SIAM J. CONTROL OPTIM.
Vol. 53, No. 4, pp. 2040–2075

DYNAMIC PROGRAMMING SUBJECT TO TOTAL VARIATION DISTANCE AMBIGUITY*

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

Abstract. The aim of this paper is to address optimality of stochastic control strategies via dynamic programming subject to total variation distance ambiguity on the conditional distribution of the controlled process. We formulate the stochastic control problem using minimax theory, in which the control minimizes the payoff while the conditional distribution, from the total variation distance set, maximizes it. First, we investigate the maximization of a linear functional on the space of probability measures on abstract spaces, among those probability measures which are within a total variation distance from a nominal probability measure, and then we give the maximizing probability measure in closed form. Second, we utilize the solution of the maximization to solve minimax stochastic control with deterministic control strategies, under a Markovian and a non-Markovian assumption, on the conditional distributions of the controlled process. The results of this part include (1) minimax optimization subject to total variation distance ambiguity constraint; (2) new dynamic programming recursions, which involve the oscillator seminorm of the value function, in addition to the standard terms; and (3) a new infinite horizon discounted dynamic programming equation, the associated contractive property, and a new policy iteration algorithm. Finally, we provide illustrative examples for both the finite and infinite horizon cases. For the infinite horizon case, we invoke the new policy iteration algorithm to compute the optimal strategies.

Key words. stochastic control, minimax, dynamic programming, total variational distance

AMS subject classifications. 90C39, 93E20, 49J35

DOI. 10.1137/140955707

1. Introduction.

Dynamic programming recursions are often employed in optimal control and decision theory to establish existence of optimal strategies, to derive necessary and sufficient optimality conditions, and to compute the optimal strategies either in closed form or via algorithms [7, 15, 21]. The cost-to-go and the corresponding dynamic programming recursion, in their general form, are functionals of the conditional distribution of the underlying state process (controlled process) given the past and present state and control processes [7]. Thus, any ambiguity of the controlled process conditional distribution will affect the optimality of the strategies. The term "ambiguity" is used to differentiate from the term "uncertainty" often used in control nomenclature to account for situations in which the true and nominal distribution (induced by models) are absolutely continuous, and hence they are defined on the same state space. This distinction is often omitted from various robust deterministic and stochastic control approaches, including minimax and risk-sensitive formulations [1, 2, 3, 4, 8, 9, 11, 13, 14, 17, 18, 20, 22]. In this paper, the class of models is described by a ball with respect to the total variation distance between the nominal distribution and the true distribution, and hence it admits distributions which are singular with respect to the nominal distribution.

*Received by the editors February 5, 2014; accepted for publication (in revised form) May 11, 2015; published electronically July 30, 2015.
http://www.siam.org/journals/sicon/53-4/95570.html

†Department of Electrical and Computer Engineering, University of Cyprus (UCY), Nicosia, Cyprus (tzortzis.ioannis@ucy.ac.cy, chadcha@ucy.ac.cy).

‡Department of Automatic Control, School of Electrical Engineering, Royal Institute of Technology (KTH), Stockholm, Sweden (themisc@kth.se).

The main objective of this paper is to investigate the effect on the cost-to-go and dynamic programming of the ambiguity in the controlled process conditional distribution, and hence on the optimal decision strategies. Specifically, we quantify the conditional distribution ambiguity of the controlled process by a ball with respect to the total variation distance metric, centered at a nominal conditional distribution, and then we derive a new dynamic programming recursion using minimax theory with two players: player I the control process and player II the conditional distribution (controlled process), opposing each other's actions. In this minimax game formulation, the objective of player I is to minimize the cost-to-go, while the objective of player II is to maximize it. The maximization over the total variation distance ball of player II is addressed by first deriving results related to the maximization of linear functionals on a subset of the space of signed measures. Utilizing these results, a new dynamic programming recursion is presented which, in addition to the standard terms, includes additional terms that codify the level of ambiguity allowed by player II with respect to the total variation distance ball. Thus, the effect of player I, the control process, is to minimize, in addition to the classical terms, the difference between the maximum and minimum values of the cost-to-go, scaled by the radius of the total variation distance ambiguity set. We treat in a unified way the finite horizon case, under both the Markovian and non-Markovian nominal controlled processes, and the infinite horizon case. For the infinite horizon case, we consider a discounted payoff and we show that the operator associated with the resulting dynamic programming equation under total variation distance ambiguity is contractive. Consequently, we derive a new policy iteration algorithm to compute the optimal strategies. Finally, we provide examples for the finite and for the infinite horizon case.

Previous related work on optimization of stochastic systems subject to total variation distance ambiguity is found in [19] for continuous time controlled diffusion processes described by Itô differential equations. However, the solution method employed in [19] is fundamentally different; it approaches the maximization problem indirectly, by employing large deviations concepts to derive the maximizing measure as a convex combination of a tilted probability measure and the nominal measure, under restrictions on the class of measures considered. The dynamic programming equation derived in [19] is limited by the assumption that the maximizing measure is absolutely continuous with respect to the nominal measure.

In this paper, our focus is to understand the effect of total variation distance ambiguity of the conditional distribution on dynamic programming from a different point of view, utilizing concepts from signed measures. Consequently, we derive a new dynamic programming recursion which depends explicitly on the radius of the total variation distance, the closed form expression of the maximizing measure, or the oscillator seminorm of the value function. One of the fundamental properties of the maximizing conditional distribution is that, as the ambiguity radius increases, the maximizing conditional distribution becomes singular with respect to the nominal distribution. The point to be made here is that the total variation distance ambiguity set admits controlled process distributions which are not necessarily defined on the same state space as the nominal controlled process distribution. In terms of robustness of the optimal policies, this additional feature is very attractive compared to minimax techniques based on relative entropy uncertainty or risk-sensitive payoffs [1, 2, 3, 4, 8, 9, 11, 13, 14, 17, 18, 20, 22], because often the true controlled distribution lies on a higher-dimensional state space compared to the nominal controlled process distribution.

The rest of the paper is organized as follows. In section 1.1, we give a high level discussion on classical dynamic programming for the Markov control model (MCM), and we present some aspects of the problems and results obtained in the paper. In section 2, we describe the abstract formulation of the minimax problem under total variation distance ambiguity, and we derive the closed form expression of the maximizing measure. In section 3, we apply the abstract setup to both the feedback control model (FCM) (e.g., non-Markov) and the MCM. We derive new dynamic programming recursions which characterize the optimality of minimax strategies. In section 3.4, we treat the infinite horizon case, where we show that the dynamic programming operator is contractive, and we develop a new policy iteration algorithm. Finally, in section 4, we present various examples to illustrate the applications of the new dynamic programming recursions.

1.1. Discussion on the main results. Next, we describe at a high level the results obtained in this paper.

1.1.1. Dynamic programming of finite horizon discounted-Markov control model. A finite horizon discounted-Markov control model (D-MCM) with deterministic strategies is a septuple

$$\text{D-MCM}: \Big(\{\mathcal{X}_i\}_{i=0}^{n}, \{\mathcal{U}_i\}_{i=0}^{n-1}, \{\mathcal{U}_i(x_i): x_i \in \mathcal{X}_i\}_{i=0}^{n-1}, \{Q_i(dx_i|x_{i-1},u_{i-1}): (x_{i-1},u_{i-1}) \in \mathcal{X}_{i-1}\times\mathcal{U}_{i-1}\}_{i=0}^{n}, \{f_i\}_{i=0}^{n-1}, h_n, \alpha\Big) \tag{1.1}$$

consisting of the following:

(a) State Space. A sequence of Polish spaces (complete separable metric spaces) $\{\mathcal{X}_i : i = 0,\dots,n\}$, which model the state space of the controlled random process $\{x_j \in \mathcal{X}_j : j = 0,\dots,n\}$.

(b) Control or Action Space. A sequence of Polish spaces $\{\mathcal{U}_i : i = 0,\dots,n-1\}$, which model the control or action set of the control random process $\{u_j \in \mathcal{U}_j : j = 0,\dots,n-1\}$.

(c) Feasible Controls or Actions. A family $\{\mathcal{U}_i(x_i) : x_i \in \mathcal{X}_i\}$ of nonempty measurable subsets $\mathcal{U}_i(x_i)$ of $\mathcal{U}_i$, where $\mathcal{U}_i(x_i)$ denotes the set of feasible controls or actions, when the controlled process is in state $x_i \in \mathcal{X}_i$, and the feasible state-actions pairs defined by $\mathbb{K}_i \triangleq \{(x_i,u_i) : x_i \in \mathcal{X}_i, u_i \in \mathcal{U}_i(x_i)\}$ are measurable subsets of $\mathcal{X}_i \times \mathcal{U}_i, i = 0,\dots,n-1$.

(d) Controlled Process Distribution. A collection of conditional distributions or stochastic kernels $Q_i(dx_i|x_{i-1},u_{i-1})$ on $\mathcal{X}_i$ given $(x_{i-1},u_{i-1}) \in \mathbb{K}_{i-1} \subseteq \mathcal{X}_{i-1}\times\mathcal{U}_{i-1}, i = 0,\dots,n$. The controlled process distribution is described by the sequence of transition probability distributions $\{Q_i(dx_i|x_{i-1},u_{i-1}) : (x_{i-1},u_{i-1}) \in \mathbb{K}_{i-1}, i = 0,\dots,n\}$.

(e) Cost-Per-Stage. A collection of nonnegative measurable functions $f_j : \mathbb{K}_j \to [0,\infty]$, called the cost-per-stage, such that $f_j(x,\cdot)$ does not take the value $+\infty$ for each $x \in \mathcal{X}_j, j = 0,\dots,n-1$. The running payoff functional is defined in terms of $\{f_j : j = 0,\dots,n-1\}$.

(f) Terminal Cost. A bounded measurable nonnegative function $h_n : \mathcal{X}_n \to [0,\infty)$ called the terminal cost. The payoff functional at the last stage is defined in terms of h_n.

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

The definition of D-MCM envisions applications of systems described by discrete-time dynamical state space models, which include random external inputs, since such

models give rise to a collection of controlled processes distributions $\{Q_i(dx_i|x_{i-1},u_{i-1}) : (x_{i-1},u_{i-1}) \in \mathbb{K}_{i-1}, i=0,\dots,n\}$. For any integer $j \geq 0$, define the product spaces by $\mathcal{X}_{0,j} \triangleq \times_{i=0}^{j}\mathcal{X}_i$ and $\mathcal{U}_{0,j-1} \triangleq \times_{i=0}^{j-1}\mathcal{U}_i$. Define the discounted sample payoff by

$$F_{0,n}^{\alpha}(x_0,u_0,x_1,u_1,\dots,x_{n-1},u_{n-1},x_n) \triangleq \sum_{j=0}^{n-1}\alpha^j f_j(x_j,u_j)+\alpha^n h_n(x_n). \tag{1.2}$$

The goal in Markov controlled optimization with deterministic strategies is to choose a control strategy or policy $g \triangleq \{g_j : j=0,1,\dots,n-1\}$, $g_j : \mathcal{X}_{0,j}\times\mathcal{U}_{0,j-1} \longrightarrow \mathcal{U}_j(x_j)$, $u_j^g = g_j(x_0^g,x_1^g,\dots,x_j^g,u_0^g,u_1^g,\dots,u_{j-1}^g)$, $j=0,1,\dots,n-1$ so as to minimize the payoff functional

$$\begin{aligned}&\mathbb{E}\left\{\sum_{j=0}^{n-1}\alpha^j f_j(x_j^g,u_j^g)+\alpha^n h_n(x_n^g)\right\}\\ &=\int_{\mathcal{X}_0\times\mathcal{X}_1\times\cdots\times\mathcal{X}_n} F_{0,n}^{\alpha}\Big(x_0,u_0^g(x_0),x_1,u_1^g(x_0,x_1),\dots,x_{n-1},\\ &\qquad\qquad u_{n-1}^g(x_0,x_1,\dots,x_{n-1}),x_n\Big)\\ &\quad Q_0(dx_0)Q_1(dx_1|x_0,u_0^g(x_0))\dots Q_n\Big(dx_n|x_{n-1},u_{n-1}^g(x_0,x_1,\dots,x_{n-1})\Big).\end{aligned} \tag{1.3}$$

Clearly, payoff (1.3) is a functional of the collection of conditional distributions $\{Q_i(\cdot|\cdot) : i=0,1,\dots,n\}$. Moreover, if this collection of distribution has countable support for each (x_{i-1},u_{i-1}), $i=0,\dots,n$, then each integral in (1.3) is reduced to a countable summation.

For $(i,x) \in \{0,1\dots,n\}\times\mathcal{X}_i$, let $V_i^0(x) \in \mathbb{R}$ represent the minimal cost-to-go or value function on the time horizon $\{i,i+1,\dots,n\}$ if the controlled process starts at state $x_i = x$ at time i, defined by

$$V_i^0(x) \triangleq \inf_{\substack{g_k\in\mathcal{U}_k(x_k)\\ k=i,\dots,n-1}} \mathbb{E}_{i,x}^g\left\{\sum_{j=i}^{n-1}\alpha^j f_j(x_j^g,u_j^g)+\alpha^n h_n(x_n^g)\right\}, \tag{1.4}$$

where $\mathbb{E}_{i,x}^g\{\cdot\}$ denotes expectation conditioned on $x_i^g = x$. A Markov property on the controlled process distributions, i.e., $Q_i(dx_i|x^{i-1},u^{i-1}) = Q_i(dx_i|x_{i-1},u_{i-1})$ for all $(x^{i-1},u^{i-1}) \in \times_{j=0}^{i-1}\mathbb{K}_j$, $i=0,1,\dots,n$, under admissible non-Markov strategies, implies that Markov control strategies are optimal [15]. Consequently, it can be shown that the value function (1.4) satisfies the following dynamic programming recursion relating the value functions $V_i^0(\cdot)$ and $V_{i+1}^0(\cdot)$ [15]:

$$V_n^0(x) = \alpha^n h_n(x), \qquad x \in \mathcal{X}_n, \tag{1.5}$$

$$V_i^0(x) = \inf_{u\in\mathcal{U}_i(x)}\left\{\alpha^i f_i(x,u)+\int_{\mathcal{X}_{i+1}} V_{i+1}^0(z)Q_{i+1}(dz|x,u)\right\}, \qquad x \in \mathcal{X}_i. \tag{1.6}$$

Since the value function $V_i^0(x)$ defined by (1.4) and the dynamic programming recursion (1.5), (1.6) depend on the complete knowledge of the collection of conditional distributions $\{Q_i(\cdot|\cdot) : i=0,\dots,n\}$, any mismatch of the collection $\{Q_i(\cdot|\cdot) : i=0,\dots,n\}$ from the true collection of conditional distributions will affect the optimality of the control strategies. Our objective is to address the impact of any ambiguity

measured by the total variation distance between the true conditional distribution and a given nominal distribution on the cost-to-go (1.4) and dynamic programming recursion (1.5), (1.6).

1.1.2. Dynamic programming of infinite horizon D-MCM. The infinite horizon D-MCM with deterministic strategies is a special case of the finite horizon D-MCM specified by a sextuple

$$\Big(\mathcal{X},\mathcal{U},\{\mathcal{U}(x):x\in\mathcal{X}\},\{Q(dz|x,u):(x,u)\in\mathcal{X}\times\mathcal{U}\},f,\alpha\Big), \tag{1.7}$$

where the elements defined under (a)–(f) are independent of time index i. That is, the state space is $\mathcal{X}$, the control or action space is $\mathcal{U}$, the feasible controls or actions is a family $\{\mathcal{U}(x):x\in\mathcal{X}\}\subset\mathcal{U}$, the controlled process distribution is a stochastic kernel $Q(\cdot|\cdot)$ on $\mathcal{X}$ given $\mathbb{K}$, where $\mathbb{K}\triangleq\{(x,u):x\in\mathcal{X},u\in\mathcal{U}(x)\}$, the cost-per-stage is a one stage cost $f:\mathbb{K}\longrightarrow[0,\infty]$, and there is no terminal cost.

The dynamic programming equation of the infinite horizon D-MCM as given by [21] is a function $v^0_\infty:\mathcal{X}\longrightarrow\mathbb{R}$ satisfying

$$v^0_\infty(x)=\inf_{u\in\mathcal{U}(x)}\left\{f(x,u)+\alpha\int_{\mathcal{X}}v^0_\infty(z)Q(dz|x,u)\right\},\qquad x\in\mathcal{X}. \tag{1.8}$$

Similar to the finite horizon D-MCM, the dynamic programming equation (1.8) depends on the conditional distribution $Q(dz|x,u)$. Hence, any ambiguity or mismatch of $Q(dz|x,u)$ from the true distribution affects the optimality of the strategies.

