10}$$

The minimax MCP is to choose a control policy $g^* \in G$ such that

$$J(g^*,Q^*,x) \triangleq \inf_{g\in G} J(g,Q^*,x) = J^*(x) \quad \forall x \in \mathcal{X}. \tag{1.11}$$

A conditional distribution Q^* that satisfies (1.10) is called a maximizing conditional distribution, a policy g^* that satisfies (1.11) is called an average cost optimal policy, and the corresponding $J^*(\cdot)$ is the minimum cost or value function of the minimax MCP. We derive the following main results.

New dynamic programming equations. In section 3 the main result is Theorem 3.7.

(1) *First dynamic programming equation.* Part (a) of Theorem 3.7 states the following. Given any $R(x) \in [0,2]$, $x \in \mathcal{X}$, if for all stationary Markov control policies $g \in G_{SM}$, the maximizing transition probability matrix $Q^*(g)$ is irreducible, then the dynamic programming equation corresponding to minimax MCP (1.11) is given by

$$J^* + V(x) = \min_{u\in\mathcal{U}} \Big\{ f(x,u) + \sum_{z\in\mathcal{X}} Q^*(z|x,u)V(z) \Big\}, \tag{1.12}$$

where $Q^*(\cdot|\cdot,\cdot)$ is the maximizing distribution defined on an optimal partition of the finite state space $\mathcal{X} = \mathcal{X}^0 \cup \mathcal{X}_0 \cup \cdots \cup \mathcal{X}_k$, $k \in \mathbb{N}$, and determined by a set of water-filling equations (3.22)–(3.23), based on the nominal distribution $Q^0(\cdot|\cdot,\cdot)$ and $R(x)$, $x \in \mathcal{X}$.

(2) *Second equivalent dynamic programming equation.* Part (b) of Theorem 3.7 states that dynamic programming equation (1.12) is equivalent to the following dynamic programming equation:

$$\begin{aligned} J^* + V(x) = \min_{u\in\mathcal{U}} \Big\{ & f(x,u) + \sum_{z\in\mathcal{X}} Q^o(z|x,u)V(z) \\ & + \frac{\alpha(x,u)}{2}\Big(\sup_{z\in\mathcal{X}} V(z) - \inf_{z\in\mathcal{Y}(x,u)} V(z)\Big) + \beta(\alpha(x,u)) \Big\} \end{aligned} \tag{1.13}$$

for some set $\mathcal{Y}(x,u) \subseteq \mathcal{X}$, and with $\beta(\cdot)$ being a nondecreasing function of $\alpha(x,u)$, both defined in Theorem 3.7(b). The new term entering in the right side of (1.13) is the weighted difference of the maximum and minimum values of the value function, over appropriate sets, which is weighted by the function $\alpha(x,u)$. In section 3.1, an example is included to illustrate that the two dynamic programming equations of Theorem 3.7 are equivalent.

A special case of (1.13) arises by considering a specific range of values of the TV parameter, and is given by

$$J^* + V(x) = \min_{u\in\mathcal{U}} \left\{ f(x,u) + \sum_{z\in\mathcal{X}} Q^o(z|x,u)V(z) + \frac{\alpha(x,u)}{2}\left(\sup_{z\in\mathcal{X}} V(z) - \inf_{z\in\mathcal{X}} V(z)\right)\right\},$$

where the last right-hand side term is the difference between the maximum and minimum values of the value function weighted by $\alpha(x,u)$. By the equivalence of dynamic programming equations (1.12) and (1.13), for this special case the maximizing distribution $Q^*(\cdot|x,u)$ in (1.13) satisfies $Q^*(\mathcal{X}^0|x,u) < 1$, $Q^*(\mathcal{X}_0|x,u) > 0$, and $Q^*(\mathcal{X}_k|x,u) = Q^0(\mathcal{X}_k|x,u)$, where $\mathcal{X}^0$ and $\mathcal{X}_k, k \in \mathbb{N}$ are to be defined in section 2.

New policy iteration algorithm and implementation. A new policy iteration algorithm is given in section 3.2, in which the policy evaluation and the policy improvement steps are performed by using the maximizing conditional distribution $Q^*(\cdot|\cdot)$ obtained under TV distance ambiguity. At the application level, we include an example in section 4 to illustrate the implementation of the new policy iteration algorithm.

2. Maximization over total variation distance ambiguity. In this section, we address the extremum problem of maximizing a linear functional subject to TV distance ambiguity for finite alphabet spaces. The results of this section will be applied in subsequent sections to solve the minimax MCP defined by (1.11) and to derive new dynamic programming equations. First, we introduce the extremum problem under investigation.

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

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

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

PROBLEM 2.1. *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 cost by*

$$\mathbb{L}_1(Q)(x,u) \triangleq \sum_{z\in\mathcal{X}} \ell(z)Q(z|x,u). \tag{2.1}$$

The objective is to find the solution of the extremum problem

$$L(R)(x,u) = \max_{Q(\cdot|x,u)\in\mathbb{B}_R(Q^o)(x,u)} \sum_{z\in\mathcal{X}} \ell(z)Q(z|x,u), \tag{2.2}$$

where

$$\mathbb{B}_R(Q^o)(x,u) \triangleq \Big\{ Q(\cdot|x,u) \in \mathbb{P}_{x,u}(\mathcal{X}) : ||Q(\cdot|x,u) - Q^o(\cdot|x,u)||_{TV} \triangleq \sum_{z\in\mathcal{X}} |Q(z|x,u) - Q^o(z|x,u)| \leq R(x)\Big\}, \quad (x,u) \in \mathcal{X}\times\mathcal{U}. \tag{2.3}$$

Problem 2.1 is a convex optimization problem on $\mathbb{P}_{x,u}(\mathcal{X})$ with the property that $L(R)$ is a nondecreasing concave function of $R(x)$ and

$$L(R)(x,u) = \sup_{Q(\cdot|x,u)\in\mathbb{P}_{x,u}(\mathcal{X}):||Q(\cdot|x,u)-Q^o(\cdot|x,u)||_{TV}=R(x)} \sum_{z\in\mathcal{X}} \ell(z)Q(z|x,u), \tag{2.4}$$

where $R(x) \leq r_{\max}(x,u)$ and $r_{\max}(x,u)$ is the smallest nonnegative number belonging to $[0,2]$ such that $L(R)$ is constant in $[r_{\max}(x,u),2]$, $(x,u)\in\mathcal{X}\times\mathcal{U}$. The proof of the above statement can be found in [12, Lemma 3.1].

The solution of Problem 2.1 is obtained by first identifying the partition of $\mathcal{X}$ into the 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 $\mathcal{X}\setminus\mathcal{X}^0$, respectively, which are achievable.

Define the maximum and minimum values of $\{\ell(x) : x\in\mathcal{X}\}$ by

$$\ell_{\max} \triangleq \max_{x\in\mathcal{X}} \ell(x), \qquad \ell_{\min} \triangleq \min_{x\in\mathcal{X}} \ell(x), \tag{2.5}$$

and their corresponding sets by $\mathcal{X}^0 \triangleq \{x\in\mathcal{X} : \ell(x)=\ell_{\max}\}$ and $\mathcal{X}_0 \triangleq \{x\in\mathcal{X} : \ell(x)=\ell_{\min}\}$, respectively. 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 $1\leq r\leq|\mathcal{X}\setminus\{\mathcal{X}^0\cup\mathcal{X}_0\}|$, define recursively the set of indices for which the sequence achieves its $(k+1)^{th}$ smallest value by

$$\mathcal{X}_k \triangleq \left\{x\in\mathcal{X}:\ell(x)=\min\left\{\ell(z):z\in\mathcal{X}\setminus\left\{\mathcal{X}^0\cup\left(\bigcup_{j=1}^{k}\mathcal{X}_{j-1}\right)\right\}\right\}\right\}, \quad k\in\{1,2,\ldots,r\}, \tag{2.6}$$

until all the elements of $\mathcal{X}$ are exhausted. Further, define the corresponding values of the sequence on these sets $\mathcal{X}_k$ by

$$\ell(\mathcal{X}_k) \triangleq \min_{x\in\mathcal{X}\setminus\mathcal{X}^0\cup(\bigcup_{j=1}^{k}\mathcal{X}_{j-1})} \ell(x), \qquad k\in\{1,2,\ldots,r\}. \tag{2.7}$$

The next theorem gives the solution to Problem 2.1.

THEOREM 2.2. *The maximum cost* (2.2) *subject to the TV distance ambiguity is given by*

$$L(R)(x,u) = \ell_{\max}Q^*(\mathcal{X}^0|x,u) + \ell_{\min}Q^*(\mathcal{X}_0|x,u) + \sum_{k=1}^{r}\ell(\mathcal{X}_k)Q^*(\mathcal{X}_k|x,u). \tag{2.8}$$

The maximizing distribution $Q^*(\cdot|x,u)$ is given by the following water-filling equations:

$$Q^*(\mathcal{X}^0|x,u) = Q^o(\mathcal{X}^0|x,u) + \frac{\alpha(x,u)}{2}, \tag{2.9a}$$

$$Q^*(\mathcal{X}_0|x,u) = \left(Q^o(\mathcal{X}_0|x,u) - \frac{\alpha(x,u)}{2}\right)^+, \tag{2.9b}$$

$$Q^*(\mathcal{X}_k|x,u) = \left(Q^o(\mathcal{X}_k|x,u) - \left(\frac{\alpha(x,u)}{2} - \sum_{j=1}^{k}Q^o(\mathcal{X}_{j-1}|x,u)\right)^+\right)^+, \tag{2.9c}$$

$$\alpha(x,u) = \min(R(x), r_{\max}(x,u)), \quad r_{\max}(x,u) = 2(1-Q^o(\mathcal{X}^0|x,u)), \tag{2.9d}$$

where $R(x)\in[0,2]$, $k\in\{1,2,\ldots,r\}$, r is the number of $\mathcal{X}_k$ sets, and $(x)^+ \triangleq \max\{0,x\}$.

Proof. See [12, Theorem 4.1] for the proof. ∎

Next, we give an alternative characterization of Theorem 2.2, which we will use to show that dynamic programming equations (1.12) and (1.13) are equivalent. Dynamic programming equation (1.13) is helpful to give a more intuitive interpretation of the ambiguity model of TV distance and to codify the impact of incorrect distribution models on the performance of the optimal decisions. We summarize the above discussion with a corollary.

COROLLARY 2.3. *The maximum cost subject to TV constraint given by* (2.8) *can be equivalently expressed as follows:*

(2.10)
$$L(R)(x,u)=\left(\max_{x\in\mathcal{X}}\ell(x)-\min_{x\in\mathcal{Y}(x,u)}\ell(x)\right)\frac{\alpha(x,u)}{2}+\sum_{z\in\mathcal{X}}\ell(z)Q^o(z|x,u)+\beta(\alpha(x,u)),$$

where $\alpha(x,u)\triangleq\min(R(x),r_{\max}(x,u))$ *and* $r_{\max}(x,u)\triangleq 2(1-Q^o(\mathcal{X}^0|x,u))$. *The set* $\mathcal{Y}(x,u)$ *and the function* $\beta(\alpha(x,u))$ *are determined as follows.*

(1) *If* $\alpha(x,u)\le 2\sum_{z\in\mathcal{X}_0}Q^o(z|x,u)$, *then* $\mathcal{Y}(x,u)=\mathcal{X}$ *and* $\beta(\alpha(x,u))=0$.

(2) *If*

$$\sum_{z\in\cup_{l=0}^{k-1}\mathcal{X}_l}Q^o(z|x,u)<\frac{\alpha(x,u)}{2}\le\sum_{z\in\cup_{l=0}^{k}\mathcal{X}_l}Q^o(z|x,u), \tag{2.11}$$

where $k\in\{1,2,\dots,r\}$, *then* $\mathcal{Y}(x,u)$ *and* $\beta(\alpha(x,u))$ *are given by*

$$\mathcal{Y}(x,u)\triangleq\mathcal{X}\setminus\{\mathcal{X}^0\cup(\cup_{l=0}^{k-1}\mathcal{X}_l)\}, \tag{2.12a}$$

$$\beta(\alpha(x,u))\triangleq\sum_{l=0}^{k-1}\left\{\sum_{z\in\mathcal{X}_l}Q^o(z|x,u)\left(\min_{z\in\mathcal{Y}(x,u)}\ell(z)-\min_{z\in\mathcal{X}_l}\ell(z)\right)\right\}. \tag{2.12b}$$

(3) *If* $\alpha(x,u)\ge r_{\max}(x,u)$, *then the solution is constant, in particular,* $L(R)(x,u)=\ell_{\max}$.

Proof. For the proof of Corollary 2.3 we distinguish the following three cases.

