Systems & Control Letters 120 (2018) 29–35

Contents lists available at ScienceDirect

Systems & Control Letters

journal homepage: www.elsevier.com/locate/sysconle

The average cost of Markov chains subject to total variation distance uncertainty☆

A.A. Malikopoulos [a,*], C.D. Charalambous [b], I. Tzortzis [b]

[a] Department of Mechanical Engineering, University of Delaware, Newark, DE 19716, USA
[b] Department of Electrical Engineering, University of Cyprus, Nicosia, Cyprus, Cyprus

ARTICLE INFO

Article history:
Received 11 June 2015
Received in revised form 17 July 2018
Accepted 20 August 2018

Keywords:
Stochastic optimal control
Controlled Markov chain
Average cost
Total variation distance

ABSTRACT

This paper addresses the problem of controlling a Markov chain so as to minimize the long-run expected average cost per unit time when the invariant distribution is unknown but we know it belongs to a given uncertain set. The mathematical model used to describe this set is the total variation distance uncertainty. We show that the equilibrium control policy, which yields higher probability to the states with low cost and lower probability to the states with the high cost, is an optimal control policy that minimizes the average cost. Recognition of such a policy may be of value in practical situations with constraints consistent to those studied here when the invariant distribution is uncertain and deriving online an optimal control policy is required.

1. Introduction

The average cost criterion is prominent as being complex to analyze compared to other optimization criteria. While many classical criteria lead to rational complete solutions, the long-run cost may not. The average cost criterion for Markov Chains (MC) with finite state and arbitrary action spaces has been extensively reported in the literature (see, *e.g.*, [1–4] and references therein). A significant amount of research has been also reported for the problem with finite state and action spaces [5–10]. Bather [11] reviewed various techniques for a controlled MC with a finite state space when there is a finite set of possible transition matrices; an example illustrated the unpredictable behavior of policy sequences derived by backward induction. He proposed a new approach based on the idea of classifying the states according to their accessibility from one another. Feinberg [12] considered four average reward criteria on discrete time Markov decision model with a finite state space, and proved the existence of persistently nearly optimal strategies in various classes of strategies for models with complete state information.

Research efforts have focused on infinite horizon, discrete-time Markov Decision Processes (MDPs) with more general state and action spaces. Hordjik [13] extended some earlier results to countable state and action spaces by introducing the *Lyapunov function* method for controlled Markov processes. Based on this method, a solution to the average cost problem can be achieved yielding an optimal control policy. Borkar [14–18] presented a convex analytic approach to address this problem in a general framework with unbounded cost by treating the control problem as a constrained optimization problem on a suitably defined closed convex set of *ergodic occupation measures*. In this work, necessary and sufficient conditions for the existence of an optimal stable stationary deterministic policy were established; moreover, Borkar provided conditions for optimality in terms of dynamic programming when an optimal stable stationary policy is known to exist. Sennott [19] introduced conditions that guarantee an optimal control policy in problems with possibly unbounded, non-negative costs. Cavazos-Cadena [20] considered denumerable state spaces and stationary control policies that induce an ergodic chain; the *value iteration* scheme was utilized to construct convergent approximations of a solution to the *optimality equation* as well as a sequence of stationary policies whose limit points are optimal. Leizarowitz and Zaslavski [21] recently addressed the problem of uniqueness and stability of optimal control policies when a complete set of unicost MDPs is endowed. The problem of minimizing the long-run expected average cost of a complex system consisting of interactive subsystems was addressed in [22]. The problem of minimizing the average cost in a controlled MC by solving a dual constrained optimization problem was addressed in [23]. It was shown that the control policy that yields higher probability to the states with low cost and lower probability to the states with the high cost is an optimal solution and it is defined as an Equilibrium Control Policy (ECP).

☆ This research was supported by the ARPAE's NEXTCAR Program, Cyprus under the award number DE–AR0000796. This support is gratefully acknowledged.

* Corresponding author.
E-mail addresses: andreas@udel.edu (A.A. Malikopoulos), chadcha@ucy.ac.cy (C.D. Charalambous), tzortzis.ioannis@ucy.ac.cy (I. Tzortzis).

https://doi.org/10.1016/j.sysconle.2018.08.003
0167-6911/

In this paper, we address the problem of controlling a MC so as to minimize the long-run expected average cost per unit time when the invariant distribution is unknown but we know it belongs to the Total Variation (TV) distance uncertainty set. We treat the stochastic optimal control problem as a dual constrained optimization problem and we show that the ECP is an optimal control policy that minimizes the average cost. Furthermore, we show that this solution is optimal for the original stochastic control problem without considering uncertainty.

This problem has become increasingly important in automotive related applications [24–27]. In particular, in hybrid electric vehicles (HEVs) implementing online an optimal control policy to distribute the power demanded by the driver optimally to the subsystems, e.g., the internal combustion engine, motor, generator, and battery, constitutes a challenging control problem and has been the object of intense study for the last two decades [28]. In this problem, we select the long-run, expected average cost per unit time criterion as we wish to optimize HEV efficiency (minimize losses) for any different driver and commute on average. However, since the driver's driving style is unknown, the invariant distribution is not known a priori but we know that it belongs to an uncertain set.

The remainder of the paper proceeds as follows: In Section 2, we introduce our notation and formulate the problem. In Section 3, we introduce the uncertainty set based on TV distance. In Section 4, we formulate the stochastic control problem and provide a solution that yields the ECP. Finally, we present an illustrative application in Section 5, and we draw concluding remarks in Section 6.