1.1.3. Dynamic programming with total variation distance ambiguity. Motivated by the above discussion, the objective of this paper is to investigate dynamic programming under ambiguity of the conditional distributions of the controlled processes

$$\Big\{Q_i(dx_i|x_{i-1},u_{i-1}):(x_{i-1},u_{i-1})\in\mathbb{K}_{i-1}\Big\},\quad i=0,\dots,n.$$

The ambiguity of the conditional distributions of the controlled process is modeled by the total variation distance. Specifically, given a collection of nominal controlled process distributions $\{Q^o_i(dx_i|x_{i-1},u_{i-1}):(x_{i-1},u_{i-1})\in\mathbb{K}_{i-1}\}$, $i=0,\dots,n$, the corresponding collection of true controlled process distributions $\{Q_i(dx_i|x_{i-1},u_{i-1}):(x_{i-1},u_{i-1})\in\mathbb{K}_{i-1}\}$, $i=0,\dots,n$, is modeled by a ball with respect to the total variation distance centered at the nominal conditional distribution having radius $R_i\in[0,2]$, $i=0,\dots,n$, defined by

$$\begin{aligned}&\mathbf{B}_{R_i}(Q^o_i)(x_{i-1},u_{i-1})\\&\quad\triangleq\Big\{Q_i(\cdot|x_{i-1},u_{i-1}):||Q_i(\cdot|x_{i-1},u_{i-1})-Q^o_i(\cdot|x_{i-1},u_{i-1})||_{TV}\le R_i\Big\}.\end{aligned}$$

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

$$||\alpha-\beta||_{TV}\triangleq\sup_{P\in\mathcal{P}(\Sigma)}\sum_{F_i\in P}|\alpha(F_i)-\beta(F_i)|,\quad\alpha,\beta\in\mathcal{M}_1(\Sigma), \tag{1.9}$$

where $\mathcal{M}_1(\Sigma)$ denotes the set of probability measures on the σ-algebra $\mathcal{B}(\Sigma)$ and $\mathcal{P}(\Sigma)$ denotes the collection of all finite partitions of Σ. Note that the distance metric (1.9) induced by the total variation norm does not require absolute continuity of

the measures $\alpha \in \mathcal{M}_1(\Sigma)$ and $\beta \in \mathcal{M}_1(\Sigma)$. Therefore, the total variation distance model of ambiguity is quite general; it includes linear, nonlinear, finite, or countable state space models, etc., since no assumptions are imposed on the structure of the stochastic control dynamical system model, which induces the collection of conditional distributions $\{Q_i(\cdot|\cdot) : i = 0, \ldots, n\}$, $\{Q_i^o(\cdot|\cdot) : i = 0, \ldots, n\}$. Given the above description of ambiguity in distribution with respect to total variation distance metric, we reformulate the value function and dynamic programming recursion via minimax theory as follows.

For $(i, x) \in \{0, 1 \ldots, n\} \times \mathcal{X}_i$, let $V_i(x) \in \mathbb{R}$ represent the minimal cost-to-go on the time horizon $\{i, i+1, \ldots, n\}$ if the state of the controlled process starts at state $x_i = x$ at time i, defined by

$$V_i(x) \triangleq \inf_{\substack{g_k \in \mathcal{U}_k(x_k) \\ k=i,\ldots,n-1}} \sup_{\substack{Q_{k+1}(\cdot|x_k,u_k) \in \mathbf{B}_{R_{k+1}}(Q^o_{k+1})(x_k,u_k) \\ k=i,\ldots,n-1}} \mathbb{E}^g_{i,x} \left\{ \sum_{j=i}^{n-1} \alpha^j f_j(x_j^g, u_j^g) + \alpha^n h_n(x_n^g) \right\},$$

where $\mathbb{E}^g_{i,x}$ denotes conditional expectation with respect to the true collection of conditional distribution $\{Q_k(\cdot|\cdot) : k = i, \ldots, n\}$. Even in the above minimax setting, the Markov property of the controlled process distribution under an admissible non-Markov strategy implies that Markov control strategies are optimal. Moreover, the value function satisfies the following dynamic programming recursion relating the value function $V_i(\cdot)$ and $V_{i+1}(\cdot)$ for all $i = 0, 1, \ldots, n-1$:

$$\begin{aligned} V_n(x) &= \alpha^n h_n(x), \quad x \in \mathcal{X}_n, \\ V_i(x) &= \inf_{u \in \mathcal{U}_i(x)} \sup_{Q_{i+1}(\cdot|x,u) \in \mathbf{B}_{R_{i+1}}(Q^o_{i+1})(x,u)} \\ &\qquad \left\{ \alpha^i f_i(x,u) + \int_{\mathcal{X}_{i+1}} V_{i+1}(z) Q_{i+1}(dz|x,u) \right\}, \ x \in \mathcal{X}_i. \end{aligned}$$

Based on this formulation, if $V_{i+1}(\cdot)$ is bounded continuous nonnegative, we show that the new dynamic programming equation is given by

$$V_n(x) = \alpha^n h_n(x), \quad x \in \mathcal{X}_n, \tag{1.10}$$

$$\begin{aligned} V_i(x) = \inf_{u \in \mathcal{U}_i(x)} \Bigg\{ & \alpha^i f_i(x,u) + \int_{\mathcal{X}_{i+1}} V_{i+1}(z) Q^o_{i+1}(dz|x,u) \\ & + \frac{R_i}{2} \left(\sup_{z \in \mathcal{X}_{i+1}} V_{i+1}(z) - \inf_{z \in \mathcal{X}_{i+1}} V_{i+1}(z) \right) \Bigg\}, \quad x \in \mathcal{X}_i. \end{aligned} \tag{1.11}$$

Note that the new term in the right side of (1.11) has the interpretation of minimizing the future ambiguity. It is the oscillator seminorm of $V_{j+1}(\cdot)$, called the global modulus of continuity of $V_{j+1}(\cdot)$, which measures the difference between the maximum and minimum values of $V_{j+1}(\cdot)$.

For the infinite horizon D-MCM, the new dynamic programming equation is given by

$$\begin{aligned} v_\infty(x) = \inf_{u \in \mathcal{U}(x)} \Bigg\{ & f(x,u) + \alpha \int_{\mathcal{X}} v_\infty(z) Q^o(dz|x,u) \\ & + \alpha \frac{R}{2} \left(\sup_{z \in \mathcal{X}} v_\infty(z) - \inf_{z \in \mathcal{X}} v_\infty(z) \right) \Bigg\}, \quad x \in \mathcal{X}. \end{aligned} \tag{1.12}$$

For finite and countable alphabet spaces $\mathcal{X}$, the integrals in the right-hand side of (1.11), (1.12) are replaced by summations.

In addition to the D-MCM, we will also discuss the general discounted feedback control model (i.e., we relax the Markovian assumption). In summary, the issues discussed and results obtained in this paper are the following: (1) formulation of finite horizon discounted stochastic optimal control subject to conditional distribution ambiguity described by total variation distance via minimax theory; (2) dynamic programming recursions for (a) nominal D-MCM and (b) the discounted-feedback control model (D-FCM), under total variation distance ambiguity on the conditional distribution of the controlled process; (3) formulation of the infinite horizon D-MCM and dynamic programming equation under conditional distribution ambiguity described by total variation distance via minimax theory; (4) characterization of the maximizing conditional distribution belonging to the total variation distance set and the corresponding new dynamic programming recursions; (5) the contraction property of the infinite horizon D-MCM dynamic programming and new policy iteration algorithm; and (6) examples for the finite and infinite horizon cases.

Relative entropy and exponential functions. Related work on modeling uncertainty in probability distribution utilizes relative entropy [2, 9, 13, 17, 20] defined by

$$H(\alpha||\beta) \triangleq \begin{cases} \int_\Sigma \log(\frac{\alpha(dx)}{\beta(dx)})\alpha(dx) & \text{if } \alpha(\cdot) \ll \beta(\cdot) \text{ and } \log\frac{\alpha}{\beta} \in L^1(\alpha), \\ +\infty & \text{otherwise,} \end{cases} \tag{1.13}$$

where $\alpha \ll \beta$ means that α is absolutely continuous with respect to β (i.e., $\beta(A) = 0$ for some measurable set A, and then $\alpha(A) = 0$). However, by Pinsker's inequality, distance in total variation of probability measures is a lower bound on relative entropy or Kullback–Leibler distance, that is,

$$||\alpha - \beta||_{TV} \leq \sqrt{2H(\alpha||\beta)}, \quad \alpha, \beta \in \mathcal{M}_1(\Sigma). \tag{1.14}$$

Hence, for any fixed $\beta \in \mathcal{M}_1(\Sigma)$, then

$$\Big\{\alpha \in \mathcal{M}_1(\Sigma) : H(\alpha||\beta) \leq \frac{r^2}{2}\Big\} \subseteq \mathbb{B}_R(\beta) \equiv \Big\{\alpha \in \mathcal{M}_1(\Sigma) : ||\alpha - \beta||_{TV} \leq r\Big\}.$$

This means that even for those measures which satisfy $\alpha \ll \beta$, the ambiguity set described by relative entropy is a subset of the much larger total variation distance ambiguity set. Moreover, by the definition of relative entropy (1.13), for any finite $r \in [0, \infty]$ and fixed $\beta \in \mathcal{M}_1(\Sigma)$, any ambiguity set described by relative entropy consists of only those measures $\alpha \in \mathcal{M}_1(\Sigma)$ which are absolutely continuous with $\beta \in \mathcal{M}_1(\Sigma)$. The relative entropy constraint set is defined by

$$\mathbf{A}_r(Q_i^o)(x_{i-1}, u_{i-1}) \triangleq \Big\{Q_i(\cdot|x_{i-1}, u_{i-1}) : H(Q_i||Q_i^0)(x_{i-1}, u_{i-1}) \leq r_i(x_{i-1})\Big\},$$

where $r_i : \mathcal{X} \longmapsto [0, \infty)$, $i = 0, 1, \ldots, n$. The minimax optimization problem subject to relative entropy constraint on the conditional distribution of the controlled process is formulated as follows:

$$\begin{aligned} J_{0,n}(\pi^*, Q_k^* : k = 0, \ldots, n) = & \inf_{\substack{u_k \in \mathcal{U}_k(x_k) \\ k=0,1,\ldots,n-1}} \sup_{\substack{Q_k(\cdot|x_{k-1}, u_{k-1}) \in \mathbf{A}_r(Q_k^o)(x_{k-1}, u_{k-1}) \\ k=0,1,\ldots,n}} \\ & \mathbb{E}_{\mathbf{Q}_\nu^\pi}\left\{\sum_{k=0}^{n-1} \alpha^k f_k(x_k^g, u_k^g) + \alpha^n h_n(x_n^g)\right\}. \end{aligned} \tag{1.15}$$

We formulate the stochastic control problem in alignment with the dynamic programming equation of total variation distance constraint as in [2, 9, 13, 17, 20]. For $(i,x) \in \{0,1,\dots,n\} \times \mathcal{X}_i$, let $V_i(x) \in \mathbb{R}$ represent the minimal cost-to-go defined by

$$V_i(x) = \inf_{\substack{u_k \in \mathcal{U}_k(x_k) \\ k=i,\dots,n-1}} \sup_{\substack{Q_{k+1}(\cdot|x_k,u_k) \in \mathbf{A}_r(Q^o_{k+1})(x_k,u_k) \\ k=i,\dots,n-1}} \mathbb{E}^g_{i,x} \left\{ \sum_{j=i}^{n-1} \alpha^j f_j(x^g_j, u^g_j) + \alpha^n h_n(x^g_n) \right\}.$$

The dynamic programming equations are given by

$$V_n(x) = \alpha^n h_n(x), \quad x \in \mathcal{X}_n,$$
$$V_i(x) = \inf_{u \in \mathcal{U}_i(x)} \sup_{Q_{i+1}(\cdot|x,u) \in \mathbf{A}_r(Q^0_{i+1})(x,u)} \left\{ \alpha^i f_i(x,u) + \int_{\mathcal{X}_{i+1}} V_{i+1}(z) Q_{i+1}(dz|x,u) \right\}.$$

By Lagrange duality theorem [16], then

$$\begin{aligned} V_i(x) = \inf_{u \in \mathcal{U}_i(x)} \inf_{s_{i+1}(x) \geq 0} \sup_{Q_{i+1}(\cdot|x,u): H(Q_{i+1}||Q^0_{i+1})(x,u) < \infty} & \\ \Big\{ \alpha^i f_i(x,u) + \int_{\mathcal{X}_{i+1}} & V_{i+1}(z) Q_{i+1}(dz|x,u) \\ - s_{i+1}(x) \big(H(Q_{i+1}||Q^0_{i+1})(x,u) - r_{i+1}(x) \big) & \Big\}, \end{aligned} \tag{1.16}$$

where $s_{i+1}(x)$ is the Lagrange multiplier. By [18] (Proposition 2.3), the supremum over $Q_{i+1}(\cdot|x,u)$ with $H(Q_{i+1}||Q^0_{i+1})(x,u) < \infty$ is attained at

$$Q^*_{i+1}(dz|x,u) = \frac{\exp\big(\frac{1}{s_{i+1}(x)} V_{i+1}(z)\big) Q^0_{i+1}(dz|x,u)}{\int_{\mathcal{X}_{i+1}} \exp\big(\frac{1}{s_{i+1}(x)} V_{i+1}(z)\big) Q^0_{i+1}(dz|x,u)}. \tag{1.17}$$

Substituting (1.17) into (1.16) yields

$$V_n(x) = \alpha^n h_n(x), \quad x \in \mathcal{X}_n, \tag{1.18}$$

$$\begin{aligned} V_i(x) = \inf_{u \in \mathcal{U}_i(x)} \inf_{s_{i+1}(x) \geq 0} & \Big\{ \alpha^i f_i(x,u) \\ & + s_{i+1}(x) \log \int_{\mathcal{X}_{i+1}} \exp\left(\frac{1}{s_{i+1}(x)} V_{i+1}(z) \right) Q^0_{i+1}(dz|x,u) \Big\} \\ & + s_{i+1}(x) r_{i+1}(x). \end{aligned} \tag{1.19}$$

The Lagrange multipliers $\inf_{s_{i+1}(x) \geq 0}\{\cdot\}$ can be found by the relative entropy constraint which holds with equality, i.e., $H(Q^*_{i+1}||Q^0_{i+1})(x,u)|_{s_{i+1}(x) = s^*_{i+1}(x)} = r_{i+1}(x)$ for $i = 0,1,\dots,n$. A further elaboration on the connections between stochastic optimal control with risk-sensitive payoff and minimax stochastic control in which the maximization is with respect to relative entropy ambiguity is found in [2, 9, 13, 17, 18, 20] (where all duality relations require that relative entropy is finite). A specific example which illustrates the differences between relative entropy and the total variation distance ambiguity is presented analytically in section 4.3.

2. Maximization with total variation distance ambiguity. In this section, we recall certain results from [19] on the maximization of a linear functional on the space of probability distributions subject to total variation distance ambiguity. We use these results to derive the maximizing probability distribution subject to total variation distance ambiguity of the controlled process.

Let (Σ, d_Σ) denote a complete, separable metric space (a Polish space), and $(\Sigma, \mathcal{B}(\Sigma))$ the corresponding measurable space, in which $\mathcal{B}(\Sigma)$ is the σ-algebra generated by open sets in Σ. Let $\mathcal{M}_1(\Sigma)$ denote the space of countably additive probability measures on $(\Sigma, \mathcal{B}(\Sigma))$. Define the spaces

$$\begin{aligned}
BC(\Sigma) &\triangleq \big\{\text{Bounded continuous functions } \ell : \Sigma \longrightarrow \mathbb{R} : ||\ell|| \triangleq \sup_{x\in\Sigma} |\ell(x)| < \infty\big\},\\
BM(\Sigma) &\triangleq \big\{\text{Bounded measurable functions } \ell : \Sigma \longrightarrow \mathbb{R} : ||\ell|| < \infty\big\},\\
C(\Sigma) &\triangleq \big\{\text{Continuous functions } \ell : \Sigma \longrightarrow \mathbb{R} : ||\ell|| < \infty\big\},\\
C^+(\Sigma) &\triangleq \big\{\ell \in C(\Sigma) : \ell \geq 0\big\},\\
BC^+(\Sigma) &\triangleq \big\{\ell \in BC(\Sigma) : \ell \geq 0\big\}, \quad BM^+(\Sigma) \triangleq \big\{\ell \in BM(\Sigma) : \ell \geq 0\big\}.
\end{aligned}$$

Clearly, $BC(\Sigma)$, $BM(\Sigma)$, $C(\Sigma)$ are Banach spaces. We present the maximizing measure for $\ell \in BC^+(\Sigma)$, although the results can be generalized to real-valued functions $\ell \in L^{\infty,+}(\Sigma, \mathcal{B}(\Sigma), \nu)$, the set of all $\mathcal{B}(\Sigma)$-measurable, nonnegative essentially bounded functions defined $\nu - a.e.$ endowed with the essential supremum norm $||\ell||_{\infty,\nu} = \nu - \text{ess } \sup_{x\in\Sigma} \ell(x)$.