(1) Consider $R(x)\in[0,2]$, $x\in\mathcal{X}$, and assume that $\alpha(x,u)$ in (2.9d) is equal to $R(x)<r_{\max}(x,u)$. Further, assume that the maximizing distribution given by (2.9a) and (2.9b) is such that $Q^*(\mathcal{X}^0|x,u)<1$ and $Q^*(\mathcal{X}_0|x,u)>0$. Then, by (2.9c) we have $Q^*(\mathcal{X}_k|x,u)=Q^o(\mathcal{X}_k|x,u)$ for all $k=1,\dots,r$. Substituting $Q^*(\cdot|x,u)$ back to (2.8) we obtain

$$\begin{aligned}L(R)(x,u)=&\ell_{\max}\left(Q^o(\mathcal{X}^0|x,u)+\frac{\alpha(x,u)}{2}\right)+\ell_{\min}\left(Q^o(\mathcal{X}_0|x,u)-\frac{\alpha(x,u)}{2}\right)\\&+\sum_{k=1}^{r}\ell(\mathcal{X}_k)Q^o(\mathcal{X}_k|x,u)\\=&(\ell_{\max}-\ell_{\min})\frac{\alpha(x,u)}{2}+\sum_{z\in\mathcal{X}}\ell(z)Q^o(z|x,u).\end{aligned}$$

When $\alpha(x,u)\le 2\sum_{z\in\mathcal{X}_0}Q^o(z|x,u)$ holds, this corresponds to (2.10) with $\mathcal{Y}(x,u)=\mathcal{X}$ and $\beta(\alpha(x,u))=0$ (i.e., item 1 of Corollary 2.3).

(2) Consider $R(x)\in[0,2]$, $x\in\mathcal{X}$, and assume that $\alpha(x,u)=R(x)<r_{\max}(x,u)$, while the maximizing distribution given by (2.9a) is such that $Q^*(\mathcal{X}^0|x,u)<1$ and

the maximizing distribution given by (2.9b)–(2.9c) is such that $Q^*(\mathcal{X}_0|x,u) = 0$ and $Q^*(\mathcal{X}_1|x,u) = 0, \ldots, Q^*(\mathcal{X}_{k-1}|x,u) = 0$ and $Q^*(\mathcal{X}_k|x,u) > 0$, for some $k \in \{1,2,\ldots,r\}$. Then by (2.9c), we have $Q^*(\mathcal{X}_j|x,u) = Q^o(\mathcal{X}_j|x,u)$ for all $j = k+1,\ldots,r$. Substituting $Q^*(\cdot|x,u)$ into (2.8), we obtain

$$\begin{aligned}
L(R)(x,u) =& \ell_{\max}\bigg(Q^o(\mathcal{X}^0|x,u) + \frac{\alpha(x,u)}{2}\bigg) + \sum_{j=k+1}^{r} \ell(\mathcal{X}_j) Q^o(\mathcal{X}_j|x,u) \\
&+ \ell(\mathcal{X}_k)\bigg(Q^o(\mathcal{X}_k|x,u) + \sum_{j=1}^{k} Q^o(\mathcal{X}_{j-1}|x,u) - \frac{\alpha(x,u)}{2}\bigg) \\
=& (\ell_{\max} - \ell(\mathcal{X}_k))\frac{\alpha(x,u)}{2} + \sum_{z\in\mathcal{X}} \ell(z) Q^o(z|x,u) \\
&+ (\ell(\mathcal{X}_k) - \ell(\mathcal{X}_{k-1}))Q^o(\mathcal{X}_{k-1}|x,u) + \cdots + (\ell(\mathcal{X}_k) - \ell(\mathcal{X}_0))Q^o(\mathcal{X}_0|x,u).
\end{aligned}$$

This corresponds to (2.10) with $\mathcal{Y}(x,u)$ and $\beta(\alpha(x,u))$ given by (2.12a) and (2.12b), respectively, when (2.11) holds for some $k \in \{1,2,\ldots,r\}$.

(3) Consider $R(x) \in [0,2]$ and assume that $\alpha(x,u) \geq r_{\max}(x,u) = 2(1 - Q^o(\mathcal{X}^0|x,u))$. Then, by (2.4), the maximum cost $L(R)(x,u)$ is constant in $[r_{\max}(x,u), 2]$ and, hence, it is enough to choose $\alpha(x,u) = r_{\max}(x,u)$. Then, by (2.9a), $Q^*(\mathcal{X}^0|x,u) = 1$, and consequently, $Q^*(\mathcal{X} \setminus \mathcal{X}^0|x,u) = 0$. Substituting the maximizing distribution $Q^*(\cdot|x,u)$ in (2.8) we obtain that $L(R)(x,u) = \ell_{\max}$. In (2.10) it can be shown, by substituting $\alpha(x,u) = r_{\max}(x,u)$ and $k = r$, that $L(R)(x,u) = \ell_{\max}$. ∎

Remark 2.4. We note the following.

- For $\mathcal{Y}(x,u) = \mathcal{X}$ and $\beta(\alpha(x,u)) = 0$, equation (2.10) gives the oscillator seminorm.
- The first term on the right side of (2.10) measures the difference between the maximum and minimum values of $\ell(x)$, $x \in \mathcal{X}$, with respect to the optimal partition of the state space $\mathcal{X}$.
- As the TV distance increases, (i.e., see (2.11)), new terms enter in the right side of (2.10), and this has the interpretation of taking into account the impact of incorrect distribution models.
- A specific application of the results summarized in Corollary 2.3, for developing a robust linear quadratic regulator subject to disturbance variability, can be found in [25].

3. Minimax stochastic control for finite state and control spaces. In this section we employ the results of section 2 to derive the new dynamic programming equations, and the corresponding policy iteration algorithm for the infinite horizon minimax MCP defined by (1.11).

Consider the problem of minimizing the finite horizon version of (1.10) defined by

$$J_n^*(x) = \inf_{g\in G} \sup_{Q(\cdot|x,u)\in \mathbf{B}_R(Q^o)(x,u)} \mathbb{E}_x^g\bigg\{\sum_{k=0}^{n-1} f(x_k,u_k)\bigg\}. \tag{3.1}$$

Let $V : \mathcal{X} \longmapsto \mathbb{R}$ denote the value function corresponding to (3.1). Then, V satisfies

the dynamic programming equation [11, 24]

$$V_n(x) = 0 \quad \forall x \in \mathcal{X}, \tag{3.2a}$$

$$V_j(x) = \inf_{u\in\mathcal{U}(x)} \sup_{Q(\cdot|x,u)\in\mathbf{B}_R(Q^o)(x,u)} \left\{ f(x,u) + \sum_{z\in\mathcal{X}} V_{j+1}(z) Q(z|x,u) \right\}. \tag{3.2b}$$

By Corollary 2.3, the solution of the inner optimization problem is given by (2.10), with $\ell(\cdot) = V_{j+1}(\cdot)$. Hence, (3.2b) is equivalent to the following dynamic programming equation:

$$\begin{aligned} V_j(x) = \inf_{u\in\mathcal{U}(x)} \Bigg\{ & f(x,u) + \sum_{z\in\mathcal{X}} V_{j+1}(z) Q^o(z|x,u) \\ & + \frac{\alpha(x,u)}{2} \Bigg(\sup_{z\in\mathcal{X}} V_{j+1}(z) - \inf_{z\in\mathcal{Y}(x,u)} V_{j+1}(z) \Bigg) + \beta_{j+1}(\alpha(x,u)) \Bigg\}, \end{aligned} \tag{3.3}$$

where $\alpha(x,u)$, $\mathcal{Y}(x,u) \subseteq \mathcal{X}$, and $\beta_{j+1}(\alpha(x,u))$ are defined in Corollary 2.3. Moreover, by applying Theorem 2.2, then (3.2b) is equivalent to

$$V_j(x) = \inf_{u\in\mathcal{U}(x)} \left\{ f(x,u) + \sum_{z\in\mathcal{X}} V_{j+1}(z) Q^*(z|x,u) \right\}, \tag{3.4}$$

where $Q^*(\cdot|x,u)$, $(x,u) \in \mathbb{K}$ is the maximizing conditional distribution given by (2.9).

Let us define $\overline{V}_j(x) = V_{n-j}(x)$. Then, from (3.2b), $\overline{V}_j(\cdot)$ satisfies

$$\overline{V}_j(x) = \inf_{u\in\mathcal{U}(x)} \sup_{Q(\cdot|x,u)\in\mathbf{B}_R(Q^o)(x,u)} \left\{ f(x,u) + \sum_{z\in\mathcal{X}} \overline{V}_{j-1}(z) Q(z|x,u) \right\}. \tag{3.5}$$

We rewrite (3.5) as follows:

$$\begin{aligned} & \overline{V}_j(x) + \frac{1}{j}\overline{V}_j(x) \\ & = \inf_{u\in\mathcal{U}(x)} \sup_{Q(\cdot|x,u)\in\mathbf{B}_R(Q^o)(x,u)} \left\{ f(x,u) + \sum_{z\in\mathcal{X}} Q(z|x,u) \left(\overline{V}_{j-1}(z) + \frac{1}{j}\overline{V}_j(x) \right) \right\}. \end{aligned} \tag{3.6}$$

Next, we introduce the following standard assumption [19].

ASSUMPTION 3.1. *There exists a pair* $(V(\cdot), J^*)$, $V : \mathcal{X} \longmapsto \mathbb{R}$ *and* $J^* \in \mathbb{R}$, *such that*

$$\lim_{j\to\infty} \left(\overline{V}_j(x) - jJ^* \right) = V(x) \quad \forall x \in \mathcal{X}. \tag{3.7}$$

Under Assumption 3.1,

$$\lim_{j\to\infty} \frac{1}{j}\overline{V}_j(x) = J^* \quad \forall x \in \mathcal{X} \tag{3.8}$$

and the limit does not depend on $x \in \mathcal{X}$. In addition, by taking the supremum with respect to $x \in \mathcal{X}$ on both sides of (3.7), by virtue of the finite cardinality of $\mathcal{X}$, we can exchange the limit and the supremum to obtain

$$\lim_{j\to\infty} \sup_{x\in\mathcal{X}} \left(\overline{V}_j(x) - jJ^* \right) = \sup_{x\in\mathcal{X}} \lim_{j\to\infty} \left(\overline{V}_j(x) - jJ^* \right) = \sup_{x\in\mathcal{X}} V(x). \tag{3.9}$$

By Assumption 3.1 and by (3.8) we have the following identities:

$$
\begin{aligned}
J^*+V(x) &= \lim_{j\to\infty}\left(\frac{1}{j}\overline{V}_j(x)+(\overline{V}_j(x)-jJ^*)\right)\\
&\overset{(a)}{=} \lim_{j\to\infty}\inf_{u\in\mathcal{U}(x)}\sup_{Q(\cdot|x,u)\in\mathbf{B}_R(Q^o)(x,u)}\left\{f(x,u)+\sum_{z\in\mathcal{X}}Q(z|x,u)\left(\overline{V}_{j-1}(z)+\frac{1}{j}\overline{V}_j(x)\right)-jJ^*\right\}\\
&\overset{(b)}{=} \lim_{j\to\infty}\inf_{u\in\mathcal{U}(x)}\left\{f(x,u)-jJ^*+\sum_{z\in\mathcal{X}}Q^o(z|x,u)\left(\overline{V}_{j-1}(z)+\frac{1}{j}\overline{V}_j(x)\right)\right.\\
&+\frac{\alpha(x,u)}{2}\left(\sup_{z\in\mathcal{X}}\left(\overline{V}_{j-1}(z)+\frac{1}{j}\overline{V}_j(x)\right)-\inf_{z\in\mathcal{Y}(x,u)}\left(\overline{V}_{j-1}(z)+\frac{1}{j}\overline{V}_j(x)\right)\right)\\
&+\sum_{l=0}^{k-1}\left(\sum_{z\in\mathcal{X}_l}Q^o(z|x,u)\left(\inf_{z\in\mathcal{Y}(x,u)}\left(\overline{V}_{j-1}(z)+\frac{1}{j}\overline{V}_j(x)\right)\right.\right.\\
&\left.\left.\left.\quad-\inf_{z\in\mathcal{X}_l}\left(\overline{V}_{j-1}(z)+\frac{1}{j}\overline{V}_j(x)\right)\right)\right)\right\}\\
&\overset{(c)}{=} \lim_{j\to\infty}\inf_{u\in\mathcal{U}(x)}\left\{f(x,u)+\sum_{z\in\mathcal{X}}Q^o(z|x,u)\left(\overline{V}_{j-1}(z)-(j-1)J^*+\frac{1}{j}\overline{V}_j(x)-J^*\right)\right.\\
&+\frac{\alpha(x,u)}{2}\left(\sup_{z\in\mathcal{X}}(\overline{V}_{j-1}(z)-(j-1)J^*)-\inf_{z\in\mathcal{Y}(x,u)}(\overline{V}_{j-1}(z)-(j-1)J^*)\right)\\
&+\sum_{l=0}^{k-1}\left(\sum_{z\in\mathcal{X}_\ell}Q^o(z|x,u)\left(\inf_{z\in\mathcal{Y}(x,u)}(\overline{V}_{j-1}(z)-(j-1)J^*)\right.\right.\\
&\left.\left.\left.\quad-\inf_{z\in\mathcal{X}_l}(\overline{V}_{j-1}(z)-(j-1)J^*)\right)\right)\right\},
\end{aligned}
$$

where (a) is obtained by using (3.6), (b) is obtained by using the equivalent formulation (3.3), and (c) is obtained by adding and subtracting $J^*+(j-1)J^*\alpha(x,u)+k(j-1)J^*$. Since $\mathcal{U}$ and $\mathcal{X}$ are of finite cardinality we can interchange the limit and the minimization and maximization operations, to arrive to the following dynamic programming equation:

$$
\begin{aligned}
J^*+V(x) &= \min_{u\in\mathcal{U}(x)}\left\{f(x,u)+\sum_{z\in\mathcal{X}}Q^o(z|x,u)V(z)\right.\\
&+\frac{\alpha(x,u)}{2}\left(\sup_{z\in\mathcal{X}}V(z)-\inf_{z\in\mathcal{Y}(x,u)}V(z)\right)\\
&\left.+\sum_{l=0}^{k-1}\left(\sum_{z\in\mathcal{X}_l}Q^o(z|x,u)\left(\inf_{z\in\mathcal{Y}(x,u)}V(z)-\inf_{z\in\mathcal{X}_l}V(z)\right)\right)\right\}\\
&= \min_{u\in\mathcal{U}(x)}\left\{f(x,u)+\sum_{z\in\mathcal{X}}Q^o(z|x,u)V(z)+\frac{\alpha(x,u)}{2}\left(\sup_{z\in\mathcal{X}}V(z)-\inf_{z\in\mathcal{Y}(x,u)}V(z)\right)\right.\\
&\left.+\beta(\alpha(x,u))\right\}. \qquad (3.10)
\end{aligned}
$$