2. Problem formulation

We consider a system that evolves according to a controlled Markov process with a finite alphabet state space $\mathcal{S}$ of finite cardinality $|\mathcal{S}| = N$, and a finite alphabet control space $\mathcal{U}$ of finite cardinality $|\mathcal{U}|$, from which control actions are chosen. The evolution of the state occurs at each of a sequence of stages $t = 0, 1, \ldots$, and it is portrayed by the sequence of the random variables X_t and U_t corresponding to the system's state and control action. In our formulation, a state-dependent constraint is incorporated; that is, for each realization of the state $X_t = i \in \mathcal{S}$, we are given a nonempty subset $\mathcal{C}(i) \subset \mathcal{U}$ of the control space, and the feasible set of state–action pairs, $\Gamma := \{(i, u)|i \in \mathcal{S} \text{ and } u \in \mathcal{C}(i)\}$. For each realization of the state $X_t = i \in \mathcal{S}$, we define the function $\phi_i : \mathcal{S} \to \mathcal{U}$ that maps the state space to the control space defined as the control law. Each sequence π of the functions ϕ_i, $\pi = \{\phi_1, \ldots, \phi_{|\mathcal{S}|}\}$, is defined as a stationary control policy of the system. Furthermore we consider a function $l : \Gamma \to \mathbb{R}_+$ called the cost function (cost-per-stage).

At each stage, the controller observes the system's state $X_t = i \in \mathcal{S}$, and an action, $U_t = \phi_i = u$, is realized from the feasible set of actions $\mathcal{C}(i)$ at this state. At the next stage t, the system transits to the state $X_{t+1} = j \in \mathcal{S}$ imposed by the conditional probability $\mathbb{P}(X_{t+1} = j|X_t = i, U_t = u)$, and a cost $l(X_t, U_t) = l(i, u)$ is incurred. After the transition to the next state has occurred, a new action is selected, and the process is repeated. The completed period of time over which the system is observed is called the *decision-making horizon* and is denoted by T. The horizon can be either finite or infinite; in this paper, we consider infinite-horizon decision-making problems.

2.1. Long-run expected average cost subject to a distance uncertainty

We consider the long-run expected average cost per unit time. The average cost criterion is considered usually for developing the power management control in HEVs or plug-in HEVs (PHEVs), where we seek to derive an optimal control policy that will optimize the efficiency of the HEV/PHEV in the long-term and not necessarily for a specific period of time [29,30]. The assumption of an infinite number of stages is never satisfied in practice. However, it is a reasonable approximation for problems involving a finite but very large number of stages.

Problem Statement P0. The minimum average cost corresponding to the optimal control policy π^* is

$$J^*(\pi^*) = \min_{\pi \in \Pi} \lim_{T \to \infty} \frac{1}{T+1} E\Big[\sum_{t=0}^{T} l(X_t, U_t)\Big]. \tag{1}$$

To guarantee that the limit in (1) exists, we impose the following assumption.

Assumption 2.1. For each stationary control policy $\pi = \{\phi_1, \phi_2, \ldots, \phi_{|\mathcal{S}|}\}$, the MC $\{X_t|t = 1, 2, \ldots\}$ has a single ergodic class.

Namely, for each stationary policy $\pi \in \Pi$, there is a unique invariant distribution (row vector)

$$\mu(\pi) = [\mu_1(\pi), \mu_2(\pi), \ldots, \mu_{|\mathcal{S}|}(\pi)],$$

such that $\mu(\pi) = \mu(\pi) \cdot P(\pi)$, with $\sum_{i \in \mathcal{S}} \mu_i(\pi) = 1$, where $P(\pi)$ is the transition probability matrix. A proof of this assertion may be found in [[31], p. 227]. Under Assumption 2.1, it is known [[32], p. 175] that

$$\lim_{T \to \infty} \frac{1}{T+1} \sum_{t=0}^{T} [P(\pi)]^t = \mathbf{1} \cdot \mu(\pi), \tag{2}$$

where $\mathbf{1} = [1, 1, \ldots, 1]^T$ is the column vector whose elements are all unity. Substituting (2) into (1) shows that long run average expected average cost per unit time, $J(\pi)$, does not depend on the initial state and is given by

$$J(\pi) = \mu(\pi) \cdot l(\pi), \tag{3}$$

where $l(\pi) = [l(1, \phi_1), l(2, \phi_2), \ldots, l(i, \phi_i), \ldots, l(\mathcal{S}, \phi_{|\mathcal{S}|})]^T$ is the column vector of the cost function. Consequently, a stationary control policy is optimal if

$$J^* = J^*(\pi^*) = \inf\{J(\pi)|\pi \in \Pi\}, \tag{4}$$

where Π is the set of the feasible control policies. To simplify notation, if the context makes it clear we do not emphasize the dependence of the average cost $J(\pi)$, invariant distribution $\mu(\pi)$, and cost function $l(\pi)$ on the control policy π, and we denote them simply by J, μ, and l.

Problem Statement P1. Our objective is to derive the optimal control policy that minimizes the long-run, expected average cost per unit time in (3), when the invariant distribution, $\mu(\pi)$, is unknown but it belongs to an uncertain set, described by the TV distance ball.

The mathematical model used to describe the uncertainty set is the TV distance developed in earlier work [33,34]. The problem of deriving an optimal control policy that minimizes the average cost can be reformulated as a dual constrained optimization problem. More specifically, we can formulate a problem to derive a control policy that minimizes the cost at each state with maximum probability, or alternatively, maximizes the probability of the states incurring minimum cost. The average cost in (3) is a linear functional on the Banach space of all bounded, continuous, real-valued functions. The existence of a family of probability measures which attain the supremum of the average cost in the general case has been discussed in [35]. The uncertainty set based on TV distance is weak*-compact and the functional weak* continuous [35]. Hence, there exist a probability measure in this set that maximizes the functional J. Since the set Γ is compact there exists a cost-per-stage that minimizes the functional J. The following section provides the solution of the above optimization problem.