From [19], we have the following. For $\ell \in BC^+(\Sigma)$, and $\mu \in \mathcal{M}_1(\Sigma)$ fixed,

$$L(\nu^*) \triangleq \sup_{||\nu-\mu||_{TV}\leq R} \int_\Sigma \ell(x)\nu(dx) = \frac{R}{2}\left\{\sup_{x\in\Sigma} \ell(x) - \inf_{x\in\Sigma} \ell(x)\right\} + \int_\Sigma \ell(x)\mu(dx), \tag{2.1}$$

where $R \in [0, 2]$, ν^* satisfies the constraint $||\xi^*||_{TV} = ||\nu^* - \mu||_{TV} = R$, it is normalized $\nu^*(\Sigma) = 1$, and $\nu^*(A) \in [0, 1]$ on any $A \in \mathcal{B}(\Sigma)$. Moreover, by defining[1]

$$\begin{aligned}
x^0 \in \Sigma^0 &\triangleq \{x \in \overline{\Sigma} : \ell(x) = \sup\{\ell(y) : y \in \Sigma\} \equiv \ell_{\max}\},\\
x_0 \in \Sigma_0 &\triangleq \{x \in \overline{\Sigma} : \ell(x) = \inf\{\ell(y) : y \in \Sigma\} \equiv \ell_{\min}\},
\end{aligned}$$

where $\overline{\Sigma}$ denotes the closure of Σ. Then, the payoff $L(\nu^*)$ can be written as

$$L(\nu^*) = \int_{\Sigma^0} \ell_{\max}\nu^*(dx) + \int_{\Sigma_0} \ell_{\min}\nu^*(dx) + \int_{\Sigma\setminus\Sigma^0\cup\Sigma_0} \ell(x)\mu(dx) \tag{2.2}$$

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

$$\begin{gathered}
\int_{\Sigma^0} \nu^*(dx) = \mu(\Sigma^0) + \frac{R}{2} \in [0, 1], \qquad \int_{\Sigma_0} \nu^*(dx) = \mu(\Sigma_0) - \frac{R}{2} \in [0, 1],\\
\nu^*(A) = \mu(A) \quad \forall A \subseteq \Sigma \setminus \Sigma^0 \cup \Sigma_0.
\end{gathered} \tag{2.3}$$

Note that if $\Sigma^0 = \Sigma_0 = \{\emptyset\}$, then $\nu(\Sigma^0) = \nu(\Sigma_0) = 0$, and $L(\nu^*) = \int_{\Sigma\setminus\Sigma^0\cup\Sigma_0} \ell(x)\mu(dx)$.

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

The second right-hand side term in (2.1) is related to the oscillator seminorm of $f \in BM(\Sigma)$, called the global modulus of continuity, and it is defined by

$$\operatorname{osc}(f) \triangleq \sup_{(x,y)\in\Sigma\times\Sigma} |f(x) - f(y)| = 2 \inf_{\beta\in\mathbb{R}} ||f - \beta|| \quad \text{for } f \in BM(\Sigma).$$

However, for $f \in BM^+(\Sigma)$, then

$$\operatorname{osc}(f) = \sup_{x\in\Sigma} |f(x)| - \inf_{x\in\Sigma} |f(x)| = \sup_{x\in\Sigma} f(x) - \inf_{x\in\Sigma} f(x).$$

Note that the above results can be extended to $f \in C^+(\Sigma)$.

The maximizing measure for finite and countable alphabet spaces. Here, we further elaborate on the form of the maximizing measures for finite and countable alphabet spaces, since we use them to analyze finite horizon D-MCM and D-FCM and infinite horizon D-MCM with finite (or countable) state and control spaces.

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

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

and its corresponding support sets by

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

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

$$\Sigma_k \triangleq \left\{ x \in \Sigma : \ell(x) = \min \left\{ \ell(\alpha) : \alpha \in \Sigma \setminus \Sigma^0 \cup \left(\bigcup_{j=1}^{k} \Sigma_{j-1} \right) \right\} \right\}, \quad k \in \{1, 2, \ldots, r\}$$

until all the elements of Σ are exhausted. Further, define the corresponding values of the sequence on sets Σ_k by

$$\ell(\Sigma_k) \triangleq \min_{x\in\Sigma\setminus\Sigma^0\cup(\bigcup_{j=1}^{k}\Sigma_{j-1})} \ell(x), \quad k \in \{1, 2, \ldots, r\},$$

where r is the number of Σ_k sets which is at most $|\Sigma \setminus \Sigma^0 \cup \Sigma_0|$. For example, when $k = 1$, $\ell(\Sigma_1) = \min_{x\in\Sigma\setminus\Sigma^0\cup\Sigma_0} \ell(x)$, when $k = 2$, $\ell(\Sigma_2) = \min_{x\in\Sigma\setminus\Sigma^0\cup\Sigma_0\cup\Sigma_1} \ell(x)$, and so on.

In [10], it is shown that the maximum payoff subject to total variation constraint is given by

$$L(\nu^*) = \ell_{\max}\nu^*(\Sigma^0) + \ell_{\min}\nu^*(\Sigma_0) + \sum_{k=1}^{r} \ell(\Sigma_k)\nu^*(\Sigma_k) \tag{2.4}$$

and that the optimal probabilities are given by the following equations (water-filling algorithms):

$$\nu^*(\Sigma^0) \triangleq \sum_{x\in\Sigma^0} \nu^*(x) = \sum_{x\in\Sigma^0} \mu(x) + \frac{\alpha}{2} \equiv \mu(\Sigma^0) + \frac{\alpha}{2}, \tag{2.5}$$

$$\nu^*(\Sigma_0) \triangleq \sum_{x\in\Sigma_0} \nu^*(x) = \left(\sum_{x\in\Sigma_0} \mu(x) - \frac{\alpha}{2}\right)^+ \equiv \left(\mu(\Sigma_0) - \frac{\alpha}{2}\right)^+, \tag{2.6}$$

$$\begin{aligned}\nu^*(\Sigma_k) \triangleq \sum_{x\in\Sigma_k} \nu^*(x) &= \left(\sum_{x\in\Sigma_k} \mu(x) - \left(\frac{\alpha}{2} - \sum_{j=1}^{k} \sum_{x\in\Sigma_{j-1}} \mu(x)\right)^+\right)^+ \\ &\equiv \left(\mu(\Sigma_k) - \left(\frac{\alpha}{2} - \sum_{j=1}^{k} \mu(\Sigma_{j-1})\right)^+\right)^+,\end{aligned} \tag{2.7}$$

$$\alpha \triangleq \min(R, R_{\max}), \quad R_{\max} \triangleq 2\left(1 - \sum_{x\in\Sigma^0} \mu(x)\right) \equiv 2(1-\mu(\Sigma^0)), \quad R \in [0,2], \tag{2.8}$$

where $k \in \{1, 2, \ldots, r\}$ and r is the number of Σ_k sets which is at most $|\Sigma \setminus \Sigma^0 \cup \Sigma_0|$.

The parameter α reinforces the intuitive notion of the total variation between the true and nominal probability distribution as having attributes similar to "physical mass." Thus, if $\alpha = R_{\max}$, then (2.5) implies that the probability "mass" on Σ^0 set is $\nu^*(\Sigma^0) = 1$ and hence $\nu^*(\Sigma \setminus \Sigma^0) = 0$. However, if $\alpha = R < R_{\max}$, then (2.5) implies that the probability mass on Σ^0 set is $\nu^*(\Sigma^0) < 1$, and hence equations (2.6)–(2.7) are employed. While $\nu^*(\Sigma_0) > 0$, (2.7) implies that $\nu^*(\Sigma_k) = \mu(\Sigma_k)$ for all $k = 1, \ldots, r$. However, if $\nu^*(\Sigma_0) = 0$, that is, all the probability mass is removed from Σ_0, then the solution is obtained by moving further into the partition using (2.7). For all $R \in [0, 2]$, the resulting solution is described via a water-filling effect.

We are now equipped with the solution of maximizing linear functionals with total variation distance ambiguity for both finite, countable alphabets and abstract alphabet spaces (Polish spaces), and therefore we are ready to apply these results to the dynamic programming recursion under ambiguity on the conditional distribution.

3. Minimax stochastic control with total variation distance ambiguity. In this section, we first introduce the general definition of the finite horizon discounted-feedback control model (D-FCM) with randomized and deterministic control policies, under total variation distance uncertainty (which includes the D-MCM introduced in section 1.1), and then we apply the characterization of the maximizing distribution of section 2 to the dynamic programming recursion. In the last section, we discuss the infinite horizon D-MCM.

Define $\mathbb{N}^n \triangleq \{0, 1, 2, \ldots, n\}, n \in \mathbb{N}$. The state space and the control space are sequences of Polish spaces $\{\mathcal{X}_j : j = 0, 1, \ldots, n\}$ and $\{\mathcal{U}_j : j = 0, 1, \ldots, n-1\}$, respectively. These spaces are associated with their corresponding measurable spaces $(\mathcal{X}_j, \mathcal{B}(\mathcal{X}_j))$ for all $j \in \mathbb{N}^n$, $(\mathcal{U}_j, \mathcal{B}(\mathcal{U}_j))$ for all $j \in \mathbb{N}^{n-1}$. Define the product spaces by $\mathcal{X}_{0,n} \triangleq \times_{i=0}^{n} \mathcal{X}_i$, $\mathcal{U}_{0,n-1} \triangleq \times_{i=0}^{n-1} \mathcal{U}_i$ and introduce their product measurable spaces, $(\mathcal{X}_{0,n}, \mathcal{B}(\mathcal{X}_{0,n}))$, $(\mathcal{U}_{0,n-1}, \mathcal{B}(\mathcal{U}_{0,n-1}))$, respectively, for $n \in \mathbb{N}^n$. The state process is denoted by $x^n \triangleq \{x_j : j = 0, 1, \ldots, n\}$, and the control process is denoted by $u^{n-1} \triangleq \{u_j : j = 0, 1, \ldots, n-1\}$. For any measurable spaces $(\mathcal{X}, \mathcal{B}(\mathcal{X})), (\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$,

the set of stochastic kernels on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ conditioned on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ is denoted by $\mathcal{Q}(\mathcal{Y}|\mathcal{X})$.

Given $(\mathcal{X}_{0,n}, \mathcal{B}(\mathcal{X}_{0,n}))$, $(\mathcal{U}_{0,n-1}, \mathcal{B}(\mathcal{U}_{0,n-1}))$, the Borel state and control or action spaces, respectively, and the initial state distribution $\nu_0(dx_0)$, we introduce the space $H_{0,n}$ of admissible observable histories by

$$H_{0,n} \triangleq \mathbb{K}_0 \times \mathbb{K}_1 \times \cdots \times \mathbb{K}_{n-1} \times \mathcal{X}_n \equiv \times_{i=0}^{n-1} \mathbb{K}_i \times \mathcal{X}_n, \quad n \in \mathbb{N}, \quad H_{0,0} = \mathcal{X}_0,$$

where $\mathbb{K}_i \triangleq \{(x_i, u_i) : x_i \in \mathcal{X}_i, u_i \in \mathcal{U}_i(x_i)\}$ denote the feasible state-action pairs for $i = 0, 1, \ldots, n-1$. A typical element $h_{0,n} \in H_{0,n}$ is a sequence of the form

$$h_{0,n} = (x_0, u_0, \ldots, x_{n-1}, u_{n-1}, x_n), \quad (x_i, u_i) \in \mathbb{K}_i, \quad i = 0, \ldots, n-1, \quad x_n \in \mathcal{X}_n.$$

Similarly, introduce

$$G_{0,n} = \mathcal{X}_0 \times \mathcal{U}_0 \times \cdots \times \mathcal{X}_{n-1} \times \mathcal{U}_{n-1} \times \mathcal{X}_n \equiv \times_{i=0}^{n-1} (\mathcal{X}_i \times \mathcal{U}_i) \times \mathcal{X}_n, \quad n \in \mathbb{N},$$
$$G_{0,0} = H_{0,0} = \mathcal{X}_0.$$

The spaces $G_{0,n}$ and $H_{0,n}$ are equipped with the natural σ-algebra $\mathcal{B}(G_{0,n})$ and $\mathcal{B}(H_{0,n})$, respectively.

Next, we give the precise definition of the discounted feedback control model.

DEFINITION 3.1. *A finite horizon D-FCM is a septuple*

$$\begin{aligned} \textit{D-FCM}: \Big(& \mathcal{X}_{0,n}, \mathcal{U}_{0,n-1}, \{\mathcal{U}_i(x_i) : x_i \in \mathcal{X}_i\}_{i=0}^{n-1}, \{Q_i(dx_i|x^{i-1}, u^{i-1}) : \\ & (x^{i-1}, u^{i-1}) \in \mathcal{X}_{0,i-1} \times \mathcal{U}_{0,i-1}\}_{i=0}^{n}, \{f_i\}_{i=0}^{n-1}, h_n, \alpha \Big) \end{aligned} \tag{3.1}$$

consisting of the items (a)–(c), (e)–(g) *of finite horizon D-MCM* (1.1), *while the controlled process distribution in* (d) *is replaced by the non-Markov collection* $\{Q_i(dx_i|x^{i-1}, u^{i-1}) : (x^{i-1}, u^{i-1}) \in \times_{j=0}^{i-1} \mathbb{K}_j\}_{i=0}^{n}$.

Next, we give the definitions of randomized, deterministic, and stationary control strategies or policies.

DEFINITION 3.2. *A randomized control strategy is a sequence* $\pi \triangleq \{\pi_0, \ldots, \pi_{n-1}\}$ *of stochastic kernels* $\pi_i(\cdot|\cdot)$ *on* $(\mathcal{U}_i, \mathcal{B}(\mathcal{U}_i))$ *conditioned on* $(H_{0,i}, \mathcal{B}(H_{0,i}))$ *(e.g.,* $\pi_i(du_i|x^i, u^{i-1})$*) satisfying*

$$\pi_i(\mathcal{U}_i(x_i)|x^i, u^{i-1}) = 1 \quad \textit{for every} \quad (x^i, u^{i-1}) \in H_{0,i}, \quad i = 0, 1, \ldots, n-1.$$

The set of all such policies is denoted by $\mathbf{\Pi}_{0,n-1}$.

A strategy $\pi \triangleq \{\pi_i : i = 0, \ldots, n-1\} \in \mathbf{\Pi}_{0,n-1}$ *is called the following:*

(a) *Deterministic feedback strategy if there exists a sequence* $g \triangleq \{g_j : j = 0, 1, \ldots, n-1\}$ *of measurable functions* $g_j : \times_{i=0}^{j-1} \mathbb{K}_i \times \mathcal{X}_j \longmapsto \mathcal{U}_j$, *such that for all* $(x^j, u^{j-1}) \in H_{0,j}$, $j \in \mathbb{N}^{n-1}$, $g_j(x_0, u_0, x_1, u_1, \ldots, x_{j-1}, u_{j-1}, x_j) \in \mathcal{U}_j(x_j)$, *and* $\pi_j(\cdot|x^j, u^{j-1})$ *assigns mass* 1 *to some point in* $\mathcal{U}_j$, *that is,*

$$\pi_i\left(A_i|x^i, u^{i-1}\right) = I_{A_i}\left(g_i\left(x^i, u^{i-1}\right)\right) \quad \forall A_i \in \mathcal{B}(\mathcal{U}_i), \quad i = 0, 1, \ldots, n-1,$$

where $I_{A_i}(\cdot)$ *is the indicator function of* $A_i \in \mathcal{B}(\mathcal{U}_i)$.
The set of deterministic feedback strategies is denoted by $\mathbf{\Pi}^{DF}_{0,n-1}$.

(b) *Deterministic Markov strategy if there exists a sequence* $g \triangleq \{g_j : j = 0, 1, \ldots, n-1\}$ *of measurable functions* $g_j : \mathcal{X}_j \to \mathcal{U}_j$ *satisfying* $g_j(x_j) \in \mathcal{U}_j(x_j)$ *for all* $x_j \in \mathcal{X}_j$, $j \in \mathbb{N}^{n-1}$, *and* $\pi_j(\cdot|x^j, u^{j-1})$ *is concentrated at* $g_j(x_j) \in \mathcal{U}_j(x_j)$ *for all* $(x^j, u^{j-1}) \in H_{0,j}$, $j \in \mathbb{N}^{n-1}$.
The set of deterministic Markov strategies is denoted by $\mathbf{\Pi}^{DM}_{0,n-1}$.

(c) *Deterministic stationary Markov strategy if there exists a measurable function* $g : \mathcal{X} \longrightarrow \mathcal{U}$ *such that* $g(x_t) \in \mathcal{U}(x_t)$ *for all* $x_t \in \mathcal{X}$, *and* $\pi_j(\cdot|x^j, u^{j-1})$ *assigns mass to some point* u_j *for all* $(x^j, u^{j-1}) \in H_{0,j}$, *e.g.,*

$$\pi_i(A_i|x^i, u^{i-1}) = I_{A_i}(g(x_i)) \quad \forall A_i \in \mathcal{B}(\mathcal{U}_i), \quad i = 0, \ldots, n-1.$$

The set of deterministic stationary Markov strategies is denoted by $\mathbf{\Pi}^{DS}_{0,n-1}$.