By Corollary 2.3, (2.10) is the solution of Problem 2.1, and hence, dynamic program-

ming equation (3.10) is equivalently expressed as follows:

$$J^* + V(x) = \min_{u\in\mathcal{U}(x)} \max_{Q(\cdot|x,u)\in\mathbf{B}_R(Q^o)(x,u)} \left\{ f(x,u) + \sum_{z\in\mathcal{X}} Q(z|x,u)V(z) \right\}. \tag{3.11}$$

Next, we state the first main theorem of this section.

THEOREM 3.2. *Suppose $\mathcal{X}$ and $\mathcal{U}$ are of finite cardinality and Assumption* 3.1 *holds. If there exists a solution (V, J^*) to the dynamic programming equation* (3.10)*, and g^* is a stationary policy such that $g^*(x)$ attains the minimum in the right-hand side of* (3.10) *for every x, then g^* is an optimal policy and J^* is the minimum average cost.*

Proof. Let $g \in G$ be any policy and $u \in \mathcal{U}(x)$. Since (V, J^*) satisfies the dynamic programming equation (3.10), which is equivalent to (3.11), and by the definition of g^*, then

$$\begin{aligned} &f(x,u)+\sum_{z\in\mathcal{X}} Q^o(z|x,u)V(z)+\frac{\alpha(x,u)}{2}\left(\max_{z\in\mathcal{X}} V(z)-\min_{z\in\mathcal{Y}(x,u)} V(z)\right)+\beta(\alpha(x,u)) \\ &= \max_{Q(\cdot|x,u)\in\mathbf{B}_R(Q^o)(x,u)} \left\{ f(x,u) + \sum_{z\in\mathcal{X}} Q(z|x,u)V(z) \right\} \\ &\geq \max_{Q(\cdot|x,g^*(x))\in\mathbf{B}_R(Q^o)(x,g^*(x))} \left\{ f(x,g^*(x)) \ + \sum_{z\in\mathcal{X}} Q(z|x,g^*(x))V(z) \right\} \\ &= J^* + V(x). \end{aligned} \tag{3.12}$$

Denoting the maximization with respect to $Q(\cdot|x,u)$ in (3.12) by $Q^*(\cdot|x,u)$ and the corresponding expectation by $\mathbb{E}^{g,Q^*}$, and taking expectation on both sides of (3.12), we have

$$\begin{aligned} \mathbb{E}^{g,Q^*}(f(x_j,u_j)) &\geq J^* + \mathbb{E}^{g,Q^*}(V(x_j)) - \mathbb{E}^{g,Q^*}\left(\sum_{z\in\mathcal{X}} Q^*(z|x_j,u_j)V(z)\right) \\ &= J^* + \mathbb{E}^{g,Q^*}(V(x_j)) - \mathbb{E}^{g,Q^*}(V(x_{j+1})). \end{aligned} \tag{3.13}$$

Then, from (1.10) we have that for all $g \in G$,

$$\begin{aligned} J(\pi) &\geq \liminf_{j\to\infty} \left(\frac{1}{j} \sum_{k=0}^{j-1} \mathbb{E}^{g,Q^*}(f(x_k,u_k)) \right) \\ &\overset{\text{(a)}}{\geq} \liminf_{j\to\infty} \left(J^* + \frac{1}{j}(\mathbb{E}^{g,Q^*}(V(x_0)) - \mathbb{E}^{g,Q^*}(V(x_j))) \right) \\ &\overset{\text{(b)}}{=} J^*, \end{aligned}$$

where (a) is obtained by using (3.13), and (b) is obtained because the last term vanishes as $j \to \infty$. Thus, $J^* \leq \inf_{g\in G} J(g,x)$. However, when g is replaced by g^*, equality holds throughout, and as a result g^* is optimal, that is, $J^* = J^*(x) = \inf_{g\in G} J(g,x)$, $g^* \in G$ is an average cost optimal policy, and J^* is the value. ∎

Remark 3.3. Theorem 3.2 extends to countable alphabet state and control spaces provided conditions are imposed to ensure lim sup can be interchanged with sup lim in (3.9).

3.1. Existence. Dynamic programming equation (3.10) and hence Theorem 3.2 are valid under Assumption 3.1. In this section, we characterize the solution of the infinite horizon minimax average cost MCM, under the standard irreducibility condition, on the transition probabilities of the controlled process. First, we introduce some notation.

Identify the state space $\mathcal{X}$ by $\mathcal{X} = \{x_1, x_2, \ldots, x_{|\mathcal{X}|}\}$ consisting of $|\mathcal{X}|$ elements. Then, any function $V : \mathcal{X} \longmapsto \mathbb{R}$ may be represented by a vector in $\mathbb{R}^{|\mathcal{X}|}$, as follows:

$$V = (\ V(x_1) \quad \cdots \quad V(x_{|\mathcal{X}|})\ )^T \in \mathbb{R}^{|\mathcal{X}|}.$$

Any stationary control policy $g \in G_{SM}$, $g : \mathcal{X} \longmapsto \mathbb{R}$, may also be identified with a $g \in \mathbb{R}^{|\mathcal{X}|}$. For any g, let $Q(g) \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ defined by $Q(g)_{ij} = P(x_{t+1} = x_i | x_t = x_j, u_t = g(x_j))$ and

$$f(g) = (\ f(x_1, g(x_1)) \quad \cdots \quad f(x_{|\mathcal{X}|}, g(x_{|\mathcal{X}|}))\ )^T \in \mathbb{R}^{|\mathcal{X}|}.$$

Let $q_0 \in \mathbb{R}^{|\mathcal{X}|}$ be defined by $q_0(x_i) \triangleq P(\{x_0 = x_i\})$, $i = 1, \ldots, |\mathcal{X}|$, and $e \stackrel{\triangle}{=} (1, \ldots, 1)^T \in \mathbb{R}^{|\mathcal{X}|}$. The maximization of the expected n-stage cost, for a fixed $q_0(x) \in \mathbb{R}^{|\mathcal{X}|}$, is given by[2]

$$J_n(g, q_0) \triangleq J_n(g, x) q_0^T(x) = \max_{Q(\cdot|x,u)\in \mathbf{B}_R(Q^o)(x,u)} \mathbb{E}^g \left\{ \sum_{k=0}^{n-1} f(x_k, u_k) \right\}$$

$$= \max_{Q(\cdot|x,u)\in \mathbf{B}_R(Q^o)(x,u)} \left\{ \sum_{k=0}^{n-1} q_0^T Q(g)^k f(g) \right\} = \max_{Q(\cdot|x,u)\in \mathbf{B}_R(Q^o)(x,u)} q_0^T \left\{ \sum_{k=0}^{n-1} Q(g)^k \right\} f(g).$$

With $Q^*(\cdot|x,u)$ denoting the maximizing conditional distribution, then

$$\max_{Q(\cdot|x,u)\in \mathbf{B}_R(Q^o)(x,u)} q_0^T \left\{ \sum_{k=0}^{n-1} Q(g)^k \right\} f(g) = q_0^T \left\{ \sum_{k=0}^{n-1} Q^*(g)^k \right\} f(g).$$

Hence, the maximizing average cost per unit-time is given by

$$J(g, q_0) = \limsup_{n\to\infty} \max_{Q(\cdot|x,u)\in \mathbf{B}_R(Q^o)(x,u)} \frac{1}{n} \mathbb{E}^g \left\{ \sum_{k=0}^{n-1} f(x_k, u_k) \right\}$$

$$= \limsup_{n\to\infty} \frac{1}{n} q_0^T \left\{ \sum_{k=0}^{n-1} Q^*(g)^k \right\} f(g). \tag{3.14}$$

Since $q_0 \in \mathbb{R}^{|\mathcal{X}|}$ and $f(g) \in \mathbb{R}^{|\mathcal{X}|}$ are independent of n, the following limit exists [19, Lemma 5.4]:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} Q^*(g)^k = Q_1^*, \tag{3.15}$$

where $Q_1^* \in \mathbb{R}_+^{|\mathcal{X}|\times|\mathcal{X}|}$ is a stochastic matrix and it is the solution of the equation $Q_1^* Q^* = Q_1^*$. In view of (3.14) and (3.15), then

$$J(g, q_0) = q_0^T Q_1^*(g) f(g). \tag{3.16}$$

[2]The notation $J_n(g, q_0)$ means that $q_0(x)$ is fixed instead of $x_0 = x$.

Next, we recall the following definition of reducible stochastic matrix from [19, page 44].

DEFINITION 3.4. *A stochastic matrix $P \in \mathbb{R}_+^{|\mathcal{X}|\times|\mathcal{X}|}$ is said to be reducible if by row and column permutations it can be placed into block upper-triangular form*

$$P = \begin{pmatrix} P_1 & P_2 \\ 0 & P_3 \end{pmatrix}, \quad \text{where } P_1, P_2 \text{ are square matrices.}$$

A stochastic matrix which is not reducible is said to be irreducible.

Note that (3.16) depends on the probability distribution q_0 of the initial state. However, if Q_1^* is assumed to be an irreducible stochastic matrix, then there exists a unique vector q such that [19, Lemma 5.7]

$$J(g, q_0) = q_0^T Q_1^*(g) f(g) = q(g)^T f(g) \equiv J(g), \tag{3.17}$$

where $q(g)$ is the unique invariant probability distribution, that is, $Q^*(g)q(g) = q(g)$, and the matrix Q_1^* has all rows equal to q. Then, the average cost per unit-time $J(g, q_0) \equiv J(g)$ is independent of the initial distribution. Hence, for the remainder of this section, we will assume that for every stationary Markov control policy $g \in G_{SM}$, the stochastic matrix $Q^*(g)$ is irreducible. The next proposition summarizes the above results.

PROPOSITION 3.5 (see [27]). *Let $g \in G_{SM}$ be a stationary Markov control policy, $g : \mathcal{X} \longmapsto \mathcal{U}$, and assume that $Q^*(g) \in \mathbb{R}_+^{|\mathcal{X}|\times|\mathcal{X}|}$ is irreducible. Then, there exists a $V(g) \in \mathbb{R}^{|\mathcal{X}|}$ such that*

$$J(g)e + V(g) = f(g) + Q^*(g)V(g). \tag{3.18}$$

Proof. See [27] for the proof. ∎

LEMMA 3.6. *Assume the following hold.*

1. *For any stationary control policy $g \in G_{SM}$, $Q^*(g) \in \mathbb{R}_+^{|\mathcal{X}|\times|\mathcal{X}|}$ is irreducible.*
2. *There exists a $g^* \in G_{SM}$ such that*

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

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

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

Proof. By Proposition 3.5(c), there exists a $V(g^*, \cdot) : \mathcal{X} \longmapsto \mathbb{R}$ and J^* such that for all $x \in \mathcal{X}$

$$J^* + V(g^*, x) = f(x, g^*(x)) + \sum_{z \in \mathcal{X}} Q^*(z|x, g^*(x)) V(g^*, z). \tag{3.19}$$

Then, for all $x \in \mathcal{X}$

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

Define $g_1 : \mathcal{X} \longmapsto \mathcal{U}$ as $g_1(x) = \text{argmin}_{u\in\mathcal{U}}\{f(x,u) + \sum_{z\in\mathcal{X}} Q^*(z|x,u)V(g^*,z)\}$. Suppose that for some $x_2 \in \mathcal{X}$ strict inequality holds in (3.19); then

$$J^* + V(g^*,x) > \min_{u\in\mathcal{U}}\left\{f(x_2,u) + \sum_{z\in\mathcal{X}} Q^*(z|x_2,u)V(g^*,z)\right\}. \tag{3.20}$$

Multiplying (3.20) by $q(g_1)(x_0) > 0$ and summing over $x_0 \in \mathcal{X}$ yields

$$\begin{aligned}
&J^* + \sum_{x_0\in\mathcal{X}} q(g_1)(x_0)V(g^*,x_0)\\
&> \min_{u\in\mathcal{U}}\left\{\sum_{x_0\in\mathcal{X}} q(g_1)(x_0)f(x_0,u) + \sum_{x_0\in\mathcal{X}} q(g_1)(x_0)\sum_{z\in\mathcal{X}} Q^*(z|x_0,u)V(g^*,z)\right\}\\
&= \sum_{x_0\in\mathcal{X}} q(g_1)(x_0)f(x_0,g_1(x_0)) + \sum_{x_0\in\mathcal{X}} q(g_1)(x_0)\sum_{z\in\mathcal{X}} Q^*(z|x_0,g_1(x_0))V(g^*,z)\\
&= J(g_1) + \sum_{z\in\mathcal{X}} q(g_1)V(g^*,z) \quad \text{by Proposition 3.5(a)}
\end{aligned}$$

which gives $J^* > J(g_1)$, contradicting assumption 2. Hence, equality holds in (3.19) for every $x \in \mathcal{X}$. ∎

Next, we state the second main theorem of this section.