According to Definition 3.2, the set of control policies is nonempty, since we have assumed existence of measurable functions $g_j : \mathbb{K}_{0,j-1} \times \mathcal{X}_j \longrightarrow \mathcal{U}_j$ such that for all $x^j, u^{j-1} \in \mathbb{K}_{0,j-1} \times \mathcal{X}_j$, $g_j(x^j, u^{j-1}) \in \mathcal{U}_j(\mathcal{X}_j)$ for all $j \in \mathbb{N}^{n-1}$. Sufficient conditions for this to hold are in general obtained via measurable selection theorems [12]. For denumerable set (countable alphabet) $\mathcal{X}_j$ endowed with the discrete topology, any function is measurable. Given a controlled process $\{Q_i(\cdot|x^{i-1}, u^{i-1}) : (x^{i-1}, u^{i-1}) \in \mathbb{K}_{0,i-1}\}_{i=0}^n$ and a randomized control process $\{\pi_i(\cdot|x^i, u^{i-1}) : (x^i, u^{i-1}) \in \mathbb{K}_{0,i-1} \times \mathcal{X}_i\}_{i=0}^n \in \mathbf{\Pi}_{0,n-1}$ and the initial probability $\nu(\cdot) \in \mathcal{M}_1(\mathcal{X}_0)$, by Ionescu–Tulceu theorem [6] there exists a unique probability measure $\mathbf{Q}^\pi_\nu$ on $(\Omega, \mathcal{F})$ defined by

$$\begin{aligned} &\mathbf{Q}^\pi_\nu(dx_0, du_0, dx_1, du_1, \ldots, dx_{n-1}, du_{n-1}, dx_n) \\ &\quad = Q_0(dx_0)\pi_0(du_0|x_0) \otimes Q_1(dx_1|x_0, u_0)\pi_1(du_1|x^1, u_0) \\ &\qquad \otimes \cdots \otimes Q_{n-1}(dx_{n-1}|x^{n-2}, u^{n-2})\pi_{n-1}(du_{n-1}|x^{n-1}, u^{n-2}) \\ &\qquad \otimes Q_n(dx_n|x^{n-1}, u^{n-1}) \end{aligned} \tag{3.2}$$

such that

$$\begin{aligned} \mathbf{Q}^\pi_\nu(x_0 \in A) &= \nu(A), \quad A \in \mathcal{B}(\mathcal{X}_0), \\ \mathbf{Q}^\pi_\nu(u_j \in B|h_{0,j}) &= \pi_j(B|h_{0,j}), \quad B \in \mathcal{B}(\mathcal{U}_j), \\ \mathbf{Q}^\pi_\nu(x_{j+1} \in C|h_{0,j}, u_j) &= Q(C|h_{0,j}, u_j), \quad C \in \mathcal{B}(\mathcal{X}_{j+1}). \end{aligned}$$

Given the sample payoff

$$F^\alpha_{0,n}(x_0, u_0, x_1, u_1, \ldots, x_{n-1}, u_{n-1}, x_n) \triangleq \sum_{j=0}^{n-1} \alpha^j f_j(x_j, u_j) + \alpha^n h_n(x_n), \tag{3.3}$$

its expectation is

$$\begin{aligned} &\mathbb{E}_{\mathbf{Q}^\pi_\nu}\left\{F^\alpha_{0,n}(x_0, u_0, x_1, u_1, \ldots, x_{n-1}, u_{n-1}, x_n)\right\} \\ &\quad = \int F^\alpha_{0,n}(x_0, u_0, x_1, u_1, \ldots, x_{n-1}, u_{n-1}, x_n) \\ &\qquad \mathbf{Q}^\pi_\nu(dx_0, du_0, dx_1, du_1, \ldots, dx_{n-1}, du_{n-1}, dx_n). \end{aligned} \tag{3.4}$$

Note that the class of randomized strategies $\mathbf{\Pi}_{0,n-1}$ embeds deterministic feedback and Markov strategies.

3.1. Variation distance ambiguity. Next, we introduce the definitions of nominal controlled process distributions (for finite horizon D-FCM and D-MCM) and their corresponding ambiguous controlled process distributions.

For each $\pi \in \mathbf{\Pi}^{DF}_{0,n-1}$, $\pi \in \mathbf{\Pi}^{DM}_{0,n-1}$, and $\pi \in \mathbf{\Pi}^{DS}_{0,n-1}$, the nominal controlled process is described by a sequence of conditional distributions as follows.

Definition 3.3 (nominal controlled process distributions). *A nominal controlled state processes* $\{x^g = x^g_0, x^g_1, \ldots, x^g_n : \pi \in \mathbf{\Pi}^{DF}_{0,n-1}, \ \pi \in \mathbf{\Pi}^{DM}_{0,n-1}, \ \text{or} \ \pi \in \mathbf{\Pi}^{DS}_{0,n-1}\}$ *corresponds to a sequence of stochastic kernels as follows:*

(a) Feedback controlled process.

$$\textit{For every} \quad A \in \mathcal{B}(\mathcal{X}_j), \quad \text{Prob}(x_j \in A | x^{j-1}, u^{j-1}) = Q_j^o(A | x^{j-1}, u^{j-1}),$$

where $Q_j^o(A|x^{j-1}, u^{j-1}) \in \mathcal{Q}(\mathcal{X}_j|\mathbb{K}_{0,j-1})$ *for all* $j \in \mathbb{N}^n$.

(b) Markov controlled process.

$$\textit{For every} \quad A \in \mathcal{B}(\mathcal{X}_j), \quad \text{Prob}(x_j \in A | x^{j-1}, u^{j-1}) = Q_j^o(A | x_{j-1}, u_{j-1}),$$

where $Q_j^o(A|x_{j-1}, u_{j-1}) \in \mathcal{Q}(\mathcal{X}_j|\mathbb{K}_{j-1})$ *for all* $j \in \mathbb{N}^n$.

(c) Stationary Markov controlled process.

$$\textit{For every} \quad A \in \mathcal{B}(\mathcal{X}), \quad \text{Prob}(x_j \in A | x^{j-1}, u^{j-1}) = Q^o(A | x_{j-1}, u_{j-1}),$$

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

The class of controlled processes is described by the sequence of stochastic kernels,

$$\{Q_j(dx_j|x^{j-1}, u^{j-1}) \in \mathcal{Q}(\mathcal{X}_j|\mathbb{K}_{0,j-1}) : j = 0, \ldots, n\}$$

belonging to a total variation distance set as follows.

DEFINITION 3.4 (class of controlled process distribution). *Given a nominal controlled process stochastic kernel of Definition* 3.3, *and* $R_i \in [0, 2], 0 \le i \le n$, *the class of controlled process stochastic kernels is defined as follows:*

(a) Class with respect to feedback nominal controlled process.
Given a fixed $Q_j^o(\cdot|x^{j-1}, u^{j-1}) \in \mathcal{Q}(\mathcal{X}_j|\mathbb{K}_{0,j-1})$, $j = 0, 1, \ldots, n$, *the class of stochastic kernels is defined by*

$$\begin{aligned}&\mathbf{B}_{R_i}(Q_i^o)(x^{i-1}, u^{i-1})\\ &\quad \triangleq \Big\{ Q_i(\cdot|x^{i-1}, u^{i-1}) \in \mathcal{Q}(\mathcal{X}_i|\mathbb{K}_{0,i-1}) :\\ &\qquad ||Q_i(\cdot|x^{i-1}, u^{i-1}) - Q_i^o(\cdot|x^{i-1}, u^{i-1})||_{TV} \le R_i \Big\}, \quad i = 0, 1, \ldots, n.\end{aligned}$$

(b) Class with respect to Markov nominal controlled process.
Given a fixed $Q_j^o(\cdot|x_{j-1}, u_{j-1}) \in \mathcal{Q}(\mathcal{X}_j|\mathbb{K}_{j-1})$, $j = 0, 1, \ldots, n$, *the class of stochastic kernels is defined by*

$$\begin{aligned}&\mathbf{B}_{R_i}(Q_i^o)(x^{i-1}, u^{i-1})\\ &\quad \triangleq \Big\{ Q_i(\cdot|x^{i-1}, u^{i-1}) \in \mathcal{Q}(\mathcal{X}_i|\mathbb{K}_{0,i-1}) :\\ &\qquad ||Q_i(\cdot|x^{i-1}, u^{i-1}) - Q_i^o(\cdot|x_{i-1}, u_{i-1})||_{TV} \le R_i \Big\}, \quad i = 0, 1, \ldots, n.\end{aligned}$$

(c) Class with respect to stationary Markov nominal controlled process.
Given a fixed $Q^o(\cdot|x_{j-1}, u_{j-1}) \in \mathcal{Q}(\mathcal{X}|\mathbb{K})$, *the class of stochastic kernels is defined by*

$$\mathbf{B}_R(Q^0)(x, u) \triangleq \Big\{ Q(\cdot|x, u) \in \mathcal{Q}(\mathcal{X}|\mathbb{K}) : ||Q(\cdot|x, u) - Q^o(\cdot|x, u)||_{TV} \le R \Big\}.$$

Note that in Definition 3.4 (a), (b), although we use the same notation $\mathbf{B}_{R_i}(Q_i^o)(x^{i-1}, u^{i-1})$, these sets are different because the nominal distribution $Q_i^o(\cdot|\cdot)$ can be of feedback or Markov form. The above model is motivated by the fact that dynamic programming involves conditional expectation with respect to the collection of conditional distributions $\{Q_i(\cdot|x^{i-1}, u^{i-1}) \in \mathcal{Q}(\mathcal{X}_i|\mathbb{K}_{0,i-1}) : i = 0, \ldots, n\}$. Therefore, any ambiguity in these distributions will affect the optimality of the strategies.

3.2. Payoff functional. For each $\pi \in \mathbf{\Pi}_{0,n-1}^{DF}$ or $\pi \in \mathbf{\Pi}_{0,n-1}^{DM}$, the discounted payoff is defined by

$$J_{0,n}(\pi, Q_i : i = 0, \ldots, n) \triangleq \mathbb{E}_{\mathbf{Q}_\nu^\pi} \left\{ \sum_{j=0}^{n-1} \alpha^j f_j(x_j, u_j) + \alpha^n h_n(x_n) \right\}, \tag{3.5}$$

where $\mathbb{E}_{\mathbf{Q}_\nu^\pi}\{\cdot\}$ denotes expectation with respect to the true joint measure $\mathbf{Q}_\nu^\pi(dx^n, du^{n-1})$ defined by (3.2) such that $Q_i(\cdot|x^{i-1}, u^{i-1}) \in \mathbf{B}_{R_i}(Q_i^o)$, $i = 0, 1, \ldots, n$ (e.g., it belongs to the total variation distance ball of Definition 3.4).

Next, we introduce assumptions so that the maximization over the class of ambiguous measures is well-defined.

Assumption 3.5. The nominal system family satisfies the following assumption: The maps $\{f_j : \mathcal{X}_j \times \mathcal{U}_j \longmapsto \mathbb{R} : j = 0, 1, \ldots, n-1\}$, $h_n : \mathcal{X}_n \longmapsto \mathbb{R}$ are bounded, continuous, and nonnegative.

Note that it is possible to relax Assumption 3.5 to lower semicontinuous nonnegative functions bounded from below.

3.3. Minimax dynamic programming for finite horizon D-FCM and D-MCM. In this section, we shall apply the results of section 2 to formulate and solve minimax stochastic control under (a) finite horizon D-FCM ambiguity and (b) finite horizon D-MCM ambiguity.

3.3.1. Dynamic programming for finite horizon D-FCM subject to ambiguity. Utilizing the above formulation, next we define the minimax stochastic control problem, where the maximization is over a total variation distance ball, centered at the nominal conditional distribution $Q_i^o(dx_i|x^{i-1}, u^{i-1}) \in \mathcal{Q}(\mathcal{X}_i|\mathbb{K}_{0,i-1})$ having radius $R_i \in [0, 2]$ for $i = 0, 1, \ldots, n$.

PROBLEM 3.6. *Given a nominal feedback controlled process of Definition* 3.3(a), *an admissible policy set* $\mathbf{\Pi}_{0,n-1}^{DF}$, *and an ambiguity class* $\mathbf{B}_{R_k}(Q_k^o)(x^{k-1}, u^{k-1}), k = 0, \ldots, n$ *of Definition* 3.4(a), *find a* $\pi^* \in \mathbf{\Pi}_{0,n-1}^{DF}$ *and a sequence of stochastic kernels* $Q_k^*(dx_k|x^{k-1}, u^{k-1}) \in \mathbf{B}_{R_k}(Q_k^o)(x^{k-1}, u^{k-1})$, $k = 0, 1, \ldots, n$ *which solve the following minimax optimization problem:*

$$\begin{aligned} J_{0,n}(\pi^*, Q_k^* : k = 0, \ldots, n) = \inf_{\pi \in \mathbf{\Pi}_{0,n-1}^{DF}} \sup_{\substack{Q_k(\cdot|x^{k-1}, u^{k-1}) \in \mathbf{B}_{R_k}(Q_k^o)(x^{k-1}, u^{k-1}) \\ k=0,1,\ldots,n}} \\ \mathbb{E}_{\mathbf{Q}_\nu^\pi} \left\{ \sum_{k=0}^{n-1} \alpha^k f_k(x_k^g, u_k^g) + \alpha^n h_n(x_n^g) \right\}. \end{aligned} \tag{3.6}$$

Next, we apply dynamic programming to characterize the solution of (3.6), by first addressing the maximization. Define the payoff associated with the maximization problem

$$J_{0,n}(\pi, Q_k^* : k = 0, \ldots, n) \triangleq \sup_{\substack{Q_k(\cdot|x^{k-1}, u^{k-1}) \in \mathbf{B}_{R_k}(Q_k^o)(x^{k-1}, u^{k-1}) \\ k=0,1,\ldots,n}} J_{0,n}(\pi, Q_k : k = 0, \ldots, n).$$

For a given $\pi \in \mathbf{\Pi}_{0,n-1}^{DF}$, which defines $\{g_j : j = 0, \ldots, n-1\}$, and $\pi_{[k,m]} \equiv u_{[k,m]}^g$, denoting the restriction of policies in $[k, m]$, $0 \le k \le m \le n-1$, define the conditional expectation taken over the events $\mathcal{G}_{0,j} \triangleq \sigma\{x_0^g, \ldots, x_j^g, u_0^g, \ldots, u_j^g\}$ maximized over the

class $\mathbf{B}_{R_k}(Q_k^o)(x^{k-1},u^{k-1})$, $k=j+1,\dots,n$, as follows [7, 15]:

$$V_j(u^g_{[j,n-1]},\mathcal{G}_{0,j}) \triangleq \sup_{\substack{Q_k(\cdot|x^{k-1},u^{k-1})\in\mathbf{B}_{R_k}(Q_k^o)(x^{k-1},u^{k-1})\\ k=j+1,\dots,n}} \mathbb{E}_{\mathbf{Q}_\nu^\pi}\left\{\sum_{k=j}^{n-1}\alpha^k f_k(x_k^g,u_k^g)+\alpha^n h_n(x_n^g)|\mathcal{G}_{0,j}\right\}, \tag{3.7}$$

where $\mathbb{E}_{\mathbf{Q}_\nu^\pi}\{\cdot|\mathcal{G}_{0,j}\}$ denotes conditional expectation with respect to $\mathcal{G}_{0,j}$ calculated on the probability measure $\mathbf{Q}_\nu^\pi$. Then, $V_j(u^g_{[j,n-1]},\mathcal{G}_{0,j})$ satisfies the following dynamic programming equation [15]:

$$V_n(\mathcal{G}_{0,n}) = \alpha^n h_n(x_n^g), \tag{3.8}$$

$$V_j(u^g_{[j,n-1]},\mathcal{G}_{0,j}) = \sup_{Q_{j+1}(\cdot|x^j,u^j)\in\mathbf{B}_{R_{j+1}}(Q^0_{j+1})(x^j,u^j)} \Big\{\mathbb{E}_{Q_{j+1}(\cdot|x^j,u^j)}\Big\{\alpha^j f_j(x_j^g,u_j^g)+V_{j+1}(u^g_{[j+1,n-1]},\mathcal{G}_{0,j+1})\Big\}\Big\}, \tag{3.9}$$

where $\mathbb{E}_{Q_{j+1}(\cdot|x^j,u^j)}\{\cdot\}$ denotes expectation with respect to $Q_{j+1}(dx_{j+1}|\mathbb{K}_{0,j})$.

Next, we present the dynamic programming recursion for the minimax problem. Let $V_j(\mathcal{G}_{0,j})$ represent the minimax payoff on the future time horizon $\{j,j+1,\dots,n\}$ at time $j\in\mathbb{N}_+^n$ defined by

$$\begin{aligned} V_j(\mathcal{G}_{0,j}) &\triangleq \inf_{\pi\in\mathbf{\Pi}^{DF}_{j,n-1}}\sup_{\substack{Q_k(\cdot|x^{k-1},u^{k-1})\in\mathbf{B}_{R_k}(Q_k^0)(x^{k-1},u^{k-1})\\ k=j+1,\dots,n}} \left\{\mathbb{E}_{\mathbf{Q}_\nu^\pi}\left\{\sum_{k=j}^{n-1}\alpha^k f_k(x_k^g,u_k^g)+\alpha^n h_n(x_n^g)|\mathcal{G}_{0,j}\right\}\right\}\\ &= \inf_{\pi\in\mathbf{\Pi}^{DF}_{j,n-1}} V_j(u^g_{[j,n-1]},\mathcal{G}_{0,j}). \end{aligned} \tag{3.10}$$

Then, by reconditioning, we obtain

$$\begin{aligned} &V_j(\mathcal{G}_{0,j})\\ &\quad\triangleq \inf_{u\in\mathcal{U}_{ad}[j,n-1]}\sup_{\substack{Q_k(\cdot|x^{k-1},u^{k-1})\in\mathbf{B}_{R_k}(Q_k^0)(x^{k-1},u^{k-1})\\ k=j+1,\dots,n}}\\ &\qquad\left\{\mathbb{E}_{\mathbf{Q}_\nu^\pi}\left\{\alpha^j f_j(x_j^g,u_j^g)+\mathbb{E}_{\mathbf{Q}_\nu^\pi}\left\{\sum_{k=j+1}^{n-1}\alpha^k f_k(x_k^g,u_k^g)+\alpha^n h_n(x_n^g)|\mathcal{G}_{0,j+1}\right\}|\mathcal{G}_{0,j}\right\}\right\}. \end{aligned} \tag{3.11}$$

Hence, we deduce the following dynamic programming recursion:

$$V_n(\mathcal{G}_{0,n}) = \alpha^n h_n(x_n^g), \tag{3.12}$$

$$V_j(\mathcal{G}_{0,j}) \triangleq \inf_{u_j\in\mathcal{U}_j(x)}\sup_{Q_{j+1}(\cdot|x^j,u^j)\in\mathbf{B}_{R_{j+1}}(Q^0_{j+1})(x^j,u^j)} \Big\{\mathbb{E}_{Q_{j+1}(\cdot|x^j,u^j)}\Big\{\alpha^j f_j(x_j^g,u_j^g)+V_{j+1}(\mathcal{G}_{0,j+1})\Big\}\Big\}. \tag{3.13}$$

By applying the results of section 2 to (3.12), (3.13), we obtain the following theorem.

THEOREM 3.7. *Suppose there exists an optimal policy for Problem* 3.6, *and assume that* $V_{j+1}(\cdot) : \mathcal{X}_{0,j+1} \times \mathcal{U}_{0,j} \longmapsto [0,\infty)$ *in* (3.10) *is bounded continuous in* $x \in \mathcal{X}_{j+1}$, $j = 0, \ldots, n-1$.

(1) *The dynamic programming recursion is given by*

$$V_n(\mathcal{G}_{0,n}) = \alpha^n h_n(x_n^g), \tag{3.14}$$

$$V_j(\mathcal{G}_{0,j}) = \inf_{u_j \in \mathcal{U}_j(x)} \Bigg\{ \mathbb{E}_{Q^o_{j+1}} \Big(\alpha^j f_j(x_j^g, u_j^g) + V_{j+1}(\mathcal{G}_{0,j+1}) | \mathcal{G}_{0,j} \Big) + \frac{R_j}{2} \Bigg(\sup_{x_{j+1} \in \mathcal{X}_{j+1}} V_{j+1}(\mathcal{G}_{0,j+1}) - \inf_{x_{j+1} \in \mathcal{X}_{j+1}} V_{j+1}(\mathcal{G}_{0,j+1}) \Bigg) \Bigg\}. \tag{3.15}$$

Moreover,

$$V_j(\mathcal{G}_{0,j}) = \inf_{u_j \in \mathcal{U}_j(x)} \mathbb{E}_{Q^*_{j+1}} \Big\{ \alpha^j f_j(x_j^g, u_j^g) + V_{j+1}(\mathcal{G}_{0,j+1}) | \mathcal{G}_{0,j} \Big\}, \tag{3.16}$$

where the optimal conditional distributions $\{Q_j^* : j = 0, 1, \ldots, n-1\}$ *are given by*

$$Q^*_{j+1}\left(\mathcal{X}^+_{j+1} | x^j, u^j\right) = Q^o_{j+1}(\mathcal{X}^+_{j+1} | x^j, u^j) + \frac{R_{j+1}}{2} \in [0,1], \quad (x^j, u^j) \in \mathbb{K}_{0,j}, \tag{3.17}$$

$$Q^*_{j+1}\left(\mathcal{X}^-_{j+1} | x^j, u^j\right) = Q^o_{j+1}(\mathcal{X}^-_{j+1} | x^j, u^j) - \frac{R_{j+1}}{2} \in [0,1], \quad (x^j, u^j) \in \mathbb{K}_{0,j}, \tag{3.18}$$

$$Q^*_{j+1}\left(A | x^j, u^j\right) = Q^o_{j+1}(A | x^j, u^j), \ \forall A \subseteq \mathcal{X}_{j+1} \backslash \mathcal{X}^+_{j+1} \cup \mathcal{X}^-_{j+1}, \quad (x^j, u^j) \in \mathbb{K}_{0,j}, \tag{3.19}$$

and[2]

$$\mathcal{X}^+_{j+1} \triangleq \Big\{ x_{j+1} \in \mathcal{X}_{j+1} : V_{j+1}(\mathcal{G}_{0,j}, x_{j+1}) = \sup \big\{ V_{j+1}(\mathcal{G}_{0,j}, x_{j+1}) : x_{j+1} \in \mathcal{X}_{j+1} \big\} \Big\}, \tag{3.20}$$

$$\mathcal{X}^-_{j+1} \triangleq \Big\{ x_{j+1} \in \mathcal{X}_{j+1} : V_{j+1}(\mathcal{G}_{0,j}, x_{j+1}) = \inf \big\{ V_{j+1}(\mathcal{G}_{0,j}, x_{j+1}) : x_{j+1} \in \mathcal{X}_{j+1} \big\} \Big\}. \tag{3.21}$$

(2) *The total payoff is given by*

$$J_{0,n}(\pi^*, Q_i^* : i = 0, \ldots, n-1) = \sup_{Q_0(\cdot) \in \mathbf{B}_{R_0}(Q^o)} \mathbb{E}_{Q_0} \Big\{ V_0(\mathcal{G}_{0,0}) \Big\}. \tag{3.22}$$

[2]Note that the notation Σ^0 and Σ_0 in section 2 is identical to the notation $\mathcal{X}^+_{j+1}$ and $\mathcal{X}^-_{j+1}$, respectively.

Proof.

(1) Consider (3.13) expressed in integral form

$$V_j(\mathcal{G}_{0,j}) = \inf_{u_j \in \mathcal{U}_j(x)} \left\{ \alpha^j f_j(x_j, u_j) + \sup_{Q_{j+1}(\cdot|x^j,u^j) \in \mathbf{B}_{R_{j+1}}(Q^o_{j+1})(x^j,u^j)} \int V_{j+1}(\mathcal{G}_{0,j}, z) Q_{j+1}(dz|x^j, u^j) \right\}. \tag{3.23}$$

By applying (2.5), we obtain (3.14), (3.15), while (3.17)–(3.21) follow as well.

(2) By evaluating (3.10) at $j = 0$, we obtain (3.22). This completes the derivation. ∎

By Theorem 3.7, the maximizing measure is given by (3.17)–(3.19), and it is a functional of the nominal measure. At this stage, we cannot claim that the maximizing measure is Markovian, and hence the optimal strategy is not necessarily Markov. Therefore, the computation of optimal strategies using non-Markov nominal controlled processes is computationally intensive. Next, we restrict the minimax formulation to Markov controlled nominal processes.

3.3.2. Dynamic programming for finite horizon D-MCM subject to ambiguity. Consider the Markov nominal controlled processes, based on Definition 3.3(b), and define

$$V_j(u^g_{[j,n-1]}, \mathcal{G}_{0,j}) \triangleq \sup_{\substack{Q_k(\cdot|x^{k-1},u^{k-1}) \in \mathbf{B}_{R_k}(Q^o_k)(x^{k-1},u^{k-1}) \\ k=j+1,\ldots,n}} \mathbb{E}_{\mathbf{Q}^\pi_\nu} \left\{ \sum_{k=j}^{n-1} \alpha^k f_k(x^g_k, u^g_k) + \alpha^n h_n(x^g_n) | \mathcal{G}_{0,j} \right\}.$$

In view of section 2, specifically, the relation between the maximizing distribution and the nominal distribution (2.1)–(2.3), which also apply to conditional distributions, we deduce that the maximization conditional distribution $Q^*_i(dx_i|x^{i-1}, u^{i-1})$ is Markovian, and hence $Q^*_i(dx_i|x^{i-1}, u^{i-1}) = Q^*_i(dx_i|x_{i-1}, u_{i-1})$ for all $(x^{i-1}, u^{i-1}) \in \mathbb{K}_{0,i-1}$. This observation can be verified by checking expressions (3.17)–(3.19). Then we define

$$V_j(u^g, x) \triangleq \sup_{\substack{Q_k(\cdot|x_{k-1},u_{k-1}) \in \mathbf{B}_{R_k}(Q^o_k)(x_{k-1},u_{k-1}) \\ k=j+1,\ldots,n}} \mathbb{E}_{\mathbf{Q}^\pi_\nu} \left\{ \sum_{k=j}^{n-1} \alpha^k f_k(x^g_k, u^g_k) + \alpha^n h_n(x^g_n) | x \right\}. \tag{3.24}$$

Utilizing the above observations, we obtain the analogue of Theorem 3.7 for finite horizon D-MCM, as follows.

Define the value function

$$V_j(x) = \inf_{\pi \in \mathbf{\Pi}^{DM}_{j,n-1}} \sup_{\substack{Q_k(\cdot|x_{k-1},u_{k-1}) \in \mathbf{B}_{R_k}(Q^o_k)(x_{k-1},u_{k-1}) \\ k=j+1,\ldots,n}} \mathbb{E}_{\mathbf{Q}^\pi_\nu} \left\{ \sum_{k=j}^{n-1} \alpha^k f_k(x^g_k, u^g_k) + \alpha^n h_n(x^g_n) | x \right\}. \tag{3.25}$$

Then we obtain the following theorem.

THEOREM 3.8. *Suppose there exists an optimal policy for Problem* 3.6 *for the class of Markov nominal controlled process of Definition* 3.4(b). *Then the following hold:*

(1) *If the infimum over feedback strategies in* (3.25) *exists, it is Markov* $\pi \in \mathbf{\Pi}_{0,n-1}^{DM}$.

(2) *The value function* $V_j(x)$ *satisfies the dynamic programming recursion*

$$V_n(x) = \alpha^n h_n(x), \quad x \in \mathcal{X}_n, \tag{3.26}$$

$$V_j(x) = \inf_{u \in \mathcal{U}_j(x)} \sup_{Q_{j+1}(\cdot|x,u) \in \mathbf{B}_{R_{j+1}}(Q_{j+1}^o)(x,u)} \mathbb{E}_{Q_{j+1}(\cdot|x,u)} \Big\{ \alpha^j f_j(x,u) + V_{j+1}(x_{j+1}) \Big\}, \quad x \in \mathcal{X}_j. \tag{3.27}$$

(3) *Assume that* $V_{j+1}(\cdot) : \mathcal{X}_{j+1} \to [0,\infty)$ *is bounded continuous in* $x \in \mathcal{X}_{j+1}$, $j = 0, \ldots, n-1$, *and then the dynamic programming recursion is given by*

$$V_n(x) = \alpha^n h_n(x), \quad x \in \mathcal{X}_n, \tag{3.28}$$

$$V_j(x) = \inf_{u \in \mathcal{U}_j(x)} \Bigg\{ \alpha^j f_j(x,u) + \int_{\mathcal{X}_{j+1}} V_{j+1}(z) Q_{j+1}^o(dz|x,u) + \frac{R_j}{2} \left(\sup_{z \in \mathcal{X}_{j+1}} V_{j+1}(z) - \inf_{z \in \mathcal{X}_{j+1}} V_{j+1}(z) \right) \Bigg\}, \quad x \in \mathcal{X}_j. \tag{3.29}$$

Moreover,

$$V_j(x) = \inf_{u \in \mathcal{U}_j(x)} \mathbb{E}_{Q_{j+1}^*} \Big\{ \alpha^j f_j(x_j^g, u_j^g) + V_{j+1}(x_{j+1}) | x_j = x \Big\}, \tag{3.30}$$

where the optimal conditional distribution $\{Q_j^*(\cdot|\cdot,\cdot) : j = 0, 1, \ldots, n-1\}$ *is given by*

$$Q_{j+1}^* \left(\mathcal{X}_{j+1}^+ | x_j, u_j \right) = Q_{j+1}^o(\mathcal{X}_{j+1}^+ | x_j, u_j) + \frac{R_{j+1}}{2} \in [0,1], \quad (x_j, u_j) \in \mathbb{K}_j, \tag{3.31}$$

$$Q_{j+1}^* \left(\mathcal{X}_{j+1}^- | x_j, u_j \right) = Q_{j+1}^o(\mathcal{X}_{j+1}^- | x_j, u_j) - \frac{R_{j+1}}{2} \in [0,1], \quad (x_j, u_j) \in \mathbb{K}_j, \tag{3.32}$$

$$Q_{j+1}^* \left(A | x_j, u_j \right) = Q_{j+1}^o(A | x_j, u_j), \ \forall A \subseteq \mathcal{X}_{j+1} \backslash \mathcal{X}_{j+1}^+ \cup \mathcal{X}_{j+1}^-, \quad (x_j, u_j) \in \mathbb{K}_j, \tag{3.33}$$

and

$$\mathcal{X}_{j+1}^+ \triangleq \Big\{ x_{j+1} \in \mathcal{X}_{j+1} : V_{j+1}(x_{j+1}) = \sup\{ V_{j+1}(x_{j+1}) : x_{j+1} \in \mathcal{X}_{j+1} \} \Big\}, \tag{3.34}$$

$$\mathcal{X}_{j+1}^- \triangleq \Big\{ x_{j+1} \in \mathcal{X}_{j+1} : V_{j+1}(x_{j+1}) = \inf\{ V_{j+1}(x_{j+1}) : x_{j+1} \in \mathcal{X}_{j+1} \} \Big\}. \tag{3.35}$$

(4) *The total minimax payoff is*

$$J_{0,n}(g^*, \{Q_i^*\}_{i=0}^n) = \sup_{Q_0(\cdot) \in \mathbf{B}_{R_0}(Q_0^o)} \mathbb{E}_{Q_0} \Big\{ V_0(x_0) \Big\}. \tag{3.36}$$

Proof.

(1) Since the nominal controlled process is Markov, from Theorem 3.7, (3.17)–(3.19) we deduce that the maximizing measure is also Markov. By the same arguments as in [15], we can show that if the infimum over $u \in \mathbf{\Pi}_{0,n-1}^{DF}$ in (3.25) exists, then it is Markov, and hence $u \in \mathbf{\Pi}_{0,n-1}^{DM}$.

(2) By reconditioning, we deduce that the value function satisfies the dynamic programming equation (3.26), (3.27).
(3) By definition, (3.27) is also equivalent to

$$V_j(x) = \inf_{u\in\mathcal{U}(x)} \Big\{ \alpha^j f_j(x,u) + \sup_{Q_{j+1}(\cdot|x,u)\in \mathbf{B}_{R_{j+1}}(Q^o_{j+1})(x,u)} \int_{\mathcal{X}_{j+1}} V_{j+1}(z) Q_{j+1}(dz|x,u) \Big\}.$$

Hence, by applying the results of section 2, we obtain (3.28)–(3.33).
(4) By evaluating (3.25) at $j = 0$, we obtain (3.36). This completes the derivation. □

REMARK 3.9. We make the following observations regarding Theorem 3.8:
(a) The dynamic programming equation (3.28), (3.29) involves in its right-hand side the oscillator seminorm of $V_{j+1}(\cdot)$.
(b) The dynamic programming recursion (3.28), (3.29) can be applied to a controlled process with continuous alphabets and to a controlled process with finite or countable alphabets, such as Markov decision models.

Next, we show that for any $j \in \mathbb{N}^{n-1}$, the minimax payoff $V_j(x) \equiv V_j^R(x)$ as a function of R_j is nondecreasing and concave.

LEMMA 3.10. *Suppose that the conditions of Theorem* 3.8 *hold and in addition* $R_j = R$, $j = 1,\dots,n$. *The minimax payoff* $V_j^R(x) \equiv V_j(x)$ *defined by* (3.25) *is a nondecreasing concave function of* R.