THEOREM 3.7. *Assume that for all stationary Markov control policies $g \in G_{SM}$, and for a given TV parameter $R(x) \in [0,2]$, $x \in \mathcal{X}$, the maximizing transition matrix $Q^*(g)$ is irreducible. Then, the following hold.*

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

$$J^* + V(x) = \min_{u\in\mathcal{U}}\left\{f(x,u) + \sum_{z\in\mathcal{X}} Q^*(z|x,u)V(z)\right\}. \tag{3.21}$$

The maximizing conditional distribution $Q^*(\cdot|x,u)$, $(x,u) \in \mathbb{K}$ is given by

$$Q^*(\mathcal{X}^0|x,u) = Q^o(\mathcal{X}^0|x,u) + \frac{\alpha(x,u)}{2}, \tag{3.22a}$$

$$Q^*(\mathcal{X}_0|x,u) = \left(Q^o(\mathcal{X}_0|x,u) - \frac{\alpha(x,u)}{2}\right)^+, \tag{3.22b}$$

$$Q^*(\mathcal{X}_k|x,u) = \left(Q^o(\mathcal{X}_k|x,u) - \left(\frac{\alpha(x,u)}{2} - \sum_{j=1}^{k} Q^o(\mathcal{X}_{k-1}|x,u)\right)^+\right)^+, \tag{3.22c}$$

$$\alpha(x,u) = \min(R(x), r_{\max}(x,u)), \quad r_{\max}(x,u) = 2(1 - Q^o(\mathcal{X}^0|x,u)), \tag{3.22d}$$

and the corresponding support sets by

(3.23a)

$$\mathcal{X}^0 \triangleq \{x \in \mathcal{X} : V(x) = \max\{V(x) : x \in \mathcal{X}\}\},$$

(3.23b)

$$\mathcal{X}_0 \triangleq \{x \in \mathcal{X} : V(x) = \min\{V(x) : x \in \mathcal{X}\}\},$$

(3.23c)

$$\mathcal{X}_k \triangleq \left\{x \in \mathcal{X} : V(x) = \min\left\{V(\alpha) : \alpha \in \mathcal{X} \setminus \left\{\mathcal{X}^0 \cup \left(\bigcup_{j=1}^{k} \mathcal{X}_{j-1}\right)\right\}\right\}\right\},$$

where $k = 1, 2, \ldots, r$, *and* r *is the number of* $\mathcal{X}_k$ *sets which is at most* $|\mathcal{X} \setminus \mathcal{X}^0 \cup \mathcal{X}_0|$.

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

$$J^* + V(x) = \min_{u \in \mathcal{U}} \left\{ f(x,u) + \sum_{z \in \mathcal{X}} Q^o(z|x,u)V(z) + \frac{\alpha(x,u)}{2}\left(\max_{z \in \mathcal{X}} V(z) - \min_{z \in \mathcal{Y}(x,u)} V(z)\right) + \beta(\alpha(x,u))\right\}, \tag{3.24}$$

where $\max_{z \in \mathcal{X}} V(z)$ *and* $\min_{z \in \mathcal{Y}} V(z)$ *denote componentwise maximum and minimum, respectively. The support sets are given by* (3.23), *and* $\alpha(x,u)$ *is given by* (3.22d). *The set* $\mathcal{Y}(x,u)$ *and the function* $\beta(\alpha(x,u))$ *are as defined in Corollary* 2.3, *with* $V(\cdot)$ *replacing* $\ell(\cdot)$.

(c) *If* $g^*(x)$ *attains the minimum in* (3.21) *or, equivalently, in* (3.24) *for every* x, *then* g^* *is an average cost optimal policy.*

(d) *The minimum average cost is* J^*.

Proof. Theorem 3.7 is obtained by combining Theorem 3.2 and Lemma 3.6 and by applying the results of section 2. ∎

The main observation is that in specific applications one may employ either dynamic programming equation (3.21) or (3.24). Next, we present a simple example to illustrate through analytic step-by-step calculations the equivalence of the two dynamic programming equations.

Example 1. Consider a stochastic control system, with state space $\mathcal{X} = \{0, 1, 2\}$ and control space $\mathcal{U} \in \{0, 1\}$. For $R(x) \in [0, 2]$, $x \in \mathcal{X}$, choose a nominal transition probability matrix, under each possible control, given by

$$Q^o(u_0) = Q^o(u_1) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \end{pmatrix}. \tag{3.25}$$

Notice that since the transition probability matrix Q^o is independent of the control actions, i.e., $Q^o(z|x,u) = Q^o(z|x)\ \forall(z,x,u)$, then the stochastic optimal control problem is not dynamic but rather a static problem. Nevertheless, we compute the solution and demonstrate the equivalence of dynamic programming equations (3.21)

and (3.24) of Theorem 3.7. To this end, we define the sample pay-off $f(x,u) \triangleq x+u$, i.e.,

$$\begin{aligned} &f(0,0)=0, \quad f(1,0)=1, \quad f(2,0)=2,\\ &f(0,1)=1, \quad f(1,1)=2, \quad f(2,1)=3. \end{aligned}$$

(1) Solution based on first dynamic programming equation: The solution of the infinite horizon average cost dynamic programming equation (3.21) subject to TV constraint is based on the maximizing distribution given by (3.22), and the identification of the support sets given by (3.23). At this point, let us assume that the support sets are known and given by $\mathcal{X}^0=\{2\}$, $\mathcal{X}_0=\{0\}$, $\mathcal{X}_1=\{1\}$, and $r=1$. In the next section, through the new policy iteration algorithm, we will give the exact procedure for calculating the optimal support sets. In addition, since the nominal transition probability matrix (3.25) is the same for all $u\in\mathcal{U}$, for notation convenience the dependence of (3.22) on the controls will be removed. Applying (3.22) for all $x\in\mathcal{X}$ we obtain the following.
(a) For $x=0$, we have that $\alpha(x=0)=\min(R(x=0), r_{max}(x=0)) \triangleq \min(R(x=0), 2(1-Q^o(\mathcal{X}^0|x=0)))=\min(R(x=0),2)=R(x=0)\ \forall R(x=0)\in[0,2]$. By (3.22a)–(3.22c) we have

$$\begin{aligned} Q^*(\mathcal{X}^0|x=0) &= Q^o(x=2|x=0)+\frac{R(x=0)}{2}=0+\frac{R(x=0)}{2},\\ Q^*(\mathcal{X}_0|x=0) &= \left(Q^o(x=0|x=0)-\frac{R(x=0)}{2}\right)^+=0,\\ Q^*(\mathcal{X}_1|x=0) &= \left(Q^o(x=1|x=0)-\left(\frac{R(x=0)}{2}-Q^o(x=0|x=0)\right)^+\right)^+\\ &= \left(1-\frac{R(x=0)}{2}\right)^+=1-\frac{R(x=0)}{2}. \end{aligned}$$

(b) For $x=1$, we have that $\alpha(x=1)=\min(R(x=1),2(1-Q^o(\mathcal{X}^0|x=1)))=\min(R(x=1),2(1-\frac{1}{2}))=\min(R(x=1),1)\ \forall R(x=1)\in[0,2]$. The maximizing distribution is calculated, precisely as in (a), to obtain

$$\begin{aligned} Q^*(\mathcal{X}^0|x=1) &= Q^o(x=2|x=1)+\frac{\min\{R(x=1),1\}}{2}=\frac{1}{2}+\frac{\min\{R(x=1),1\}}{2},\\ Q^*(\mathcal{X}_0|x=1) &= \left(Q^o(x=0|x=1)-\frac{\min\{R(x=1),1\}}{2}\right)^+\\ &= \left(0-\frac{\min\{R(x=1),1\}}{2}\right)^+=0,\\ Q^*(\mathcal{X}_1|x=1) &= \left(Q^o(x=1|x=1)-\left(\frac{\min\{R(x=1),1\}}{2}-Q^o(x=0|x=1)\right)^+\right)^+\\ &= \frac{1}{2}-\frac{\min\{R(x=1),1\}}{2}. \end{aligned}$$

(c) For $x=2$, then $\alpha(x=2)=\min\{R(x=2),2(1-Q^o(\mathcal{X}^+|x=2))\}=\min\{R(x=2),2\}=R(x=2)\ \forall R(x=2)\in[0,2]$. The maximizing distribution is calculated by

following the same procedure as in (a) and (b) to obtain

$$\begin{aligned}
Q^*(\mathcal{X}^0|x=2) &= Q^o(x=2|x=2) + \frac{R(x=2)}{2} = \frac{R(x=2)}{2},\\
Q^*(\mathcal{X}_0|x=2) &= \left(Q^o(x=0|x=2) - \frac{R(x=2)}{2}\right)^+ = \left(1 - \frac{R(x=2)}{2}\right)^+\\
&= 1 - \frac{R(x=2)}{2},\\
Q^*(\mathcal{X}_1|x=2) &= \left(Q^o(x=1|x=2) - \left(\frac{R(x=2)}{2} - Q^o(x=0|x=2)\right)^+\right)^+\\
&= \left(0 - \left(\frac{R(x=2)}{2} - 1\right)^+\right)^+ = 0.
\end{aligned}$$

By (a), (b), and (c), the maximizing transition probability matrix is given by

(3.26)
$$Q^*(u_0) = Q^*(u_1) = \begin{pmatrix} 0 & 1 - \frac{R(x=0)}{2} & \frac{R(x=0)}{2} \\ 0 & \frac{1}{2} - \frac{\min\{R(x=1),1\}}{2} & \frac{1}{2} + \frac{\min\{R(x=1),1\}}{2} \\ 1 - \frac{R(x=2)}{2} & 0 & \frac{R(x=2)}{2} \end{pmatrix},$$

which depends on $R(x) \in [0,2]$, $x \in \mathcal{X}$. Assume $R(x) = R = \frac{1}{2}\ \forall x \in \mathcal{X}$. Next, we show under this assumption that dynamic programming equations (3.21) and (3.24) provide the same solution.

Applying dynamic programming equation (3.21), for the value of $R = \frac{1}{2}$, we immediately get that $g(x=0) = g(x=1) = g(x=2) = 0$ and

$$J^* + V(0) = \frac{3}{4}V(1) + \frac{1}{4}V(2), \qquad J^* + V(1) = 1 + \frac{1}{4}V(1) + \frac{3}{4}V(2),$$
$$J^* + V(2) = 2 + \frac{3}{4}V(0) + \frac{1}{4}V(2).$$

The optimal solution is given by

$$J^* = \frac{11}{10}, \qquad V(1) = \frac{16}{15} + V(0), \qquad V(2) = \frac{6}{5} + V(0). \tag{3.27}$$

(2) Solution based on second equivalent dynamic programming equation: Next, we employ the second equivalent dynamic programming equation given by (3.24), with $R(x) = R = \frac{1}{2}\ \forall x \in \mathcal{X}$.

(a) For $x = 0$, we already showed in (1a) that $\alpha(x=0) = R\ \forall R \in [0,2]$. Clearly, when $k = 1$, (2.11) holds, i.e.,

$$Q^o(z=0|x=0) = 0 \le \frac{\alpha(x=0)}{2} = \frac{1}{4} \le Q^o(z=0|x=0) + Q^o(z=1|x=0) = 1,$$

and hence, by (2.12) we have that

$$\begin{aligned}
\mathcal{Y}(x=0) &= \mathcal{X} \setminus \{\mathcal{X}^0 \cup \mathcal{X}_0\} = \mathcal{X}_1,\\
\beta(\alpha(x=0)) &= \sum_{z \in \mathcal{X}_0} Q^o(z|x=0)\left(\min_{z \in \mathcal{Y}(x=0)} V(z) - \min_{z \in \mathcal{X}_0} V(z)\right) = 0.
\end{aligned}$$

(b) For $x = 1$, we already showed in (1b) that $\alpha(x = 1) = \min(R, 1)\ \forall R \in [0, 2]$. Clearly, when $k = 1$, (2.11) holds, i.e.,

$$Q^o(z = 0|x = 1) = 0 \leq \frac{\alpha(x = 1)}{2} = \frac{1}{4} \leq Q^o(z = 0|x = 1) + Q^o(z = 1|x = 1) = \frac{1}{2},$$

and hence, by (2.12) we have that

$$\begin{aligned}\mathcal{Y}(x = 1) &= \mathcal{X} \setminus \{\mathcal{X}^0 \cup \mathcal{X}_0\} = \mathcal{X}_1,\\ \beta(\alpha(x = 1)) &= \sum_{z\in\mathcal{X}_0} Q^o(z|x = 1)\left(\min_{z\in\mathcal{Y}(x=1)} V(z) - \min_{z\in\mathcal{X}_0} V(z)\right) = 0.\end{aligned}$$

(c) For $x = 2$, we already showed in (1c) that $\alpha(x = 2) = \min(R, 2) = R\ \forall R \in [0, 2]$, and since $\alpha(x = 2) = \frac{1}{2} < 2\sum_{z\in\mathcal{X}_0} Q^o(z|x = 2) = 2$ we have that $\mathcal{Y}(x = 2) = \mathcal{X}$ and $\beta(\alpha(x = 2)) = 0$.

Applying the second equivalent dynamic programming equation (3.24) we immediately get that $g(x = 0) = g(x = 1) = g(x = 2) = 0$ and

$$\begin{aligned}J^* + V(0) &= V(1) + \frac{\alpha(x = 0)}{2}\left(\max_{z\in\mathcal{X}} V(z) - \min_{z\in\mathcal{Y}(x=0)} V(z)\right) + \beta(\alpha(x = 0))\\ &= V(1) + \frac{1}{4}(V(2) - V(1)),\\ J^* + V(1) &= 1 + \frac{1}{2}V(1) + \frac{1}{2}V(2) + \frac{\alpha(x = 1)}{2}\left(\max_{z\in\mathcal{X}} V(z) - \min_{z\in\mathcal{Y}(x=1)} V(z)\right)\\ &\quad + \beta(\alpha(x = 1)) = 1 + \frac{1}{2}V(1) + \frac{1}{2}V(2) + \frac{1}{4}(V(2) - V(1)),\\ J^* + V(2) &= 2 + V(0) + \frac{\alpha(x = 2)}{2}\left(\max_{z\in\mathcal{X}} V(z) - \min_{z\in\mathcal{Y}(x=2)} V(z)\right) + \beta(\alpha(x = 2))\\ &= 2 + V(0) + \frac{1}{4}(V(2) - V(0)).\end{aligned}$$

Solving the above set of equations we obtain

$$J^* = \frac{11}{10}, \qquad V(1) = \frac{16}{15} + V(0), \qquad V(2) = \frac{6}{5} + V(0). \tag{3.28}$$

Notice that (3.27) and (3.28) are identical as expected.

3.2. Policy iteration algorithm. In this section, we provide a modified version of the classical policy iteration algorithm for average cost dynamic programming [19, 27]. From part (a) of Theorem 3.7, policy evaluation and improvement steps of the policy iteration algorithm must be performed using the maximizing conditional distribution Q^*, obtained under TV distance ambiguity constraint, defined on the support sets $\mathcal{X}^0$, $\mathcal{X}_0$, and $\mathcal{X}_k$, for all $k = 1, \ldots, r$. Moreover, one needs to guarantee that for the given TV parameter $R(x)$, $x \in \mathcal{X}$, the corresponding maximizing matrix Q^* is irreducible;[3] otherwise, the policy iteration algorithm may not be sufficient to give the optimal policy and the minimum cost.

[3]The irreducibility check can be easily accomplished either by constructing the zero-pattern of Q^* or by computing the reachability matrix.

ALGORITHM 3.8. *(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,* $R(x) \in [0,2]$, $x \in \mathcal{X}$ *TV parameter,* $\mathcal{P}_{\mathcal{X}}$ *family of partitions of the state space* $\mathcal{X}$ *in the sense of* (3.23), n *number of all possible partitions,* $m = 0$ *iteration index,* $g_0 : \mathcal{X} \longmapsto \mathcal{U}$ *arbitrary stationary Markov control policy.*

1. *(policy evaluation) For all* $P(i) \in \mathcal{P}_{\mathcal{X}}$, $i = 1, 2, \ldots, n$, *calculate* $Q^{P(i)}(g_m)$ *using* (3.22), *and solve*

$$J_{Q^{P(i)}}(g_m)e + V_{Q^{P(i)}}(g_m) = f(g_m) + Q^{P(i)}(g_m)V_{Q^{P(i)}}(g_m) \tag{3.29}$$

for $J_{Q^{P(i)}}(g_m) \in \mathbb{R}$ *and* $V_{Q^{P(i)}}(g_m) \in \mathbb{R}^{|\mathcal{X}|}$. *Identify the support sets of* (3.29) *using* (3.23) *and let* $S^{P(i)}$ *denote the grouping of these sets. Calculate*

$$\mathcal{L}^{P(i)}(g_m) = Q^{P(i)}(g_m)V_{Q^{P(i)}}(g_m) \quad \forall i = 1, 2, \ldots, n. \tag{3.30}$$

If

$$P(i) = \arg\max_{P\in\mathcal{P}_{\mathcal{X}}} \mathcal{L}^{P}(g_m) \quad \textit{and} \quad P(i) \textit{ is consistent with } S^{P(i)}, \tag{3.31}$$

let $P^*(g_m) = P(i)$, $Q^*(g_m) = Q^{P^*}(g_m)$, $V_{Q^*}(g_m) = V_{Q^{P^*}}(g_m)$, $J_{Q^*}(g_m) = J_{Q^{P^*}}(g_m)$, *and proceed to step* 2.

2. *(policy improvement) Let*

$$g_{m+1} = \arg\min_{g\in\mathbb{R}^{|\mathcal{X}|}} \{f(g) + Q^*(g)V_{Q^*}(g_m)\}. \tag{3.32}$$

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

The policy iteration algorithm is valid under the assumption that for all stationary Markov control policies $g \in G_{SM}$, the maximizing transition probability matrix $Q^*(g) \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ is irreducible. Then by [5], the policy iteration algorithm yields an optimal policy in a finite number of iterations. Moreover, by the equivalence of dynamic programming equations (3.21) and (3.24), Algorithm 3.8 may be rewritten in terms of part (b) of Theorem 3.7.

Next, we apply the policy evaluation step (step 1) of Algorithm 3.8 to determine analytically the optimal support sets of Example 1.

TABLE 3.1
Family of partitions of the state space $\mathcal{X}$.

Partition	$\mathcal{X}^0$	$\mathcal{X}_0$	$\mathcal{X}_1$
$P(1)$	$\{0\}$	$\{1\}$	$\{2\}$
$P(2)$	$\{0\}$	$\{2\}$	$\{1\}$
$P(3)$	$\{1\}$	$\{0\}$	$\{2\}$
$P(4)$	$\{1\}$	$\{2\}$	$\{0\}$
$P(5)$	$\{2\}$	$\{0\}$	$\{1\}$
$P(6)$	$\{2\}$	$\{1\}$	$\{0\}$
$P(7)$	$\{0,1\}$	$\{2\}$	-
$P(8)$	$\{0,2\}$	$\{1\}$	-
$P(9)$	$\{1,2\}$	$\{0\}$	-
$P(10)$	$\{0\}$	$\{1,2\}$	-
$P(11)$	$\{1\}$	$\{0,2\}$	-
$P(12)$	$\{2\}$	$\{0,1\}$	-

Example 1 (continuing from p. 2859). Let $\mathcal{P}_{\mathcal{X}}$ denote the family of partitions of the state space $\mathcal{X}$ in the sense of (3.23), i.e., $\mathcal{P}_{\mathcal{X}} = \{P(i) : i = 1, 2, \ldots, 12\}$, where $P(i)$ stands for partition i as shown in Table 3.1. For illustration purposes let us consider partition $P(5)$, i.e., $\mathcal{X}^0 = \{2\}$, $\mathcal{X}_0 = \{0\}$, $\mathcal{X}_1 = \{1\}$. By applying (3.22) with $Q^{P(5)}$ replacing Q^*, we obtain

$$Q^{P(5)} = \frac{1}{4}\begin{pmatrix} 0 & 3 & 1 \\ 0 & 1 & 3 \\ 3 & 0 & 1 \end{pmatrix} \quad \left(\text{see (3.26) with } R = \frac{1}{2}\right). \tag{3.33}$$

Moreover, by solving dynamic programming equation (3.29), then $J^{P(5)} = \frac{11}{10}$, and $V^{P(5)} = [0 \ \frac{16}{15} \ \frac{6}{5}]^T$. Next, identify the support sets using (3.23), and let $S^{P(5)}$ denote the grouping of these sets, i.e., $S^{P(5)} : \{\mathcal{X}^0 = \{2\},\ \mathcal{X}_0 = \{0\},\ \mathcal{X}_1 = \{1\}\}$. Calculating (3.30), we obtain $\mathcal{L}^{P(5)} = Q^{P(5)} V_{Q^{P(5)}} = [\frac{11}{10} \ \frac{7}{6} \ \frac{3}{10}]^T$. Repeating the same calculations for all possible partitions $P \in \mathcal{P}_{\mathcal{X}}$, one can construct Table 3.2. Clearly, $P(5)$ is the partition which satisfies both (3.30) and (3.31).

In section 4, we illustrate the implementation of Algorithm 3.8 through an example.

3.2.1. Discussion. Part (a) of Theorem 3.7 indicates that under a stationary Markov control policy $g \in G_{SM}$, and for an irreducible maximizing distribution $Q^*(\cdot|x,u)$, there exists a solution to the dynamic programming equation (3.21). Moreover, the maximizing distribution $Q^*(\cdot|x,u)$, which is given by (3.22), is calculated based on the support sets (3.23), the nominal distribution $Q^o(\cdot|x,u)$, and the value of $R(x) \in [0,2]$, $x \in \mathcal{X}$. Hence, in order to apply the policy iteration algorithm, for each possible partition $P(i) \in \mathcal{P}(\mathcal{X})$, one needs to know in advance that for a given TV parameter and a nominal distribution $Q^o(\cdot|x,u)$, the maximizing distribution $Q^*(\cdot|x,u)$ is irreducible. Otherwise, for a given partition of the state space $\mathcal{X}$, the policy iteration algorithm may not be sufficient to give the optimal policy and the minimum cost. In particular, as we show next, if the irreducibility condition is not satisfied, then the policy iteration algorithm need not have a unique solution.

As an example (borrowed from [22]), consider a stochastic control system with state space $\mathcal{X} = \{1,2,3\}$ and control set $\mathcal{U} = \{u_1, u_2\}$. Let us assume that the stochastic control system is characterized by a nominal distribution $Q^o(u_1)$ and $Q^o(u_2)$ under controls u_1 and u_2, respectively. Given a TV parameter $R(x) \in [0,2]$, $x \in \mathcal{X}$, and an arbitrary selected partition $\bar{P}$ from the family of partitions $\mathcal{P}_{\mathcal{X}}$ of the state space $\mathcal{X}$, we assume that the maximizing distribution under controls u_1 and u_2 is given by

$$Q^{\bar{P}}(u_1) = \frac{1}{9}\begin{pmatrix} 0 & 5 & 4 \\ 0 & 9 & 0 \\ 0 & 0 & 9 \end{pmatrix}, \quad Q^{\bar{P}}(u_2) = \frac{1}{9}\begin{pmatrix} 2 & 7 & 0 \\ 3 & 6 & 0 \\ 8 & 0 & 1 \end{pmatrix}. \tag{3.34}$$

The cost function under each state and action is given by

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

Clearly, for this control system the transition probability matrix $Q^{\bar{P}}(\cdot)$, under both controls, is reducible, since the system under controls u_1 and u_2 contains more than one communication class.[4] In particular, the transition diagram for (3.34) is as shown

[4]States i and j belong to the same communication class if and only if each of these states can reach and be reached by the other.

TABLE 3.2
Solution of Example 1.