Proof. Consider two values for $R^1, R^2 \in \mathbb{R}^+$ such that $0 \le R^1 \le R^2$. Since

$$\mathbf{B}_{R^1}(Q^o_k)(x_{k-1},u_{k-1}) \subseteq \mathbf{B}_{R^2}(Q^o_k)(x_{k-1},u_{k-1}),$$

then for every $Q_k(\cdot,x_{k-1},u_{k-1}) \in \mathbf{B}_{R^1}(Q^o_k)(x_{k-1},u_{k-1})$ we have $Q_k(\cdot,x_{k-1},u_{k-1}) \in B_{R^2}(Q^o_k)(x_{k-1},u_{k-1})$, $k = j+1,\dots,n-1$. Hence, $V_j^{R^1}(x) \le V_j^{R^2}(x)$ and thus $V_j^R(x)$ is a nondecreasing function of $R \in \mathbb{R}^+$.

Next, for a fixed $\pi \in \Pi^{DM}_{j,n-1}$, consider two points (R^1, V_j^{π,R^1}), (R^2, V_j^{π,R^2}) such that $\{Q^1_k(\cdot|x_{k-1},u_{k-1}) : k = j+1,\dots,n\}$ achieves the supremum in (3.24) for R^1, and $\{Q^2_k(\cdot|x_{k-1},u_{k-1}) : k = j+1,\dots,n\}$ achieves the supremum in (3.24) for R^2. Then,

$$\begin{aligned} &||Q^1_k(\cdot|x_{k-1},u_{k-1}) - Q^o_k(\cdot|x_{k-1},u_{k-1})||_{TV} \le R^1, \quad k = j+1,\dots,n-1,\\ &||Q^2_k(\cdot|x_{k-1},u_{k-1}) - Q^o_k(\cdot|x_{k-1},u_{k-1})||_{TV} \le R^2, \quad k = j+1,\dots,n-1. \end{aligned}$$

For any $\lambda \in (0,1)$, we have

$$\begin{aligned} &||\lambda Q^1_k(\cdot|x_{k-1},u_{k-1}) + (1-\lambda)Q^2_k(\cdot|x_{k-1},u_{k-1}) - Q^o_k(\cdot|x_{k-1},u_{k-1})||_{TV}\\ &\le \lambda||Q^1_k(\cdot|x_{k-1},u_{k-1}) - Q^o_k(\cdot|x_{k-1},u_{k-1})||_{TV} + (1-\lambda)||Q^2_k(\cdot|x_{k-1},u_{k-1})\\ &\quad - Q^o_k(\cdot|x_{k-1},u_{k-1})||_{TV} \le \lambda R^1 + (1-\lambda)R^2, \qquad k = j+1,\dots,n. \end{aligned} \tag{3.37}$$

Define $Q^*_k(\cdot|x_{k-1},u_{k-1}) \triangleq \lambda Q^1_k(\cdot|x_{k-1},u_{k-1}) + (1-\lambda)Q^2_k(\cdot|x_{k-1},u_{k-1})$, $R = \lambda R^1 + (1-\lambda)R^2$. By (3.37), $Q^*_k \in \mathbf{B}_R(Q^o_k)(x_{k-1},u_{k-1})$, $k = j+1,\dots,n$. Define the unique

probability measure

$$Q^*_{j+1,n}(dx^n||u^n) \triangleq \lambda \otimes_{k=j+1}^{n} Q^1_k(dx_k|x_{k-1},u_{k-1}) + (1-\lambda) \otimes_{k=j+1}^{n} Q^2_k(dx_k|x_{k-1},u_{k-1}).$$

Then,

$$V_j^{\pi,R}(x) \geq \int \left(\sum_{k=j}^{n-1} f_k(x_k,u_k) + h_n(x_n) \right) Q^*_{j+1,n}(dx^n||u^n)).$$

Hence,

$$\begin{aligned} V_j^{\pi,R}(x_j) &= \text{RHS of (3.24)} \\ &\geq \lambda \int \left(\sum_{k=j}^{n-1} f_k(x_k,u_k) + h_n(x_n) \right) \otimes_{k=j+1}^{n} Q^1_k(dx_k|x_{k-1},u_{k-1}) \\ &\quad + (1-\lambda) \int \left(\sum_{k=j}^{n-1} f_k(x_k,u_k) + h_n(x_n) \right) \otimes_{k=j+1}^{n} Q^2_k(dx_k|x_{k-1},u_{k-1}) \\ &= \lambda V_j^{\pi,R^1}(x_j) + (1-\lambda) V_j^{\pi,R^2}(x_j), \quad j = 0,\dots,n-1. \end{aligned}$$

Hence, for any $\pi \in \mathbf{\Pi}^{DM}_{j,n-1}$, $V_j^{\pi,R}(x_j)$ is a concave function of R, and thus it is also concave for the $\pi \in \mathbf{\Pi}^{DM}_{j,n-1}$, which achieve the infimum in (3.25). □

This concavity property of the payoff is also verified in the examples presented in section 4.

REMARK 3.11. *The previous results apply to randomized strategies as well.*

3.4. Minimax dynamic programming for infinite horizon D-MCM subject to ambiguity. In this section, we consider the infinite horizon version of the finite horizon D-MCM, and we derive similar results. In addition, we show that the operator associated with the dynamic programming equation is contractive, and we introduce a new policy iteration algorithm.

Consider the problem of minimizing the finite horizon cost

$$\sup_{\substack{Q_k(\cdot|x,u)\in \mathbf{B}_{R_k}(Q^o_k)(x,u) \\ k=0,1,\dots,n}} \mathbb{E}_{\mathbf{Q}^\pi_\nu} \left\{ \sum_{j=0}^{n-1} \alpha^j f(x^g_j,u^g_j) \right\} \tag{3.38}$$

with $0 < \alpha < 1$. By Theorem 3.8, the value function of (3.38), denoted by $V_j(x)$, $j = 0,\dots,n$, $x \in \mathcal{X}_j$, satisfies the dynamic programming equations (3.28), (3.29) with $h_n = 0$, $R_j = R$, $\mathcal{X}_j = \mathcal{X}$, $\mathcal{U}_j = \mathcal{U}$, $\mathcal{U}_j(x) = \mathcal{U}(x)$, and $Q^o_j(\cdot|\cdot) = Q^o(\cdot|\cdot)$. Define $v_i(x) = \alpha^{i-n} V_{n-i}(x)$, where $0 \leq i \leq n$ is the time to go (see [21]). Then,

$$v_0(x) = 0, \tag{3.39}$$

$$\begin{aligned} v_i(x) = \inf_{u\in\mathcal{U}(x)} \Bigg\{ & f(x,u) + \alpha \int_{\mathcal{X}} v_{i-1}(z) Q^o(dz|x,u) \\ & + \alpha \frac{R}{2} \left(\sup_{z\in\mathcal{X}} v_{i-1}(z) - \inf_{z\in\mathcal{X}} v_{i-1}(z) \right) \Bigg\}. \end{aligned} \tag{3.40}$$

In contrast with the finite horizon case, the one given by (3.39)–(3.40) proceeds from lower to higher values of indices i. The dynamic programming for the discounted cost

$$\mathbb{E}_{\mathbf{Q}_\nu^\pi}\left\{\sum_{j=0}^{\infty}\alpha^j f(x_j^g,u_j^g)\right\} \tag{3.41}$$

is given by

$$v_\infty(x)=\inf_{u\in\mathcal{U}(x)}\left\{f(x,u)+\alpha\int_{\mathcal{X}}v_\infty(z)Q^o(dz|x,u)+\alpha\frac{R}{2}\left(\sup_{z\in\mathcal{X}}v_\infty(z)-\inf_{z\in\mathcal{X}}v_\infty(z)\right)\right\}. \tag{3.42}$$

The maximizing conditional distribution is

$$Q^*\left(\mathcal{X}^+|x,u\right)=Q^o(\mathcal{X}^+|x,u)+\frac{R}{2}\in[0,1],\quad (x,u)\in\mathbb{K}, \tag{3.43}$$

$$Q^*\left(\mathcal{X}^-|x,u\right)=Q^o(\mathcal{X}^-|x,u)-\frac{R}{2}\in[0,1],\quad (x,u)\in\mathbb{K}, \tag{3.44}$$

$$Q^*\left(A|x,u\right)=Q^o(A|x,u)\quad \forall A\subseteq\mathcal{X}\setminus\mathcal{X}^+\cup\mathcal{X}^-,\quad (x,u)\in\mathbb{K}, \tag{3.45}$$

where

$$\mathcal{X}^+\triangleq\left\{x\in\mathcal{X}:V(x)=\sup\{V(x):x\in\mathcal{X}\}\right\}, \tag{3.46}$$

$$\mathcal{X}^-\triangleq\left\{x\in\mathcal{X}:V(x)=\inf\{V(x):x\in\mathcal{X}\}\right\}. \tag{3.47}$$

Next, we show that the operator in the right-hand side of (3.42) is contractive.

LEMMA 3.12. *Let L be the class of all measurable functions $V:\mathcal{X}\longrightarrow\mathbb{R}$ with finite norm $||V||\triangleq\max_{x\in\mathcal{X}}|V(x)|$, and $T:L\longmapsto L$ defined by*

$$(TV)(x)=\inf_{u\in\mathcal{U}(x)}\left\{f(x,u)+\alpha\int_{\mathcal{X}}V(z)Q^o(dz|x,u)+\alpha\frac{R}{2}\left(\sup_{z\in\mathcal{X}}V(z)-\inf_{z\in\mathcal{X}}V(z)\right)\right\}. \tag{3.48}$$

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

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

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

Let

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

Then,

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

where (a) is obtained by applying (2.1) with $\ell \equiv \alpha V_1$, $\nu^*(\cdot) \equiv Q^{V_1}(\cdot|\cdot)$, $\mu(\cdot) \equiv Q^o(\cdot|\cdot)$, and (b) is obtained by first applying (2.1) as in (a) with Q^{V_2} and then replacing Q^{V_2} by Q^{V_1} which is suboptimal and hence the upper bound. By reversing the roles of V_1 and V_2, we get $(TV_2)(x) - (TV_1)(x) \le \alpha ||V_2 - V_1||$. Hence, $|(TV_1)(x) - (TV_2)(x)| \le \alpha ||V_1 - V_2||$ for all $x \in \mathcal{X}$, and

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

which implies that the operator $T : L \longmapsto L$ is a contraction. ∎

Utilizing Lemma 3.12, we obtain the following theorem, which is analogous to the classical result given in [21].

THEOREM 3.13. *Assume that* $v_\infty \in BC^+(\mathcal{X})$ *and* $\sup_{z\in\mathcal{X}} v_\infty(z)$, $\inf_{z\in\mathcal{X}} v_\infty(z)$ *are finite.*

(1) *The dynamic programming equation*

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

has a unique solution.

(2) *Moreover,*

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

(3) *The mapping T defined by*

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

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

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

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

Proof.

(1) This follows from [21, Theorem 6.3.6, part (a)].

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

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

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

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

and so

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

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

(3) This follows from Lemma 3.12.

(4) This follows from [21, Theorem 6.3.6, part (b)]. ∎

3.4.1. Policy iteration algorithm. Next, we present a modified version of the classical policy iteration algorithm [15]. From part 4 of Theorem 3.8, the policy improvement and policy evaluation steps of a policy iteration algorithm must be performed using the maximizing conditional distribution obtained under total variation distance ambiguity constraint. Hence, in addition to the classical case, in which the policy improvement and evaluation steps are performed using the nominal conditional distribution, here, under the assumption that $f(\cdot)$ is bounded and nonnegative, by invoking the results developed in earlier sections we propose a modified algorithm which is expected to converge to a stationary policy in a finite number of iterations,

since both state space $\mathcal{X}$ and control space $\mathcal{U}$ are finite sets, and at each iteration a better stationary policy will be obtained.

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

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

Write $z \leq y$ if $z(i) \leq y(i)$ for all $i \in \mathbb{Z}^n \triangleq \{1, 2, \ldots, n\}$, and $z < y$ if $z \leq y$ and $z \neq y$. For a stationary control law g, let

$$f(g) = \begin{pmatrix} f(x_1, g(x_1)) & \cdots & f(x_n, g(x_n)) \end{pmatrix}^T$$

and define each entry of the transition matrix $Q^o(g) \in \mathbb{R}^{n \times n}$ by $Q^o_{ij}(g) = Q^o(x_j|x_i, g(x_i)) \equiv Q^{g,o}(x_i|x_j)$. Rewrite (3.48) (with $\sup_{z \in \mathcal{X}} V(z)$ denoting componentwise supremum, and similarly for the infimum) as

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

which by Theorem 3.8 is equivalent to

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

where $Q^*(g) \in \mathbb{R}^{n \times n}$ and is given by (3.43)–(3.45). Note that the minimization is taken componentwise, i.e., $g(x_1)$ is the minimum of the first component of $f(g) + \alpha Q^*(g)V$ and so on. For each stationary policy g, define $T(g) : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ by

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

Then, $T(g)$ is a contraction mapping on the space of bounded continuous functions to itself, and from Theorem 3.13 it follows that

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

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

ALGORITHM 3.14 (policy iteration). *Consider the notation above.*

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

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

Identify the support sets using (3.46)–(3.47) *and the analogue of Σ_k of section* 2, *and construct the matrix $Q^*(g_0)$ using* (3.43)–(3.45). *Solve the equation*

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

1. *For $m = m + 1$ while $\min_{g \in \mathbb{R}^n} \{f(g) + \alpha Q^*(g)V_{Q^*}(g_{m-1})\} < V_{Q^*}(g_{m-1})$, do the following:*
   (a) *(Policy improvement) Let $g_m \in \mathbb{R}^n$ be such that*

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

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

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

Identify the support sets using (3.46)–(3.47) *and construct the matrix* $Q^*(g_m)$ *using* (3.43)–(3.45)*. Solve the equation*

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

2. *Set* $g^* = g_m$.

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

4. Examples. In section 4.1, we illustrate an application of the finite horizon minimax problem to the well-known machine replacement example, and in section 4.2 we illustrate an application of the infinite horizon minimax problem for discounted cost by employing the policy iteration algorithm. In section 4.3, we illustrate an application of the finite horizon minimax problem to the inventory control example by considering as a constraint the total variation distance between the nominal and the true probability distributions. Comparisons are included for the case for which the constraint is replaced by the relative entropy between the nominal and the true probability distributions.

4.1. Finite horizon MCM—machine replacement example. Consider a machine replacement example inspired by [5]. Specifically, we have a machine that is either running or is broken down. If it runs throughout one week, it makes a profit of €100 for that week. If it fails during the week, the profit is zero for that week. If it is running at the beginning of the week and we perform preventive maintenance, the probability that it will fail during the week is 0.4. If we do not perform such maintenance, the probability of failure is 0.7. The maintenance cost is set to €20. When the machine is broken down at the start of the week, it may either be repaired at a cost of €40, in which case it will fail during the week with a probability of 0.4, or it may be replaced at a cost of €150 by a new machine that is guaranteed to run through its first week of operation. Assume that after $N > 1$ weeks the machine, irrespective of its state, is scrapped with no cost.

The system dynamics is of the form $x_{k+1} = f_k(x_k, u_k, w_k)$, $k = 0, 1, \ldots, N-1$, where the state x_k is an element of a space $S_k = \{\mathbf{R}, \mathbf{B}\}$, $\mathbf{R}$ = machine running, $\mathbf{B}$ = machine broken down, the control u_k is an element of a space $U_k(x_k)$, $U_k(\mathbf{R}) = \{m, nm\}$, m = maintenance, nm = no maintenance, $U_k(\mathbf{B}) = \{r, s\}$, r = repair, s = replace. The random disturbance has a nominal conditional distribution $w_k \sim \mu(\cdot|x_k, u_k)$.

Such a system can be described in terms of the discrete-time system equation $x_{k+1} = w_k$, where the nominal probability distribution of w_k is given by

$$\begin{aligned}
\mu(w_k = \mathbf{R}|x_k = \mathbf{R}, u_k = m) &= 0.6, & \mu(w_k = \mathbf{B}|x_k = \mathbf{R}, u_k = m) &= 0.4,\\
\mu(w_k = \mathbf{R}|x_k = \mathbf{R}, u_k = nm) &= 0.3, & \mu(w_k = \mathbf{B}|x_k = \mathbf{R}, u_k = nm) &= 0.7,\\
\mu(w_k = \mathbf{R}|x_k = \mathbf{B}, u_k = r) &= 0.6, & \mu(w_k = \mathbf{B}|x_k = \mathbf{B}, u_k = r) &= 0.4,\\
\mu(w_k = \mathbf{R}|x_k = \mathbf{B}, u_k = s) &= 1, & \mu(w_k = \mathbf{B}|x_k = \mathbf{B}, u_k = s) &= 0
\end{aligned}$$

and the input costs C_u are given by the following: if $u = m$, then $C_m = €20$, if $u = nm$, then $C_{nm} = €0$, if $u = r$, then $C_r = €40$, and if $u = s$, then $C_s = €150$.

The cost per stage is $g_k(x_k,u_k,w_k)=C_{u_k}$ if $w_k=\mathbf{R}$, and $g_k(x_k,u_k,w_k)=C_{u_k}+100$ if $w_k=\mathbf{B}$. Since it is assumed that after N weeks the machine, irrespective of its state, is scrapped without incurring any cost, the terminal cost is $g_N(\mathbf{R})=g_N(\mathbf{B})=0$.

The dynamic programming algorithm for the minimax problem subject to total variation distance uncertainty is given by

$$V_N(x_N)=0, \tag{4.1}$$

$$\begin{aligned} V_k(x_k) &= \min_{u_k\in U_k(x_k)}\ \max_{\nu(dw_k|x_k,u_k):||\nu(\cdot|x_k,u_k)-\mu(\cdot|x_k,u_k)||_{TV}\le R} \\ &\quad \mathbb{E}\Big\{g_k(x_k,u_k,w_k)+V_{k+1}(f(x_k,u_k,w_k))\Big\} \\ &= \min_{u_k\in U_k(x_k)}\ \max_{\nu(dw_k|x_k,u_k):||\nu(\cdot|x_k,u_k)-\mu(\cdot|x_k,u_k)||_{TV}\le R} \mathbb{E}\Big\{\ell_k(x_k,u_k,w_k)\Big\}, \end{aligned} \tag{4.2}$$

where $\ell_k(x_k,u_k,w_k)=g_k(x_k,u_k,w_k)+V_{k+1}(w_k)$, $k=0,1,\dots,N-1$. To address the maximization problem in (4.2), for each $k=0,1,\dots,N-1$, $x_k\in\{\mathbf{R},\mathbf{B}\}$, and $u_k\in\{m,nm,r,s\}$, define the maximum and minimum values of $\ell(x_k,u_k,w_k)$ by

$$\ell_{\max}(x_k,u_k)\triangleq\max_{w_k\in\{\mathbf{R},\mathbf{B}\}}\ell(x_k,u_k,w_k),\qquad \ell_{\min}(x_k,u_k)\triangleq\min_{w_k\in\{\mathbf{R},\mathbf{B}\}}\ell(x_k,u_k,w_k)$$

and its corresponding support sets by $\Sigma^0=\{w_k\in\{\mathbf{R},\mathbf{B}\}:\ell(x_k,u_k,w_k)=\ell_{\max}(x_k,u_k)\}$, and $\Sigma_0=\{w_k\in\{\mathbf{R},\mathbf{B}\}:\ell(x_k,u_k,w_k)=\ell_{\min}(x_k,u_k)\}$. By employing (2.5), the maximizing conditional probability distribution of the random parameter w_k is given by

$$\alpha=\min\left(\frac{R}{2},1-\mu(\Sigma^0|x_k,u_k)\right), \tag{4.3a}$$

$$\nu^*(\Sigma^0|x_k,u_k)=\mu(\Sigma^0|x_k,u_k)+\alpha,\qquad \nu^*(\Sigma_0|x_k,u_k)=\Big(\mu(\Sigma_0|x_k,u_k)-\alpha\Big)^+. \tag{4.3b}$$

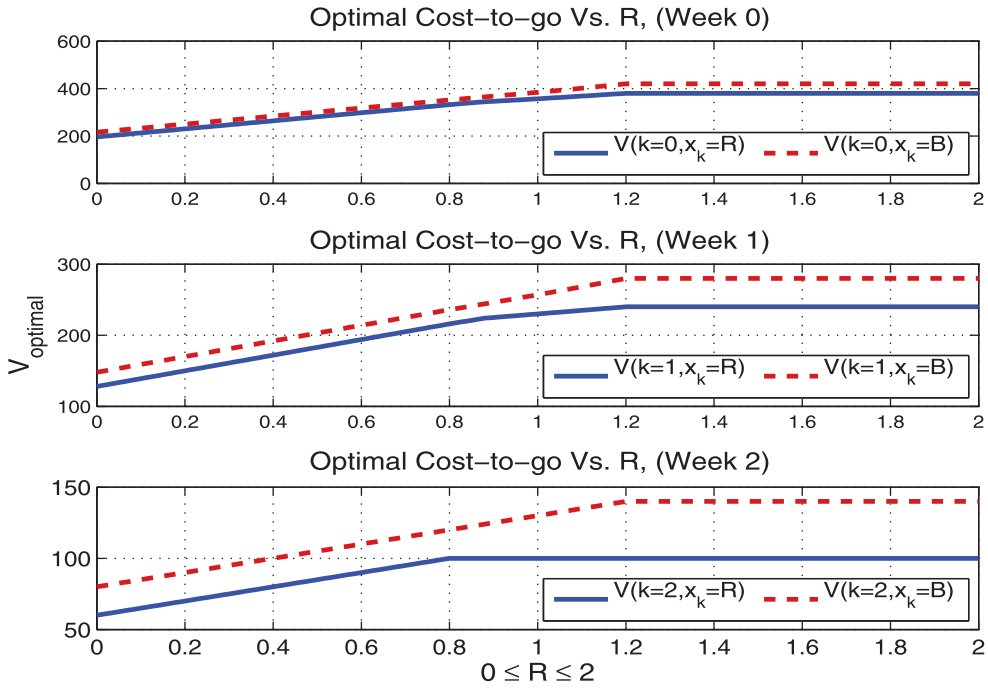
Based on this formulation, the dynamic programming equation is given by

$$V_N(x_N)=0, \tag{4.4}$$

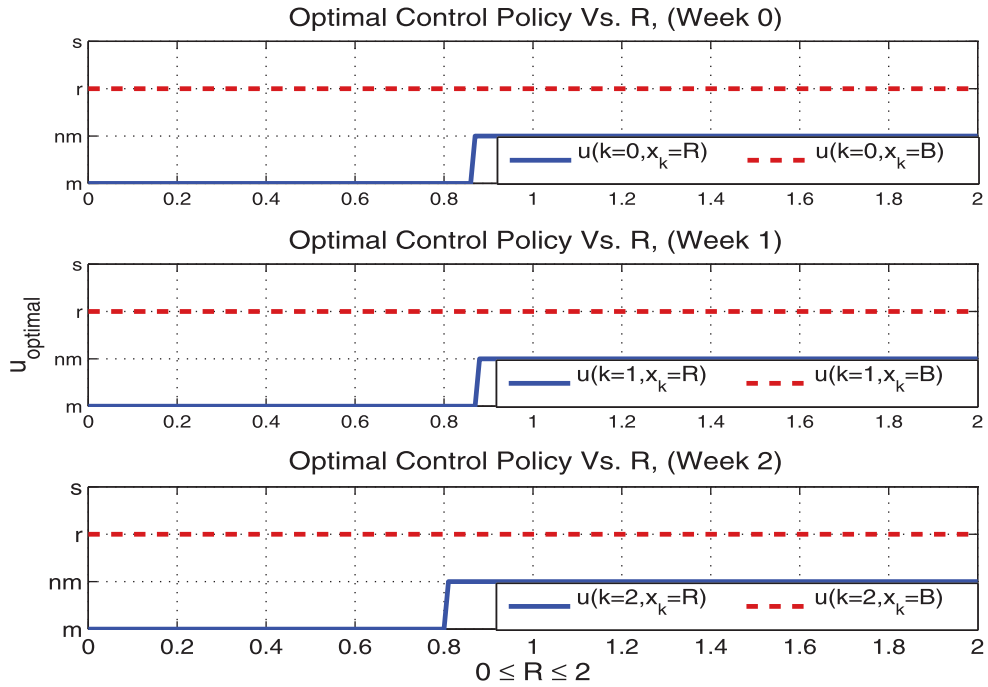
$$V_k(x_k)=\min_{u_k\in U_k(x_k)}\mathbb{E}_{\nu^*(\cdot|\cdot,\cdot)}\Big\{g_k(x_k,u_k,w_k)+V_{k+1}(f(x_k,u_k,w_k))\Big\}. \tag{4.5}$$

We assume that the planning horizon is $N=3$. The optimal cost-to-go and the optimal control policy, for each week and each possible state, as a function of $R\in[0,2]$ are illustrated in Figure 1. Clearly, Figure 1(a) depicts that the optimal cost-to-go is a nondecreasing concave function of R as stated in Lemma 3.10.

In addition, the optimum solution for two possible values of R and for each week results in optimal control policies as depicted in Table 1. By setting $R=0$, we choose to calculate the optimal control policy when the true conditional probability $\nu(\cdot|x_k,u_k)=\mu(\cdot|x_k,u_k)$, $k=0,1,2$. This corresponds to the classical dynamic programming algorithm. By setting $R=0.85$, we choose to calculate the optimal control policy when the true conditional distribution $\nu(\cdot|x_k,u_k)\neq\mu(\cdot|x_k,u_k)$, $k=0,1,2$. Taking into consideration the maximization (that is, by setting $R>0$) the dynamic programming algorithm results in optimal control policies which are more robust with respect to uncertainty, but with the sacrifice of low present and future costs. In cases in which we need to balance the desire for low costs with the undesirability of scenarios with high uncertainty, we must choose the appropriate value of R by using Figure 1(a).



(a) Optimal cost-to-go

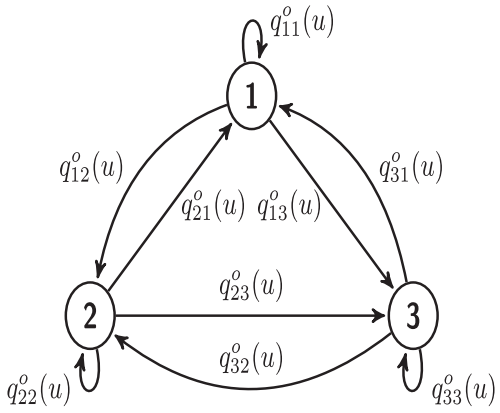


(b) Optimal control policy ("m" = maintenance, "nm" = no maintenace, "r" = repair, "s" = replace)

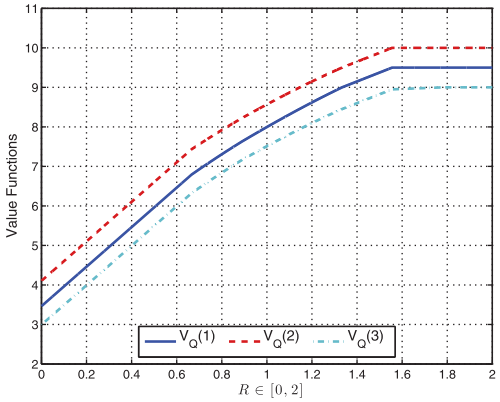
FIG. 1. *Machine replacement example.*

TABLE 1
Dynamic programming algorithm results.

		Week 0		Week 1		Week 2	
	State	Cost-to-go	Optimal Policy	Cost-to-go	Optimal Policy	Cost-to-go	Optimal Policy
$R = 0$	**R**	196	m	128	m	60	m
	B	216	r	148	r	80	r
$R = 0.85$	**R**	340	m	221	m	100	nm
	B	360	r	241	r	122	r



(a) Transition probability graph



(b) Optimal value function

FIG. 2. *Infinite horizon D-MCM.*

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

Assume that the nominal transition probabilities are given under controls u_1 and u_2 by

$$Q^o(u_1) = \frac{1}{9}\begin{pmatrix} 3 & 1 & 5 \\ 4 & 2 & 3 \\ 1 & 6 & 2 \end{pmatrix}, \quad Q^o(u_2) = \frac{1}{9}\begin{pmatrix} 1 & 2 & 6 \\ 4 & 2 & 3 \\ 4 & 1 & 4 \end{pmatrix}, \tag{4.6}$$

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

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

Using policy iteration of section 3.3, with initial policies $g_0(1) = u_1$, $g_0(2) = u_2$, $g_0(3) = u_2$, the algorithm converges to the following optimal policy and value after two iterations:

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

Figure 2(b) depicts the optimal value functions for all possible values of R and shows that the value functions are nondecreasing and concave functions of total variation parameter R.

4.3. Finite horizon MCM—inventory control example. Consider an inventory control example inspired by [5]. Specifically, an optimal inventory ordering policy of a quantity of a certain item at each of the N periods must be found so as to meet a stochastic demand. Let us denote

- x_k, stock available at the beginning of the kth period;
- u_k, stock ordered at the beginning of the kth period;
- w_k, demand during kth period with given probability distribution;
- h, holding cost per unit item remaining unsold at the end of the kth period;
- c, cost per unit stock ordered;
- p, shortage cost per unit demand unfilled.

The random disturbance at time k, w_k may depend on values of x_k and u_k but not on values of prior disturbances $w_0, \ldots, w_{k-1}$. Excess demand is backlogged and filled as soon as additional inventory becomes available. Inventory and demand are nonnegative integers variables. Thus, we assume a nominal system given by

$$x_{k+1} = \max(0, x_k + u_k - w_k) \tag{4.7}$$

and a total sample payoff over N periods given by

$$\sum_{k=0}^{N-1} \left(cu_k + h\max(0, x_k + u_k - w_k) + p\max(0, w_k - x_k - u_k)\right).$$

We further assume that w_k is independent and identically distributed according to $\mu_{w_k}(\cdot) = \mu_w(\cdot)$. We formulate the problem as a minimax optimization of the expected cost as follows:

$$\min_{u_k \in U_k(x_k)} \max_{\substack{\nu_{w_k}(\cdot):||\nu_{w_k}(\cdot)-\mu_w(\cdot)||_{TV} \le R \\ k=0,\ldots,N-1}} \mathbb{E}\left\{\sum_{k=0}^{N-1} \left(cu_k + h\max(0, x_k + u_k - w_k) + p\max(0, w_k - x_k - u_k)\right)\right\}. \tag{4.8}$$

Assume the following:

- The nominal and the true distribution of $\{w_k : k = 0, 1, \ldots, N-1\}$ is $\mu_{w_k}(\cdot) = \mu_w(\cdot)$ and $\nu_{w_k}(\cdot)$, respectively, $k = 0, 1, \ldots, N-1$.

- The maximum capacity $(x_k + u_k)$ for stock is two units.
- The planning horizon $N = 3$ periods.
- The holding cost h and the ordering cost c are both one unit.
- The shortage cost p is three units.
- The demand w_k has a nominal probability distribution given by $\mu_w(w_k = 0) = 0.2$, $\mu_w(w_k = 1) = 0.7$, and $\mu_w(w_k = 2) = 0.1$, $k = 0, 1, \ldots, N-1$.

Dynamic programming subject to total variation distance constraint. The dynamic programming algorithm for the minimax problem subject to total variation distance uncertainty is given by

$$V_N(x_N) = 0,$$

$$\begin{aligned} V_k(x_k) = & \min_{0\le u_k\le 2-x_k} \max_{\nu_{w_k}(\cdot):||\nu_{w_k}(\cdot)-\mu_w(\cdot)||_{TV}\le R} \\ & \mathbb{E}\Big\{u_k + \max(0, x_k + u_k - w_k) + 3\max(0, w_k - x_k - u_k) \\ & \qquad + V_{k+1}(\max(0, x_k + u_k - w_k))\Big\} \qquad (4.10) \\ = & \min_{0\le u_k\le 2-x_k} \max_{\nu_{w_k}(\cdot):||\nu_{w_k}(\cdot)-\mu_w(\cdot)||_{TV}\le R} \mathbb{E}\Big\{\ell_k(x_k, u_k, w_k)\Big\}, \end{aligned} \qquad (4.9)$$

where

$$\begin{aligned} \ell_k(x_k, u_k, w_k) = & u_k + \max(0, x_k + u_k - w_k) \\ & + 3\max(0, w_k - x_k - u_k) + V_{k+1}(\max(0, x_k + u_k - w_k)). \end{aligned}$$

To address the maximization problem in (4.10), for each $k = 0, 1, \ldots, N-1$, $x_k \in \{0, 1, 2\}$ and $0 \le u_k \le 2-x_k$, define the maximum and minimum values of $\ell(x_k, u_k, w_k)$ by $\ell_{\max}(x_k, u_k) \triangleq \max_{w_k\in\{0,1,2\}} \ell(x_k, u_k, w_k)$ and $\ell_{\min}(x_k, u_k) \triangleq \min_{w_k\in\{0,1,2\}} \ell(x_k, u_k, w_k)$, respectively. Its corresponding support sets are given by

$$\begin{aligned} \Sigma^0 &= \Big\{w_k \in \{0, 1, 2\} : \ell(x_k, u_k, w_k) = \ell_{\max}(x_k, u_k)\Big\}, \\ \Sigma_0 &= \Big\{w_k \in \{0, 1, 2\} : \ell(x_k, u_k, w_k) = \ell_{\min}(x_k, u_k)\Big\}. \end{aligned}$$