	$Q^{P(i)}$	$J_{Q^{P(i)}}$	$V_{Q^{P(i)}}$	$S^{P(i)}$	$\mathcal{L}^{P(i)}$
$P(1)$	$\frac{1}{4}\begin{bmatrix}1&3&0\\1&1&2\\4&0&0\end{bmatrix}$	4/5	$\begin{bmatrix}0\\16/15\\6/5\end{bmatrix}$	$\mathcal{X}^0=\{2\}$ $\mathcal{X}_0=\{0\}$ $\mathcal{X}_1=\{1\}$	$\begin{bmatrix}4/5\\13/15\\0\end{bmatrix}$
$P(2)$	$\frac{1}{4}\begin{bmatrix}1&3&0\\1&2&1\\4&0&0\end{bmatrix}$	18/23	$\begin{bmatrix}0\\24/23\\28/23\end{bmatrix}$	$\mathcal{X}^0=\{2\}$ $\mathcal{X}_0=\{0\}$ $\mathcal{X}_1=\{1\}$	$\begin{bmatrix}18/23\\19/23\\0\end{bmatrix}$
$P(3)$	$\frac{1}{4}\begin{bmatrix}0&4&0\\0&3&1\\3&1&0\end{bmatrix}$	24/23	$\begin{bmatrix}0\\24/23\\28/23\end{bmatrix}$	$\mathcal{X}^0=\{2\}$ $\mathcal{X}_0=\{0\}$ $\mathcal{X}_1=\{1\}$	$\begin{bmatrix}24/23\\25/23\\6/23\end{bmatrix}$
$P(4)$	$\frac{1}{4}\begin{bmatrix}0&4&0\\0&3&1\\3&1&0\end{bmatrix}$	24/23	$\begin{bmatrix}0\\24/23\\28/23\end{bmatrix}$	$\mathcal{X}^0=\{2\}$ $\mathcal{X}_0=\{0\}$ $\mathcal{X}_1=\{1\}$	$\begin{bmatrix}24/23\\25/23\\6/23\end{bmatrix}$
$P(5)$	$\frac{1}{4}\begin{bmatrix}0&3&1\\0&1&3\\3&0&1\end{bmatrix}$	11/10	$\begin{bmatrix}0\\16/15\\6/5\end{bmatrix}$	$\mathcal{X}^0=\{2\}$ $\mathcal{X}_0=\{0\}$ $\mathcal{X}_1=\{1\}$	$\begin{bmatrix}11/10\\7/6\\3/10\end{bmatrix}$
$P(6)$	$\frac{1}{4}\begin{bmatrix}0&3&1\\0&1&3\\3&0&1\end{bmatrix}$	11/10	$\begin{bmatrix}0\\16/15\\6/5\end{bmatrix}$	$\mathcal{X}^0=\{2\}$ $\mathcal{X}_0=\{0\}$ $\mathcal{X}_1=\{1\}$	$\begin{bmatrix}11/10\\7/6\\3/10\end{bmatrix}$
$P(7)$	$\frac{1}{4}\begin{bmatrix}0&4&0\\\frac{1}{2}&\frac{5}{2}&1\\4&0&0\end{bmatrix}$	12/13	$\begin{bmatrix}0\\12/13\\14/13\end{bmatrix}$	$\mathcal{X}^0=\{2\}$ $\mathcal{X}_0=\{0\}$ $\mathcal{X}_1=\{1\}$	$\begin{bmatrix}12/13\\11/13\\0\end{bmatrix}$
$P(8)$	$\frac{1}{4}\begin{bmatrix}\frac{1}{2}&3&\frac{1}{2}\\\frac{1}{2}&1&\frac{5}{2}\\4&0&0\end{bmatrix}$	10/11	$\begin{bmatrix}0\\34/33\\12/11\end{bmatrix}$	$\mathcal{X}^0=\{2\}$ $\mathcal{X}_0=\{0\}$ $\mathcal{X}_1=\{1\}$	$\begin{bmatrix}10/11\\31/33\\0\end{bmatrix}$
$P(9)$	$\frac{1}{4}\begin{bmatrix}0&4&0\\0&2&2\\3&\frac{1}{2}&\frac{1}{2}\end{bmatrix}$	15/14	$\begin{bmatrix}0\\15/14\\17/14\end{bmatrix}$	$\mathcal{X}^0=\{2\}$ $\mathcal{X}_0=\{0\}$ $\mathcal{X}_1=\{1\}$	$\begin{bmatrix}15/14\\8/7\\2/7\end{bmatrix}$
$P(10)$	$\frac{1}{4}\begin{bmatrix}1&3&0\\1&\frac{3}{2}&\frac{3}{2}\\4&0&0\end{bmatrix}$	42/53	$\begin{bmatrix}0\\56/53\\64/53\end{bmatrix}$	$\mathcal{X}^0=\{2\}$ $\mathcal{X}_0=\{0\}$ $\mathcal{X}_1=\{1\}$	$\begin{bmatrix}42/53\\45/53\\0\end{bmatrix}$
$P(11)$	$\frac{1}{4}\begin{bmatrix}0&4&0\\0&3&1\\3&1&0\end{bmatrix}$	24/23	$\begin{bmatrix}0\\24/23\\28/23\end{bmatrix}$	$\mathcal{X}^0=\{2\}$ $\mathcal{X}_0=\{0\}$ $\mathcal{X}_1=\{1\}$	$\begin{bmatrix}24/23\\25/23\\6/23\end{bmatrix}$
$P(12)$	$\frac{1}{4}\begin{bmatrix}0&3&1\\0&1&3\\3&0&1\end{bmatrix}$	11/10	$\begin{bmatrix}0\\16/15\\6/5\end{bmatrix}$	$\mathcal{X}^0=\{2\}$ $\mathcal{X}_0=\{0\}$ $\mathcal{X}_1=\{1\}$	$\begin{bmatrix}11/10\\7/6\\3/10\end{bmatrix}$

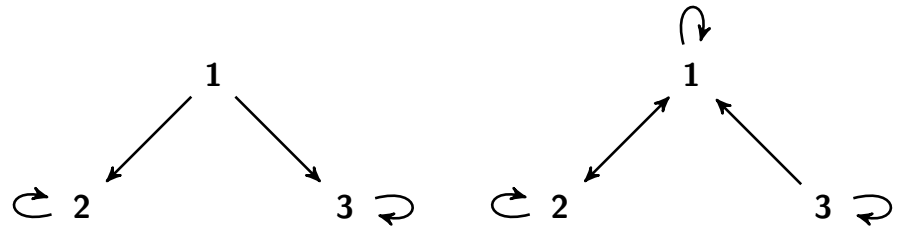
in Figure 3.1, where the zero-pattern matrix, under controls u_1 and u_2, is given by[5]

$$Z^{\bar{P}}(u_1)=\begin{pmatrix}0&1&1\\0&1&0\\0&0&1\end{pmatrix},\quad Z^{\bar{P}}(u_2)=\begin{pmatrix}1&1&0\\1&1&0\\1&0&1\end{pmatrix}. \tag{3.35}$$

Using the policy iteration algorithm, Algorithm 3.8, with initial policies $g_0(1)=g_0(2)=g_0(3)=u_1$, the optimality equation (3.29) for this system may be written as

$$J_{Q^{\bar{P}}(g_0)}\begin{pmatrix}1\\1\\1\end{pmatrix}+V_{Q^{\bar{P}}}(g_0)=\begin{pmatrix}2\\1\\3\end{pmatrix}+\frac{1}{9}\begin{pmatrix}0&5&4\\0&9&0\\0&0&9\end{pmatrix}V_{Q^{\bar{P}}}(g_0),$$

[5]The zero-pattern is constructed by setting each positive element of a transition matrix to 1 and all other elements to 0.



(a) Zero-pattern $Z^{\bar{P}}(u_1)$ with three communication classes $\{1\}$, $\{2\}$, and $\{3\}$.

(b) Zero-pattern $Z^{\bar{P}}(u_2)$ with two communication classes $\{1,2\}$ and $\{3\}$.

FIG. 3.1. *Diagram for the zero-pattern of the transition matrix under controls u_1 and u_2.*

and hence

$$\begin{aligned}
&J_{Q^{\bar{P}}} + V_{Q^{\bar{P}}}(g_0,1) = 2 + \frac{5}{9}V_{Q^{\bar{P}}}(g_0,2) + \frac{4}{9}V_{Q^{\bar{P}}}(g_0,3),\\
&J_{Q^{\bar{P}}} + V_{Q^{\bar{P}}}(g_0,2) = 1 + V_{Q^{\bar{P}}}(g_0,2) \Longrightarrow J_{Q^{\bar{P}}} = 1,\\
&J_{Q^{\bar{P}}} + V_{Q^{\bar{P}}}(g_0,3) = 3 + V_{Q^{\bar{P}}}(g_0,3) \Longrightarrow J_{Q^{\bar{P}}} = 3.
\end{aligned}$$

The second and third equations show that the system for the partition $\bar{P} \in \mathcal{P}_{\mathcal{X}}$ is inconsistent. Hence, in our proposed policy iteration algorithm for solving the minimax stochastic control problem with average cost, the above scenario of reducible maximizing distributions $Q^*(\cdot|x,u)$ is excluded, since as we showed it might lead to inconsistencies in our system.

To circumvent the irreducibility assumption on the maximizing distribution, one may proceed one step further and characterize the existence of optimal strategies by

$$\begin{aligned}
J^*(x) = \min_{u\in\mathcal{U}}\Bigg\{ &\sum_{z\in\mathcal{X}} Q^o(z|x,u)V(z) + \alpha(x,u)\Bigg(\max_{z\in\mathcal{X}} V(z) - \min_{z\in\mathcal{Y}(x,u)} V(z)\Bigg)\\
&+ \beta(\alpha(x,u))\Bigg\},\\
J^*(x) + V(x) = \min_{u\in\mathcal{U}}\Bigg\{ &f(x,u) + \sum_{z\in\mathcal{X}} Q^o(z|x,u)V(z)\\
&+ \alpha(x,u)\Bigg(\max_{z\in\mathcal{X}} V(z) - \min_{z\in\mathcal{Y}(x,u)} V(z)\Bigg) + \beta(\alpha(x,u))\Bigg\},
\end{aligned}$$

where the set $\mathcal{Y}(x,u)$ and the function $\beta(\alpha(x,u))$ are specified in part (b) of Theorem 3.7. Note that the pair of generalized dynamic programming equations solves the minimax MCP, without imposing irreducibility of the conditional distribution of the controlled process.

4. Example—infinite horizon minimax MDP. In this section we illustrate an application of the infinite horizon minimax problem for average cost, by considering a stochastic control system with state space $\mathcal{X} = \{0,1,2\}$ 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

$$Q^o(u_1) = \frac{1}{9}\begin{pmatrix} 3 & 1 & 5\\ 4 & 2 & 3\\ 1 & 6 & 2\end{pmatrix}, \quad Q^o(u_2) = \frac{1}{9}\begin{pmatrix} 1 & 2 & 6\\ 4 & 2 & 3\\ 4 & 1 & 4\end{pmatrix}, \tag{4.1}$$

the TV distance radius is $R(x) = R = 6/9\ \forall x \in \mathcal{X}$, and the cost function under each state and action is $f(0,u_1) = 2$, $f(1,u_1) = 1$, $f(2,u_1) = 3$, $f(0,u_2) = 0.5$, $f(1,u_2) = 3$, and $f(2,u_2) = 0$. To obtain an optimal stationary policy of the infinite horizon minimax problem for average cost, the policy iteration algorithm, Algorithm 3.8, is applied. Let $g_0 : \mathcal{X} \longmapsto \mathcal{U}$ be $g_0(0) = u_1$, $g_0(1) = u_2$, $g_0(2) = u_2$.