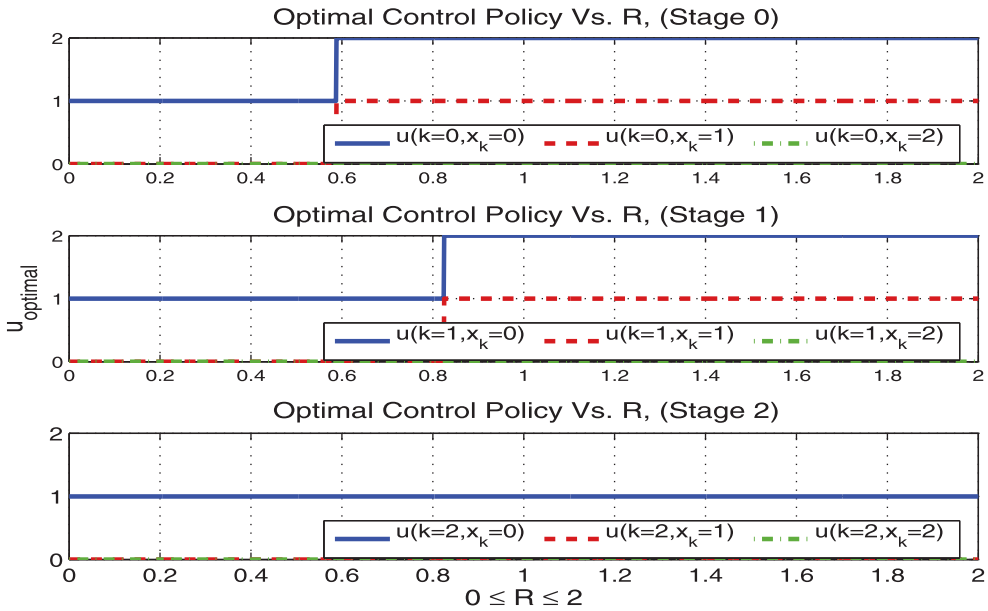
For all remaining sequence $\{\ell(x_k, u_k, w_k) : w_k \in \{0, 1, 2\} \setminus \Sigma^0 \cup \Sigma_0\}$ and for $1 \le r \le |\{0, 1, 2\} \setminus \Sigma^0 \cup \Sigma_0|$, define recursively the set of indices for which $\ell(x_k, u_k, u_k)$ achieves its $(j + 1)$th smallest value by

$$\begin{aligned} \Sigma_j \triangleq & \Bigg\{w_k\in\{0, 1, 2\} : \ell(x_k, u_k, w_k) \\ & = \min\Bigg\{\ell(x_k, u_k, \alpha_k) : \alpha_k \in \{0, 1, 2\} \setminus \Sigma^0 \cup \Bigg(\bigcup_{i=1}^{j} \Sigma_{i-1}\Bigg)\Bigg\}\Bigg\}, \quad j \in \{1, 2, \ldots, r\} \end{aligned}$$

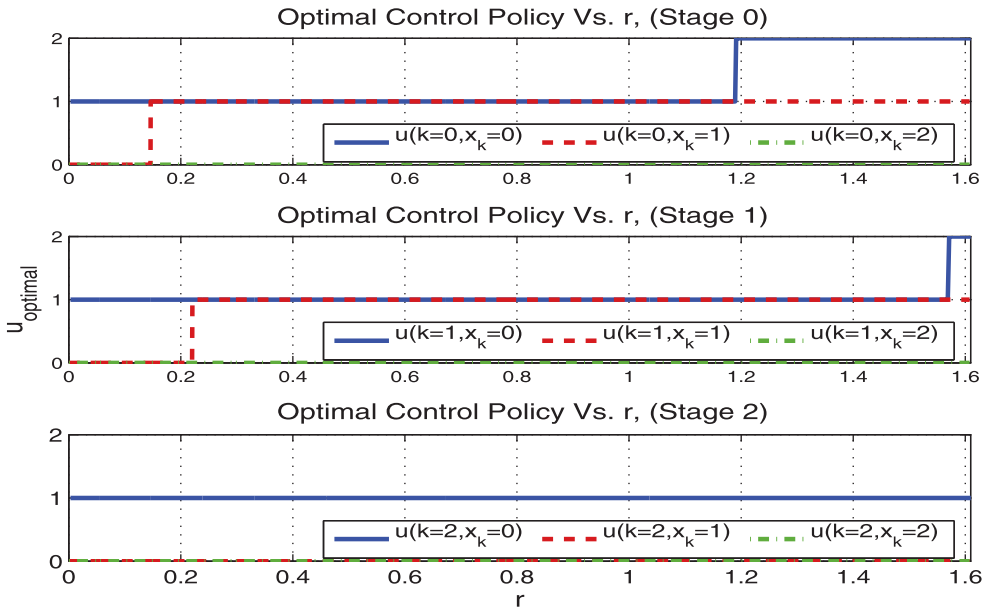
until all the elements of $\{0, 1, 2\}$ are exhausted. Further, define

$$\ell_{\Sigma_j}(x_k, u_k) \triangleq \min_{w_k\in\{0,1,2\}\setminus\Sigma^0\cup(\bigcup_{i=1}^{j}\Sigma_{i-1})} \ell(x_k, u_k, w_k), \quad j \in \{1, 2, \ldots, r\}.$$

Once we identify the support sets and the corresponding values of the sequence $\ell(x_k, u_k, w_k)$ on these sets, we employ (2.4) and (2.5)–(2.8) to calculate the maximizing distribution $\nu^*_{w_k}(\cdot)$ and the extremum solution of $\max_{\nu_{w_k}(\cdot):||\nu_{w_k}(\cdot)-\mu_w(\cdot)||_{TV}\le R}$



(a) Under Total Variation as a Constraint



(b) Under Relative Entropy as a Constraint

FIG. 3. *Optimal control policy of the inventory control example.*

$\mathbb{E}\{\ell(x_k, u_k, w_k)\}$. Finally, by employing (4.9)–(4.10), the optimal cost-to-go and hence the optimal ordering policy are obtained. The optimal control policy, for each period and for each possible state, as a function of $R \in [0, 2]$, is illustrated in Figure 3(a).

Dynamic programming subject to the relative entropy constraint. The dynamic programming algorithm for the minimax problem subject to the relative

entropy constraint is given by

$$
\begin{aligned}
V_N(x_N) &= 0,\\
V_k(x_k) &= \min_{0\le u_k\le 2-x_k}\ \min_{s(x_k)\ge 0}\\
&\quad \Bigg\{ s(x_k)\log \mathbb{E}\\
&\qquad \Bigg\{ \exp\Bigg(\frac{1}{s(x_k)}\big(u_k+\max(0,x_k+u_k-w_k)+3\max(0,w_k-x_k-u_k)\\
&\qquad\qquad + V_{k+1}(\max(0,x_k+u_k-w_k))\big)\Bigg)\Bigg\} + s(x_k)r(x_k)\Bigg\}.
\end{aligned}
$$

The above dynamic programming equations are obtained by slightly modifying dynamic programming equations (1.19)–(1.19), since the cost of the inventory control example is also a function of the demand w_k. The problem with relative entropy is a convex optimization problem, and the maximization of the cost over the relative entropy is a concave nondecreasing function of $r(x) \in [0, r_{\max})$ for all x where $r_{\max}$ can be computed. In addition, since the ambiguity set described by relative entropy is a subset of the much larger total variation ambiguity set,[3] lower values of the optimal cost-to-go are obtained compared to the ones obtained under total variation ambiguity.

Figure 3(b) depicts the optimal control policy for each period and for each possible state, as a function of the relative entropy constraint. Figures 4 and 5 depict a realization of the inventory control example under the resulting optimal control policy for three possible scenarios: (i) without ambiguity,[4] (ii) with ambiguity based on total variation distance, and (iii) with ambiguity based on relative entropy. In particular, the comparison is performed by first choosing the maximizing distribution $\nu^* = [0.1 \quad 0.3 \quad 0.6]$ and by calculating the total variation and the relative entropy parameters. The resulting total variation parameter R is equal to one, while the resulting relative entropy parameter r is equal to 0.75. Then by extracting the optimal control policies from Figure 3(a) and (b) (for the corresponding value of total variation and relative entropy parameter), and by selecting the stock available x_k, and the demand w_k, for each period as shown in Figures 4 and 5, it is clear that optimal control policy under total variation distance ambiguity is more robust with respect to optimal control policies with no ambiguity and with relative entropy ambiguity in which excess demand is lost. Similar behavior is observed for other choices of parameters. In conclusion, the dynamic programming based on relative entropy has the disadvantage that it does not admit distributions which are singular with respect to the nominal distribution, and this rules out the cases in which the nominal systems are simplified versions of the true systems. This is in contrast to the dynamic programming based on total variation distance.

5. Conclusions. In this paper, we examined the optimality of stochastic control strategies via dynamic programming, when the ambiguity class is described by the total variation distance between the conditional distribution of the controlled process and the nominal conditional distribution. The problem is formulated using minimax strategies in which the control process seeks to minimize the payoff while

[3] See Pinsker's inequality (1.14).

[4] This scenario corresponds to the classical dynamic programming; see Figure 3(a) or Figure 3(b) for $R = 0$ and $r = 0$, respectively.

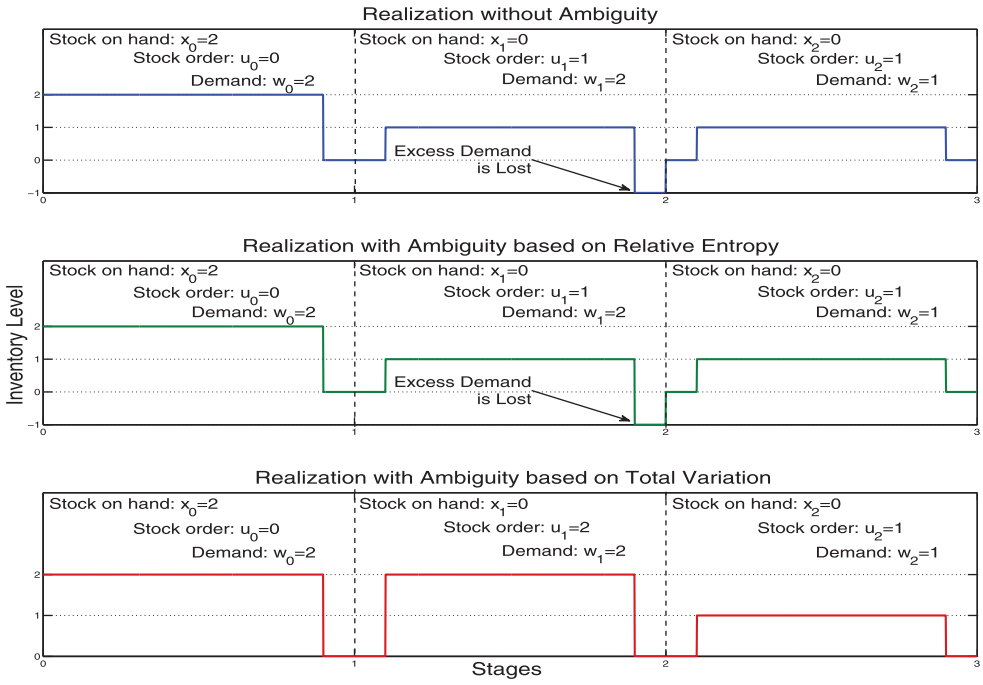


FIG. 4. *Realization of the inventory control example under the resulting optimal policy.*

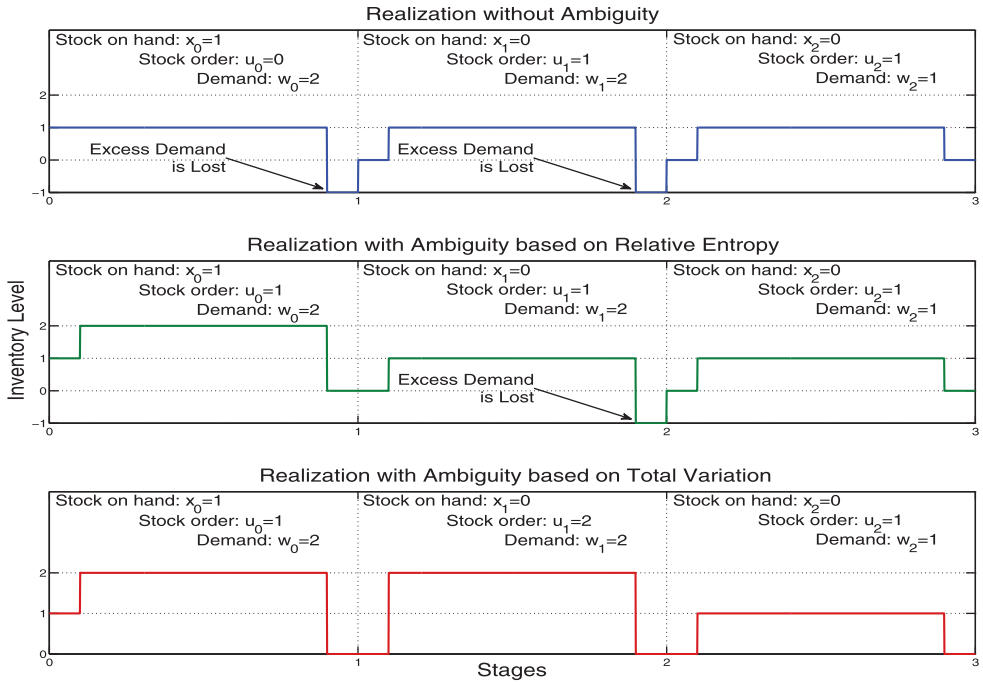


FIG. 5. *Realization of the inventory control example under the resulting optimal policy.*

the controlled process seeks to maximize it over the total variation ambiguity class. By using concepts from signed measures, a closed form expression of the maximizing measure is derived. It is then employed to obtain a new dynamic programming recursion which, in addition to the standard terms, includes the oscillator seminorm of the value function, while for the infinite horizon case a new discounted dynamic programming equation is obtained. It is shown that the dynamic programming operator is contractive, and a new policy iteration algorithm is developed for computing the optimal stochastic control strategies. Finally, we illustrate through examples the applications of our results.

Acknowledgment. The authors are grateful to the reviewer who suggested including comparisons with relative entropy ambiguity.

REFERENCES

[1] N.U. Ahmed, *Linear and Nonlinear Filtering for Scientists and Engineers*, World Scientific, Singapore, 1999.

[2] J.S. Baras and M. Rabi, *Maximum entropy models, dynamic games, and robust output feedback control for automata*, in Proceedings of the 44th IEEE Conference on Decision and Control, and the European Control Conference, Seville, Spain, 2005.

[3] T.S. Basar and P. Bernhard, *H-infinity Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*, Collect. Syst. Compl., 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.

[5] D.P. Bertsekas, *Dynamic Programming and Optimal Control*, Athena Scientific, Nashua, NH, 2005.

[6] D.P. Bertsekas and S.E. Shreve, *Stochastic Optimal Control: The Discrete-Time Case*, Athena Scientific, Nashua, NH, 2007.

[7] P.E. Caines, *Linear Stochastic Systems*, Wiley, New York, 1988.

[8] C.D. Charalambous and J. Hibey, *Minimum principle for partially observable nonlinear risk-sensitive control problems using measure-valued decompositions*, Stoch. Stoch. Rep., 57 (1996), pp. 247–288.

[9] 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.

[10] 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.

[11] R.J. Elliott, L. Aggoun, and J.B. Moore, *Hidden Markov Models: Estimation and Control*, Springer, New York, 1995.

[12] O. Hernandez-Lerma and J.B. Lasserre, *Discrete-time Markov control processes: Basic optimality criteria*, in Applications of Mathematics Stochastic Modelling and Applied Probability, Springer, New York, 1996.

[13] 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.

[14] P.R. Kumar and J.H. Van Schuppen, *On the optimal control of stochastic systems with an exponential-of-integral performance index*, J. Math. Anal. Appl., 80 (1981), pp. 312–332.

[15] P.R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*, Prentice-Hall, Englewood Cliffs, NJ, 1986.

[16] D.G. Luenberger, *Optimization by Vector Space Methods*, Professional Series, Wiley, New York, 1968.

[17] I.R. Petersen, M.R. James, and P. Dupuis, *Minimax optimal control of stochastic uncertain systems with relative entropy constraints*, IEEE Trans. Automat. Control, 45 (2000), pp. 398–412.

[18] P. Dai Pra, L. Meneghini, and W.J. Runggaldier, *Connections between stochastic control and dynamic games*, Math. Control Signals Systems, 9 (1996), pp. 303–326.

[19] F. Rezaei, C.D. Charalambous, and N.U. Ahmed, *Optimal control of uncertain stochastic systems subject to total variation distance uncertainty*, SIAM J. Control Optim., 50 (2012), pp. 2683–2725.

[20] V.A. Ugrinovskii and I.R. Petersen, *Finite horizon minimax optimal control of stochastic partially observed time varying uncertain systems*, Math. Control Signals Systems, 12 (1999), pp. 1–23.

[21] J.H. van Schuppen, *Mathematical control and system theory of discrete-time stochastic systems*, preprint, 2014.

[22] P. Whittle, *A risk-sensitive maximum principle: The case of imperfect state observations*, IEEE Trans. Automat. Control, 36 (1991), pp. 793–801.