A. Let $m = 0$.

1. (policy evaluation) The results of the policy evaluation step for all partitions of the state space $\mathcal{X}$ in the sense of (3.23), i.e., $P(i) \in \mathcal{P}_{\mathcal{X}}$, $i = 1, 2, \ldots, 12$, are summarized in Table 4.1. For illustration purposes, here we show analytically the calculations for partition $P(4)$, i.e., $\mathcal{X}^0 = \{1\}$, $\mathcal{X}_0 = \{2\}$, and $\mathcal{X}_1 = \{0\}$ (see Table 3.1). From (3.22d), $r_{\max}(x,u) = 2(1 - Q^o(\mathcal{X}^0|x,u))$, i.e.,

$$r_{\max}(0,u_1) = 2(1 - Q^o(x=1|x=0,u_1)) = 2(1 - q^o_{01}(u_1)) = 2\left(1 - \frac{1}{9}\right) = \frac{16}{9},$$
$$r_{\max}(1,u_1) = 2(1 - Q^o(x=1|x=1,u_1)) = 2(1 - q^o_{11}(u_1)) = 2\left(1 - \frac{2}{9}\right) = \frac{14}{9},$$
$$r_{\max}(2,u_1) = 2(1 - Q^o(x=1|x=2,u_1)) = 2(1 - q^o_{21}(u_1)) = 2\left(1 - \frac{6}{9}\right) = \frac{6}{9},$$

and $\alpha(x,u) = \min(R(x), r_{\max}(x,u))$, where $R(x) = R = 6/9\ \forall x \in \mathcal{X}$, i.e., $\alpha(0,u_1) = \min(R, r_{\max}(0,u_1)) = \min(\frac{6}{9}, \frac{16}{9}) = \frac{6}{9}$. Similarly, $\alpha(1,u_1) = \frac{6}{9}$, and $\alpha(2,u_1) = \frac{6}{9}$. Following a similar procedure, $r_{\max}(x,u_2) = [\frac{14}{9}\ \frac{14}{9}\ \frac{16}{9}]$ and, hence, $\alpha(x,u_2) = [\frac{6}{9}\ \frac{6}{9}\ \frac{6}{9}]$. From (3.22a)–(3.22c),

$$\begin{aligned} &Q^{P(4)}(g_0) \\ &= \begin{pmatrix} \left(q^o_{00}(u_1) - (\frac{\alpha(0,u_1)}{2} - q^o_{02}(u_1))^+\right)^+ & q^o_{01}(u_1) + \frac{\alpha(0,u_1)}{2} & \left(q^o_{02}(u_1) - \frac{\alpha(0,u_1)}{2}\right)^+ \\ \left(q^o_{10}(u_2) - (\frac{\alpha(1,u_2)}{2} - q^o_{12}(u_2))^+\right)^+ & q^o_{11}(u_2) + \frac{\alpha(1,u_2)}{2} & \left(q^o_{12}(u_2) - \frac{\alpha(1,u_2)}{2}\right)^+ \\ \left(q^o_{20}(u_2) - (\frac{\alpha(2,u_2)}{2} - q^o_{22}(u_2))^+\right)^+ & q^o_{21}(u_2) + \frac{\alpha(2,u_2)}{2} & \left(q^o_{22}(u_2) - \frac{\alpha(2,u_2)}{2}\right)^+ \end{pmatrix} \\ &= \frac{1}{9}\begin{pmatrix} 3 & 4 & 2 \\ 4 & 5 & 0 \\ 4 & 4 & 1 \end{pmatrix}. \end{aligned}$$

Next, we proceed to solve $J_{Q^{P(4)}}(g_0)e + V_{Q^{P(4)}}(g_0) = f(g_0) + Q^{P(4)}(g_0)V_{Q^{P(4)}}(g_0)$ for $J_{Q^{P(4)}}(g_0) \in \mathbb{R}$ and $V_{Q^{P(4)}}(g_0) \in \mathbb{R}^3$, which is given by

$$J_{Q^{P(4)}}(g_0)\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} V_{Q^{P(4)}}(g_0,0) \\ V_{Q^{P(4)}}(g_0,1) \\ V_{Q^{P(4)}}(g_0,2) \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix} + \frac{1}{9}\begin{pmatrix} 3 & 4 & 2 \\ 4 & 5 & 0 \\ 4 & 4 & 1 \end{pmatrix}\begin{pmatrix} V_{Q^{P(4)}}(g_0,0) \\ V_{Q^{P(4)}}(g_0,1) \\ V_{Q^{P(4)}}(g_0,2) \end{pmatrix}.$$

Since $V_{Q^{P(4)}}(g_0)$ is uniquely determined up to an additive constant, let $V_{Q^{P(4)}}(g_0,2) = 0$. The solution is

$$\begin{pmatrix} V_{Q^{P(4)}}(g_0,0) \\ V_{Q^{P(4)}}(g_0,1) \\ V_{Q^{P(4)}}(g_0,2) \end{pmatrix} = \begin{pmatrix} 9/5 \\ 27/8 \\ 0 \end{pmatrix}, \quad J_{Q^*}(g_0) = 2.3.$$

Using (3.23), we identify the support sets and we let $S^{P(4)}$ denote the grouping of these sets, i.e., $S^{P(4)} : \{\mathcal{X}^0 = \{1\},\ \mathcal{X}_0 = \{2\},\ \mathcal{X}_1 = \{0\}\}$. Calculating (3.30), we obtain $\mathcal{L}^{P(4)} = Q^{P(4)}V_{Q^{P(4)}} = [\frac{21}{10}\ \frac{107}{40}\ \frac{25}{10}]^T$. The calculations for all possible partitions

$P(i) \in \mathcal{P}_{\mathcal{X}}$, $i = 1, \ldots, 12$, are as shown in Table 4.1. Since partition $P(4) \in \mathcal{P}_{\mathcal{X}}$ is the one which satisfies both (3.30) and (3.31), we let $P^*(g_0) = P(4)$, $Q^*(g_0) = Q^{P(4)}(g_0)$, $V_{Q^*}(g_0) = V_{Q^{P(4)}}(g_0)$, and $J_{Q^*}(g_0) = J_{Q^{P(4)}}(g_0)$.

2. (policy improvement) Let $g_1 = \operatorname{argmin}_{g \in \mathbb{R}^3}\{f(g) + Q^*(g)V_{Q^*}(g_0)\}$, where

(4.2)
$$Q^*(u_1)=Q^{P(4)}(u_1) = \frac{1}{9}\begin{pmatrix} 3 & 4 & 2 \\ 4 & 5 & 0 \\ 0 & 9 & 0 \end{pmatrix} \quad \text{and} \quad Q^*(u_2)=Q^{P(4)}(u_2) = \frac{1}{9}\begin{pmatrix} 1 & 5 & 3 \\ 4 & 5 & 0 \\ 4 & 4 & 1 \end{pmatrix}.$$

Then

$$\begin{aligned}
g_1(0) &= \operatorname{argmin}\{f(0,u_1) + q^*_{00}(u_1)V_{Q^*}(g_0,0)+q^*_{01}(u_1)V_{Q^*}(g_0,1)+q^*_{02}(u_1)V_{Q^*}(g_0,2),\\
&\qquad\qquad f(0,u_2) + q^*_{00}(u_2)V_{Q^*}(g_0,0)+q^*_{01}(u_2)V_{Q^*}(g_0,1)+q^*_{02}(u_2)V_{Q^*}(g_0,2)\}\\
&= \operatorname{argmin}\{4.1, 2.575\},\\
g_1(1) &= \operatorname{argmin}\{f(1,u_1) + q^*_{10}(u_1)V_{Q^*}(g_0,0)+q^*_{11}(u_1)V_{Q^*}(g_0,1)+q^*_{12}(u_1)V_{Q^*}(g_0,2),\\
&\qquad\qquad f(1,u_2) + q^*_{10}(u_2)V_{Q^*}(g_0,0)+q^*_{11}(u_2)V_{Q^*}(g_0,1)+q^*_{12}(u_2)V_{Q^*}(g_0,2)\}\\
&= \operatorname{argmin}\{3.675, 5.675\},\\
g_1(2) &= \operatorname{argmin}\{f(2,u_1) + q^*_{20}(u_1)V_{Q^*}(g_0,0)+q^*_{21}(u_1)V_{Q^*}(g_0,1)+q^*_{22}(u_1)V_{Q^*}(g_0,2),\\
&\qquad\qquad f(2,u_2) + q^*_{20}(u_2)V_{Q^*}(g_0,0)+q^*_{21}(u_2)V_{Q^*}(g_0,1)+q^*_{22}(u_2)V_{Q^*}(g_0,2)\}\\
&= \operatorname{argmin}\{6.375, 2.3\}.
\end{aligned}$$

Thus, $g_1(0) = u_2$, $g_1(1) = u_1$, and $g_1(2) = u_2$.

3. Since, $g_1 \neq g_0$, let $m = 1$ and return to step 2.

B. Let $m = 1$.

1. (policy evaluation) Following the same calculations as in $m = 0$ and by using $g_1 : \mathcal{X} \longmapsto \mathcal{U}$, the results of the policy evaluation step for all partitions $P(i) \in \mathcal{P}_{\mathcal{X}}$, $i = 1, \ldots, 12$, are summarized in Table 4.2. Since $P(4) \in \mathcal{P}_{\mathcal{X}}$ is the partition which satisfies both (3.30) and (3.31), we let $P^*(g_1) = P(4)$, and

$$Q^*(g_1) = Q^{P(4)}(g_1) = \frac{1}{9}\begin{pmatrix} 1 & 5 & 3 \\ 4 & 5 & 0 \\ 4 & 4 & 1 \end{pmatrix}$$

and

$$\begin{pmatrix} V_{Q^*}(g_1,0) \\ V_{Q^*}(g_1,1) \\ V_{Q^*}(g_1,2) \end{pmatrix} = \begin{pmatrix} V_{Q^{P(4)}}(g_1,0) \\ V_{Q^{P(4)}}(g_1,1) \\ V_{Q^{P(4)}}(g_1,2) \end{pmatrix} = \begin{pmatrix} 15/32 \\ 9/8 \\ 0 \end{pmatrix}, \quad J_{Q^*}(g_1)=J_{Q^{P(4)}}(g_1)=0.708.$$

2. (policy improvement) Let $g_2 = \operatorname{argmin}_{g \in \mathbb{R}^3}\{f(g) + Q^*(g)V_{Q^*}(g_1)\}$. Since $P^*(g_1) = P^*(g_0) = P(4)$, then the maximizing transition probability matrices under

TABLE 4.1
Results of the policy evaluation step for $m = 0$.

	$Q^{P(i)}$	$J_{Q^{P(i)}}$	$V_{Q^{P(i)}}$	$S^{P(i)}$	$\mathcal{L}^{P(i)}$
$P(1)$	$\frac{1}{9}\begin{pmatrix}6&0&3\\7&0&2\\7&0&2\end{pmatrix}$	Nonirreducible transition probability matrix			
$P(2)$	$\frac{1}{9}\begin{pmatrix}6&1&2\\7&2&0\\7&1&1\end{pmatrix}$	71/40	$\begin{pmatrix}9/5\\27/8\\0\end{pmatrix}$	$\mathcal{X}^0=\{1\}$ $\mathcal{X}_0=\{2\}$ $\mathcal{X}_1=\{0\}$	$\begin{pmatrix}63/40\\43/20\\71/40\end{pmatrix}$
$P(3)$	$\frac{1}{9}\begin{pmatrix}0&4&5\\1&5&3\\1&4&4\end{pmatrix}$	17/10	$\begin{pmatrix}9/5\\27/8\\0\end{pmatrix}$	$\mathcal{X}^0=\{1\}$ $\mathcal{X}_0=\{2\}$ $\mathcal{X}_1=\{0\}$	$\begin{pmatrix}3/2\\83/40\\17/10\end{pmatrix}$
$P(4)$	$\frac{1}{9}\begin{pmatrix}3&4&2\\4&5&0\\4&4&1\end{pmatrix}$	23/10	$\begin{pmatrix}9/5\\27/8\\0\end{pmatrix}$	$\mathcal{X}^0=\{1\}$ $\mathcal{X}_0=\{2\}$ $\mathcal{X}_1=\{0\}$	$\begin{pmatrix}21/10\\107/40\\25/10\end{pmatrix}$
$P(5)$	$\frac{1}{9}\begin{pmatrix}0&1&8\\1&2&6\\1&1&7\end{pmatrix}$	23/40	$\begin{pmatrix}9/5\\27/8\\0\end{pmatrix}$	$\mathcal{X}^0=\{1\}$ $\mathcal{X}_0=\{2\}$ $\mathcal{X}_1=\{0\}$	$\begin{pmatrix}3/8\\19/20\\23/40\end{pmatrix}$
$P(6)$	$\frac{1}{9}\begin{pmatrix}1&0&8\\3&0&6\\2&0&7\end{pmatrix}$	Nonirreducible transition probability matrix			
$P(7)$	$\frac{1}{9}\begin{pmatrix}4.5&2.5&2\\5.5&3.5&0\\5.5&2.5&1\end{pmatrix}$	153/80	$\begin{pmatrix}9/5\\117/40\\0\end{pmatrix}$	$\mathcal{X}^0=\{1\}$ $\mathcal{X}_0=\{2\}$ $\mathcal{X}_1=\{0\}$	$\begin{pmatrix}137/80\\147/80\\153/80\end{pmatrix}$
$P(8)$	$\frac{1}{9}\begin{pmatrix}3.5&0&5.5\\5&0&4\\4.5&0&4.5\end{pmatrix}$	Nonirreducible transition probability matrix			
$P(9)$	$\frac{1}{9}\begin{pmatrix}0&2.5&6.5\\1&3.5&4.5\\1&2.5&5.5\end{pmatrix}$	91/80	$\begin{pmatrix}9/5\\27/8\\0\end{pmatrix}$	$\mathcal{X}^0=\{1\}$ $\mathcal{X}_0=\{2\}$ $\mathcal{X}_1=\{0\}$	$\begin{pmatrix}15/16\\121/80\\91/80\end{pmatrix}$
$P(10)$	$\frac{1}{9}\begin{pmatrix}6&0&3\\7&0.5&1.5\\7&0&2\end{pmatrix}$	Nonirreducible transition probability matrix			
$P(11)$	$\frac{1}{9}\begin{pmatrix}1.5&4&3.5\\2.5&5&1.5\\2.5&4&2.5\end{pmatrix}$	2	$\begin{pmatrix}9/5\\27/8\\0\end{pmatrix}$	$\mathcal{X}^0=\{1\}$ $\mathcal{X}_0=\{2\}$ $\mathcal{X}_1=\{0\}$	$\begin{pmatrix}9/5\\19/8\\2\end{pmatrix}$
$P(12)$	$\frac{1}{9}\begin{pmatrix}1&0&8\\2.5&0.5&6\\2&0&7\end{pmatrix}$	Nonirreducible transition probability matrix			

each possible control $Q^*(u_1)$ and $Q^*(u_2)$ are given by (4.2). Then,

$$\begin{aligned}
g_2(0) &= \operatorname{argmin}\{f(0,u_1) + q^*_{00}(u_1)V_{Q^*}(g_1,0)+q^*_{01}(u_1)V_{Q^*}(g_1,1)+q^*_{02}(u_1)V_{Q^*}(g_1,2),\\
&\qquad\qquad f(0,u_2) + q^*_{00}(u_2)V_{Q^*}(g_1,0)+q^*_{01}(u_2)V_{Q^*}(g_1,1)+q^*_{02}(u_2)V_{Q^*}(g_1,2)\}\\
&= \operatorname{argmin}\{2.65, 1.17\},\\
g_2(1) &= \operatorname{argmin}\{f(1,u_1) + q^*_{10}(u_1)V_{Q^*}(g_1,0)+q^*_{11}(u_1)V_{Q^*}(g_1,1)+q^*_{12}(u_1)V_{Q^*}(g_1,2),\\
&\qquad\qquad f(1,u_2) + q^*_{10}(u_2)V_{Q^*}(g_1,0)+q^*_{11}(u_2)V_{Q^*}(g_1,1)+q^*_{12}(u_2)V_{Q^*}(g_1,2)\}\\
&= \operatorname{argmin}\{1.833, 3.833\},\\
g_2(2) &= \operatorname{argmin}\{f(2,u_1) + q^*_{20}(u_1)V_{Q^*}(g_1,0)+q^*_{21}(u_1)V_{Q^*}(g_1,1)+q^*_{22}(u_1)V_{Q^*}(g_1,2),\\
&\qquad\qquad f(2,u_2) + q^*_{20}(u_2)V_{Q^*}(g_1,0)+q^*_{21}(u_2)V_{Q^*}(g_1,1)+q^*_{22}(u_2)V_{Q^*}(g_1,2)\}\\
&= \operatorname{argmin}\{4.125, 0.7083\}.
\end{aligned}$$

Thus, $g_2(0) = u_2$, $g_2(1) = u_1$, and $g_2(2) = u_2$.

3. Since, $g_2 = g_1$, then $g^* = g_1$ is an optimal policy with J_{Q^*}=0.708, and V_{Q^*}=[15/32 9/8 0].

TABLE 4.2
Results of the policy evaluation step for $m = 1$.

	$Q^{P(i)}$	$J_{Q^{P(i)}}$	$V_{Q^{P(i)}}$	$S^{P(i)}$	$\mathcal{L}^{P(i)}$
$P(1)$	$\frac{1}{9}\begin{pmatrix}4&0&5\\7&0&2\\7&0&2\end{pmatrix}$	Nonirreducible transition probability matrix			
$P(2)$	$\frac{1}{9}\begin{pmatrix}4&2&3\\7&2&0\\7&1&1\end{pmatrix}$	47/96	$\begin{pmatrix}15/32\\9/8\\0\end{pmatrix}$	$\mathcal{X}^0=\{1\}$ $\mathcal{X}_0=\{2\}$ $\mathcal{X}_1=\{0\}$	$\begin{pmatrix}11/24\\59/96\\47/96\end{pmatrix}$
$P(3)$	$\frac{1}{9}\begin{pmatrix}0&5&4\\1&5&3\\1&4&4\end{pmatrix}$	9/16	$\begin{pmatrix}9/16\\9/8\\0\end{pmatrix}$	$\mathcal{X}^0=\{1\}$ $\mathcal{X}_0=\{2\}$ $\mathcal{X}_1=\{0\}$	$\begin{pmatrix}5/8\\11/16\\9/16\end{pmatrix}$
$P(4)$	$\frac{1}{9}\begin{pmatrix}1&5&3\\4&5&0\\4&4&1\end{pmatrix}$	17/24	$\begin{pmatrix}15/32\\9/8\\0\end{pmatrix}$	$\mathcal{X}^0=\{1\}$ $\mathcal{X}_0=\{2\}$ $\mathcal{X}_1=\{0\}$	$\begin{pmatrix}65/96\\5/6\\17/24\end{pmatrix}$
$P(5)$	$\frac{1}{9}\begin{pmatrix}0&0&9\\1&2&6\\1&1&7\end{pmatrix}$	13/80	$\begin{pmatrix}27/80\\9/8\\0\end{pmatrix}$	$\mathcal{X}^0=\{1\}$ $\mathcal{X}_0=\{2\}$ $\mathcal{X}_1=\{0\}$	$\begin{pmatrix}0\\23/80\\13/80\end{pmatrix}$
$P(6)$	$\frac{1}{9}\begin{pmatrix}0&0&9\\3&0&6\\2&0&7\end{pmatrix}$	Nonirreducible transition probability matrix			
$P(7)$	$\frac{1}{9}\begin{pmatrix}2.5&3.5&3\\5.5&3.5&0\\5.5&2.5&1\end{pmatrix}$	241/432	$\begin{pmatrix}11/24\\1\\0\end{pmatrix}$	$\mathcal{X}^0=\{1\}$ $\mathcal{X}_0=\{2\}$ $\mathcal{X}_1=\{0\}$	$\begin{pmatrix}223/432\\241/432\\241/432\end{pmatrix}$
$P(8)$	$\frac{1}{9}\begin{pmatrix}2&0&7\\5&0&4\\4.5&0&4.5\end{pmatrix}$	Nonirreducible transition probability matrix			
$P(9)$	$\frac{1}{9}\begin{pmatrix}0&2.5&6.5\\1&3.5&4.5\\1&2.5&5.5\end{pmatrix}$	29/80	$\begin{pmatrix}9/20\\9/8\\0\end{pmatrix}$	$\mathcal{X}^0=\{1\}$ $\mathcal{X}_0=\{2\}$ $\mathcal{X}_1=\{0\}$	$\begin{pmatrix}5/16\\39/80\\29/80\end{pmatrix}$
$P(10)$	$\frac{1}{9}\begin{pmatrix}4&0.5&4.5\\7&0.5&1.5\\7&0&2\end{pmatrix}$	133/408	$\begin{pmatrix}57/136\\18/17\\0\end{pmatrix}$	$\mathcal{X}^0=\{1\}$ $\mathcal{X}_0=\{2\}$ $\mathcal{X}_1=\{0\}$	$\begin{pmatrix}25/102\\157/408\\133/408\end{pmatrix}$
$P(11)$	$\frac{1}{9}\begin{pmatrix}0&5&4\\2.5&5&1.5\\2.5&4&2.5\end{pmatrix}$	117/184	$\begin{pmatrix}45/92\\9/8\\0\end{pmatrix}$	$\mathcal{X}^0=\{1\}$ $\mathcal{X}_0=\{2\}$ $\mathcal{X}_1=\{0\}$	$\begin{pmatrix}5/8\\35/46\\117/184\end{pmatrix}$
$P(12)$	$\frac{1}{9}\begin{pmatrix}0&0&9\\2.5&0.5&6\\2&0&7\end{pmatrix}$	Nonirreducible transition probability matrix			

5. Conclusions. In this work, we examined the optimality of the minimax MDP via dynamic programming on an infinite horizon, when the ambiguity class is described by the TV distance between the conditional distribution of the true controlled process and the conditional distribution of a nominal controlled process. As optimality criterion we considered the average cost per unit-time. Under the assumption that for every stationary Markov control policy the maximizing stochastic matrix is irreducible, we derived a new dynamic programming equation and a new policy iteration algorithm. Finally, an application of our recommended policy iteration algorithm is shown via an illustrative example.

REFERENCES

[1] A. ARAPOSTATHIS, V. S. BORKAR, E. FERNÁANDEZ-GAUCHERAND, M. K. GHOSH, AND S. I. MARCUS, *Discrete-time controlled Markov processes with average cost criterion: A survey*, SIAM J. Control Optim., 31 (1993), pp. 282–344, https://doi.org/10.1137/0331018.

[2] J. 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, 2005, pp. 1043–1049.
[3] T. Basar and P. Bernhard, *H-∞ Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*, Collection Systèmes complexes, Birkhäuser, Basel, 1995.
[4] A. Bensoussan and R. Elliot, *A finite dimensional risk-sensitive control problem*, SIAM J. Control Optim., 33 (1995), pp. 1834–1846, https://doi.org/10.1137/S0363012993255879.
[5] D. Bertsekas, *Dynamic Programming and Stochastic Control*, Academic Press, New York, 1976.
[6] V. S. Borkar, *On minimum cost per unit time control of Markov chains*, SIAM J. Control Optim., 22 (1984), pp. 965–978, https://doi.org/10.1137/0322062.
[7] V. S. Borkar, *Control of Markov chains with long-run average cost criterion: The dynamic programming equations*, SIAM J. Control Optim., 27 (1989), pp. 642–657, https://doi.org/10.1137/0327034.
[8] P. E. Caines, *Linear Stochastic Systems*, John Wiley & Sons, New York, 1988.
[9] C. D. Charalambous and J. Hibey, *Minimum principle for partially observable nonlinear risk-sensitive control problems using measure-valued decompositions*, Stochastics Stochastics Rep., 57 (1996), pp. 247–288.
[10] C. D. Charalambous and F. Rezaei, *Stochastic uncertain systems subject to relative entropy constraints: Induced norms and monotonicity properties of minimax games*, IEEE Trans. Automat. Control, 52 (2007), pp. 647–663.
[11] C. D. Charalambous, I. Tzortzis, and T. Charalambous, *Dynamic programming with total variational distance uncertainty*, in Proceedings of the 51st IEEE Conference on Decision and Control (CDC), IEEE, 2012, pp. 1909–1914.
[12] C. D. Charalambous, I. Tzortzis, S. Loyka, and T. Charalambous, *Extremum problems with total variation distance and their applications*, IEEE Trans. Automat. Control, 59 (2014), pp. 2353–2368.
[13] T. Cover and J. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, 1991.
[14] N. Dunford and J. Schwartz, *Linear Operators: General Theory*, Interscience Publishers, New York, 1958.
[15] P. Dupuis and R. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, John Wiley & Sons, New York, 1997.
[16] A. Gibbs and F. Su, *On choosing and bounding probability metrics*, Internat. Statist. Rev., 70 (2002), pp. 419–435.
[17] O. Hernandez-Lerma and J. B. Lasserre, *Discrete-time Markov Control Processes: Basic Optimality Criteria*, Appl. Math. (N.Y.) 30, Springer-Verlag, New York, 1996.
[18] M. James, J. Baras, and R. Elliot, *Risk-sensitive control and dynamic games for partially observed discrete-time nonlinear systems*, IEEE Trans. Automat. Control, 39 (1994), pp. 780–792.
[19] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*, Prentice Hall, Upper Saddle River, NJ, 1986.
[20] S. Mannor, O. Mebel, and H. Xu, *Robust MDPs with k-rectangular uncertainty*, Math. Oper. Res., 41 (2016), pp. 1484–1509, https://doi.org/10.1287/moor.2016.0786.
[21] I. Petersen, M. James, and P. Dupuis, *Minimax optimal control of stochastic uncertain systems with relative entropy constraints*, IEEE Trans. Automat. Control, 45 (2000), pp. 398–412.
[22] M. L. Puterman, *Markov Decision Processes*, John Wiley & Sons, New York, 1994.
[23] L. I. Sennott, *Another set of conditions for average optimality in Markov control processes*, Systems Control Lett., 24 (1995), pp. 147–151.
[24] I. Tzortzis, C. D. Charalambous, and T. Charalambous, *Dynamic Programming subject to total variation distance ambiguity*, SIAM J. Control Optim., 53 (2015), pp. 2040–2075, https://doi.org/10.1137/140955707.
[25] I. Tzortzis, C. D. Charalambous, T. Charalambous, C. K. Kourtellaris, and C. N. Hadjicostis, *Robust linear quadratic regulator for uncertain systems*, in Proceedings of the 55th IEEE Conference on Decision and Control (CDC), IEEE, 2016, pp. 1515–1520.
[26] V. Ugrinovskii and I. Petersen, *Finite horizon minimax optimal control of stochastic partially observed time varying uncertain systems*, Math. Control Signals Systems, 12 (1999), pp. 1–23.
[27] J. H. Van Schuppen, *Mathematical Control and System Theory of Discrete-Time Stochastic Systems*, preprint, 2017.

[28] H. Xu and S. Mannor, *Distributionally robust Markov decision processes*, Math. Oper. Res., 37 (2012), pp. 288–300.
[29] I. Yang, *A convex optimization approach to distributionally robust Markov decision processes with Wasserstein distance*, IEEE Control Syst. Lett., 1 (2017), pp. 164–169.
[30] P. Yu and H. Xu, *Distributionally robust counterpart in Markov decision processes*, IEEE Trans. Automat. Control, 61 (2016), pp. 2538–2543.